The Metalog Distributions

By Thomas W. Keelin. October, 2016. See http://www.metalogdistributions.com for data and Excel implementation. The *Decision-Analysis*-published version of this manuscript be obtained from http://dx.doi.org/10.1287/deca.2016.0338.

Abstract

The metalog distributions constitute a new system of continuous univariate probability distributions designed for flexibility, simplicity, and ease/speed of use in practice. The system is comprised of unbounded, semi-bounded, and bounded distributions, each of which offers nearly unlimited shape flexibility compared to Pearson, Johnson, and other traditional systems of distributions. Explicit shape-flexibility comparisons are provided. Unlike other distributions that require non-linear optimization for parameter estimation, the metalog quantile functions and PDFs have simple closed-form expressions that are quantile-parameterized linearly by CDF data. Applications in fish biology and hydrology show how metalogs may aid data and distribution research by imposing fewer shape constraints than other commonly used distributions. Applications in decision analysis show how the metalog system can be specified with three assessed quantiles, how it facilities Monte Carlo simulation, and how applying it aided an actual decision that would have been made wrongly based on commonly-used discrete methods.

1. Introduction

In economics, business, engineering, science and other fields, continuous uncertainties frequently arise that are not easily- or well-characterized by previously-named continuous probability distributions. Frequently, there is data available from measurements, assessments, derivations, simulations or other sources that characterize the range of an uncertainty. But the underlying process that generated this data is either unknown or fails to lend itself to convenient derivation of equations that appropriately characterize the probability density (PDF), cumulative (CDF) or quantile distribution functions.

Desiring a continuous probability distribution but lacking appropriate functional forms, some analysts have attempted to "fit" their data to previously-named distributions, often with less-than-satisfactory results. For example, one may attempt to derive the parameters of a normal distribution from a given set of CDF data, but the resulting normal distribution will never be a satisfactory representation if the data itself is indicative of a skewed or bounded distribution, of which the normal is neither. While fitting the same data set to the parameters of a beta distribution may yield a beta distribution with appropriate skewness, the resulting beta distribution may not be satisfactory if the data itself is representative of an unbounded or semi-bounded distribution, which the beta is not. Moreover, such fitting involves considerable effort and complexity since such probability distributions are often non-linear in their parameters, lack a closed-form expression, or both.

Moreover, among a set of previously-named distribution that have bounds that match natural bounds of the data, it may be unclear which of many distributions to select. The choice of distribution can be

important because it inherently imposes shape constraints that may or may not appropriately represent the data and the process that generated it. In such cases, one needs a distribution that has flexibility far beyond that of traditional distributions -- one that enables "the data to speak for itself" in contrast to imposing unexamined and possibly inappropriate shape constraints on that data. While this need applies to a wide range of empirically generated frequency data, it can be especially acute when a probability distribution is used to represent state-of-information (or belief-based) data as is common in decision analysis and in an increasingly wide range of other modern applications of probability.

When there are many continuous uncertainties with very different characteristics to represent, as is often the case in decision analysis, it may be simply impractical to attempt to find a continuous representation tailored to each uncertainty using traditional methods. So decision analysts often resort to using discrete (e.g. three branch) representations. These have multiple shortcomings including that they artificially cut off the tails and introduce undue lumpiness into the analysis.

Desiring a continuous probability distribution but lacking appropriate functional forms, other analysts have resorted to sorting their data into buckets to develop histograms, which have the advantage of being able to represent the shape and location of most any continuous uncertainty. However, histogram development also involves effort and complexity, often includes an arbitrary choice of bucket limits, and inherently results in a lumpy stair-step display rather than a smooth PDF. Maximum entropy methods (Abbas 2003), which strive to add no information beyond the data, similarly result in either a stair-step or piecewise linear PDF. When knowledge of smoothness is present in addition to the data, such formulations are less than ideal.

For applications that require probabilistic (Monte Carlo) simulation, the situation of having data but not continuous distribution functions is even more challenging and complicated. Sampling directly from the data itself (discrete sampling) is not satisfactory if one believes there are gaps, lack of sufficient tail representation, or other shortcomings in the data. Sampling from bucketed data (histograms) requires programming of the buckets and is inherently lumpy. Moreover, even if an appropriate continuous distribution has been identified (e.g. by a data "fit" to its parameters), most continuous CDF's cannot be solved analytically for their inverse-CDF (quantile function), which is required for simulation. So look-up tables or non-linear programming must be employed for each sample.

The metalog family of distributions can solve all these problems, and it has been proven effective and easy to use in practice. The metalog distributions can effectively represent a wide range of continuous probability distributions -- whether skewed or symmetric, bounded, semi-bounded, or unbounded. Scaling constants that determine its shape and location are uniquely determined by a convenient linear transformation of CDF data. In contrast to other continuous distributions, there is no need for non-linear optimization to fit parameters to the data. In addition, the metalog's simple, algebraic closed forms are easy to program, making it easy to replace lumpy, stair-step, or piecewise linear PDF displays with smooth, continuous ones.

For simulation applications, the metalog distributions enable the calculation of a sample from a uniformly distributed random number according to a simple, algebraic equation, thereby displacing any need to use a look-up tables or non-linear optimization for the calculation of each sample. Moreover, over a wide range of applications, the results of the simulation can be conveniently and accurately represented by a metalog, compressing what may otherwise require thousands of data points into a simple closed-form distributional representation.

For direct probability assessments in decision analysis and other Bayesian applications, the metalog distributions provide a convenient way to translate CDF data into smooth, continuous, closed-from distribution functions that can be used for real-time feedback to experts about the implications of their probability assessments -- free from the confines of other continuous distributions that have more limited flexibility. In practice, we have found that the resulting metalog often yields a more accurate and authentic representation of expert beliefs than the data itself.

The unbounded metalog distribution is a Quantile-Parameterized Distribution (QPD), (Keelin and Powley, 2011), and might be regarded as an easier-to-use and more-broadly –applicable successor to the Simple Q Normal distribution introduced in that paper. Like the Simple Q Normal, the metalog distribution can effectively represent a wide range of unbounded continuous probability distributions. The metalog, however, has several advantages: an unlimited number of terms rather than just four (enabling more flexible distributional representations); closed-form, smooth (continuously differentiable) quantile-function and PDF expressions – obviating any need for lookup tables; closed-form analytic expressions for its central moments; and closed-form analytic transforms that conveniently express probability distributions that are semi-bounded or bounded – while retaining the unbounded metalog's flexibility, smoothness, and ease-of-parameterization properties.

The remainder of this paper is organized as follows. Section 2 provides overview of the strengths and weaknesses of existing families of flexible distributions, desiderata and engineering methods for developing new flexible distributions, and how these methods have been applied previously. Section 3 applies a novel combination of these methods to develop the unbounded metalog distribution, and shows how its flexibility compares with corresponding distributions from previous distribution families, including those of Pearson and Johnson. Section 4 shows how the flexibility of unbounded the metalog along with its linear quantile-parameterization can be propagated into the domain of semi-bounded and bounded distributions. The flexibility of these semi-bounded and bounded metalogs is analyzed and compared with corresponding Pearson and Johnson distributions, among others. Section 5 further illustrates the flexibility of the metalog distributions by showing how well they approximate a wide range of existing distributions. Section 6 presents applications. Applications in fish biology and hydrology show how metalogs may aid data and distribution research by imposing fewer shape constraints than other commonly used distributions. Applications in decision analysis show how the metalog system can be specified with three assessed quantiles, how it facilities Monte Carlo simulation, and how applying it aided an actual decision that would have been made wrongly based on commonly-used discrete methods. At the end of Section 6, we provide

guidelines for distribution selection within the metalog system, using the previous applications as examples. Section 7 offers conclusions and suggested directions for future research.

2. Literature Review and Motivation

2.1 Types of Probability Distributions

For context, we divide probability distributions into three types -- Type I, Type II and Type III. Type I distributions can be derived from an underlying *probability model*, from which they gain much of their appeal and legitimacy. For example, the normal distribution was originally derived as a limiting case of the previously-known binomial distribution (De Moivre, 1756) and is also the limiting shape for various central limit theorems. Similarly, the exponential distribution can be derived as the probability distribution of waiting times between events governed by a Poisson process. The shape of a Type I distribution is determined largely or entirely by its underlying probability model. For example, the normal distribution has one location parameter μ and one scale parameter σ, but no shape parameters. The exponential distribution has a single scale parameter λ, but no shape parameters. Such shape restrictions make Type I distributions an excellent choice for practical use whenever the situation fits the probability model, and especially so when empirical data that would otherwise characterize the distribution are sparse or unreliable.

Type II probability distributions gain their appeal and legitimacy less from an underlying probability model and more from their ability to represent *specific* probabilistic data or processes that are not known to correspond to an existing Type I model. Most commonly they are "generalizations" of other previously identified distributions, formed by adding one or more parameters that enable a good fit to the specific (ad hoc) data under consideration. For example, Mead (1965) generalized the logit-normal distribution (proposed previously by Johnson, 1949) by adding a parameter that provides flexibility to fit an empirical distribution of carrot-root diameters. Theodossiou (1994) developed skewed version of a generalized student-t distribution on the basis that it provided a better representation of financial data (e.g. log daily returns of market-traded stocks) than previously available distributions. Theodossiou's distribution is itself a generalization of a previously generalized student-t distribution (McDonald and Newey, 1988). By now, Type II distributions published in the literature may number in the dozens or hundreds. Johnson, Kotz, and Balakrishnan (1994) detail many Type I distributions and Type II generalizations.

Type III distributions gain their appeal and legitimacy from being as *broadly* applicable as possible. Unlike Type II distributions designed to match a specific class or classes of empirical data, Type III distributions would ideally match most *any* set of data. This ideal includes, but is not limited to, effectively representing data consistent with the numerous Type I and Type II distributions. Moreover, with the success and resurgence of the Bayesian revolution (McGrayne, 2011) and the evolution of the theory and practice of decision analysis (Howard (1968, 2015), Raiffa (1968), Keeney and Raiffa (1992), Spetzler et. al. 2015,

among others), this ideal includes effectively representing Bayesian priors and other state-of-information-based (or belief-based) distributions over a very wide range of probabilistic data.

2.2 Type III Families of Distributions

Since no single, universally-applicable distribution has yet been found, Type III probability distributions have typically been developed as "systems" or "families" distributions. Within a given family, criteria are provided to enable practitioners to pick which particular distribution to use and how to estimate its parameters from data. The metalog system introduced by this paper is such a family of distributions.

In his book on families of distributions, Ord (1972) lamented that keeping track of "the wide-ranging and rapidly-expanding literature (on families of distributions) is probably a hopeless task." This is even more the case now – more than forty years later. So, for this paper, we shall content ourselves with discussion of a few well-known systems of distributions -- specifically, the Pearson (1895, 1901, 1916), Johnson (1949), Tadikamalla and Johnson (1982) systems. We shall also discuss the general family of quantile-parameterized distributions (QPDs), Keelin and Powley (2011), because the unbounded metalog is one of these. A more complete discussion of Type III systems distributions can be found in Ord (1972) and Johnson, Kotz, and Balakrishnan (1994).

2.3 Type III Desiderata: Flexibility, Simplicity, Ease/Speed of Use

Johnson (1949) identified several criteria for judging the desirability of any Type III system of distributions, including his own. In this view, Type-I considerations are less important than practical-use considerations such as flexibility, simplicity, and ease of use. Similar criteria have been adopted and employed subsequently by Mead (1965) and Johnson, et. al, (1994), among others.

Flexibility

Flexibility is the ability of the family to represent a wide range of probabilistic data whatever may be its source or rationale. Since any distribution can be easily modified via linear transformation to accommodate changes in location and scale, shape flexibility, in contrast to location and scale, is key. To maximize shape flexibility in probability distribution design, one must eschew Type I considerations that limit flexibility. However, such Type I considerations may play useful a role for interpreting special cases of a more general and flexible distribution.

Flexibility also includes the ability to match natural bounds, if any. For example, distances, times, volumes, and other such variables often have a natural lower bound (zero) and no specific upper bound. Percentages of a population or frequencies of occurrence typically have both a lower bound (zero) and an upper bound (one). Other variables, such as bi-directional error measurements or deviations from a point, may be naturally unbounded both high and low.

Simplicity

Simplicity refers to the simplicity of functional form of the PDF and CDF and/or quantile function, ease of algebraic manipulation, and ease of interpretation. For example, we consider closed-form algebraic expressions to be simpler than those that include limits, integrals, statistical functions like Beta and Gamma, look-up tables, or implicitly defined functions that require iteration.

Ease/Speed of Use

Two critical components of ease of use are ease of distribution selection and ease of parameter estimation. Absent Type I considerations, the literature provides incomplete guidance for distribution selection. For example, suppose that a practitioner has a specific set of empirical data that she wishes to represent with a continuous probability distribution. She knows this her data has a natural lower bound of zero, no natural upper bound, and that it is right-skewed "sort of like a lognormal". There are, however, many distributions that look "sort of like a lognormal." Beyond the lognormal itself, these include the gamma, inverse gamma, chi square, log gamma, log Pearson Type III, log logistic, Burr, Rayleigh, and Weibull, among others. Which should she choose?

Once she has selected a potentially suitable distribution, she cannot know whether she has a good fit until she estimates the parameters of that distribution from her data and views the result. While many good parameter-estimation methods are available, there is no one method that is generally applicable and easy to use in all cases. In most cases, such methods need to be tailored to the particular mathematical form of the distribution under consideration and, even then, may require a non-trivial multi-variable non-linear optimization that can be solved only by iteration within distribution-specific constraints[1]. For this reason, a large literature has evolved to address distribution-specific parameter estimation[2].

Today's Requirements

Beyond ease of distribution selection and parameter estimation, ease of use depends on purpose and context. At the time of Johnson's 1949 paper, before the advent of modern computers, ease of use included having readily available distribution tables, as had been published for the normal. Today this is much different. An easy-to-use family of distributions should be easy to program (or already be pre-programmed) within the most widely used analytic processing and charting environment[3]. Once programmed, it should be fast to input data, fast and easy to estimate parameters, fast to calculate, and fast to produce interpretable results.

[1] See, for example, Thessidiou (1994)
[2] Johnson, et. al., (1994), Volumes 1 and 2, provide an excellent summary and extensive literature references for parameter estimation for a wide range of distributions.
[3] Today this is Excel.

Today, the requirements for flexibility, simplicity, and especially ease/speed-of-use are critical and can make the difference between use and non-use in practice. Decades ago, a practitioner might have had days, weeks, or months to select an appropriate distribution and to develop an accurate fit to empirical or assessed data for that distribution. In contrast, in today's professional practice of decision analysis, once data has been assessed, a practitioner might have an hour or less to devote to developing, programming, and estimating parameters for a dozen continuous uncertainties with widely-divergent shape and bounds characteristics. Distribution selection and parameter estimation must be fast, seamless, and largely without need for manual intervention over a wide range of data. Moreover, such a practitioner would need to be able to make convenient, rapid adjustments to these distributions to incorporate new information or other changes in state-of-information-based expert data and/or sensitivity analyses. Once formed, the resulting distributions need to be convenient for use in Monte Carlo simulation and ideally without the need for look-up tables or iteration.

If any of these desiderata are not met, a decision analyst might well abandon continuous distributions altogether in favor of discrete approximations, despite their limitations of artificially cutting off the tails and introducing undue lumpiness into the analysis. This particularly challenging environment with respect to flexibility, simplicity, and ease/speed of use motivated our development of the metalog family.

2.4 Engineering Design of Probability Distributions

When designing Type II or Type III probability distributions to best accomplish desiderata as described above, one faces a wide range of choices. These are summarized in a strategy table[4,5] in Table 1. The first row in each column identifies a key decision and subsequent rows identify specific options that are available for that decision. Table 1 is not meant to cover all possible cases, but rather is intended to be illustrative of key choices that have been made by previous researchers and to provide context for understanding the metalog family. It is also intended to provide a point of reference for future researchers who wish to develop new probability distributions or systems of distributions.

As shown in this table, when designing Type II or Type III probability distributions, it is common to start with a particular form of a particular base distribution, to modify it with a particular method, to develop a method to estimate its parameters, and to provide guidance for selection of which distribution to use. Commonly-used base distributions include the normal[6,7,8], logistic[9,10], and student t[11,12]. Commonly-modified forms -- any of which fully specify a probability distribution -- include the probability density

[4] Howard and Abbas (2015), pp. 775-776
[5] Spetzler, Winter, and Meyer (2016), pp. 56-59
[6] Edgeworth (1896, 1907)
[7] Pearson (1895, 1905, 1916)
[8] Johnson (1949)
[9] Tadikamalla and Johnson (1982)
[10] Balakrishnan (1992)
[11] McDonald and Newey (1988)
[12] Theodossiou (1994)

Table 1. Strategy Table for Engineering Probability Distributions

Base Distribution	Form Modified	Modification Method	Parameter Estimation	Distribution Selection
normal[6,7,8]	probability density function (PDF)[6,7]	parameter addition[17,18,19]	method of moments[7]	match moments[7]
logistic[9,10]	cumulative distribution function (CDF)[13]	parameter substitution[7]	maximum likelihood[22]	match bounds
student t[11,12]	quantile function (inverse CDF)[14,15]	trans-formation[8,9,20]	probability-weighted[23]- and L-moments[24]	…
…	characteristic function[16]	series expansion[6,21]	quantile parameterization[20,25]	

Function[6,7], cumulative distribution function[13], quantile function[14,15], and characteristic function[16]. Commonly used modification methods include parameter addition[17,18,19], parameter substitution (substituting an expression for one or more parameters)[7], transformation[8,9, 20] , and series expansion[6,21]. Commonly used parameter estimation methods include the method of moments[7], method of maximum likelihood[22], probability-weighted moments[23], L-moments[24], and quantile-parameterization[20,25,]. For distribution selection within a family, the traditional method has been to select a distribution capable of matching the moments[7] of frequency data. But, given sufficient flexibility to match moments, one can also select a distribution based on natural bounds or other criteria.

[13] Burr (1942)
[14] Karvanen (2006)
[15] Keelin and Powley (2011)
[16] Ord (1972), pp 26-29.
[17] Mead (1965)
[18] McDonald and Newey (1988)
[19] Theodossiou (1994)
[20] Hadlock and Bickel (2016)
[21] Johnson, Kotz, and Balakrishnan (1994) and Ord (1972) provide perspectives on Gram-Charlier, Edgeworth, and other series expansions.
[22] Aldrich (1997) chronicles the development of maximum likelihood by RA Fisher during 1912-1921.
[23] Greenwood, et. al. (1979)
[24] Hosking (1998)
[25] Keelin and Powley (2011)

To provide context for the metalog family, we now show how previous researchers developed families of Type III distributions by making a coordinated set of choices across the columns of Table 1. We also cite strengths and limitations of these families.

The first family of continuous distributions was developed by Karl Pearson[7]. In Pearson's time, more and more people, Pearson among them, were recognizing that the normal distribution was not the universal "end-all" of continuous probability distributions. Specifically, it had become increasingly evident that many probabilistic data sets, survival data for example, exhibited skewness and kurtosis characteristics that the normal distribution could neither explain nor represent. So Pearson set out to develop a system of continuous distributions with variable skewness and kurtosis characteristics.

In terms of Table 1, he selected the normal as his base distribution, the differential equation that characterizes the normal density function as the form to modify, and parameter substitution as his modification method. Specifically, he substituted a quadratic function of the random variable X for the otherwise-constant variance (σ^2) in the denominator of this differential equation. This substitution effectively introduced variable skewness and kurtosis parameters into his system. Depending on the values of these parameters, Pearson's generalized-normal-density differential equation has a dozen solutions (Ord, 1972). These include the normal, beta, uniform, exponential, gamma, chi-square, F, student-t, and Cauchy distributions, among others.

As shown in Figure 1[26], Pearson's system was the first to collectively cover the entire accessible[27] space of combinations of third and fourth central moments. Zero-flexibility distributions show up as points in this diagram. These include the normal, uniform, logistic, Gumbel, and exponential. The flexibility range of triangular distributions is limited to a short line segment as shown. In contrast, bounded Pearson distributions (the beta) are sufficiently flexible to cover the entire accessible area above the Pearson-3 line[28]. Unbounded Pearson distributions (Pearson 4 and student-t) cover the area below the Pearson-5 line. Because they are symmetrical, t-distributions with various degrees of freedom (df) show up as points on the vertical axis. The area between the Pearson-3 and -5 lines and inclusive of them, is the flexibility range for semi-bounded Pearson distributions (gamma, chi square, F, inverse gamma, and inverse chi square).

So while there is at least one Pearson distribution available for each point in Figure 1, Pearson's system offers zero flexibility for choosing boundedness at a given point. For example, if a practitioner needs a semi-bounded distribution with a combination of skewness and kurtosis that is either above the Pearson 3 line or below the Pearson 5 line, there is no Pearson distribution that satisfies this need. Moreover, given a

[26] Figure 1 is the format traditionally used to display the flexibility of families of continuous distributions. See Ord (1972); Johnson (1949); Johnson, Kotz, and Balakrishnan (1994); and Tadikamalla and Johnson (1982), among others. The horizontal axis measures skewness in terms of the square of the standardized skewness while the vertical axis is standardized kurtosis. This standardization ensures that β_1 and β_2 are location- and scale-independent. See Section 3.4 below for precise definitions.

[27] "accessible" in this context refers to the area below the "upper limit for all distributions" line in Figure 1.

[28] "Pearson 3", "Pearson 4", etc. are synonymous with the terms "Pearson Type III", "Pearson Type IV", etc. as commonly used elsewhere in the literature.

particular combination of skewness and kurtosis, the Pearson system has zero flexibility to match higher-order moments. This follows from observing that Pearson introduced only two additional parameters into the normal distribution. Finally, Pearson's skewed unbounded distribution (the Pearson 4) is so difficult to use that now, a century later, researchers are still looking for practical ways to do.[29]

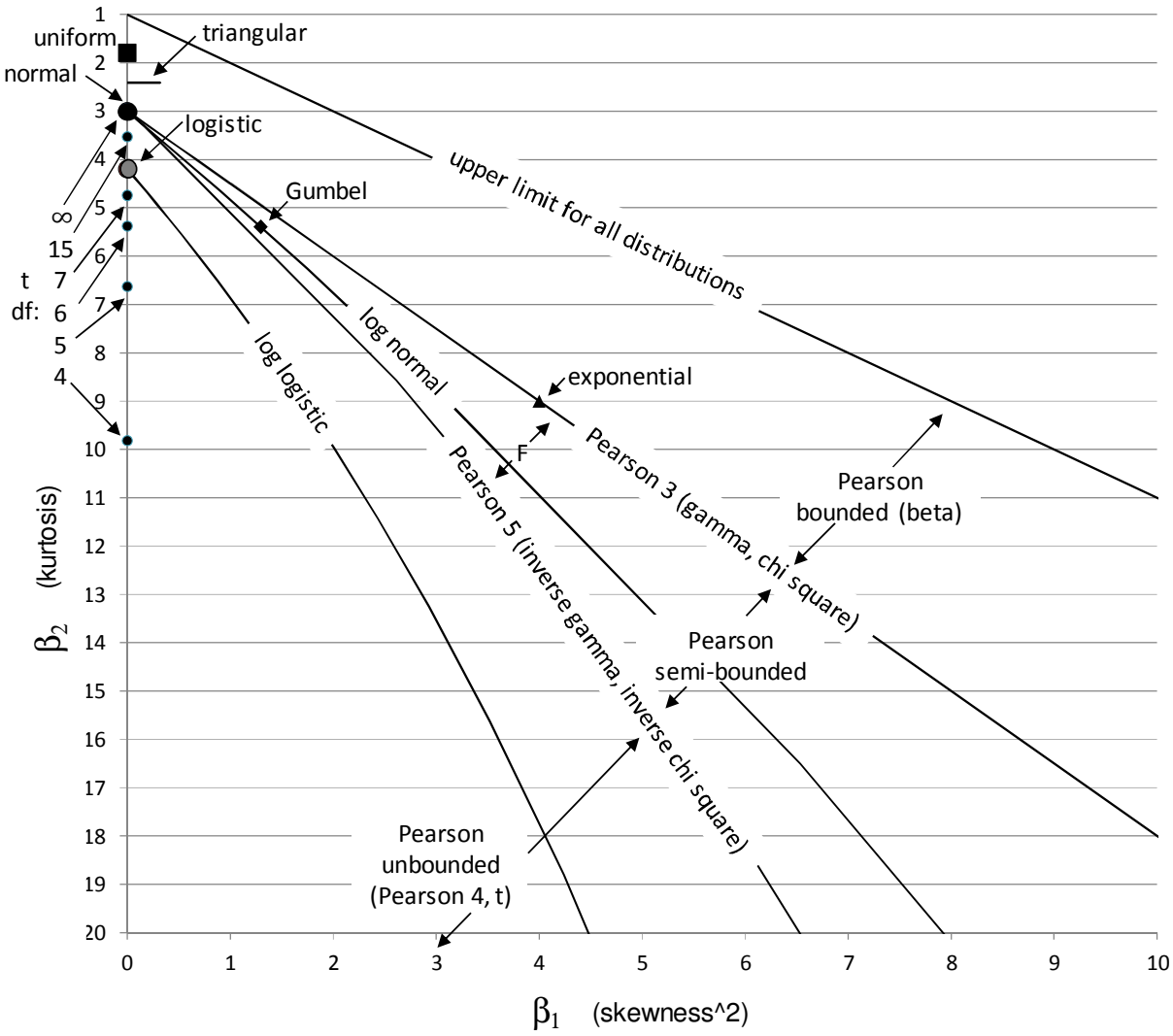
The Johnson (1949) and Tadikamalla and Johnson (1982) families of distributions have similar limitations. In terms of Table 1, Johnson (1949) selected the normal as his base distribution and transformed it using log, logit, and hyperbolic-sine transformations to produce his "S" family of distributions that, like Pearson's family, covers the entire accessible space of Figure 1. However, the only semi-bounded distribution within that family is the lognormal, which is limited to the lognormal line. All S distributions above that line are bounded, and all below it are unbounded. Tadikamalla and Johnson's (1982) "L" family is similar except that it takes the logistic in place of the normal as its base distribution. Semi-bounded distributions within the L family are limited to the log-logistic line, while all L distributions above it are bounded and below it are unbounded. Moreover, all distributions within both of these families have two or fewer shape parameters, implying that, like Pearson's family, these later families have no flexibility to match higher order moments.

Other noteworthy families of distributions are based on series expansion. Best known are the Edgeworth and Gram-Charlier series-expansions of the normal density function. While in theory these expansions have flexibility to match higher order moments, they tend to be limited to modest areas in β_1- β_2 plane by difficulty of parameter estimation and other practical considerations[21].

In contrast, as presented below, the metalog family provides a choice of boundedness for a wide range of combinations of skewness and kurtosis, flexibility to match higher order moments, and a straight-forward method for parameter estimation.

[29] Nagahara (1999). Cheng (2011).

Figure 1. Flexibility and Bounds Limitations of Pearson Distributions



3. The Unbounded Metalog Distribution

3.1 A Generalized Logistic Distribution

In terms of Table 1, our development of the metalog family starts with the logistic as a base distribution, introduces modifications to its quantile function, and uses three of the Table-1 modification methods -- parameter substitution, transformation, and series expansion.

Among its Type I interpretations, the logistic is the limiting distribution of the midrange sample (average of largest and smallest random samples) as sample size approaches infinity. We chose it as a base distribution, however, not because of its Type I interpretations, but because of its simple closed form

expressions for PDF, CDF, and quantile function; smoothness and symmetry; infinite differentiability in closed form; tail behavior that is "in between" the lighter-tailed beta and normal distributions and the heavier-tailed student t distributions, and its wide range of fully-investigated and well-known properties[30].

In terms of which form to modify, we have chosen the quantile function. Like Burr (1942), we prefer to start with a closed-form CDF or quantile function because, assuming differentiability, either one can be easily differentiated to find the PDF. In contrast, starting with the PDF often leads to a form that cannot be conveniently integrated to find the CDF or quantile function. We have chosen to modify the quantile function in particular because, in contrast to the CDF, it expresses the value x of a random variable as a function of probability y, thereby having the simplicity of being scale-independent of x and also guaranteeing ease of use in Monte Carlo simulation[31]. Moreover, the logistic quantile function in particular is linear in its parameters, and thus is already a QPD[32] prior to any modification. The logistic quantile function is

$$\mu + s \ln\left(\frac{y}{1-y}\right) \qquad \text{for } 0 < y < 1 \qquad (1)$$

where μ is the mean, median and mode, and s is proportional to standard deviation $\sigma = s\,\pi / \sqrt{3}$.

For modification method, we use a combination of parameter-substitution (following Pearson's lead) and series expansion, where a_i's are real constants.

$$\mu = a_1 + a_4(y - 0.5) + a_5(y-0.5)^2 + a_7(y-0.5)^3 + a_9(y-0.5)^4 + \ldots \qquad (2)$$
$$s = a_2 + a_3(y-0.5) + a_6(y-0.5)^2 + a_8(y-0.5)^3 + a_{10}(y-0.5)^4 + \ldots \qquad (3)$$

Substituting these series expansions for the parameters μ and s is easily interpreted. Note that the unmodified logistic distribution (1) is smooth, symmetric, unimodal, and unbounded. Imagine how its shape might change if the otherwise-constant μ and s were to change systematically. For example, given a systematically increasing standard-deviation parameter as one moves from left to right it, is natural to visualize that a right-skewed distribution would result. Alternatively, if the standard deviation parameter decreases when moving from left to right, one might visualize that a left skewed distribution would result. A range of such distributions is shown in Figure 2.

Similarly, one can envision that increasing μ from left to right would make a distribution fatter in the middle and therefore have lighter tails. And by systematically decreasing it as one moves from left to right, the

[30] Balakrishnan (1992)
[31] In Monte Carlo simulation via the inverse transform method, uniformly distributed random samples of y can simply be inserted into a closed-form quantile function to yield corresponding samples of x. This is trivially simple for closed-form quantile functions in contrast to the non-linear optimization or look-up tables typically required otherwise.
[32] Keelin and Powley (2011) provide definitions, moments derivation, linear parameter estimation, and other QPD properties that we further build upon in this paper.

distribution would become thinner (or spikier) in the middle with correspondingly heavier tails. A range of such distributions is shown in Figure 3.

Regarding (2) and (3), our choice of an unlimited number of series-expansion terms for modifying μ and s might be envisioned to provide nearly unlimited shape flexibility, the specifics of which we explore in Section 3.5.

Substituting (2) and (3) into the logistic quantile function (1) yields a generalized logistic quantile function, where n is the total number of series terms in use:

$$M_n(y) = a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3(y-0.5)\ln\left(\frac{y}{1-y}\right) + a_4(y-0.5) + \ldots \qquad (4)$$

In order for $M_n(y)$ to be a valid quantile function of a continuous distribution, it must be strictly increasing as a function of y. That is, $\frac{d}{dy}[M_n(y)] > 0$ for all $y \in (0,1)$. Applying this requirement to (4) leads to a feasibility condition on the constants a_i.

$$\frac{a_2}{y\,(1-y)} + a_3\left(\frac{y-0.5}{y(1-y)} + \ln\left(\frac{y}{1-y}\right)\right) + a_4 + \ldots > 0 \qquad \text{for all } y \in (0,1) \qquad (5)$$

For example, if a_i = 0 for all i ≥ 3, then a_2 must be positive in order for this condition to hold. Since (4) reduces to (1) in this case, the requirement that a_2 be positive is equivalent to requiring that the standard deviation be positive, which must be true for any probability distribution. (5) is the generalization of this requirement that corresponds to the generalized quantile function (4). Any set of constants $\boldsymbol{a} = (a_1, \ldots, a_n)$ that satisfies (5) we shall henceforth call *feasible*.

The order of the terms in (2), (3) and (4) is somewhat arbitrary and could be changed without loss of generality. We chose the order such that the first term would be the median; the second term would be a base shape (the logistic) that subsequent terms modify; the third term would primarily modify skewness; the fourth term would primarily modify kurtosis; and subsequent terms would alternate in further refining the s and μ parameters in (3) and (2) respectively. The third and fourth terms could be reversed if one wanted, for example, the third term to modify kurtosis and the fourth term to modify skewness. This would be useful in a situation where n = 3 and it is known from a priori considerations that a symmetric distribution with variable kurtosis properties is appropriate.

Since (4) is linear in the constants $\boldsymbol{a} = (a_1, \ldots, a_n)$, so can be the parameter estimation of these constants. Given a set of m distinct CDF data points $(\boldsymbol{x}, \boldsymbol{y})$ where $\boldsymbol{x} = (x_1, \ldots, x_m)$, $\boldsymbol{y} = (y_1, \ldots, y_m)$, the constants are related to the data by a set of linear equations:

Figure 2. Skewed distributions produced by systematically varying the standard-deviation parameter of a logistic distribution

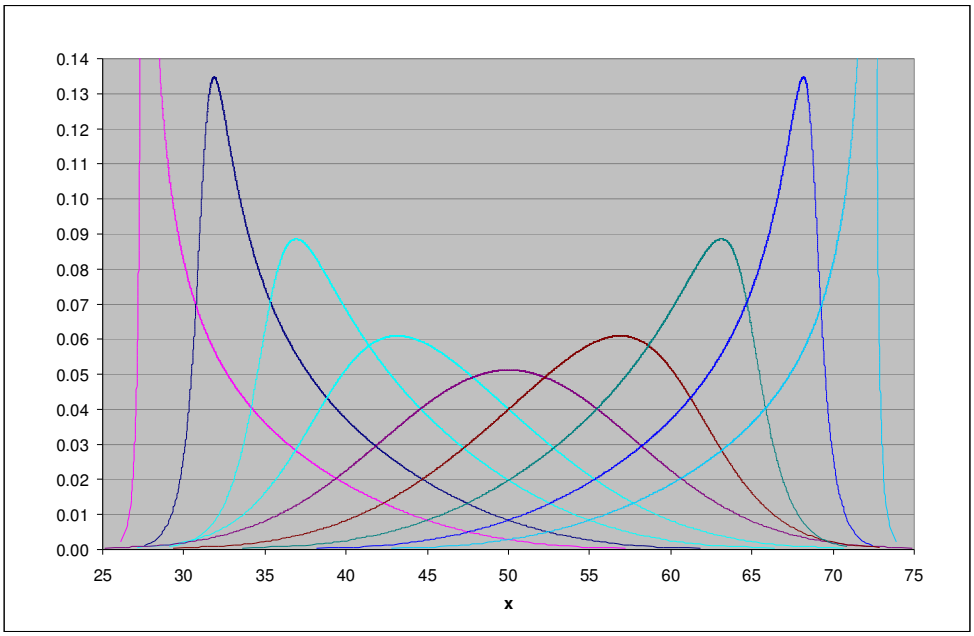
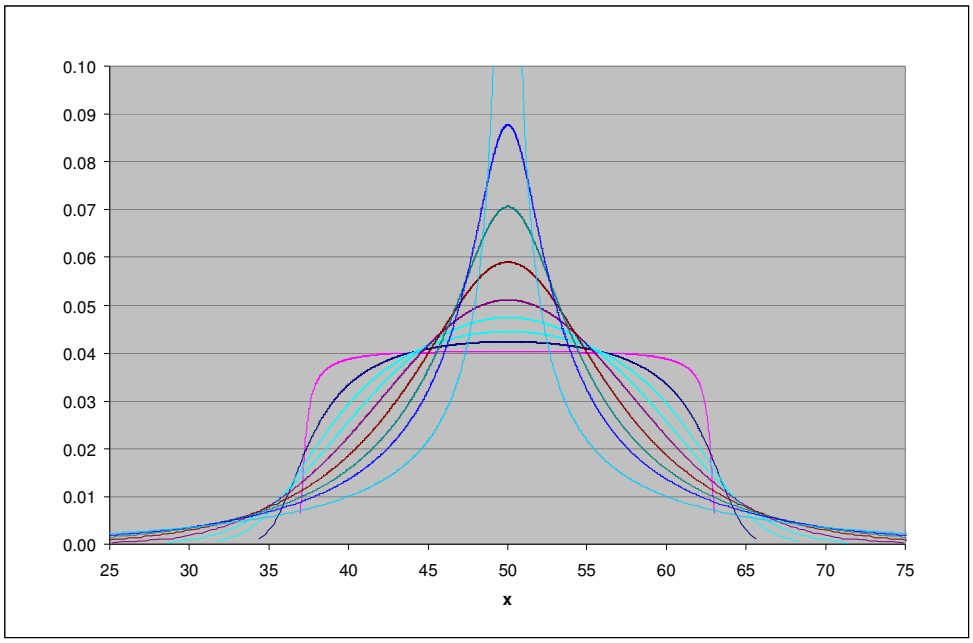


Figure 3. Symmetric distributions produced by systematically varying the mean parameter of a logistic distribution



$$x_1 = a_1 + a_2 \ln\left(\frac{y_1}{1-y_1}\right) + a_3(y_1 - 0.5)\, ln\left(\frac{y_1}{1-y_1}\right) + a_4(y_1 - 0.5) + \ldots$$
$$x_2 = a_1 + a_2 \ln\left(\frac{y_2}{1-y_2}\right) + a_3(y_2 - 0.5)\, ln\left(\frac{y_2}{1-y_2}\right) + a_4(y_2 - 0.5) + \ldots$$
$$\vdots$$
$$x_m = a_1 + a_2 \ln\left(\frac{y_m}{1-y_m}\right) + a_3(y_m - 0.5)\, ln\left(\frac{y_m}{1-y_m}\right) + a_4(y_m - 0.5) + \ldots$$

Equivalently, $\boldsymbol{x}$ = $\mathbf{Y}\boldsymbol{a}$, where $\boldsymbol{x}$ and $\boldsymbol{a}$ are column vectors and **Y** is the m x n matrix

$$\mathbf{Y} = \begin{bmatrix} 1 & \ln\left(\frac{y_1}{1-y_1}\right) & (y_1 - 0.5)\ln\left(\frac{y_1}{1-y_1}\right) & (y_1 - 0.5)\ldots \\ & & \vdots & \\ 1 & \ln\left(\frac{y_m}{1-y_m}\right) & (y_m - 0.5)\ln\left(\frac{y_m}{1-y_m}\right) & (y_m - 0.5)\ldots \end{bmatrix}$$

If m=n and **Y** is invertible, then $\boldsymbol{a}$ is uniquely determined by $\boldsymbol{a}$ = $\mathbf{Y}^{-1}\boldsymbol{x}$. If m≥n and **Y** has rank of at least n, then $\boldsymbol{a}$ is can be conveniently estimated using the familiar linear least squares equation $\boldsymbol{a}$ = $[\mathbf{Y}^T\mathbf{Y}]^{-1}\mathbf{Y}^T\boldsymbol{x}$, which reduces to $\boldsymbol{a}$ = $\mathbf{Y}^{-1}\boldsymbol{x}$ when m=n.[33] As such, this parameter estimation method can be interpreted as the maximum likelihood estimator if a Gaussian noise model is assumed. Note that it scales directly with n, the number of series terms in use. The size of the matrix to be inverted is n x n regardless of the number of data points m.

These observations give rise to the following definitions and formalizations.

3.2 Meta Distributions

We use the term "meta-distribution" to reference the class of a probability distributions that generalize a base distribution by substituting for one or more of its parameters an unlimited number of shape parameters. In doing so, the shape of a meta-distribution "goes beyond" the shape of the base distribution with considerable added flexibility. To be useful, a meta-distribution must also be associated with a practical method for estimating its parameters.

The generalized logistic distribution above is one specific example of a meta-distribution, which we formally define below as the "metalog" distribution. The term "metalog" is short for "meta-logistic".

Whenever the functional form of a base distribution is linear in its parameters, as is true for the quantile function of the logistic distribution, one can employ the same theoretical development method as above to create a new meta-distribution. For example, a meta-normal distribution can be developed by replacing (1) with the normal quantile function

[33] Keelin and Powley, 2011, also includes a weighted least squares formulation as an option for providing additional shape flexibility.

$\mu + \sigma\,\Phi^{-1}(y)$ where Φ is standard normal CDF and 0 < y < 1.

If one then substitutes series expansions like (2) and (3) for μ and σ, the "meta-normal" follows from the same subsequent development as in Section 3.1. Similarly, one could develop meta-Gumbel and meta-exponential distributions – since these too possess quantile functions that are linear in their parameters.

Such meta-distributions defined with respect to quantile functions, including the metalog, are generally Quantile Parameterized Distributions (QPDs) as defined by Keelin and Powley (2011). The Simple Q Normal distribution used for illustration in that paper is akin to the first several terms of the meta-normal.

Our initial explorations of the meta-normal distribution show that its flexibility properties are similar to those of the metalog, which we discuss below. For this paper, we have chosen to develop the metalog rather than the meta-normal because of its simple closed-form expression and greater ease of use compared to the meta-normal, which requires non-closed form look-up tables. For many practical applications, either would suffice.

3.3 The Metalog Distribution

We define the metalog distribution by formalizing the generalized logistic distribution of Section 3.1. Note that we have subsumed the linear-least-squares solution for $\boldsymbol{a}$ within the following definition in order to express the metalog, consistent with practical needs, as a function of its quantile parameters $(\boldsymbol{x}, \boldsymbol{y})$.

Definition 1. Metalog Quantile Function. The metalog quantile function with n terms is

$$M_n(y; \boldsymbol{x}, \boldsymbol{y}) = \tag{6}$$

$$a_1 + a_2 \ln\left(\frac{y}{1-y}\right) \qquad n = 2$$

$$a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3 (y - 0.5)\, ln\left(\frac{y}{1-y}\right) \qquad n = 3$$

$$a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3 (y - 0.5)\, ln\left(\frac{y}{1-y}\right) + a_4 (y - 0.5) \qquad n = 4$$

$$M_{n-1} + a_n (y - 0.5)^{\frac{n-1}{2}} \qquad \text{for odd n } \geq 5$$

$$M_{n-1} + a_n (y - 0.5)^{\frac{n}{2}-1} \ln\left(\frac{y}{1-y}\right) \qquad \text{for even n} \geq 6$$

where y is cumulative probability, $0 < y < 1$. $\boldsymbol{x} = (x_1, \ldots, x_m)$ and $\boldsymbol{y} = (y_1, \ldots, y_m)$ are column vectors of length $m >= n$ consisting of the x and y coordinates of CDF data, $0 < y_i < 1$ for each y_i, and at least n of the y_i's are distinct. The column vector of scaling constants $\boldsymbol{a} = (a_1, \ldots, a_n)$ is given by

$$\boldsymbol{a} = [\mathbf{Y}_n^T \mathbf{Y}_n]^{-1} \mathbf{Y}_n^T \boldsymbol{x}\,, \tag{7}$$ [34]

where $\mathbf{Y}_n^T$ is the transpose of $\mathbf{Y}_n$, and the m x n matrix $\mathbf{Y}_n$ is

[34] In the special case of m = n, (7) reduces to $\boldsymbol{a} = \mathbf{Y}_n^{-1} \boldsymbol{x}$.

$\mathbf{Y}_n =$ (8)

$$\begin{bmatrix} 1 & \ln\left(\frac{y_1}{1-y_1}\right) \\ & \vdots \\ 1 & \ln\left(\frac{y_m}{1-y_m}\right) \end{bmatrix} \qquad n = 2$$

$$\begin{bmatrix} 1 & \ln\left(\frac{y_1}{1-y_1}\right) & (y_1 - 0.5)\ln\left(\frac{y_1}{1-y_1}\right) \\ & & \vdots \\ 1 & \ln\left(\frac{y_m}{1-y_m}\right) & (y_m - 0.5)\ln\left(\frac{y_m}{1-y_m}\right) \end{bmatrix} \qquad n = 3$$

$$\begin{bmatrix} 1 & \ln\left(\frac{y_1}{1-y_1}\right) & (y_1 - 0.5)\ln\left(\frac{y_1}{1-y_1}\right) & (y_1 - 0.5) \\ & & \vdots & \\ 1 & \ln\left(\frac{y_m}{1-y_m}\right) & (y_m - 0.5)\ln\left(\frac{y_m}{1-y_m}\right) & (y_m - 0.5) \end{bmatrix} \qquad n = 4$$

$$\begin{bmatrix} \boldsymbol{Y}_{n-1} & \begin{matrix} (y_1 - 0.5)^{\frac{n-1}{2}} \\ \vdots \\ (y_m - 0.5)^{\frac{n-1}{2}} \end{matrix} \end{bmatrix} \qquad \text{for odd n } \geq 5$$

$$\begin{bmatrix} \boldsymbol{Y}_{n-1} & \begin{matrix} (y_1 - 0.5)^{\frac{n}{2}-1}\ln(\frac{y_1}{1-y_1}) \\ \vdots \\ (y_m - 0.5)^{\frac{n}{2}-1}\ln(\frac{y_m}{1-y_m}) \end{matrix} \end{bmatrix} \qquad \text{for even n } \geq 6.$$

Definition 2. Metalog PDF. Differentiating (6) with respect to y and inverting the result yields the metalog probability density function (PDF)[35]:

$m_n(y) =$ (9)

$$\frac{y\,(1-y)}{a_2} \qquad n = 2$$

$$\frac{1}{\left[\frac{a_2}{y\,(1-y)} + a_3\left(\frac{y-0.5}{y(1-y)} + \ln\left(\frac{y}{1-y}\right)\right)\right]} \qquad n = 3$$

$$\frac{1}{\left[\frac{a_2}{y\,(1-y)} + a_3\left(\frac{y-0.5}{y(1-y)} + \ln\left(\frac{y}{1-y}\right)\right) + a_4\,)\right]} \qquad n = 4$$

$$\left[(m_{n-1}(y))^{-1} + a_n\left(\frac{n-1}{2}\right)(\mathrm{y} - 0.5)^{\frac{n-3}{2}}\right]^{-1} \qquad \text{for odd } n \geq 5$$

$$\left[(m_{n-1}(y))^{-1} + a_n\left(\frac{(y-0.5)^{\frac{n}{2}-1}}{y(1-y)} + \left(\frac{n}{2} - 1\right)(\mathrm{y} - 0.5)^{\frac{n}{2}-2}\ln\left(\frac{y}{1-y}\right)\right)\right]^{-1} \qquad \text{for even } n \geq 6.$$

[35] For proof that this method yields the PDF, see Keelin and Powley (2011).

Note that the PDF $m_n(y)$ is expressed as a function of cumulative probability y. To plot this PDF as is customary, with values of random variable X on the horizontal axis, use $M_n(y)$ on the horizontal axis, $m_n(y)$ on the vertical axis, and vary $y \in (0,1)$ to produce the corresponding values on both axes.

For (6) and (9) to be a valid probability distribution, the matrix $\mathbf{Y}_n^T \mathbf{Y}_n$ must be invertible and the constants $\boldsymbol{a}$ must be feasible. Since (6) is a QPD, invertibility is guaranteed in all but pathological cases. [36]

Regarding feasibility, note that $m_n(y)$ is the reciprocal of the feasibility expression on the left hand side of (5). Since this expression is positive if and only if its reciprocal is positive, it follows that the feasibility condition (5) can be restated stated as

$$m_n(y) > 0 \qquad \text{for all } y \in (0,1) \qquad (10)$$

That is, $\boldsymbol{a}$ is feasible if and only if $m_n(y)$ is everywhere positive, and for any feasible $\boldsymbol{a}$, $m_n(y)$ is the probability density function that corresponds to (6).

Note that we have placed no constraints on the data $(\boldsymbol{x}, \boldsymbol{y})$. As such, there is no guarantee that any particular data set will lead to feasibility. Indeed, many data sets will not. If in doubt, feasibility must be checked according to (5) or (10). In practice, this means computing or plotting $m_n(y)$ and ensuring that the result is positive over all $y \in (0,1)$. If so, then $\boldsymbol{a}$ is feasible and $m_n(y)$ is a valid probability density function. Later in this paper, we provide closed-form constraints on the data $(\boldsymbol{x}, \boldsymbol{y})$ that ensure feasibility for the case of n = 3. Any data set $(\boldsymbol{x}, \boldsymbol{y})$ that yields feasible constants $\boldsymbol{a}$ we shall henceforth call feasible.

Given feasibility, certain special cases of these constants can be readily interpreted. In all cases, a_1 is the median as is evident from observing that all subsequent terms are zero when y = 0.5. Constants a_i for $i \geq 2$ determine shape. When $a_2 > 0$ and a_i = 0 for all $i \geq 3$, (6) is a logistic distribution exactly, with a_2 being directly proportional to the standard deviation -- as is obvious by comparison with (1) . When a_i = 0 for $i \geq 4$, a_3 primarily controls skewness. Increasing a_3 from zero results in an increasingly right-skewed distribution, while increasingly negative values of a_3 result in an increasingly left-skewed distribution. When $a_4 > 0$ and a_2 = 0, a_3 = 0, and a_i = 0 for $i \geq 5$, (6) reduces to a linear function of y, which means that it is a uniform distribution exactly. More generally, when $a_2 > 0$, $a_3 = 0$, and a_i = 0 for $i \geq 5$, a_4 determines kurtosis. Increasing a_4 from zero reduces kurtosis, resulting in a symmetric distribution that is fatter than a logistic in its mid-range with correspondingly lighter tails (e.g. more like a normal or symmetric beta distribution than a logistic.) Reducing a_4 from zero into increasingly negative values increases kurtosis, producing a distribution that is narrower than a logistic in its mid-range with correspondingly heavier tails (e.g. more like a student-t distribution with eight or fewer degrees of freedom).

Generally, the metalog, like the logistic, is unbounded. However, it is bounded in the special case that a_i = 0 for all $i \in \{2, 3, all\ even\ numbers \geq 6\}$. This is evident from observing that this is the particular set of

[36] "If such a (pathological) case were to occur, a small perturbation would solve the problem. In practical applications, we have never encountered a case where (the matrix that needs to be inverted) is singular." (Keelin and Powley, 2011, p. 212)

a_i's that multiplies the unbounded expression $ln\left(\frac{y}{1-y}\right)$ in (6). If all these a_i's are zero, then only bounded terms remain. Table 2 summarizes the above interpretations.

Table 2. Interpreting Metalog Constants

Constants	Interpretations
a_1	location, median
k * { a_i for all i >= 2}, where k > 0	k is a scale parameter
a_i for all i >= 2	shape
$a_2 > 0$, $a_i = 0$ for all i >= 3	M_n is a logistic distribution
$a_4 > 0$, $a_i = 0$ for all i ∈ {2,3, integers > 4}	M_n is a uniform distribution
$a_2 > 0$, $a_4 > 0$, and $a_i = 0$ for i ∈ {3, integers >= 5}. a_2 and a_4 need not sum to 1.	M_n is a mixture of logistic and uniform distributions, where a_1 is the mean and median of both. M_n is unimodal and symmetric. In Figures 1 and 4, M_n plots to the vertical line segment from (0, 1.8) to (0, 4.2).
$a_2 > 0$, $a_4 < 0$, $a_4 / a_2 >= -4$, and $a_i = 0$ for all i ∈ {3, integers >= 5}.	M_n is unimodal and symmetric. In Figures 1 and 4, M_n plots to the vertical line segment from (0, 4.2) to (0, 17.2).
$a_2 > 0$, $-1.67 < a_3 / a_2 < 1.67$, and $a_i = 0$ for all i >= 4	M_n is unimodal and right-skewed if $a_3 > 0$, unimodal and left-skewed if $a_3 < 0$. In Figure 4, M_n plots to the "3-term metalog" line segment from (0, 4.2) to (4.29, 8.58).
$a_i = 0$ for all $i \in \{2, 3, all\ even\ numbers \geq 6\}$	M_n is bounded
$a_i \neq 0$ for any $i \in \{2, 3, all\ even\ numbers \geq 6\}$	M_n is unbounded

3.4 Metalog Moments

We use traditional notation for moments of the n-term metalog distribution M_n:

$\mu'_{k,n}$	k^{th} moment	
$\mu_{k,n}$	k^{th} central moment	
σ_n	standard deviation = $\mu_{2,n}^{1/2}$	
β_1	square of standardized skewness = $(\mu_{3,n}/\sigma_n^3)^2$	(horizontal axis of Figures 1, 4, 6, 7)
β_2	standardized kurtosis = $\mu_{4,n}/\sigma_n^4$	(vertical axis of Figures 1, 4, 6, 7)

Since the metalog is a QPD, then as shown by Keelin and Powley (2011), it's k^{th} moment is given simply by the integral of the k^{th} power of the quantile function

$$\mu'_{k,n} = \int_{y=0}^{1} [M_n(y; \boldsymbol{x}, \boldsymbol{y})]^k \, dy$$

For n = 5 terms, this integral yields an explicit expression in closed form for the mean

$$\mu'_{1,5} = a_1 + \frac{a_3}{2} + \frac{a_5}{12} \quad \text{(mean)}$$

from which it follows that the k[th] central moment for the 5-term metalog is given by

$$\mu_{k,5} = \int_{y=0}^{1} [\, M_5(y; \boldsymbol{x}, \boldsymbol{y}) - (a_1 + \frac{a_3}{2} + \frac{a_5}{12}) \,]^k \ \mathrm{dy}$$

Though tedious to solve by hand, this integral can be shown to yield the following central moments of M_5 as closed-form polynomial expressions of the a_i's.

$$\mu_{2,5} = \frac{1}{3}\pi^2 a_2^2 + \left(\frac{1}{12} + \frac{\pi^2}{36}\right) a_3^2 + a_2 a_4 + \frac{a_4^2}{12} + \frac{a_3 a_5}{12} + \frac{a_5^2}{180} \quad \text{(variance)}$$

$$\mu_{3,5} = \pi^2 a_2^2 a_3 + \frac{1}{24}\pi^2 a_3^3 + \frac{1}{2} a_2 a_3 a_4 + \frac{1}{6}\pi^2 a_2 a_3 a_4 + \frac{1}{8} a_3 a_4^2 + a_2^2 a_5 + \frac{1}{24} a_3^2 a_5 + \frac{1}{180}\pi^2 a_3^2 a_5 + \frac{1}{4} a_2 a_4 a_5 + \frac{1}{60} a_4^2 a_5 + \frac{1}{120} a_3 a_5^2 + \frac{a_5^3}{3780} \quad \text{(skewness)}$$

$$\mu_{4,5} = \frac{7}{15}\pi^4 a_2^4 + \frac{3}{2}\pi^2 a_2^2 a_3^2 + \frac{7}{30}\pi^4 a_2^2 a_3^2 + \frac{a_3^4}{80} + \frac{1}{24}\pi^2 a_3^4 + \frac{7\pi^4 a_3^4}{1200} + 2\pi^2 a_2^3 a_4 + \frac{1}{2} a_2 a_3^2 a_4 + \frac{2}{3}\pi^2 a_2 a_3^2 a_4 + 2 a_2^2 a_4^2 + \frac{1}{6}\pi^2 a_2^2 a_4^2 + \frac{1}{8} a_3^2 a_4^2 + \frac{1}{40}\pi^2 a_3^2 a_4^2 + \frac{1}{3} a_2 a_4^3 + \frac{a_4^4}{80} + a_2^2 a_3 a_5 + \frac{1}{2}\pi^2 a_2^2 a_3 a_5 + \frac{1}{24} a_3^3 a_5 + \frac{1}{40}\pi^2 a_3^3 a_5 + \frac{5}{6} a_2 a_3 a_4 a_5 + \frac{2}{45}\pi^2 a_2 a_3 a_4 a_5 + \frac{3}{40} a_3 a_4^2 a_5 + \frac{1}{6} a_2^2 a_5^2 + \frac{1}{90}\pi^2 a_2^2 a_5^2 + \frac{1}{45} a_3^2 a_5^2 + \frac{11\pi^2 a_3^2 a_5^2}{7560} + \frac{1}{15} a_2 a_4 a_5^2 + \frac{11 a_4^2 a_5^2}{2520} + \frac{1}{420} a_3 a_5^3 + \frac{a_5^4}{15120} \quad \text{(kurtosis)}$$

As k and n increase, the number of polynomial terms increases, but within a pattern that continues with the k[th] central moment of the n-term metalog being a closed-form k[th] order polynomial of the a_i's. For example, the 9[th] central moment of the 5-term metalog $\mu_{9,5}$ has a closed form expression that consists of a 9[th] order polynomial in the a_i's with 297 terms. The 4[th] central moment of the 10-term metalog $\mu_{4,10}$ has 474 terms. These central moments are available from the author upon request. For all such central moments $\mu_{k,n}$, the central moments of $\mu_{k,j}$ where j < n can be calculated from $\mu_{k,n}$ simply by setting $a_i = 0$ for all $i > j$.

Given central moments in closed form, corresponding closed-form cumulants can also be calculated. Thus, the cumulants of the sum of independent (irrelevant[37]) metalog-distributed random variables can be expressed in closed form as the sum of the cumulants of these random variables.

[37] According to Howard and Abbas (2015)

3.5 Metalog Shape Flexibility

The shape flexibility of the metalog expands with the number of terms in use. As shown in Figure 4, for n=2, the metalog reduces to a logistic distribution and thus to the single point (0, 4.2). For n=3, metalog shape flexibility expands from a point to a line segment as shown. This line segment contains the full range of shapes shown in Figure 5.

For n=4, the metalog shape flexibility further expands to include all of area within "4-term metalog" envelope[38]. This area encompasses many common distributions including normal, uniform, triangular, logistic, exponential, Gumbel, and student t distributions with 4 or more degrees of freedom. Within the 4-term metalog envelope, the Pearson family offers unbounded distributions only below the Pearson-5 line. In contrast, the 4-term metalog offers unbounded distributions for a significant portion of the Pearson semi-bounded area and a significant portion (primarily unimodal) of the Pearson bounded area. Similarly, the 4-term metalog offers substantial additional unbounded flexibility compared to the areas below the lognormal and log-logistic lines, which are the upper limits respectively for unbounded Johnson S and L distributions.

There are certain relatively extreme skewness-kurtosis combinations that unbounded members of these other Type III families can represent that the 4-term metalog cannot. These include student t distributions with 3 or fewer degrees of freedom, and other distributions outside of the envelope.

However, with 5 or more terms, the metalog can represent multi-modal shapes and 5th or higher-order moments. In addition, the metalog's (β_1, β_2) coverage expands further. For example, with 10-terms, the metalog can reasonably represent student-t distributions with 3 or 2 degrees of freedom, the latter of which corresponds to (β_1, β_2) = (0,131). The metalog cannot effectively represent the Cauchy distribution (student t with one degree of freedom), all the moments of which are infinite.

[38] Since the metalog is parameterized by data rather than moments, we derived the metalog flexibility limits in Figure 4 by varying $\boldsymbol{a} = (a_1, \dots, a_n)$ over its feasible range and deriving the corresponding (β_1, β_2) feasible range from the moments expressions in Section 3.4. This process was enhanced by Keelin and Powley's (2011) proof that the set of feasible $\boldsymbol{a} = (a_1, \dots, a_n)$ is convex.

Figure 4. Shape flexibility for 2-4 term metalog distributions

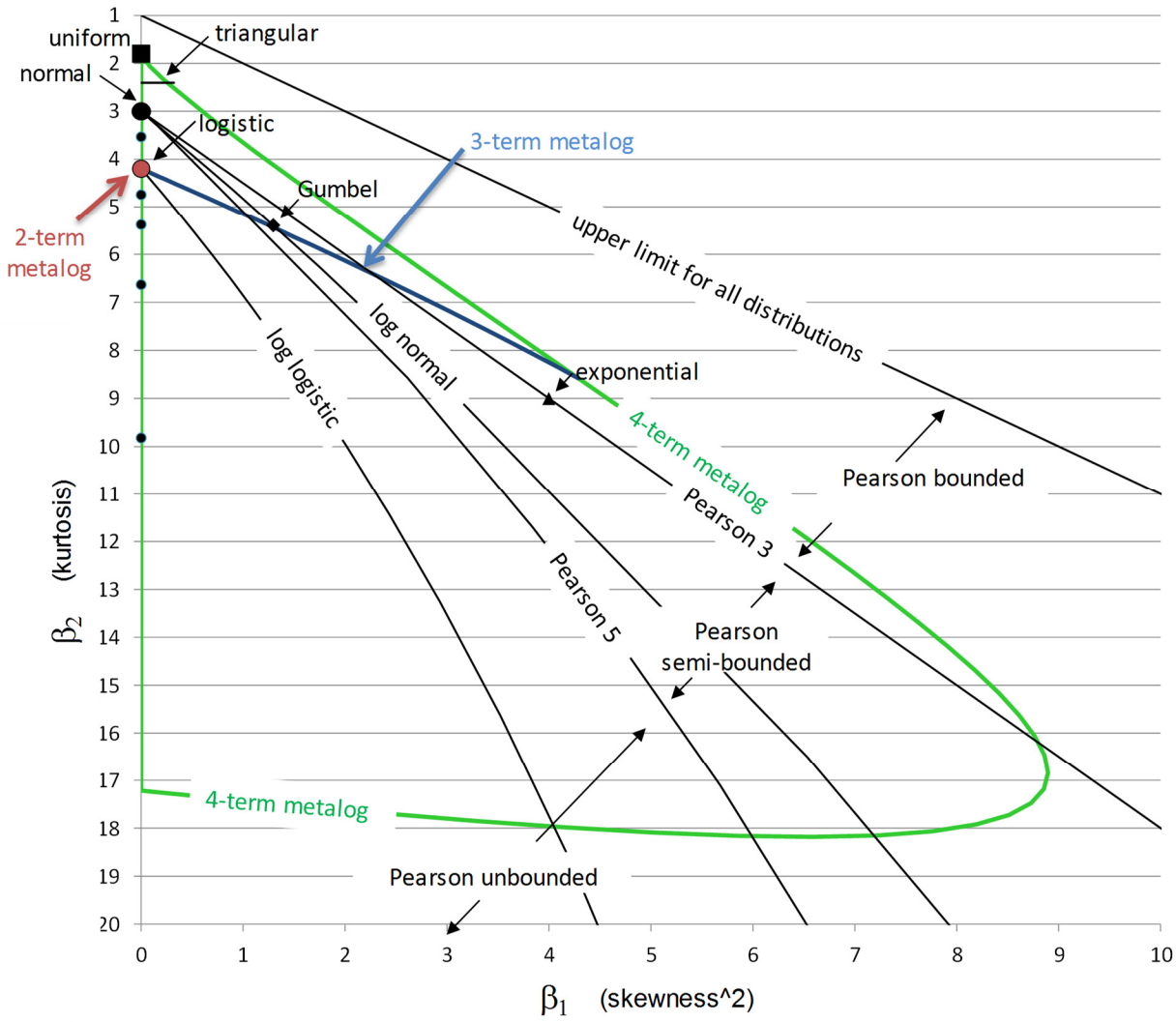
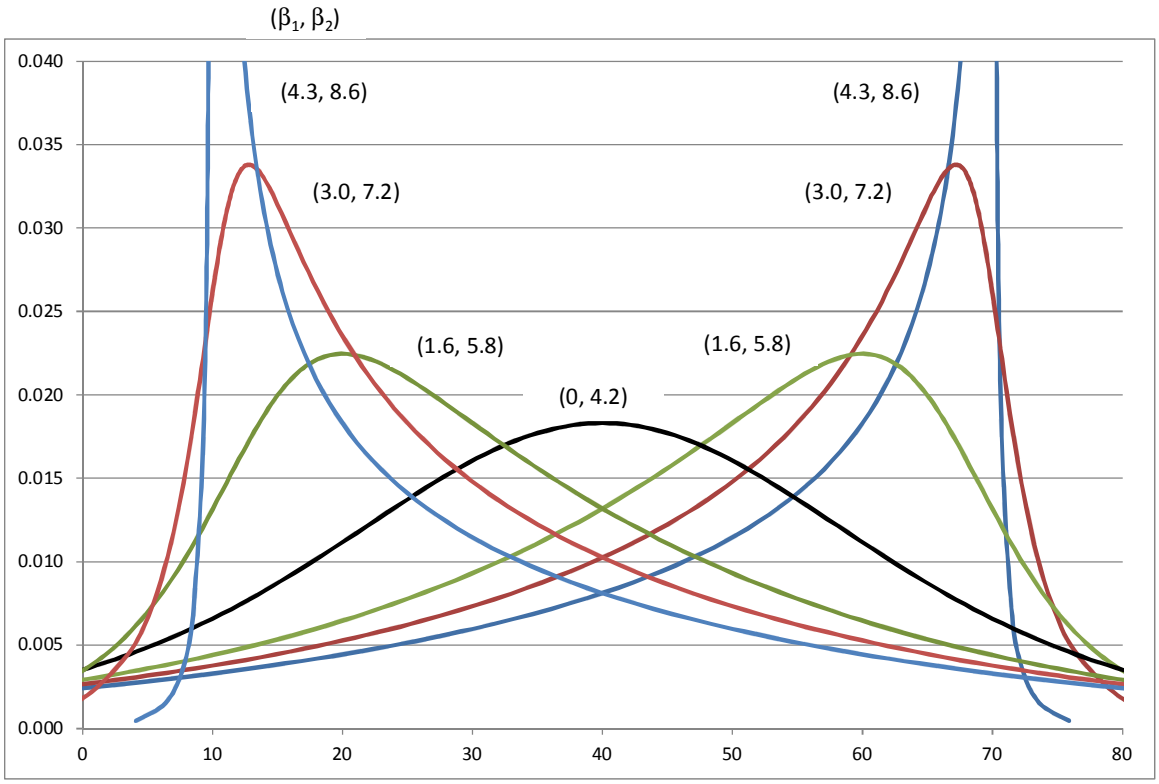


Figure 5. Range of shapes for 3-term metalog



4. Bounded and Semi-Bounded Metalogs

In many cases, one knows from a priori considerations that a distribution of interest is either semi-bounded or bounded. For example, uncertainties involving sizes, weights, and distances might naturally have a lower bound of zero and no definite upper bound. Uncertainties that involve fractions of a population are typically are bounded between zero and 100%. For such cases, it is desirable to have flexible, simple, easy-to-use distributions with bounds that can be specified a priori.

We now develop such distributions. In terms of Table 1, we use the metalog quantile function (6) as a base distribution and modify it using the method of transformation. This approach effectively propagates metalog shape flexibility forward into the domain of semi-bounded and bounded distributions. It also preserves the closed-form simplicity of (6) as well as the ease-of-use associated with linear quantile-parameterization.

Specifically, we use log and logit transformations, respectively, to produce semi-bounded and bounded members of the metalog family. These well-known transformations have been used previously for a similar purpose by Johnson (1949) and Tadikamalla and Johnson (1982).

4.1 Semi-Bounded Metalog (Log Metalog) Distribution

Suppose that $z = \ln(x - b_l)$ is metalog-distributed according to (6), where b_l is a known lower bound for x. Setting $\ln(x - b_l)$ equal to (6) and solving for x yields the log metalog quantile function with n terms:

$$M_n^{log}(y; \boldsymbol{x}, \boldsymbol{y}, b_l) = b_l + e^{M_n(y)} \qquad 0 < y < 1 \qquad (11)$$
$$= b_l \qquad y = 0$$

where $\boldsymbol{x} = (x_1, \ldots, x_m),\ m \geq n,$ each $x_i > b_l,\ \boldsymbol{y} = (y_1, \ldots, y_m),\ 0 < y_i < 1$ for each y_i, at least n of the y_i's are distinct, $\boldsymbol{z} = (\ln(x_1 - b_l), \ldots, \ln(x_m - b_l))$ is a column vector, $\mathbf{Y}_n$ is (8) and

$$\boldsymbol{a} = [\mathbf{Y}_n^T \mathbf{Y}_n]^{-1} \mathbf{Y}_n^T \boldsymbol{z} \qquad (12)$$

Differentiating (11) with respect to y and inverting the result yields the log metalog PDF:

$$m_n^{log}(y) = m_n(y)\, e^{-M_n(y)} \qquad 0 < y < 1 \qquad (13)$$
$$= 0 \qquad y = 0$$

where $m_n(y)$ is (9) and $M_n(y)$ is (6). The log metalog feasibility condition is $m_n^{log}(y) > 0$ for all $y \in (0, 1)$. Since the quantity $e^{-M_n(y)}$ is always positive, this condition is equivalent to (10). Some interpretations of log metalog constants are provided in Table 3.

Table 3. Interpreting Log Metalog Constants

Constants	Interpretations
b_l	location, lower bound
a_1	scale
a_i for all i >= 2	shape
$a_2 > 0$, $a_i = 0$ for all i >= 3	M_n^{log} is a log-logistic distribution, also known in economics as the Fisk distribution
$a_4 > 0$, $a_i = 0$ for all i ∈ {2,3, integers > 4}	M_n^{log} is a log-uniform distribution (i.e. ln(x- b_l) is uniformly distributed)

Similarly, for representations that have a known upper bound b_u and no lower bound, the transform $z = -\ln(b_u - x)$ yields a corresponding negative-log (nlog) quantile function and PDF

$$M_n^{nlog}(y; \boldsymbol{x}, \boldsymbol{y}, b_u) = b_u - e^{-M_n(y)} \qquad 0 < y < 1$$
$$= b_u \qquad y = 1$$

$$m_n^{nlog}(y) = m_n(y)\, e^{M_n(y)} \qquad 0 < y < 1$$
$$= 0 \qquad y = 1$$

where $\boldsymbol{x} = (x_1, \ldots, x_m),$ each $x_i < b_u,\ \boldsymbol{z} = (-\ln(b_u - x_1), \ldots, -\ln(b_u - x_m)),\ \boldsymbol{y} = (y_1, \ldots, y_m),\ 0 < y_i < 1$ for each y_i, and (12) determines $\boldsymbol{a}$.

4.2 Semi-Bounded Metalog Shape Flexibility

Like the metalog, log metalog shape flexibility expands with the number of terms in use. However, the addition of a lower-bound parameter b_l increases the shape dimensionality by one for each value of n. For example, the 2-term metalog is a point in the (β_1, β_2) plot and the 3-term metalog is a line segment. In contrast, the 2-term log metalog is a line in the (β_1, β_2) plot and the 3-term log metalog is an area. Effectively, this means that for any given number of terms n, the log metalog is more flexible than the metalog.

As shown in Figure 6, flexibility of the 2-term log metalog is simply that of the log-logistic line. Equivalently, this is the flexibility of the Fisk distribution in economics, which has been used in to represent survival data. The 3-term metalog increases this flexibility to cover the area between the upper and lower limits shown. The 4-term log metalog covers the expanded limits between the upper and lower 4-term lines shown. Unlike the "4-term metalog envelope" in Figure 4, these upper and lower limits extend indefinitely down and to the right corresponding to indefinitely larger values for β_1 and β_2 . From Figure 6, it is evident that this 4-term semi-bounded metalog offers far more flexibility than the Pearson semi-bounded distributions. In addition, it offers far more flexibility than the semi-bounded Johnson S and L distributions, which are limited to the log-normal and log-logistic lines respectively.

With 5 or more terms, the log metalog's (β_1, β_2) coverage expands further, providing a compelling option for representing a wide range of semi-bounded distributions. In addition, additional terms provide additional flexibility to match 5th and higher-order moments.

4.3 Bounded Metalog (Logit Metalog) Distribution

The logit metalog distribution is useful for representations that have known lower and upper bounds, b_l and b_u respectively, where $b_u > b_l$. The logit metalog distribution is the metalog transform that corresponds to $z = \text{logit}(x) = \ln\left(\frac{x-b_l}{b_u-x}\right)$ being metalog-distributed. Setting $\ln\left(\frac{x-b_l}{b_u-x}\right)$ equal to (6) and solving for x yields the logit metalog quantile function with *n* terms:

$$M_n^{logit}(y; \boldsymbol{x}, \boldsymbol{y}, b_l, b_u) = \frac{b_l + b_u e^{M_n(y)}}{1+e^{M_n(y)}} \qquad 0 < y < 1 \qquad (14)$$

$$= b_l \qquad y = 0$$
$$= b_u \qquad y = 1$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)$, $b_l < x_i < b_u$ for each x_i, $\boldsymbol{y} = (y_1, \ldots, y_m)$, $0 < y_i < 1$ for each y_i, $\boldsymbol{z} = \left(\ln\left(\frac{x_1-b_l}{b_u-x_1}\right), \ldots, \ln\left(\frac{x_m-b_l}{b_u-x_m}\right)\right)$ and (12) determines $\boldsymbol{a}$. Differentiating (14) with respect to y and inverting the result yields the logit metalog PDF:

$$m_n^{logit}(y) = m_n(y)\frac{\left(1+e^{M_n(y)}\right)^2}{(b_u - b_l)e^{M_n(y)}} \qquad 0 < y < 1 \qquad (17)$$
$$= 0 \qquad y = 0 \text{ or } y = 1$$

where $m_n(y)$ is (9) and $M_n(y)$ is (6). The logit metalog feasibility condition is $m_n^{logit}(y) > 0$ for all $y \in (0,1)$. Since the quantity $\frac{(1+e^{M_n(y)})^2}{(b_u - b_l)e^{M_n(y)}}$ is always positive, this condition is equivalent to (10). Some interpretations of logit metalog constants are provided in Table 4.

Figure 6. Shape flexibility for 2-4 term semi-bounded metalog distributions

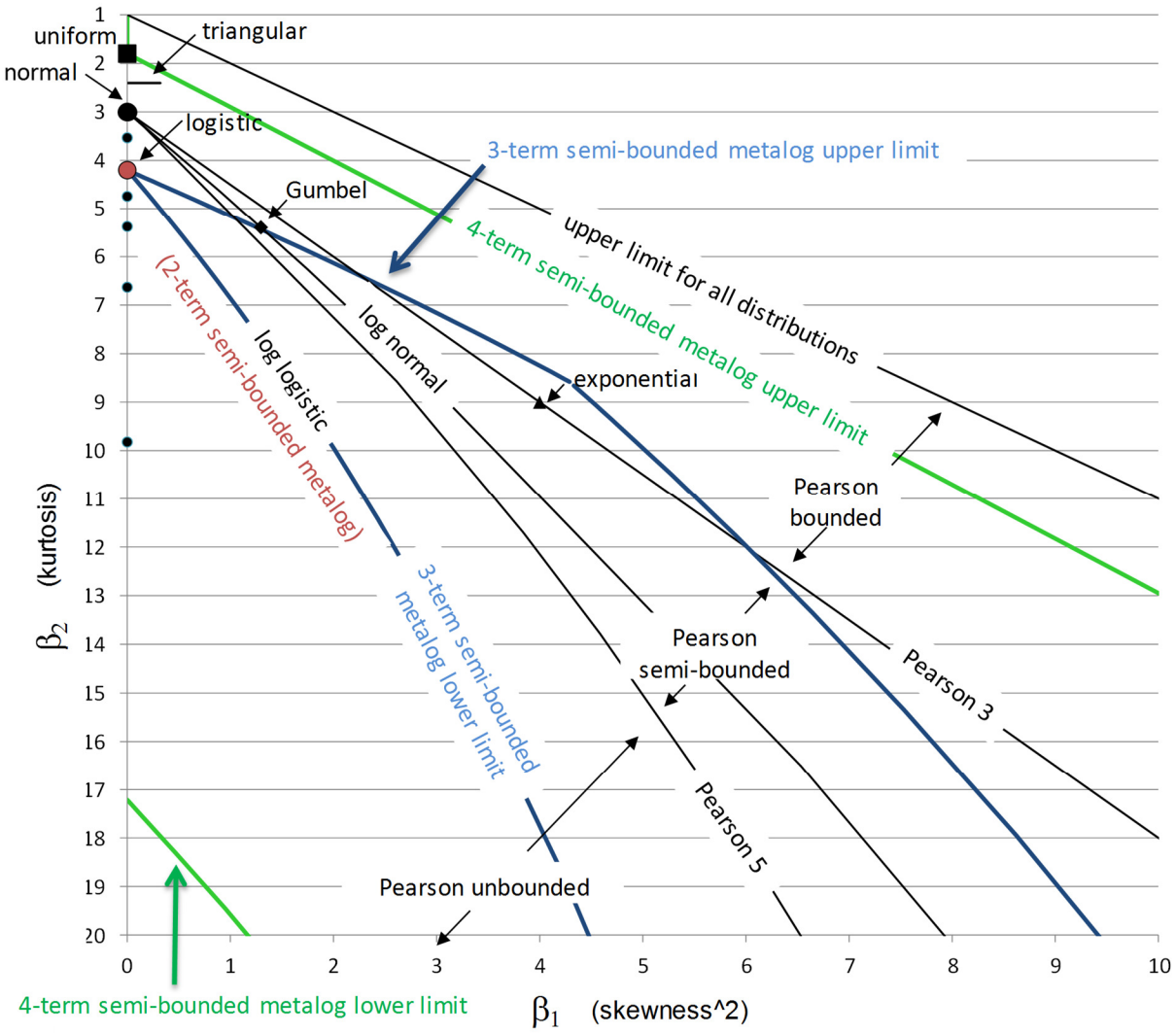


Table 4. Interpreting Logit Metalog Constants

b_l and b_u	location, lower and upper bound
b_u - b_l where $b_u > b_l$	scale
a_i for all i >=1	Shape
$a_2 > 0$, $a_i = 0$ for all i >= 3	M_n^b is a logit-logistic distribution[39], also known as the Tadikamalla and Johnson LB distribution[40]
$a_1 = 0$, $0 < a_2 < 1$, $a_i = 0$ for all i >= 3	M_n^b is a unimodal logit-logistic distribution
$a_1 = 0$, $a_2 = 1$, $a_i = 0$ for all i >= 3	M_n^b is a uniform distribution
$a_1 = 0$, $a_2 > 1$, $a_i = 0$ for all i >= 3	M_n^b is a U-shaped, symmetric logit-logistic distribution

4.4 Bounded Metalog Shape Flexibility

Like the metalog and log metalog, logit metalog shape flexibility expands with the number of terms in use. However, the presence of an upper-bound parameter in addition to a lower bound parameter increases the shape dimensionality for any value of n by two relative to the metalog and by one relative to the log metalog. For example, the 2-term metalog is a point in the (β_1, β_2) plot and the 3-term metalog is a line segment. In contrast, the 2-term logit metalog is a area in the (β_1, β_2) plot and the 3-term logit metalog is a broader area plus flexibility to match a 5th moment. Effectively, this means that for any given number of terms n, the logit metalog is more flexible than either the metalog or log metalog.

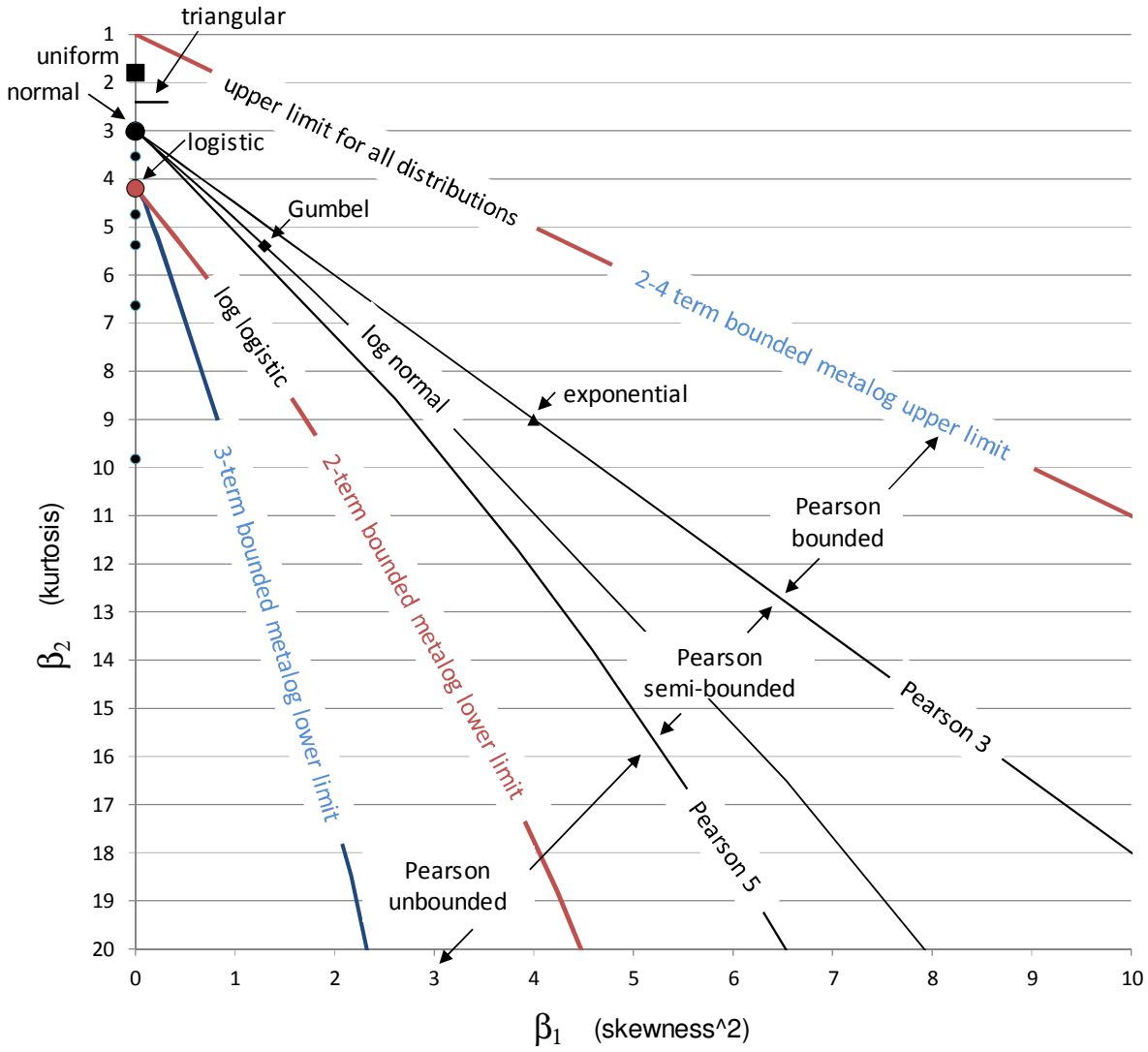
As shown in Table 4, the two-term logit metalog is also known as the Tadikamalla and Johnson LB distribution. As shown in Figure 7, the flexibility of this distribution is the entire accessible area down to and including the log-logistic line. The 3-term logit metalog increases this flexibility to cover the entire accessible area down to and including the "3-term bounded metalog lower limit". The 4-term logit metalog covers the entire accessible display area shown in Figure 7. It's lower limit includes the following points that are below that display area: (0,21), (0.1,29), (0.4,40), (1,52), (1.8,70),(3.05,95),(4.8,135), and (10.5,330).

Like the upper and lower limits in Figure 6, the upper and lower limits in Figure 7 extend indefinitely down and to the right. Thus, it is evident that this 4-term bounded metalog offers far more flexibility than the Pearson bounded distributions. In addition, it offers far more flexibility than the Johnson S and L bounded distributions, which are limited to the areas above the log-normal and log-logistic lines respectively.

With 5 or more terms, the logit metalog's (β_1, β_2) coverage expands further, providing a compelling option for representing a wide range of bounded distributions. In addition, additional terms provide additional flexibility to match 5th and higher-order moments.

[39] Wang and Rennolls (2005)
[40] Tadikamalla and Johnson1(1982). Balakrishnan (1992).

Figure 7. Shape flexibility for 2-3 term bounded metalog distributions



5. Metalog vs. Alternative Representations of Traditional Distributions

When the CDF data (x, y) is from a known source distribution, there would ordinarily be no need to represent this CDF data with a metalog. However, metalog representations of CDF data from previously-named source distributions may provide insight about the range of effectiveness and limitations of metalog representations and about metalog performance compared to alternatives. The alternatives we consider include a three-branch discrete approximation with 30%, 40%, and 30% probabilities assigned to the 10%, 50%, and 90% quantiles. They also include a range of QPDs, including the normal, the Simple Q Normal (Keelin and Powley, 2011), the logistic, and metalog distributions with various numbers of terms.

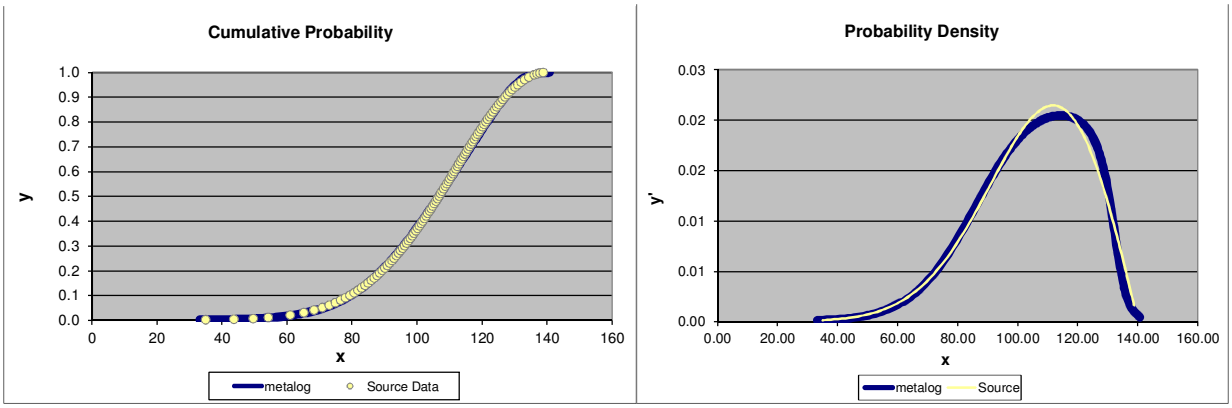
The figures and tables below compare these alternatives based on CDF data taken from a wide range of source distributions. In each case, we use 105 points from the CDF of the source distribution to

parameterize the metalog and alternative representations. These 105 points correspond to $y = \left(\frac{1}{1000}, \frac{3}{1000}, \frac{6}{1000}, \frac{10}{1000}, \frac{20}{1000}, \ldots, \frac{980}{1000}, \frac{990}{1000}, \frac{994}{1000}, \frac{997}{1000}, \frac{999}{1000}\right)$. For each y_i the corresponding x_i is the inverse CDF of the source distribution. For source distributions with known upper and/or lower bounds, we used the corresponding log or logit metalog.

5.1 Unbounded source distributions

For example, Figure 8 illustrates how M_5 approximates a particular extreme value distribution ($\mu = 100, \sigma = 20, eta = -0.5$). Visually, the metalog CDF is virtually indistinct from that of the extreme value source distribution, and the PDF's are very similar. To measure the accuracy of this approximation, we use the Kolmogorov-Smirnoff (K-S) distance (maximum cumulative-probability deviation on the CDFs). For convenience, we measure this as the maximum over the 105 points defined above. In this case, the K-S distance is 0.009, which means that the difference between the source-distribution and M_5 CDFs is everywhere less than 1% probability.

Figure 8. M_5 representation of an extreme value distribution



Source: extreme value ($\mu = 100, \sigma = 20, eta = -0.5$)

Table 5 shows this K-S distance for a range of unbounded source distributions and approximation methods. Based on the rankings at the bottom of this table, M_4 and M_5 are better that the other approximation methods and M_5 is best overall.

5. Accuracy of Various Approximations for Unbounded Source Distributions

Source Distribution	K-S Distance							
	Approximation Method							
	Discrete*	QPD						
	p: 30-40-30 q: 10-50-90	Normal	Simple Q-Normal	Logistic	Metalog M_2	Metalog M_3	Metalog M_4	Metalog M_5
Normal (μ=50, σ=15)	0.200	0.000	0.000	0.035	0.035	0.035	0.006	0.006
Logistic (μ=40, s=4.6)	0.200	0.032	0.009	0.000	0.000	0.000	0.000	0.000
Student t (df=6)	0.200	0.043	0.019	0.012	0.012	0.012	0.008	0.008
Extreme Value (μ=100,σ=20, ε=-0.5)	0.200	0.064	0.020	0.093	0.093	0.070	0.017	0.009
Extreme Value (μ=100,σ=20, ε=-0.2)	0.200	0.027	0.004	0.056	0.056	0.047	0.008	0.008
Extreme Value (μ=100,σ=20, ε=-0.025)	0.200	0.102	0.039	0.111	0.111	0.036	0.028	0.006
Maximum	0.200	0.102	0.039	0.111	0.111	0.070	0.028	0.009
Average	0.200	0.045	0.015	0.051	0.051	0.033	0.011	0.006
Rank based on lowest Maximum	8	5	3	6	6	4	2	1
Rank based on lowest Average	8	5	3	6	6	4	2	1

* Approximation is bounded, whereas source distribution is unbounded.

5.2 Semi-bounded source distributions

For a range of semi-bounded source distributions, we similarly compare the log metalog to other approximation methods. Table 6 shows the results. The log metalog approximations with 3-5 terms generally rank better than the other methods. In addition, the log metalog approximations have the same bounds as the source distributions, whereas the other approximation methods (discrete, normal, simple Q-normal, and logistic) do not.

5.3 Bounded source distributions

For a range of bounded source distributions, we similarly compare the logit metalog to other approximation methods. Table 7 shows the results. The logit metalog approximations with 3-5 terms generally rank better than the other methods. In addition, the logit metalog approximations have the same high and low bounds as the source distributions, whereas the other approximation methods do not.
While most of the source distributions in Table 3 are unimodal, note that Beta (α=0.8, β =0.9) and Beta (α =0.9, β =0.9) are bimodal (U-shaped) and are represented by the logit metalog with a high degree of accuracy (K-S distance <= 0.001). In addition, note that non-smooth PDFs (uniform and triangular) are well represented (K-S distance <= 0.003).

Table 6. Accuracy of Various Approximations for Semi-Bounded Source Distributions

	K-S Distance							
Source Distribution	Approximation Method							
	Discrete*** p: 30-40-30 q: 10-50-90	Normal**	Simple Q-Normal**	Logistic**	QPD Log Metalog M_2^{log}	M_3^{log}	M_4^{log}	M_5^{log}
Lognormal (μ=0, σ=0.5)	0.200	0.130	0.068	0.140	0.035	0.035	0.006	0.006
Lognormal (μ=0, σ=0.3)	0.200	0.078	0.026	0.092	0.035	0.035	0.006	0.006
Lognormal (μ=0, σ=0.15)	0.200	0.039	0.012	0.060	0.035	0.035	0.006	0.006
Weibull (λ=3, κ=3)	0.200	0.023	0.009	0.058	0.103	0.037	0.022	0.006
Weibull (λ=7, κ=7)	0.200	0.044	0.009	0.066	0.103	0.037	0.022	0.006
Gamma (κ=4, θ=2)	0.200	0.088	0.029	0.106	0.062	0.038	0.011	0.006
Gamma (κ=2, θ=2)	0.200	0.124	0.056	0.142	0.078	0.038	0.015	0.006
Inverse Gamma (α=3, β=1)	0.200	0.240	*	0.245	0.068	0.038	0.012	0.006
Inverse Gamma (α=5, β=0.5)	0.200	0.174	0.149	0.179	0.059	0.038	0.010	0.006
Exponential (λ=0.5)	0.200	0.174	0.130	0.193	0.103	0.037	0.022	0.006
Chi-Squared (df=3)	0.200	0.143	0.077	0.161	0.087	0.038	0.017	0.006
Chi-Squared (df=6)	0.200	0.101	0.038	0.119	0.068	0.038	0.012	0.006
Inverse Chi-Squared (df=3)	0.200	0.388	*	0.394	0.087	0.038	0.017	0.006
Inverse Chi-Squared (df=6)	0.200	0.240	*	0.245	0.068	0.038	0.012	0.006
F (df1=1, df2=1)	0.200	0.621	*	0.623	0.020	0.020	0.001	0.001
F (df1=15, df2=30)	0.200	0.106	0.045	0.118	0.039	0.033	0.007	0.006
Maximum	0.200	0.621	0.149	0.623	0.103	0.038	0.022	0.006
Average	0.200	0.170	0.054	0.184	0.066	0.036	0.013	0.006
Rank based on lowest Maximum	6	7	5	8	4	3	2	1
Rank based on lowest Average	8	6	4	7	5	3	2	1

* Approximation method does not yield a valid probability distribution.

** Approximation is unbounded whereas source distribution is semi-bounded.

***Approximation is bounded whereas source distribution is semi-bounded. In addition, low bound of approximation does not correspond to low bound of source distribution.

5.4 **Increased accuracy with higher-order terms**

Increasing the number of terms beyond 5 further increases accuracy. For example, Figure 9 shows how the 5-term metalog approximation of the extreme value distribution in Figure 8 becomes nearly exact when using 10 terms. Similar increased accuracy can be observed across the entire range of source distributions considered previously. Specifically, Table 8 shows how accuracy increases with each additional term as the number of terms increases from five to ten.

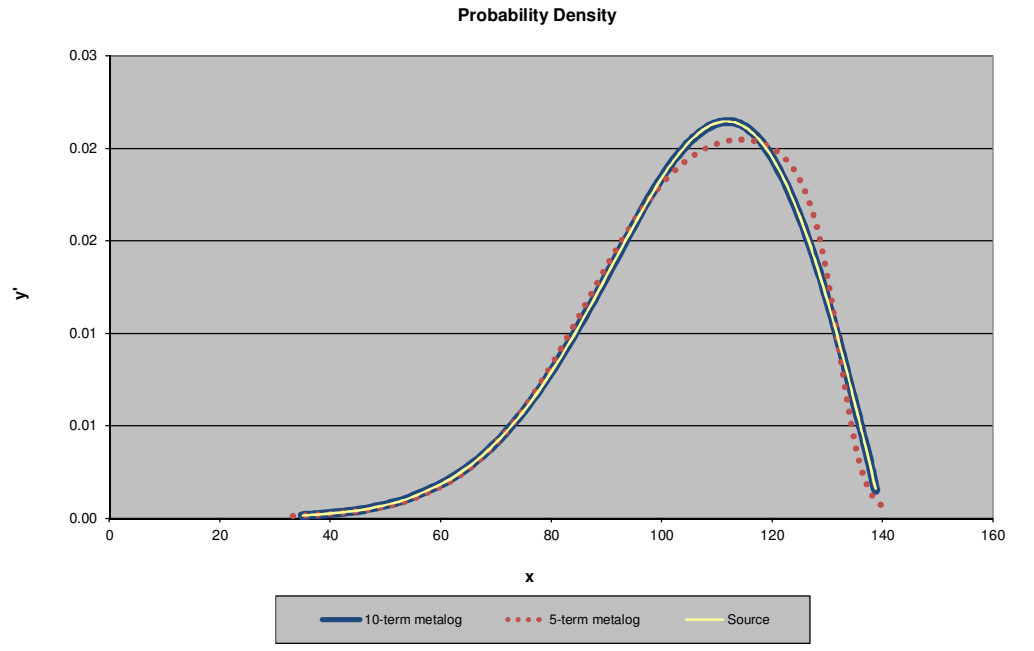
Based on Tables 5-8, we observe that the metalog distributions are capable of closely approximating a wide range of traditional distributions, and typically do so with greater accuracy than other practical alternatives.

Table 7. Accuracy of Various Approximations for Bounded Source Distributions

Source Distribution	K-S Distance							
	Approximation Method							
	Discrete*** p: 30-40-30 q: 10-50-90	Normal**	Simple Q-Normal**	Logistic**	QPD Logit Metalog M_2^{logit}	QPD Logit Metalog M_3^{logit}	QPD Logit Metalog M_4^{logit}	QPD Logit Metalog M_5^{logit}
Beta (α=3.5, β=3.5)	0.200	0.029	0.005	0.066	0.024	0.024	0.004	0.004
Beta (α=9, β=3.5)	0.200	0.054	0.012	0.084	0.044	0.031	0.008	0.005
Beta (α=0.8, β=0.9)	0.200	0.106	*	0.146	0.013	0.005	0.002	0.001
Beta (α=60, β=1.5)	0.200	0.138	0.069	0.157	0.085	0.037	0.017	0.006
Beta (α=1.2, β=1.2)	0.200	0.076	0.004	0.115	0.005	0.005	0.001	0.001
Beta (α=0.9, β=0.9)	0.200	0.095	*	0.135	0.003	0.003	0.000	0.000
Uniform (A=1, B=1)	0.200	0.088	0.000	0.127	0.000	0.000	0.000	0.000
Triangular (A=5,B=20, C=25)	0.200	0.077	0.016	0.112	0.033	0.019	0.009	0.003
Maximum	0.200	0.138	0.069	0.157	0.085	0.037	0.017	0.006
Average	0.200	0.083	0.018	0.118	0.026	0.016	0.005	0.002
Rank based on lowest Maximum	8	6	4	7	5	3	2	1
Rank based on lowest Average	8	6	4	7	5	3	2	1

* Approximation method does not yield a valid probability distribution.
** Approximation is unbounded whereas source distribution is bounded.
*** Bounds of approximation do not correspond to bounds of source distribution.

Figure 9. How Ten Terms Increases Accuracy Compared to Five



Source: extreme value ($\mu = 100, \sigma = 20, eta = -0.5$)

Table 8. How Additional Terms Increase Accuracy

	K-S Distance					
Unbounded Source Distributions	Metalog M_5	M_6	M_7	M_8	M_9	M_{10}
Normal (μ=50, σ=15)	0.006	0.002	0.001	0.001	0.001	0.000
Logistic (μ=40, s=4.6)	0.000	0.000	0.000	0.000	0.000	0.000
Student t (df=6)	0.008	0.004	0.002	0.002	0.002	0.001
Extreme Value (μ=100,σ=20, ε=-0.5)	0.009	0.002	0.001	0.001	0.001	0.000
Extreme Value (μ=100,σ=20, ε=-0.2)	0.008	0.003	0.002	0.001	0.001	0.000
Extreme Value (μ=100,σ=20, ε=-0.025)	0.006	0.005	0.005	0.001	0.000	0.000
Maximum	0.009	0.005	0.005	0.002	0.002	0.001
Average	0.006	0.003	0.002	0.001	0.001	0.000
Rank based on lowest Maximum	6	4	5	3	2	1
Rank based on lowest Average	6	5	4	3	2	1
Semi-Bounded Source Distributions	Log Metalog M_5^{log}	M_6^{log}	M_7^{log}	M_8^{log}	M_9^{log}	M_{10}^{log}
Lognormal (μ=0, σ=0.5)	0.006	0.002	0.001	0.001	0.001	0.000
Lognormal (μ=0, σ=0.3)	0.006	0.002	0.001	0.001	0.001	0.000
Lognormal (μ=0, σ=0.15)	0.006	0.002	0.001	0.001	0.001	0.000
Weibull (λ=3, κ=3)	0.006	0.004	0.003	0.001	0.000	0.000
Weibull (λ=7, κ=7)	0.006	0.004	0.003	0.001	0.000	0.000
Gamma (κ=4, θ=2)	0.006	0.002	0.002	0.001	0.000	0.000
Gamma (κ=2, θ=2)	0.006	0.003	0.002	0.001	0.000	0.000
Inverse Gamma (α=3, β=1)	0.006	0.002	0.002	0.001	0.000	0.000
Inverse Gamma (α=5, β=0.5)	0.006	0.002	0.001	0.001	0.000	0.000
Exponential (λ=0.5)	0.006	0.004	0.003	0.001	0.000	0.000
Chi-Squared (df=3)	0.006	0.003	0.003	0.001	0.000	0.000
Chi-Squared (df=6)	0.006	0.002	0.002	0.001	0.000	0.000
Inverse Chi-Squared (df=3)	0.006	0.003	0.003	0.001	0.000	0.000
Inverse Chi-Squared (df=6)	0.006	0.002	0.002	0.001	0.000	0.000
F (df1=1, df2=1)	0.001	0.000	0.000	0.000	0.000	0.000
F (df1=15, df2=30)	0.006	0.002	0.001	0.000	0.000	0.000
Maximum	0.006	0.004	0.003	0.001	0.001	0.000
Average	0.006	0.002	0.002	0.001	0.000	0.000
Rank based on lowest Maximum	6	5	4	3	2	1
Rank based on lowest Average	6	5	4	3	2	1

Table 8. How Additional Terms Increase Accuracy (continued)

	K-S Distance					
Bounded Source Distributions	Logit Metalog					
	M_5^{logit}	M_6^{logit}	M_7^{logit}	M_8^{logit}	M_9^{logit}	M_{10}^{logit}
Beta (α=3.5, β=3.5)	0.004	0.001	0.000	0.000	0.000	0.000
Beta (α=9, β=3.5)	0.005	0.002	0.001	0.000	0.000	0.000
Beta (α=0.8, β=0.9)	0.001	0.000	0.000	0.000	0.000	0.000
Beta (α=60, β=1.5)	0.006	0.003	0.003	0.001	0.000	0.000
Beta (α=1.2, β=1.2)	0.001	0.000	0.000	0.000	0.000	0.000
Beta (α=0.9, β=0.9)	0.000	0.000	0.000	0.000	0.000	0.000
Uniform (A=1, B=1)	0.000	0.000	0.000	0.000	0.000	0.000
Triangular (A=5,B=20, C=25)	0.003	0.003	0.002	0.002	0.001	0.001
Maximum	0.006	0.003	0.003	0.002	0.001	0.001
Average	0.002	0.001	0.001	0.000	0.000	0.000
Rank based on lowest Maximum	6	5	4	3	2	1
Rank based on lowest Average	6	5	4	3	2	1

6. Applications

We now turn to two applications. The first illustrates how the metalog system can produce insight about frequency data that would not be possible using traditional distributions. Thereby, it becomes evident that the metalog system offers a new vehicle for data and distribution research. The second example, decision

analysis, shows an actual decision that would that would have been made wrongly if the decision-makers had relied on 3-branch discrete approximations (as commonly used in decision analysis) instead of metalog-based continuous representations. As part of the decision-analysis application, we develop simplified expressions in terms of assessed quantiles for the metalog system for the special case of n=3.

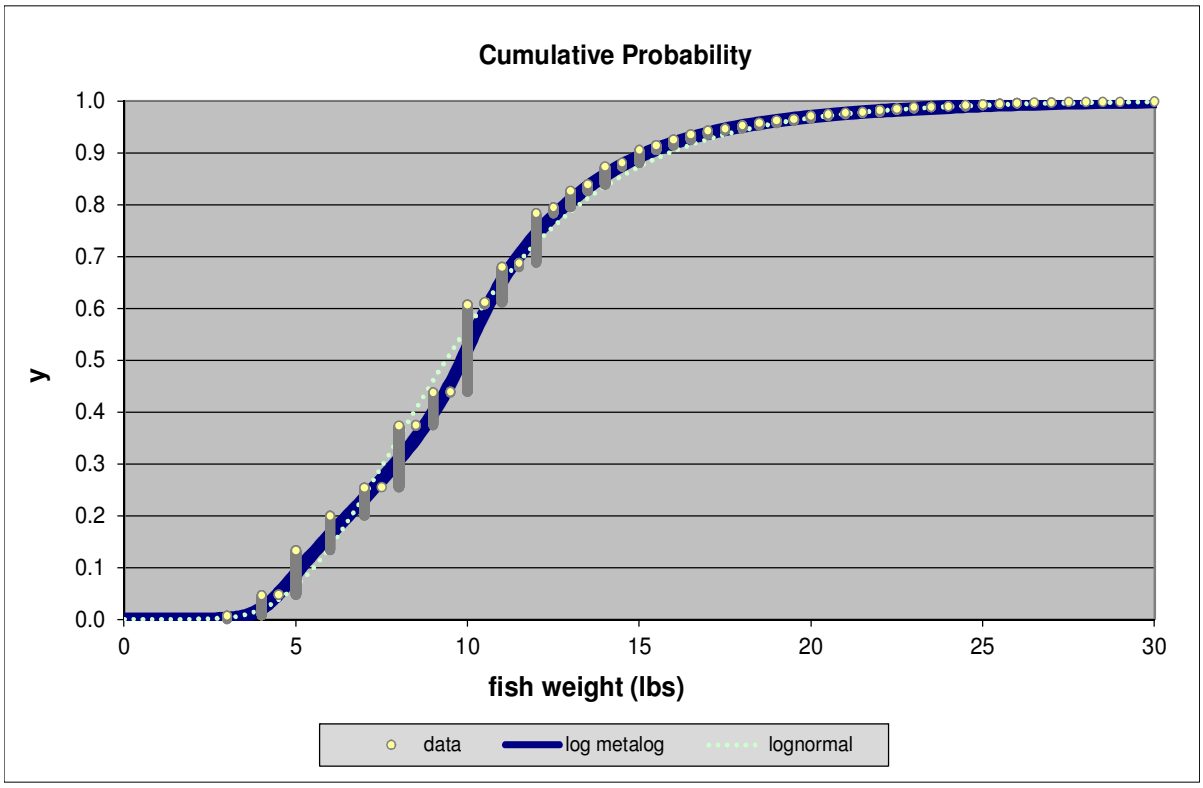
6.1 Application 1: Data and Distribution Research

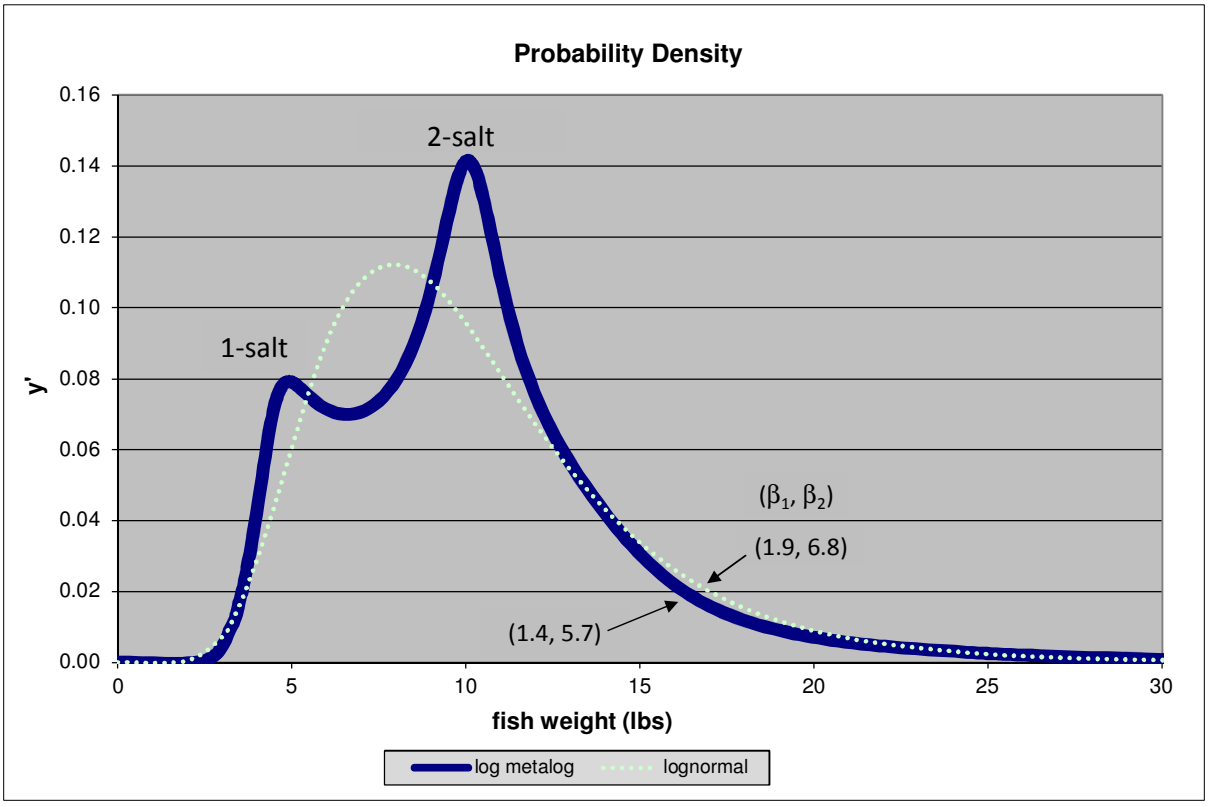
Our data and distributions research examples are based on real data from the disparate fields of fish biology and hydrology. Both show how metalog flexibility can aid data and distribution research by generating insight that might not otherwise emerge.

Fish Biology

Metalog distributions can mold themselves to the data with fewer unexamined shape constraints compared to other distribution systems such as the Pearson or Johnson. To illustrate, we consider the weight distribution of steelhead trout in the Babine River in northern British Columbia. A fly fishing lodge on that river has kept meticulous records of the weight of every fish landed by clients or staff over many

Figure 10. How the Metalog System Can Aid Data and Distribution Research





years. Specifically, during 2006-2010, 3,474 steelhead trout were caught and released. The recorded data for the weights of these fish in are plotted in Figure 10. This plot also shows two alternative distributions that could be used to represent that data. One is the lognormal, a shape which is representative of multiple other 1-2 shape parameter distributions (such as the log-logistic, gamma, log-Pearson 3, and F) that might typically be used in such a case. The other is a the ten-term log metalog M_{10}^{log} with b_l=0. Note that both CDF's appear to reasonably approximate the CDF data. However, the corresponding log metalog PDF shows a clear bi-modal pattern in the data, which the lognormal and other similar distributions lack the flexibility to represent.

The population of steelhead in the river when this lodge is open, during the fall of each year, consists of fish that are returning up river to spawn after having lived in salt water. Those fish returning from salt water to spawn for the first time are called "1-salt" fish. After spawning, these fish typically return to salt water, gain additional weight in ocean-rich feeding grounds, and then come back up the river some years later to spawn for a second time, becoming "2-salt" fish. A few very-large steelhead are "3-salt" or "4-salt" fish. One might reasonably consider that the modes of the log metalog PDF in Figure 10 may be indicative of the "1=salt" and "2-salt" fish populations respectively. Both the relative population sizes and weight differences between "1-salt" and "2-salt" fish are unsolved research questions in fish biology. It is apparent that the log metalog representation may shed some light on both questions. More broadly, by telling a more nuanced story about the data than alternative distributions, the metalog system may open new avenues for data and distribution research.

Hydrology

When a Type I interpretation of data is available, it is natural to use a corresponding Type I distribution. The advantage of this approach is that it constrains shape to that which is consistent with the Type I model, and relatively few data are needed to parameterize that model. A disadvantage is that the data may generated by a process that does not exactly correspond to the assumptions of the model, and therefore may have a legitimately different shape than the model predicts. If Type I shape constraints go unexamined, erroneous conclusions might result. In contrast, the flexibility of the metalog system allows "the data to speak for itself" with fewer unexamined shape constraints compared to other distribution families. Thus, it can be compared to various Type I representations of that same data.

In hydrology for example, it is common to compute maximum annual river stream flows and gauge heights for each year as the maximum of the 365 daily observations for that year. These measures are important for decisions such as bridge design, high-water mitigation, and river regulations. Even though there is typically autocorrelation among such observations, one might nevertheless try an extreme value distribution to represent such data given that this distribution has a simple Type I interpretation as the limiting distribution of a large number of such i.i.d samples. In Figure 11, we consider 95 years (1920-2014) of maximum annual gauge-height data as reported by the US Geological Survey for the Williamson River

(below Sprague River) near Chiloquin, Oregon[41]. Comparing log-metalog (with b_l=0) and extreme value representations of this data, we observe that the CDF's are similar. In addition, the extreme value PDF shows a shape that would commonly be attributed to this data, not only by the extreme value distribution but also by the lognormal, log Pearson 3, log-logistic, and other distributions commonly used to represent such data in hydrology. But by molding itself more closely to the data than possible with such other distributions, the log metalog PDF tells a somewhat different story: a lower mode and a "flat region" of equally likely values above that mode. To a knowledgeable expert, this deviation of the data from typically-assumed shapes might suggest systematic interpretations that would otherwise be masked by assuming a Type 1 model that may not appropriately apply.

6.2 Application 2: Decision Analysis

For decision analysis applications, it is common to use three assessed quantiles that correspond, for example, to probabilities of 0.1, 0.5, and 0.9. In this section, we show how the metalog system simplifies for such special cases. Then we apply it within an actual decision analysis.

SPT-parameterization of the metalog system

Definition 3 (Symmetric-percentile triplet)[42]. Metalog parameters $(\boldsymbol{x}, \boldsymbol{y})$ are a symmetric-percentile triplet (SPT) when they can be expressed as $\boldsymbol{y} = (\alpha, 0.5, 1\text{-}\alpha)$ and $\boldsymbol{x} = (q_\alpha, q_{0.5}, q_{1-\alpha})$ for some $\alpha \in (0, 0.5)$ and $q_\alpha < q_{0.5} < q_{1\text{-}\alpha}$.

This is often the case in decision analysis when, for example, 10-50-90 quantiles $(q_{0.1}, q_{0.5}, q_{0.9})$ are encoded from an expert and correspond to the 0.1, 0.5, and 0.9 probabilities on the CDF. We begin with the SPT-parameterized metalog distribution (SPT metalog) and then extend the results to develop the SPT-parameterized log- and logit-metalogs.

Proposition 1 (SPT unbounded metalog constants): Given that random variable X is metalog distributed and given a feasible SPT $\boldsymbol{x} = (q_\alpha, q_{0.5}, q_{1-\alpha})$, the metalog constants $\mathbf{a} = (a_1, a_2, a_3)$ can be expressed directly as

$$a_1 = q_{0.5}$$

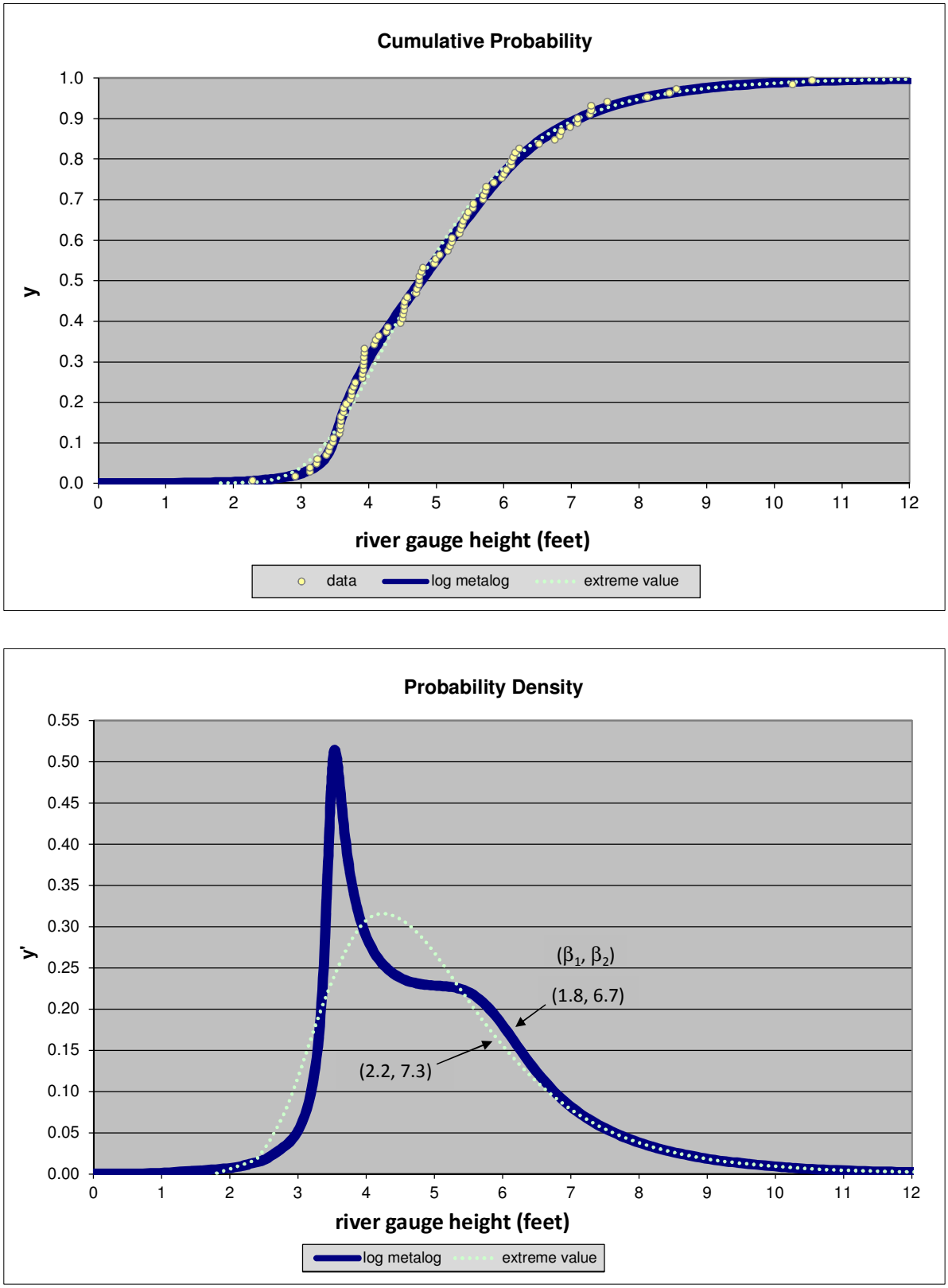
$$a_2 = \frac{1}{2}\left[\ln\left(\frac{1-\alpha}{\alpha}\right)\right]^{-1}(q_{1-\alpha} - q_\alpha) \quad (18)$$

$$a_3 = \left[(1-2\alpha)\ln\left(\frac{1-\alpha}{\alpha}\right)\right]^{-1}(1-2r)(q_{1-\alpha} - q_\alpha), \quad \text{where } r = \frac{q_{0.5} - q_\alpha}{q_{1-\alpha} - q_\alpha}. \quad (19)$$

[41] This data is available from United States Geological Survey website and from www.metalogdistributions.com.

[42] Hadlock and Bickel (2016) originally defined SPTs to parameterize Johnson Quantile-Parameterized distributions (J-QPDs). Our definition of SPT is the same, and we use it simplify parameterization of the metalog system for the special case of n=m=3. See Hadlock and Bickel for a J-QPD alternative to the SPT-parameterized metalog system presented in this section.

Figure 11. How the Metalog System Can Illuminate Unexamined Shape Constraints



Proof: For m=n=3, (7) reduces to $\boldsymbol{a} = \mathbf{Y}_3^{-1}\boldsymbol{x}$. Given that the 2nd element of $\mathbf{y}$ is 0.5, the 2nd row of $\mathbf{Y}_3$ reduces to (1, 0, 0). Inverting $\mathbf{Y}_3$ under this condition, post-multiplying by column vector $\boldsymbol{x}$, and substituting in the definition of r in (19) yields the above expressions. □

The importance of Proposition 1 is that the metalog constants $\mathbf{a}$ can be expressed directly in terms of the quantile assessments (q_α, $q_{0.5}$, $q_{1-\alpha}$). a_1 is simply the median, as is true for all metalog distributions. a_2 is proportional to the $q_{1-\alpha}$- q_α quantile range. For example when $\alpha = 0.1$, a_2 is 1/(2 ln 9) = 0.23 times the 10-90 quantile range. a_3, which controls skewness, is also proportional to the $q_{1-\alpha}$- q_α quantile range. We define r to mark the location of the median within this $q_{1-\alpha}$- q_α range. If the median is the mid-point of this range, then $= \frac{1}{2}$, $a_3 = 0$, and the 3-term metalog reduces to a symmetric logistic distribution. If the median is closer to q_α then $r < \frac{1}{2}$, a_3 is positive and the distribution is right-skewed accordingly. If the median is closer to $q_{1-\alpha}$, then a_3 is negative and the distribution is left-skewed.

There is a feasibility limit as to how much skewness and kurtosis can be represented with an SPT-parameterized metalog. Since there is a one-to-one correspondence between $\mathbf{a}$ and $\boldsymbol{x}$ in Proposition 1, this limit is just the "3-term-metalog" line segment shown in Figure 4, and the range of feasible shapes for SPT metalogs is as shown in Figure 5. Intuitively, the 3-term metalog, whether SPT-parameterized or more generally, can represent any shape from symmetric to roughly the skewness of the exponential distribution[43]. Quantitatively, this limit is determined in closed form for the SPT metalog by the following proposition.

Proposition 2 (SPT unbounded metalog feasibility): Any given SPT $\boldsymbol{x} = (q_\alpha, q_{0.5}, q_{1-\alpha})$ is a feasible parameterization of the metalog distribution if and only if

$$k_\alpha < r < 1 - k_\alpha, \quad \text{where } r = \frac{q_{0.5} - q_\alpha}{q_{1-\alpha} - q_\alpha} \text{ and } k_\alpha = \frac{1}{2}\left[1 - 1.66711\left(\frac{1}{2} - \alpha\right)\right] \qquad (20)$$

For 10-50-90 quantiles (α = 0.1), a close approximation to this expression is

$$\frac{1}{6} \leq r \leq \frac{5}{6} \qquad (21)$$

Proof: For n=3, the feasibility condition (5) reduces to

$$\frac{a_2}{y\,(1-y)} + a_3\left(\frac{y-0.5}{y(1-y)} + \ln\left(\frac{y}{1-y}\right)\right) > 0 \qquad \text{for all } y \in (0, 1) \qquad (22)$$

Consider three cases: $y \in (0, 0.5), y = 0.5, \text{and } y \in (0.5, 1.0)$. The feasibility condition is satisfied if and only if it is satisfied for all three cases. For $y = 0.5$, the second case, (22) reduces to $a_2 > 0$, which is obviously true by (18) since by definition $q_\alpha < q_{1-\alpha}$ and $0 < \alpha < 0.5$. Given $a_2 > 0$, then, for the first case, (22) can be expressed as

$$\frac{a_3}{a_2 C(y)} < 1 \text{ for all } y \in (0, 0.5), \quad \text{where } C(y) = -\left[y - 0.5 + y(1-y)ln\left(\frac{y}{1-y}\right)\right]^{-1}$$

[43] Note that in Figure 4 the exponential distribution with (β_1,β_2) = (4.0, 9.0) is very close to the end of the 3-term metalog line segment (4.3, 8.6). So conceptually we can use the exponential distribution as close proxy for the 3-term metalog skewness limit.

Since $C(y) > 0$ everywhere in this interval, the feasibility condition for this case becomes

$$\frac{a_3}{a_2} < k_0, \qquad \text{where } k_0 = \min_{y \in (0,0.5)} C(y) = 1.66711$$

Similarly, the feasibility condition for the third case is

$$\frac{a_3}{a_2} > k_1, \qquad \text{where } k_1 = \max_{y \in (0.5,1.0)} C(y) = -1.66711 = -k_0$$

Thus, (20) is satisfied if and only if $-k_0 < \frac{a_3}{a_2} < k_0$. Substituting (18) and (19) for a_2 and a_3 in this expression, defining $k_\alpha = \frac{1}{2}\left[1 - 1.66711\left(\frac{1}{2} - \alpha\right)\right]$, and simplifying yields (20). Applying (20) for $\alpha = 0.1$ yields $0.166578 \le r \le 0.833442$, of which (21) is a close approximation. □

The importance of Proposition 2 is that the feasibility of the SPT $\mathbf{x} = (q_\alpha, q_{0.5}, q_{1-\alpha})$ can readily be checked prior to any further calculations. If (20) or (21) is satisfied, then $\mathbf{x}$ is feasible, as it will always be over the range of shapes shown in Figure 5. If $\mathbf{x}$ is not feasible then adding one or more data points (n = m ≥ 4) would provide greater flexibility as shown in Figure 4.

Proposition 3 (SPT semi-bounded metalog): Given that $\ln(x - b_l)$ is metalog-distributed and given a feasible SPT $\mathbf{x} = (q_\alpha, q_{50}, q_{1-\alpha})$ with known lower bound b_l, the log metalog constants $\mathbf{a} = (a_1, a_2, a_3)$ can be expressed directly as

$$a_1 = \ln(\gamma_{0.5})$$

$$a_2 = \frac{1}{2}\left[\ln\left(\frac{1-\alpha}{\alpha}\right)\right]^{-1} \ln\left[\frac{\gamma_{1-\alpha}}{\gamma_\alpha}\right]$$

$$a_3 = \left[(1-2\alpha)\ln\left(\frac{1-\alpha}{\alpha}\right)\right]^{-1} \ln\left(\frac{\gamma_{1-\alpha}\gamma_\alpha}{\gamma_{0.5}{}^2}\right)$$

where $\gamma_\alpha = q_\alpha - b_l$, $\gamma_{0.5} = q_{0.5} - b_l$, $\gamma_{1-\alpha} = q_{1-\alpha} - b_l$, and k_α is as in (20). Moreover, $\mathbf{x}$ is feasible if and only if

$$b_l + \gamma_\alpha^{1-k_\alpha}\gamma_{1-\alpha}^{k_\alpha} < q_{0.5} < b_l + \gamma_\alpha^{k_\alpha}\gamma_{1-\alpha}^{1-k_\alpha}$$

Proof: For the log metalog, $\ln(x - b_l)$ is metalog distributed. In Proposition 1, substitute $\ln(\gamma_\alpha)$, $\ln(\gamma_{0.5})$, and $\ln(\gamma_{1-\alpha})$, for q_α , $q_{0.5}$, and $q_{1-\alpha}$ respectively. The above expressions for the log metalog constants follow from algebraic simplification. In (20), substitute $\ln(\gamma_\alpha)$, $\ln(q_{0.5} - b_l)$, and $\ln(\gamma_{1-\alpha})$, for q_α , $q_{0.5}$, and $q_{1-\alpha}$ respectively. The above expression for the log metalog feasibility condition follows from solving the resulting equation for $q_{0.5}$. □

The importance of Proposition 3 is that the log metalog constants and feasibility condition can be expressed directly in terms of the quantile assessments $(q_\alpha, q_{0.5}, q_{1-\alpha})$ and lower bound b_l. The feasible range of flexibility for the log metalog parameterized by an SPT is same as the "3-term log metalog" region in Figure 6, which also extends beyond the plot indefinitely down and to the right. Thus, the shape flexibility of an

SPT-parameterized log-metalog is inclusive of that of the SPT-parameterized metalog, but includes significant additional area as well.

Proposition 4 (SPT bounded metalog): Given that $\ln\left(\frac{x-b_l}{b_u-x}\right)$ is metalog-distributed and given a feasible SPT **x** = $(q_\alpha, q_{50}, q_{1-\alpha})$ with known lower upper and bounds b_l and b_u, the logit metalog constants **a** = (a_1, a_2, a_3) can be expressed directly as

$$a_1 = \ln(\gamma_{0.5})$$

$$a_2 = \frac{1}{2}\left[\ln\left(\frac{1-\alpha}{\alpha}\right)\right]^{-1} \ln\left[\frac{\gamma_{1-\alpha}}{\gamma_\alpha}\right]$$

$$a_3 = \left[(1-2\alpha)\ln\left(\frac{1-\alpha}{\alpha}\right)\right]^{-1} \ln\left(\frac{\gamma_{1-\alpha}\gamma_\alpha}{\gamma_{0.5}{}^2}\right)$$

where $\gamma_\alpha = \frac{q_\alpha - b_l}{b_u - q_\alpha}$, $\gamma_{0.5} = \frac{q_{0.5} - b_l}{b_u - q_{0.5}}$, $\gamma_{1-\alpha} = \frac{q_{1-\alpha} - b_l}{b_u - q_{1-\alpha}}$, and k_α is as in (20). Moreover, **x** is feasible if and only if

$$\left[b_l + b_u\gamma_\alpha^{1-k_\alpha}\gamma_{1-\alpha}^{k_\alpha}\right]\left[1+\gamma_\alpha^{1-k_\alpha}\gamma_{1-\alpha}^{k_\alpha}\right]^{-1} < q_{0.5} < \left[b_l + b_u\gamma_\alpha^{k_\alpha}\gamma_{1-\alpha}^{1-k_\alpha}\right]\left[1+\gamma_\alpha^{k_\alpha}\gamma_{1-\alpha}^{1-k_\alpha}\right]^{-1}$$

Proof: For the logit metalog, $z = \ln\left(\frac{x-b_l}{b_u-x}\right)$ is metalog distributed. In Proposition 1, substitute $\ln(\gamma_\alpha)$, $\ln(\gamma_{50})$, and $\ln(\gamma_{1-\alpha})$ for q_α , $q_{0.5}$, and $q_{1-\alpha}$ respectively. The resulting equations are identical those in Proposition 3, so the logit metalog constants follow from the same algebraic simplification as in the proof of Proposition 3. To prove the logit metalog feasibility condition, substitute $\ln(\gamma_\alpha)$, $\ln(\frac{q_{0.5}-b_l}{b_u-q_{0.5}})$, and $\ln(\gamma_{1-\alpha})$ for q_α , $q_{0.5}$, and $q_{1-\alpha}$ in (20). The above logit metalog feasibility condition follows from solving the resulting expression for $q_{0.5}$. □

The importance of Proposition 4 is that the logit metalog constants and feasibility condition can be expressed directly in terms of the quantile assessments (q_α,$q_{0.5}$, $q_{1-\alpha}$) and lower and upper bounds b_l and b_u. The feasible range of flexibility for the SPT-parameterized logit metalog is same as the "3-term logit metalog" region in Figure 7, which also extends beyond the plot indefinitely down and to the right. Comparing the feasible "3-term" ranges in Figures 4, 6, and 7, it is apparent that the shape flexibility of the SPT-parameterized logit metalog is far greater than that of the SPT-parameterized metalog and log metalog distributions.

Bidding decision example

As one illustration of the value of SPT-parameterization of the metalog family of distributions, we offer an example of an actual decision analysis in which a wrong decision would have been made if the decision-makers had relied on a commonly-used 3-branch discrete representation of continuous uncertainties instead of a metalog-system continuous representation.

The decision was how much to bid for a portfolio of 259 troubled real-estate assets, which a financial institution had offered for sale via public auction. These assets were of different geographies, sizes, and

types including single-family, multi-family, commercial, and land. To varying degrees, the value of each asset involved considerable uncertainty and complexity concerning current and future real estate values, occupancy and leases, potential tenant negotiations, local regulations, and, in some cases, bankruptcy or other litigation.

To help determine how much to bid for the portfolio and how one might monetize its various assets, a potential bidder wished to see a probability distribution over the value of the portfolio, which would be the sum of the values of the 259 individual assets. So he engaged a team of experts to assess the value of each asset. Their assignment included visiting each property, discussing comparables with local real estate agents and other knowledgeable parties, and undertaking independent research concerning any issues that would affect that asset's current or future value. As an overall summary of their conclusions, the potential bidder requested a probabilistic range of low, medium, and high scenarios for each asset. For each scenario, the team assessed a projected cash flow over time and translated this cash flow into a net present value (NPV). The low scenario was defined as the NPV such that, from the experts' perspective, there was a 10% chance that the ultimate realized NPV would be lower than this amount. The high NPV was defined such that there was a 90% chance that the ultimate realized NPV would be lower than this amount and a 10% chance that it would exceed it. The median scenario was defined such is it was equally likely that the actual realized NPV would be higher or lower than this amount. The expert's analyses and assessments resulted in the range of values for each asset as shown in Table 9.

Table 9. Range of Uncertainty in Asset Value ($ '000's)

Asset	10% probability that realized value is less than …	50% probability that realized value is less than …	90% probability that realized value is less than …
1	$ 18,150	$ 21,133	$ 22,625
2	$ 10,465	$ 11,362	$ 12,408
3	$ 15,781	$ 16,908	$ 18,260
4	$ 4,234	$ 4,422	$ 4,610
5	$ 2,629	$ 2,979	$ 3,295
6	$ 13,945	$ 14,875	$ 16,176
⋮	⋮	⋮	⋮
259	$ 3,500	$ 4,000	$ 4,500
Total		$ 185,348	

It was apparent from this data that some assets were worth far more than others. Some asset distributions were narrow while others were wide. Some asset distributions were symmetric, while others were skewed-left and still others were skewed-right. In addition, while some of the asset-level uncertainty was probabilistically independent of (irrelevant to[44]) that of other assets, the team judged that there was a degree of positive correlation among these assets due to their common dependence on the future economy and, in particular, on the future health of global and local real estate markets.

[44] Howard and Abbas (2015).

To calculate a probability distribution over the value of the portfolio, the team used a modified form of Monte Carlo simulation in which they had induced what they believed was an appropriate level of positive correlation across assets. For many of the assets, the team judged the correlation coefficient with the market to be about 80%. For other assets, especially those in litigation, the team believed the correlation with market to be negligible. The value of the portfolio for each simulation trial was the sum of the (appropriately-correlated-with-market) simulated values for each asset for that trial.

When performing the simulation, the team initially performed a discrete simulation -- using only the discrete values in Table 9 for each asset. They followed a commonly-used decision-analysis approach of assigning probabilities of 30%, 40% and 30% respectively to the low, median, and high discrete scenarios for each asset (see Bickel, Lake, and Lehman, 2011) and summing the results across assets for each simulation trial. Doing this for 1,000 simulation trials resulted in the CDF data labeled "Discrete Simulation Data" in Figure 12. To gain further insight into this distribution, they calculated the corresponding log metalog distribution M_5 parameterized by this data and plotted the results. These results are labeled "Discrete Simulation Metalog" in Figures 12 and 13.

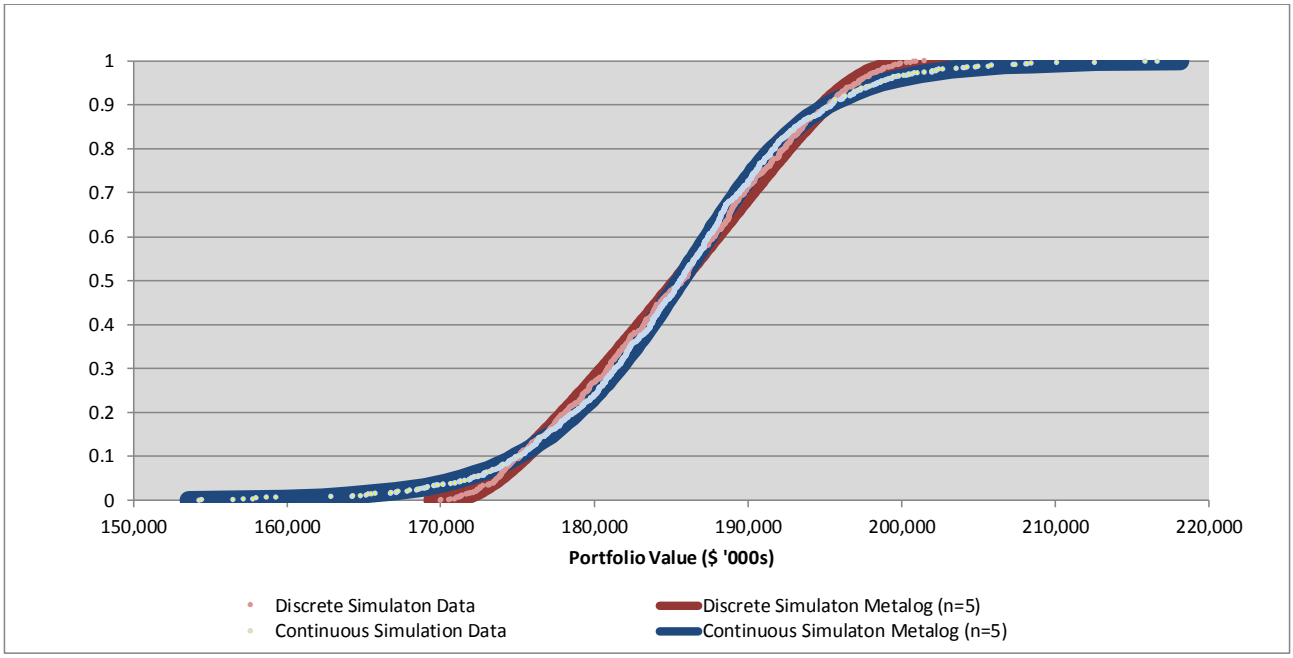
Considering Figure 13, the team felt that the discrete-simulation tails were too narrow – even though this simulation had taken correlation into account. While the median portfolio value of about $185,000,000 seemed to make sense, the near-zero probability that realized portfolio value would be less than $170,000,000 did not. They felt based on their experience that the low end of the distribution should be lower. Similarly, they felt that the high end should be higher.

The team then ran the same simulation using metalog (continuous) representations of the data in Table 9. Using the SPT assessments in Table 9, the team parameterized the 3-term metalog accordingly for each asset. Figure 14 shows the result of this calculation for the Asset 1 in Table 9. When reviewing such asset-level distributions prior to simulation, they noted that the 10-50-90 quantiles for each distribution corresponded exactly to the 10-50-90 value assessments in Table 9, and that these distributions appeared to have appropriate right- or left-skewness. They further noted that the low, median and high values appeared reasonable. Intuitively, they felt these asset-level continuous distributions were a more accurate representation of asset-level uncertainty than the three discrete scenarios.

They further observed that none of the 259 assessed 10-50-90 ranges violated feasibility conditions in Proposition 2. Rerunning the (similarly correlated) portfolio simulation based on continuous (metalog-represented) asset-level uncertainties yielded the "Continuous Simulation Data" shown in Figure 12, and the corresponding "Continuous Simulation Metalog" in Figures 12 and 13. The continuous simulation showed wider tails and a narrower mid-range. The lower end of the distribution visibly extended below $160,000,000, which made sense to the team.

Similarly, the high end now extending above $210,000,000 also made sense. After further reflection and analysis, the team concluded that the continuous simulation was a more accurate and authentic representation of the uncertainty in portfolio value than the discrete simulation. The discrete simulation, they reasoned, arbitrarily cut off the tails of the asset-level distributions prior to simulation (no values outside the low-high range were considered), so it was not surprising that the sum over 259 assets had resulted in artificially short tails as well.

Figure 12. Cumulative Distribution Functions Over Portfolio Value



Based on clarity and confidence gained through such analysis, the decision-makers chose to submit a bid for this portfolio of assets and subsequently won the auction. Had they relied only on the discrete representation in Figures 12 and 13, they would have overbid. The portfolio value ultimately realized several years later was about $180,000,000 -- just slightly less than their prior median.

To date, professional decision analysts have used metalog distributions to represent thousands of uncertainties over dozens of applications across many fields, including life-sciences asset valuations, loan asset valuations, real-estate asset valuations, environmental studies of fish migration, river stream flows, and a wide range of portfolios of such items. Like the team valuing the portfolio of troubled real-estate assets in the above example, such teams have generally concluded that treating continuous uncertainties as continuous and discrete uncertainties as discrete yields more authentic probabilistic results than discretizing all uncertainties from the outset. The metalog system enables practitioners to do this easily and conveniently.

Figure 13. Probability Density Functions Over Portfolio Value

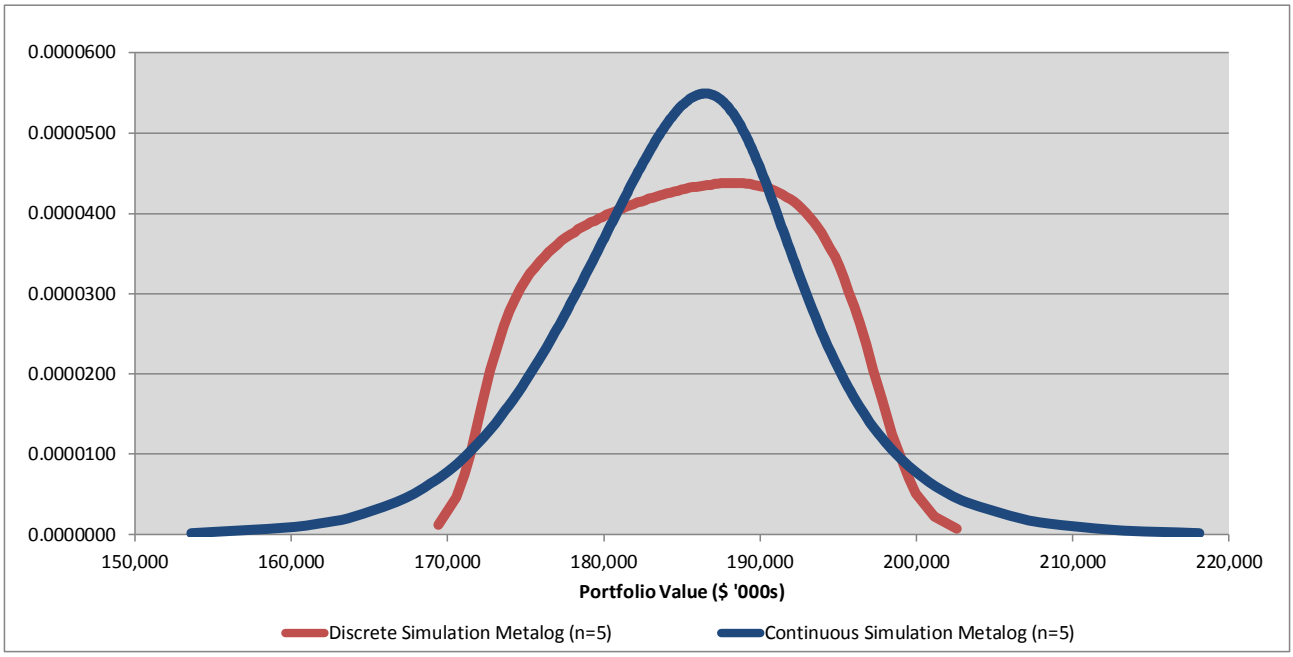
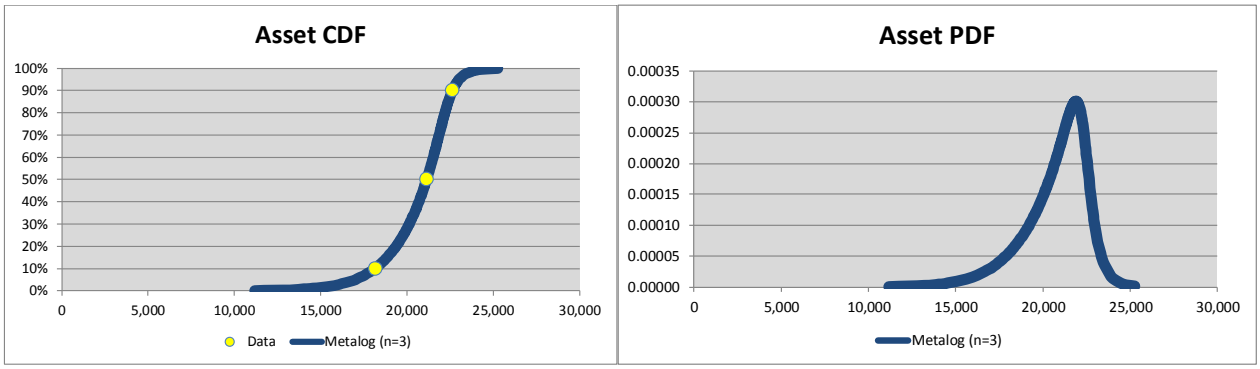


Figure 14. Metalog Distribution for Asset 1



6.3 Distribution Selection within Metalog System

Given input data $(\boldsymbol{x}, \boldsymbol{y})$ that one wishes to represent with a continuous probability distribution, which metalog should one select and how many terms should one use for that selection? As with any distribution selection that is not purely Type-1 driven, this is ultimately a matter of judgement. We now offer several guidelines and tools to help aid this judgement.

With respect to choosing among unbounded, semi-bounded, and bounded distributions, the traditional basis of choice for the Pearson and Johnson systems is to match 3rd and 4th central moments of the data

with a corresponding distribution from Figure 1. However, given a moments-based selection within the Pearson and Johnson systems, this approach has the disadvantage that it offers no choice of boundedness. In contrast, as shown in Figures 4-7, the metalog family offers a wide range of flexibility for each of its unbounded, semi-bounded, and bounded options. So as a starting point per Table 1, we suggest selecting the metalog, log metalog, or logit metalog according to whether the distribution of interest is naturally unbounded, semi-bounded, or bounded.

How many terms to use depends significantly on purpose and context. For example, in decision analysis applications with three assessed data points (m=3), it is natural to use three terms (n=3). In this case, for any feasible data, the metalog CDF will pass through these data exactly as illustrated in Figure 14. More generally, the metalog distributions will pass through the data exactly whenever n=m and the data is feasible, so it makes sense to start with n=m when this result is desired.

In the case of tens or even thousands of data points (e.g. of empirical or simulation data), an exact fit is generally neither desired nor practical. In such cases, one may wish to use: A) relatively few terms (e.g. n=3-6) if a smooth representation is desired, or B) a larger number of terms (e.g. n=7-15) if one is engaged in data or distribution research, or C) the n that maximizes some closeness-of-fit criterion such as K-S distance. In the case of B) or C), one must take care not to overfit[45] the data – as is potentially possible with any linear least squares application with variable number of terms.

To aid such considerations, we have found the "metalog panel" to be a useful tool. As shown in Figures 15 and 16, the metalog panel arrays density functions for a range of n for a given set of data parameters $(\boldsymbol{x}, \boldsymbol{y})$.

Figure 15 shows the array of log-metalog density functions for n=2 to n=16 that correspond to fish biology data in Figure 10. Figure 16 is a similar representation of the for the hydrology data in Figure 11. In both Figures 15 and 16, it is evident how the log metalog increasingly molds itself to the shape of the data and eventually stabilizes its shape as n increases. Blank cells in these figures correspond to the data being infeasible for that choice of n.

From a Bayesian perspective, the choice of n ultimately corresponds to a declaration of "yes, that's what I mean" by a decision-maker or expert. That is, the resulting distribution authentically represents his beliefs.

[45] Among others, Hawkins (2004) and Draper and Smith (1998) provide perspectives on overfitting and rules of thumb for dealing with it.

Figure 15. Metalog Panel for Fish Biology Data

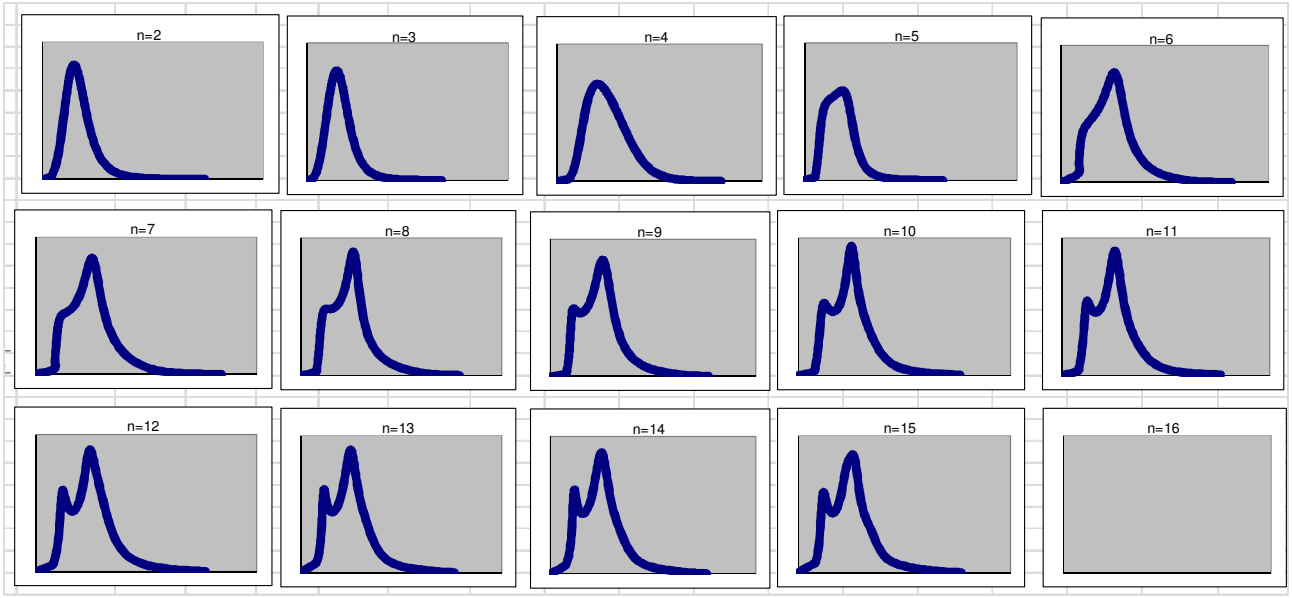
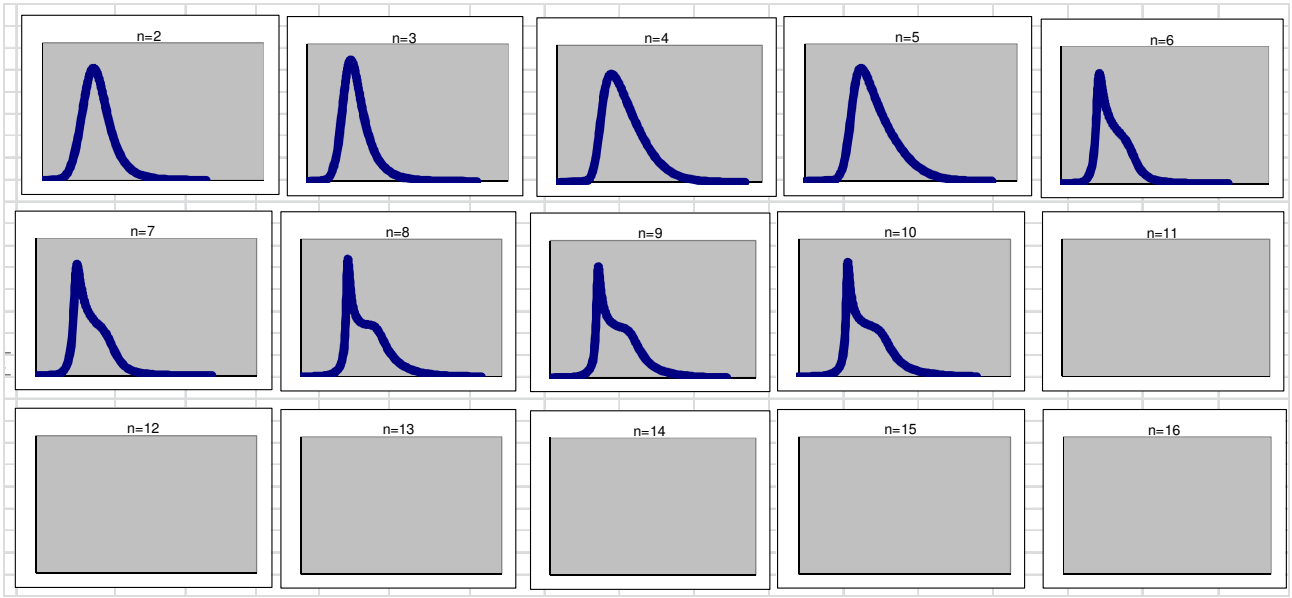


Figure 16. Metalog Panel for Hydrology Data



7. Conclusions

This paper introduces the metalog distributions, a system of continuous univariate probability distributions designed for flexibility, simplicity, and ease/speed of use in practice. While the metalog system offers unbounded, semi-bounded, and bounded distributions that broadly achieve these goals and that compare favorably with previous systems, it also suggests several areas for further research.

First, one can envision various improvements to the metalog system. These include for example characterizing the full range of metalog-system flexibility, including for five or more terms in the β_1- β_2 plane and for the ability to match fifth and higher-order central moments. In addition, it might be useful to extend to four or more terms an expression of the constants and feasibility conditions that we developed for up to three terms Section 6.2.

Second, as suggested in Section 3.2, other "meta" distributions can be developed by applying the methodology of Section 3.1 to other base distributions such as the normal, Gumbel, and exponential. While this research appears to be straight-forward, it has not been done yet, and it may well yield new systems of quantile-parameterized distributions that have certain advantages relative to the metalog.

Third and more broadly, there is a need for new distribution systems that may result from a different combination of choices or the addition of new choices to Table 1. These might include quantile-parameterized systems without feasibility conditions, with additional flexibility for given levels of feasibility, or with flexibility to represent infinite-moments distributions like the Cauchy.

Future research contributions notwithstanding, we believe the metalog system as presented in this paper is ready for use in practice – for any situation in which CDF data is known and a flexible, simple, and easy-to-use continuous probability distribution is needed to represent that data.

References

Abbas, A. E. 2003. "Entropy Methods for Univariate Distributions in Decision Analysis", C. Williams, ed. Bayesian Inference and Maximum Entropy Methods Sci. Engrg.: 22nd International Workshop, American Institute of Physics, Melville, NY, 339–349.

Aldrich, J., 1997. RA Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, *12*(3), pp.162-176.

Balakrishnan, N., 1992. *Handbook of the logistic distribution*. Marcel Dekker.

Bickel, J. E., Lake, L.W., and Lehman, J, 2011. "Discretization, Simulation, and Swanson's (Inaccurate) Mean", SPE Economics and Management, July 2011, pp. 128-140.

Burr, I.W., 1942. Cumulative frequency functions. *The Annals of mathematical statistics*, *13*(2), pp.215-232.

Cheng, R., 2011, December. Using pearson type IV and other Cinderella distributions in simulation. In *Proceedings of the Winter Simulation Conference* (pp. 457-468). Winter Simulation Conference.

De Moivre, A., 1756. *The doctrine of chances: or, A method of calculating the probabilities of events in play* (Vol. 1). Chelsea Publishing Company.

Draper, Norman R.; Smith, Harry (1998). *Applied regression analysis, 3rd Edition*. New York: Wiley. ISBN 978-0471170822.

Edgeworth, F.Y., 1896. XI. The asymmetrical probability-curve. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*,*41*(249), pp.90-99.

Edgeworth, F.Y., 1907. On the representation of statistical frequency by a series. *Journal of the Royal Statistical Society*, *70*(1), pp.102-106.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C. and Wallis, J.R., 1979. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, *15*(5), pp.1049-1054.

Hadlock and Bickel, 2016. Johnson Quantile-Parameterized Distributions, *Decision Analysis*, (tbd) …

Hawkins, D.M., 2004. The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), pp.1-12.

Hosking, J.R., 1998. *L-Moments*. John Wiley & Sons, Inc.

Howard, R.A., 1968. The foundations of decision analysis. *Systems Science and Cybernetics, IEEE Transactions on*, *4*(3), pp.211-219.

Howard, R.A. and Abbas, A.E., 2015. *Foundations of decision analysis*. Prentice Hall.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. *Biometrika*, *36*(1/2), pp.149-176.

Johnson, N.L., Kotz, S. and Balakrishnan, N., 1994. Continuous univariate distributions, vols. 1 and 2, John Wiley & Sons. *New York*.

Karvanen, J., 2006. Estimation of quantile mixtures via L-moments and trimmed L-moments. *Computational Statistics & Data Analysis*, *51*(2), pp.947-959.

Keelin, T.W. and Powley, B.W., 2011. Quantile-parameterized distributions.*Decision Analysis*, *8*(3), pp.206-219.

Keeney, R.L. and Raiffa, H., 1993. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

McGrayne, S.B., 2011. *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.

McDonald, J.B. and Newey, W.K., 1988. Partially adaptive estimation of regression models via the generalized t distribution. *Econometric theory*,*4*(03), pp.428-457.

Mead, R., 1965. A generalised logit-normal distribution. *Biometrics*, *21*(3), pp.721-732.

Nagahara, Y., 1999. The PDF and CF of Pearson type IV distributions and the ML estimation of the parameters. *Statistics & probability letters*, *43*(3), pp.251-264.

Ord, J.K., 1972. *Families of frequency distributions* (Vol. 30). London: Griffin.

Pearson, K., 1893. Contributions to the Mathematical Theory of Evolution.*Proceedings of the Royal Society of London*, *54*(326-330), pp.329-333.

Pearson, K., 1895. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A*, *186*, pp.343-414.

Pearson, K., 1901. Mathematical contributions to the theory of evolution. X. Supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *197*, pp.443-459.

Pearson, K., 1916. Mathematical contributions to the theory of evolution. XIX. Second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *216*, pp.429-457.

Raiffa, H., 1968. Decision analysis: introductory lectures on choices under uncertainty.

Spetzler, C.S. and Stael von Holstein, C.A.S., 1975. Exceptional paper-probability encoding in decision analysis. *Management science*, *22*(3), pp.340-358.

Spetzler, C., Winter, H. and Meyer, J., 2016. *Decision Quality: Value Creation from Better Business Decisions*. John Wiley & Sons.

Tadikamalla, P.R. and Johnson, N.L., 1982. Systems of frequency curves generated by transformations of logistic variables. *Biometrika*, *69*(2), pp.461-465.

Theodossiou, P., 1998. Financial data and the skewed generalized t distribution. *Management Science*, *44*(12-part-1), pp.1650-1661.

Wang, M. and Rennolls, K., 2005. Tree diameter distribution modelling: introducing the logit logistic distribution. *Canadian Journal of Forest Research*, *35*(6), pp.1305-1313.

Key Words: metalog, decision analysis, continuous probability, quantile-parameterized distributions, logistic distribution, continuous univariate distributions, Pearson distributions, Johnson distributions, flexible probability distributions, engineering design of probability distributions.

Acknowledgements

With great appreciation, the author wishes to acknowledge the associate editor and reviewers of this journal for their excellent comments and suggestions; Michael Mischke-Reeds for his encouragement and vetting of the metalog distributions in practice; Robin Arnold for helping gather the fish biology data and develop the name “metalog;” Brad Powley for his thoughtful encouragement and suggestions; Ron Howard for stimulating my interest in this topic as my thesis advisor forty years ago; and my many friends at Strategic Decisions Group for contributing to a collegial and client-needs-based environment over decades that ultimately helped motivate this work.

About The Author

Tom Keelin (“The Metalog Distributions”) has combined a career in decision analysis practice with innovations to advance the field. As Chairman of Millennial Capital, LLC, he has served as general partner for multiple successful real estate funds. He leads strategic decision-making for acquisitions, operations, dispositions, and portfolio management – using decision-analytic, modeling and probabilistic-simulation capabilities. Tom is also a founder and Managing Partner of Keelin Reeds Partners, a management

consulting firm that provides strategy and decision analytic services. In that role, he has developed asset valuation, portfolio management, and business-development-deal-terms methodologies that have enabled greater success for dozens of client companies. In both roles, he recognized the need for better continuous-uncertainty representations, and developed and published new probability distributions accordingly. Previously, as a Managing Director of the Strategic Decision Group, he led the client work for and co-authored the Harvard Business Review article "How SmithKline Beecham Makes Better Resource Allocation Decisions" (Mar-Apr '98). Through that work, he and his colleagues invented the portfolio-management standard which subsequently was adopted widely across life-sciences industry. Earlier, with Decision Focus, Inc, Tom developed the Over/Under Capacity Planning model, which effectively addressed demand uncertainty in electric power system planning and was widely adopted by many utilities and regulatory commissions over the next decade. Tom is a Fellow of the Society of Decision Professionals, and a founder and director of the Decision Education Foundation, a not-for-profit organization that helps youth learn good decision skills for life. Tom holds three degrees from Stanford University: BA in Economics and MS and PhD in Engineering-Economic Systems. Email: tomk@keelinreeds.com.