Scott Nestler PhD, CAP, PStat is principal data scientist and optimisation lead at Sumer Sports, LLC.

Tom Keelin PhD is a managing partner of Keelin Reeds Partners, a Fellow of the Society of Decision Professionals, and a founder and director of the Decision Education Foundation.

Introducing the metalog distributions

Scott Nestler and **Tom Keelin** provide an overview of a new family of distributions, originally developed for the decision analysis field, that may have virtually unlimited applicability in any field

The metalog distribution (metalogdistributions.com) is a new continuous probability distribution that has nearly universal shape and bounds flexibility. Over a wide range, it can mimic virtually all traditional distributions and fit data more accurately. Developed by Tom Keelin in 2016 as a generalisation of the logistic distribution, "metalog" is short for "meta-logistic".[1] Software resources for metalog distributions are available in R, Python, Excel, and other tools (metalogdistributions.com/software.html).

Fitting distributions to elicited data

Let us take a simple example. Decision analysts commonly elicit three quantiles (e.g., 0.1, 0.5, and 0.9) from an expert and then fit a continuous probability distribution to these points. Perhaps the quantity being estimated represents the time to complete a task, or the cost of a particular component in a system. A problem faced by decision analysts is that classical probability distributions, even the very flexible beta distribution, typically lack sufficient shape flexibility to exactly run through all three points, which is desirable to authentically represent the expert's view. The metalog distribution was originally designed to overcome this limitation. Figure 1 shows one such metalog distribution, which, within its feasibility range, will pass exactly through any three points.

A brief history lesson

One might think of the history of probability distributions as a progression of individual developments towards greater shape and bounds flexibility for fitting to data. The normal distribution (1756) laid the foundation for much of the development of classical statistics. In contrast, Bayes' theorem (1763) was the basis for state-of-information, belief-based probability representations. Because belief-based probabilities can take on any shape and may have natural bounds, probability distributions flexible enough to accommodate both were needed. Moreover, many empirical and experimental data sets exhibited shapes that could not be well matched by the normal or other continuous distributions. So began the search for continuous probability distributions with flexible shapes and bounds. The metalog is one such distribution. It has a single set of simple closed-form equations and parameter estimation with ordinary least squares. Moreover, the metalog is ideal for simulation because uniform random samples are converted directly into samples of the variable of interest x by the metalog's closed-form quantile function. See the box on page 33 for technical details.

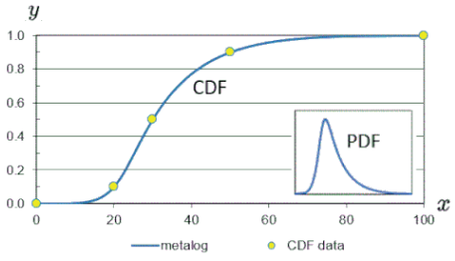

FIGURE 1: Bounded metalog parameterised with cumulative distribution function (CDF) data (20,0.1), (30,0.5), (50,0.9) and bounds (0,0) and (100,1).

The metalog distribution can be used to model data from many different domains or applications, whether it is an elicited, simulated, or empirical source

A general-purpose distribution

Traditional distributions like the normal, lognormal, exponential, and beta each have their special purposes. For example, by the central limit theorem, the normal distribution is the limiting shape of a sum of identically distributed random variables. Similarly, the exponential distribution can be derived as the probability distribution of the time between events in a Poisson process (a process in which events occur continuously and independently at a constant average rate). In contrast, the metalog is a general-purpose, more universal distribution that, as shown by blue curves in Figure 2, can closely approximate all these shapes (yellow curves) and that can take on virtually any other shape as well. The metalog is also universal in terms of its boundedness. As explained in the box (page 33), metalogs can be unbounded, semi-bounded, or bounded, where numerical upper and/or lower bounds, if any, may be specified as appropriate. For example, since the time between events in Poisson process cannot be negative, both the exponential distribution and the metalog approximating it in Figure 2(c) have a lower bound of zero. And since the normal ▶

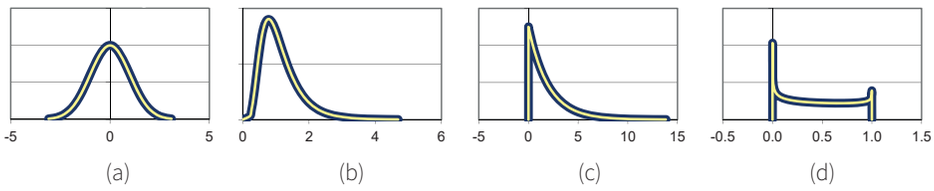

FIGURE 2 The metalog distribution (blue curves) as a close approximation to traditional distributions (yellow curves): (a) normal ($\mu = 0$, $\sigma = 1$); (b) lognormal ($\mu = 0$, $\sigma = 0.5$); (c) exponential ($\lambda = 0.5$); (d) beta ($\alpha = 0.8$, $\beta = 0.9$).

By telling a more nuanced story about the data than less flexible and less accurate alternatives, metalog distributions may open new avenues for data research and exploration

distribution in Figure 2(a) is unbounded, so is the metalog that approximates it.

Wide range of applications

The metalog distribution can be used to model data from many different domains or applications, whether it is an elicited, simulated, or empirical source. Here are examples from various areas to date: astronomy (risk of asteroid impacts), cybersecurity (risk of incidents), eliciting and combining expert opinions (fertility rates, Statistics Canada), hydrology (probability of river gauge heights), portfolio management (value of new products), and simulation (for both input and output distributions). One application of particular interest to statisticians is enabling closed-form representations of known distributions for which the CDF has no closed-form expression, such as the sum of independent identically distributed lognormal distributions.[2]

In fish biology, the metalog has been used to represent the empirical distribution of steelhead trout weights shown in Figure 3.[1] While a lognormal distribution, which might typically be used by fish biologists to fit such data, provides a reasonable fit, the 10-term metalog provides a more nuanced view. The population of steelhead in the river consists of fish returning upriver 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 rich ocean 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.

Both the relative population sizes and weight differences between 1-salt and 2-salt steelhead are unsolved research questions in fish biology. Could the two modes in the more nuanced metalog distribution correspond to 1-salt and 2-salt fish, respectively? If this bimodality were to be verified by further research (by methods such as training and test data sets, expanded time periods, alternative time-period segmentations, analysis of such data from other rivers, or various statistical methods), it would shed significant light on these unsolved research questions. Our (less ambitious) purpose here is simply to illustrate the use of metalogs as a data exploration and visualisation tool that may provide insights worthy of further exploration.

Note that this insight would have been missed if the lognormal distribution (or a similar unimodal distribution) had been assumed *a priori*. Moreover, the histogram shape for this data set depends on arbitrary bin-width and bin-location settings. While it shows two modes in Figure 3, it is unimodal under other reasonable settings. Thus, by telling a more nuanced story about the data than less flexible and less accurate alternatives, metalog distributions may open new avenues for data research and exploration.

Guidelines for selecting the number of metalog terms

Keep in mind that metalogs provide the flexibility to choose any number of terms and that using more terms means more shape flexibility. Selecting the number of terms involves a trade-off between parsimony (fewer terms) and fit accuracy (more terms). Like any other linear regression curve-fitting, it is possible to overfit metalogs by using too many terms or

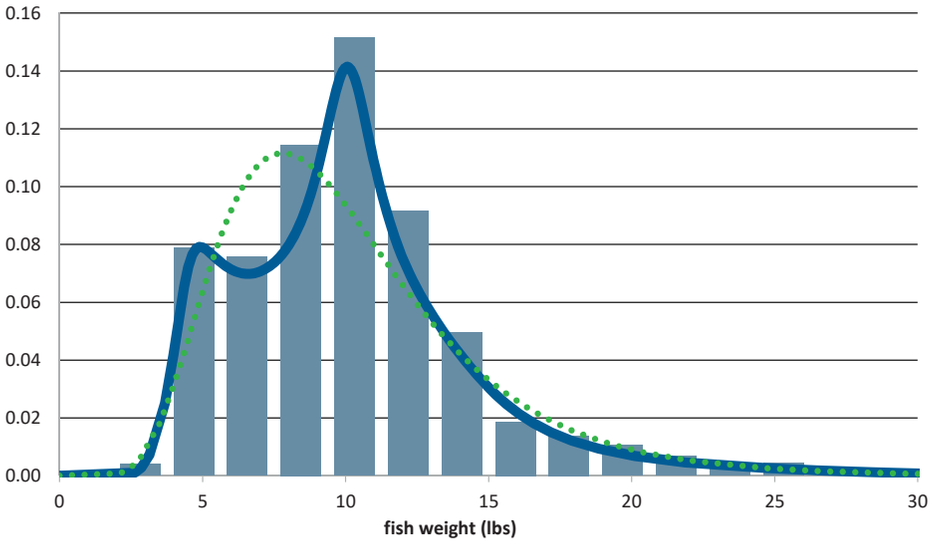

FIGURE 3 Empirical distribution of the weight of 3,474 steelhead trout caught and released on the Babine River, British Columbia, during 2006–2010. Overlaid on the data histogram (light blue bars) are the best-fit lognormal distribution (green dots) and the 10-term metalog distribution (blue curve) fitted with linear regression. The 10-term metalog captures the bimodality inherent in the data whereas the lognormal does not.

Technical details and notable properties

The metalog distribution is a generalisation of the logistic distribution. The logistic quantile function is given by

$$x = Q(y) = \mu + s \ln\left(\frac{y}{1-y}\right)$$

where $0 < y < 1$ is cumulative probability and μ and s are parameters that control location and scale, respectively. The metalog quantile function is defined by substituting power series expansions in y for these parameters:

$$\mu = a_1 + a_4(y - 0.5) + a_5(y - 0.5)^2 + \ldots$$

and

$$s = a_2 + a_3(y - 0.5) + a_6(y - 0.5)^2 + \ldots,$$

where the a-coefficients are constants that determine its location, scale, and shape.[1]

Like a Taylor series, the metalog quantile function may have any number of terms k. Each additional term adds shape flexibility: a k-term metalog has $k - 2$ shape parameters. By increasing the number of terms, the metalog has been shown to have virtually unlimited shape flexibility.[3]

Since the metalog quantile function is differentiable, it has a simple closed-form probability density function (PDF), the shape of which also depends on the a-coefficients.[1] To be a feasible probability distribution, the a-coefficients must be such that this PDF is positive for all y.

Note that by design the metalog quantile function is linear in the a-coefficients. By implication, these coefficients can be determined from data in closed form by linear regression.[1] Moreover, Bayesian linear regression[4] can be used to update, in light of new data, a metalog-distributed long-run frequency distribution over a variable of interest according to Bayes' theorem in closed form.[3]

Simple transformations of the metalog quantile function yield semi-bounded and bounded metalog distributions,[1] where the user can set upper and/or lower bounds as appropriate. Beyond honouring such user-specified bounds, these metalog transforms retain the properties of virtually unlimited shape flexibility and determination of a-coefficients by linear regression.[1]

underfit by using too few. Generally, it is best to use the smallest number of terms that appropriately fit your data. While three terms is sufficient for many applications, more can be useful. For example, we used 10-term metalogs to closely match the traditional distributions in Figure 2 and to capture the bimodality that may be evident from context in Figure 3.

You may also opt to use goodness-of-fit criteria such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC) to identify the optimal number of terms, just as you would with traditional distributions. For example, when applied to the fish weight data in Figure 3, the AIC ranking of metalog distributions from 2 to 16 terms along with a wide range of classical distributions identifies the 11-term metalog as the best fit to this data. A similar BIC ranking identifies the 10-term metalog (shown in Figure 3) as the best fit.

When to use and not use metalogs

If you have data (empirical, elicited, or simulated) and wish to find a continuous distribution to accurately represent that data, you may be best off starting with a metalog. The reason is that the metalog will fit almost any data set more accurately than traditional distributions and more easily (with linear regression; see box above).

The metalog will fit almost any data set more accurately than traditional distributions and more easily

The key exception is if you know (or are willing to assume), *a priori*, that your data was generated from a special-purpose distribution, in which case you might simply opt to use the latter.

There are some extreme data sets that even metalogs with up to 16 terms do not fit accurately. For example, for data sets from mixed discrete–continuous distributions or multimodal distributions with zero probability density between modes, such higher-order metalogs may be infeasible. For data sets with extremely fat tails (e.g., from a Cauchy distribution), an impractically large number of terms may be required to fit the tails accurately.

A distribution for all data

Analysts will often find themselves looking at a histogram and descriptive statistics for a data set and asking questions like "What kind of distribution should I fit to this?", "Does it look normal?", and "Should I use maximum likelihood estimation to determine the parameters?" It is not unusual to try a few different options and still feel unsatisfied with the fit of the distribution. Rather than this "hunt and peck" approach to choosing a distribution, we would encourage analysts to give metalogs a try. ■

References

1. Keelin, T. (2016) The metalog distributions. *Decision Analysis*, **13**(4), 243–277.

2. Keelin, T. W., Chrisman, L. and Savage, S. L. (2019) The metalog distributions and extremely accurate sums of lognormals in closed form. In *2019 Winter Simulation Conference (WSC)* (pp. 3074–3085). Piscataway, NJ: IEEE.

3. Keelin, T.W. and Howard, R.A., 2021. The metalog distributions: Virtually unlimited shape flexibility, combining expert opinion in closed form, and Bayesian updating in closed form. osf.io/xdg5e/

4. Bayesian linear regression. Wikipedia. Accessed 12/02/2021.