Keelin Reeds Partners
770 Menlo Ave., Ste 230
Menlo Park, CA 94025
650.465.4800 phone
tomk@keelinreeds.com
www.keelinreeds.com

The Metalog Distributions

Probability Distributions to Represent *Any* Continuous Uncertainty

Invited Lecture at Stanford University

Department of Management Science and Engineering

Tom Keelin

February 28, 2017

Metalog Topics

- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

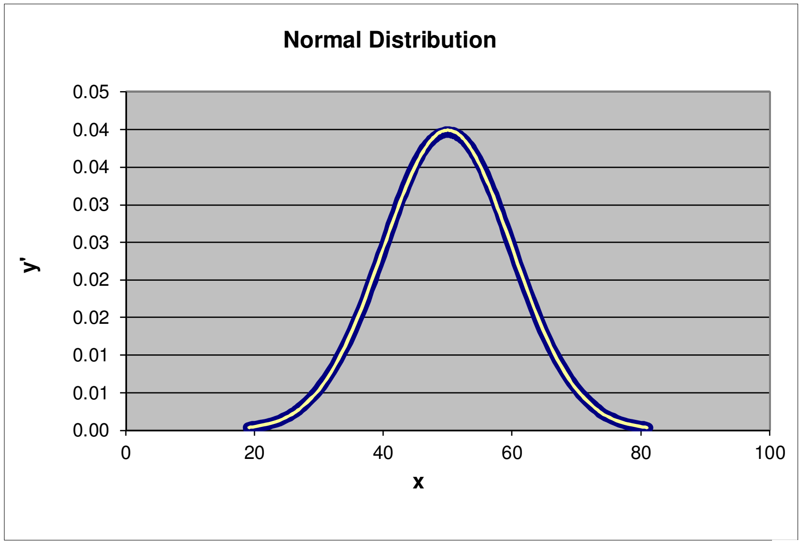
A Short History of Continuous Probability Distributions

interpretation of probability	1700–1800	1800–1900	1900–2000	2000–2100
state of information (including frequency)	foundation laid: continuous probabilities can legitimately take on any shape (Bayes,1763. Further developed by Laplace late 1700's.)			
frequency only (classical statistics)	normal distribution published (DeMoivre,1756)		normal distribution modified for skewness/ kurtosis flexibility (Edgeworth 1896, 1907. Pearson, 1895,1901, 1916)	

1700 1800 1900 2000 2100

Year

What did Pearson do?



normal pdf: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

differential equation: $\frac{f'(x)}{f(x)} = -\frac{x-\mu}{\sigma^2}$

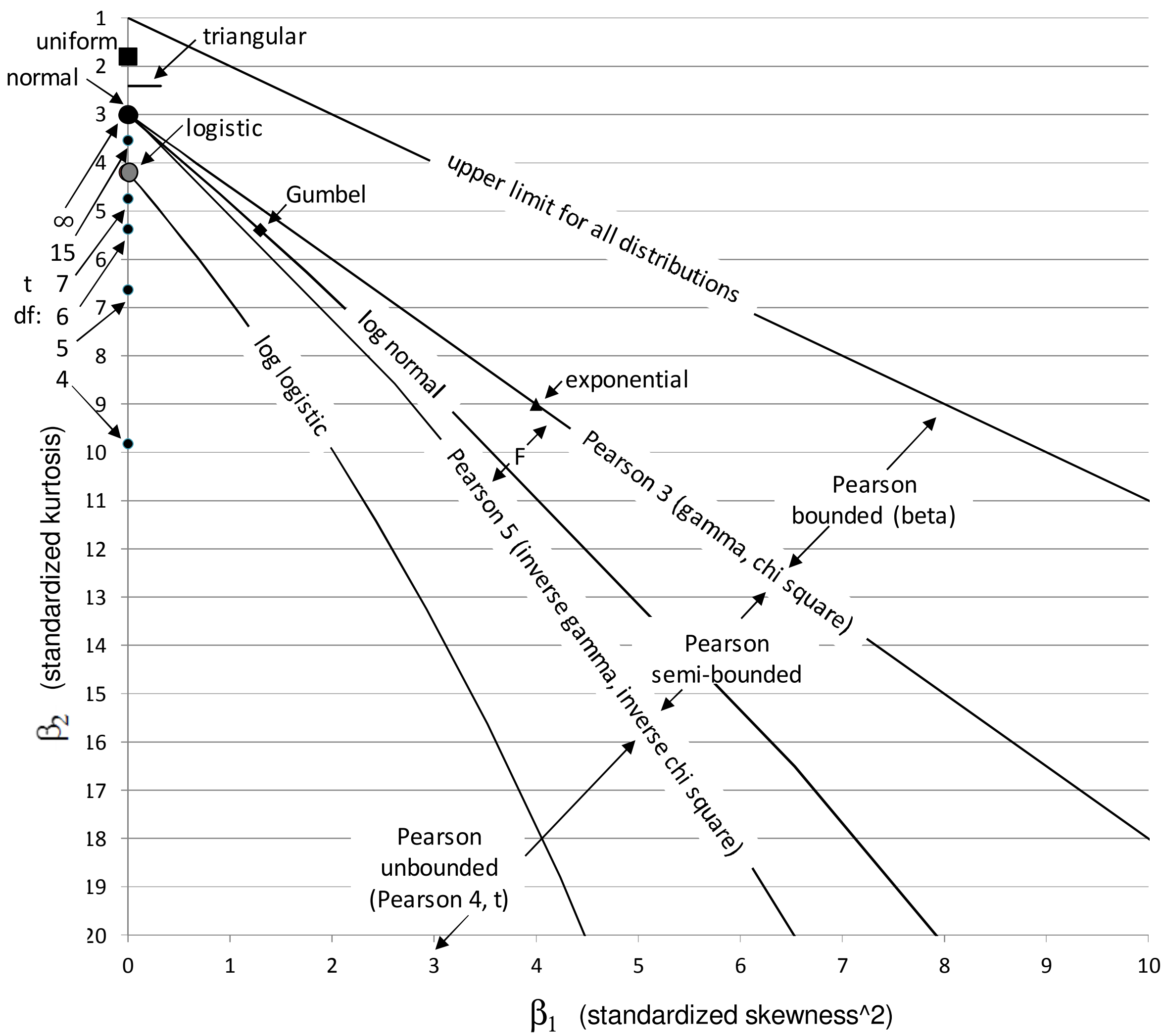
Pearson's modification: $\frac{f'(x)}{f(x)} = \frac{x-\mu}{b_0 + b_1 x + b_2 x^2}$

12 solutions to Pearson's differential equation

- beta distribution (type I)
- beta prime distribution (type VI)
- Cauchy distribution (type IV)
- chi-squared distribution (type III)
- continuous uniform distribution (limit of type I)
- exponential distribution (type III)
- gamma distribution (type III)
- *F*-distribution (type VI)
- inverse-chi-squared distribution (type V)
- inverse-gamma distribution (type V)
- normal distribution (limit of type I, III, IV, V, or VI)
- Student's *t*-distribution (type VII, which is the non-skewed subtype of type IV)

Strengths and Shortcomings of the Pearson System

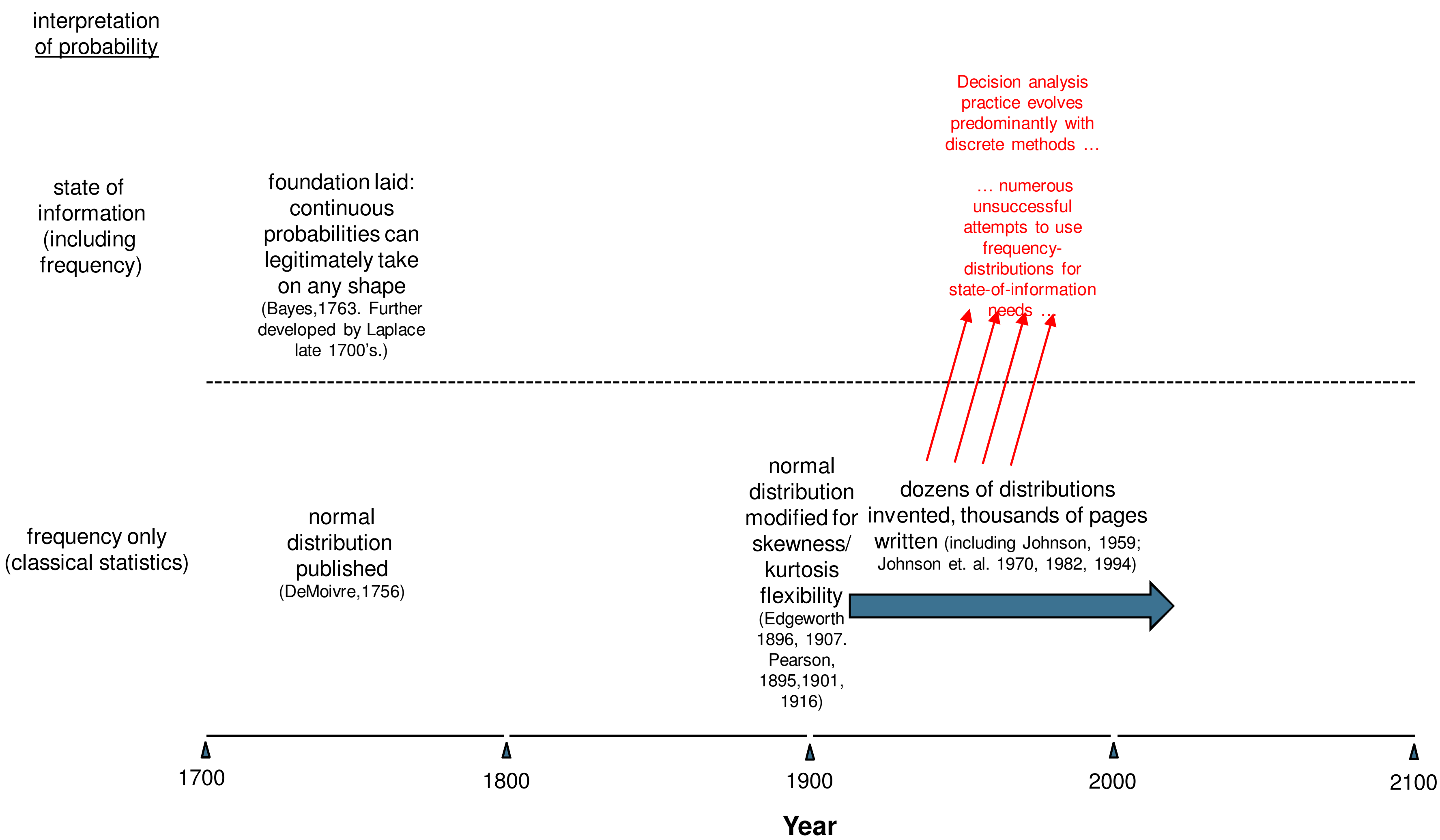
Flexibility: Can match any combination of skewness and kurtosis



Shortcomings:

- Limited to 2 shape parameters
- Given a point (β_1, β_2), Pearson and system offers
  - ❖ zero choice of boundedness
  - ❖ zero ability to match 5th or higher-order moments
- A dozen functional forms, some of which are duplicative, with incomplete guidance for which to use.
- Parameter estimation can require non-linear optimization (with situation-specific manual intervention).

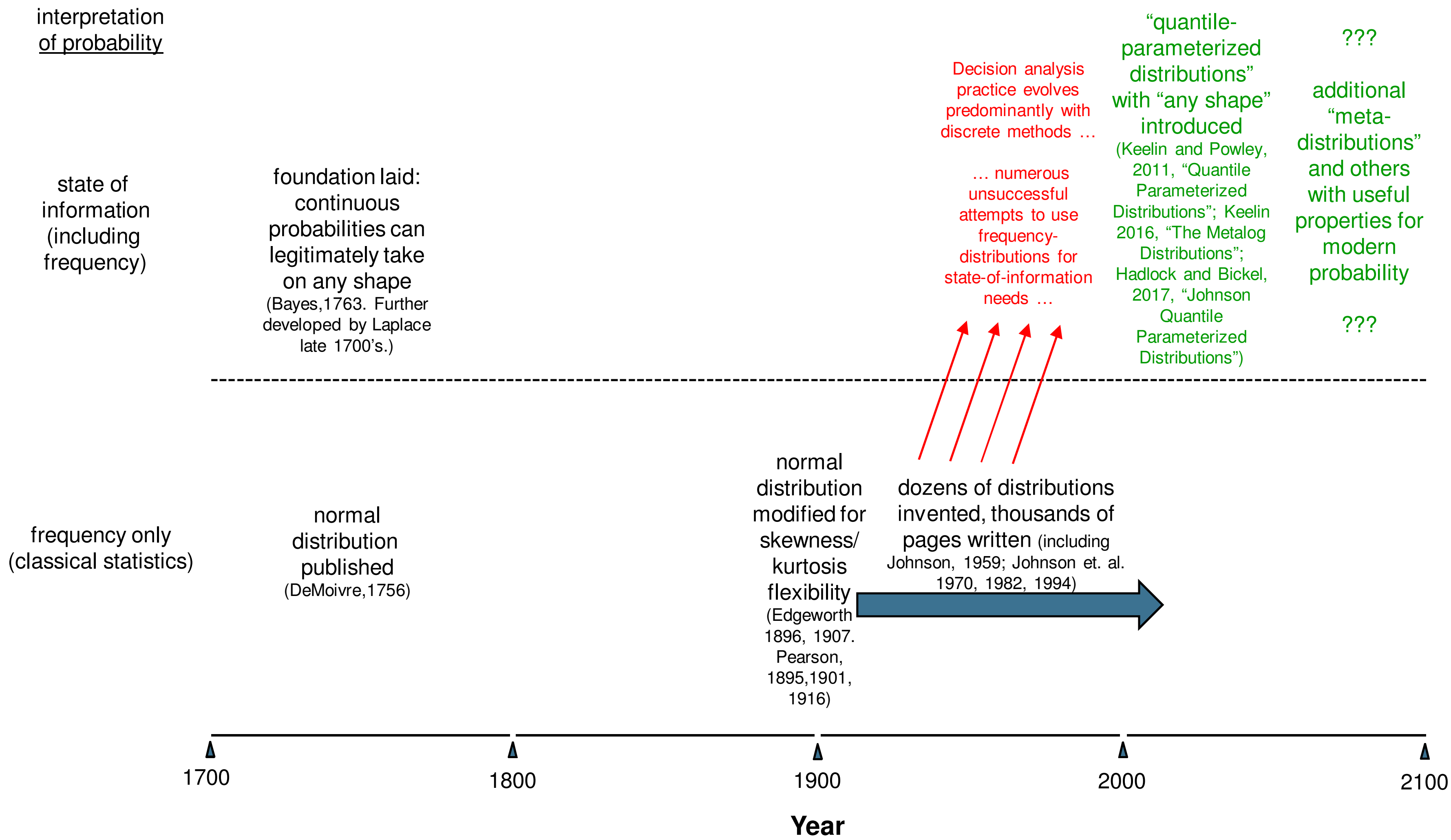
A Short History of Continuous Probability Distributions



Personal Journey to a New Family of Probability Distributions

- Early days at Stanford
  - "It's not easy to invent a new probability distribution."
- Decision analysis (DA) with discrete methods -- first 25 years of professional practice
  - Continuous distributions were desirable but largely impractical.
- DA with simulation – next 15 years. Continuous distributions
  - Computationally tractable (in a few cases)
  - Otherwise impractical (encoding, parameter estimation, lack of flexibility)
- 2009 light-bulb moment: why not <u>invent</u> continuous distributions that meet the needs of modern ("state-of-information") probability applications?
  - White-board sketches (starting with how to add skewness to the Normal distribution and parameterize it with 10/50/90 assessments) led to …

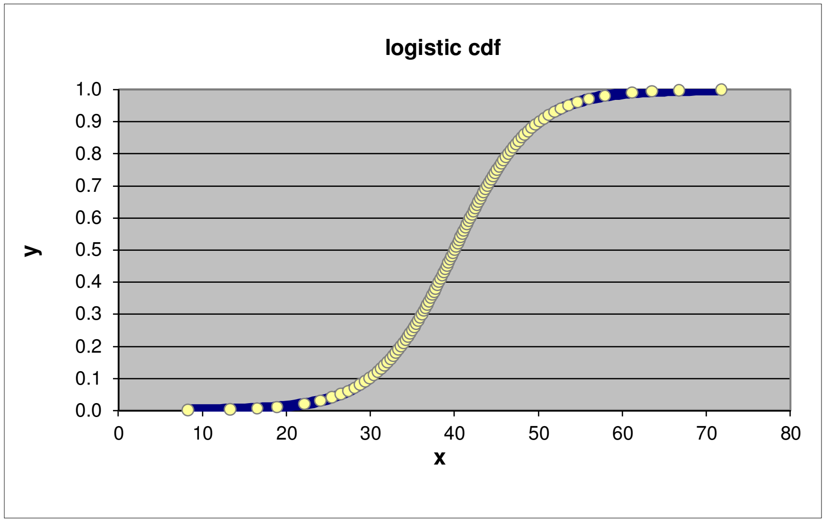
A Short History of Continuous Probability Distributions

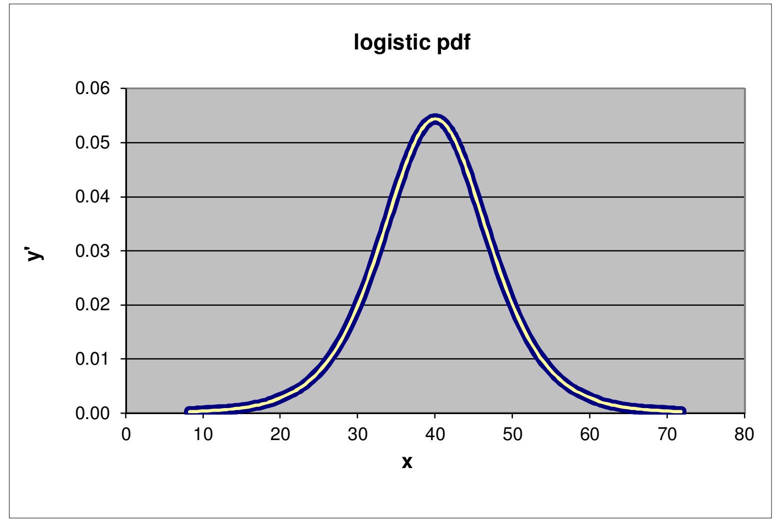


Metalog Topics

- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

Logistic Distribution





cdf: $y = \dfrac{1}{1+e^{-\frac{x-\mu}{s}}}$

pdf: $y' = \dfrac{e^{\frac{x-\mu}{s}}}{s\left(1+e^{\frac{x-\mu}{s}}\right)^2}$

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

pdf: $y' = \dfrac{y\,(1-y)}{s}$

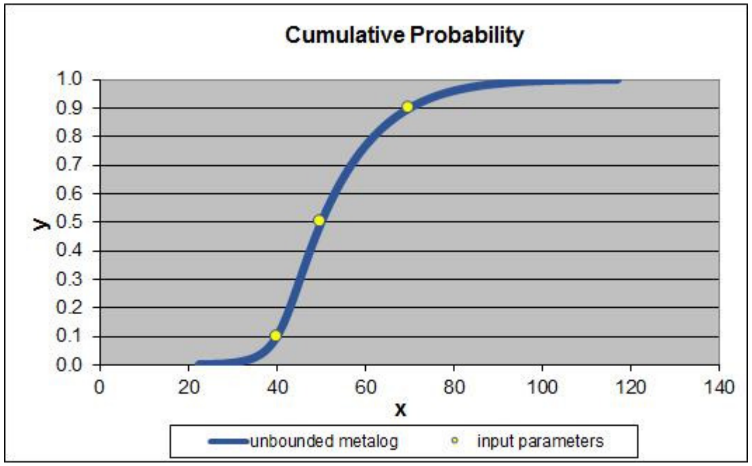
where: $\sigma = s\,\pi / \sqrt{3}$

"metalog" is short for "meta logistic"

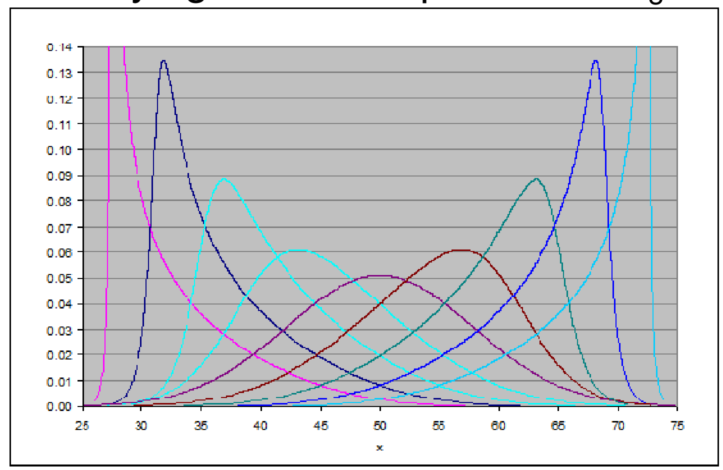
What's this?

$$x = \overbrace{a_1 + a_2 \ln\left(\frac{y}{1-y}\right)}^{\text{logistic distribution}} + \overbrace{a_3 (y-0.5)\ln\left(\frac{y}{1-y}\right)}^{\text{skewness term}} \equiv M_3(y)$$

"3-term metalog distribution"



… varying skewness parameter a_3

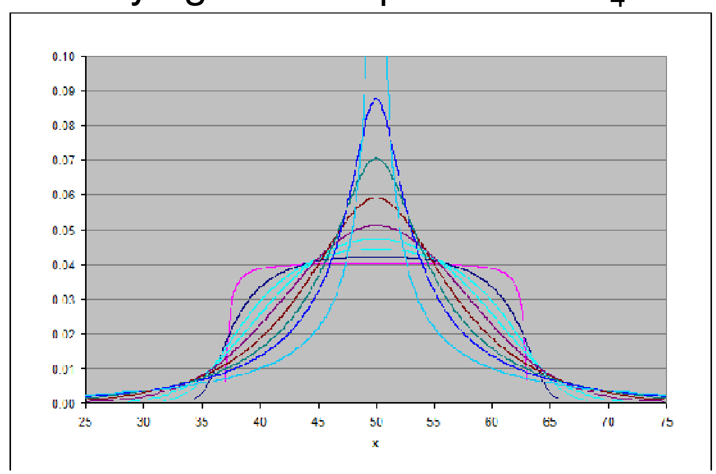


add a 4th term …

$$x = a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3 (y-0.5)\ln\left(\frac{y}{1-y}\right) \underbrace{+ a_4 (y-0.5)}_{\text{kurtosis term}} \equiv M_4(y)$$

"4-term metalog distribution"

… varying kurtosis parameter a_4



… generalizes to any number of terms $M_n(y)$

Metalog constants a_i are determined linearly from CDF data.

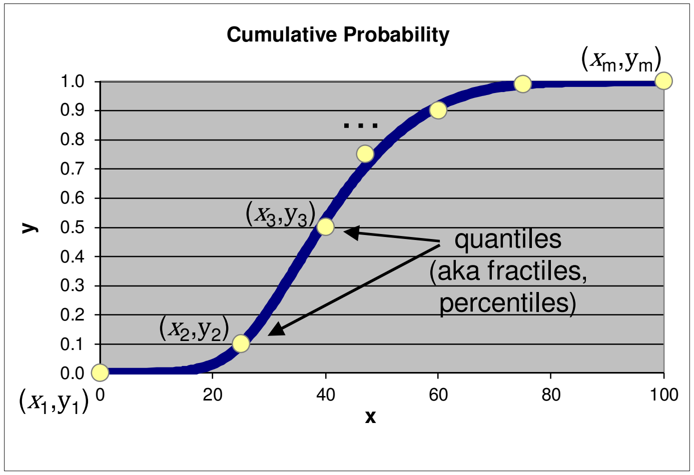
Given CDF data points $(\boldsymbol{x}, \boldsymbol{y})$ where $\boldsymbol{x} = (x_1, \ldots, x_m)$, $\boldsymbol{y} = (y_1, \ldots, y_m)$, and m>=n, the constants $\boldsymbol{a} = (a_1, \ldots, a_n)$ are related to the data by a set of linear equations:

$$x_1 = a_1 + a_2 \ln\left(\frac{y_1}{1-y_1}\right) + a_3(y_1 - 0.5)\ ln\left(\frac{y_1}{1-y_1}\right) + a_4(y_1 - 0.5) + \ldots$$

$$x_2 = a_1 + a_2 \ln\left(\frac{y_2}{1-y_2}\right) + a_3(y_2 - 0.5)\ ln\left(\frac{y_2}{1-y_2}\right) + a_4(y_2 - 0.5) + \ldots$$

$$\vdots$$

$$x_m = a_1 + a_2 \ln\left(\frac{y_m}{1-y_m}\right) + a_3(y_m - 0.5)\ ln\left(\frac{y_m}{1-y_m}\right) + a_4(y_m - 0.5) + \ldots$$



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

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

invertibility guaranteed except in pathological cases

case 1 If $\mathbf{Y}$ is invertible and m=n, then $\boldsymbol{a}$ is uniquely determined by $\boldsymbol{a} = \mathbf{Y}^{-1}\boldsymbol{x}$.

case 2 If m>n and $\mathbf{Y}$ has rank of at least n, then $\boldsymbol{a}$ can be estimated by ordinary least squares $\boldsymbol{a} = [\mathbf{Y}^T \mathbf{Y}]^{-1} \mathbf{Y}^T \boldsymbol{x}$.

$\boldsymbol{a} = [\mathbf{Y}^T \mathbf{Y}]^{-1} \mathbf{Y}^T \boldsymbol{x}$ *

works either way

Feasibility of $(\boldsymbol{x}, \boldsymbol{y})$: $M_n(y)$ is strictly increasing. Equivalently, density function $m_n(y)$ is positive over 0<y<1.

Metalog moments are closed-form polynomials of the a_i's.

For example, for the 4-term metalog

metalog mean: $a_1 + \frac{a_3}{2}$

metalog variance: $\frac{1}{3}\pi^2 a_2^2 + \left(\frac{1}{12} + \frac{\pi^2}{36}\right) a_3^2 + a_2 a_4 + \frac{a_4^2}{12}$

metalog 3rd central moment:

$$\pi^2 a_2^2 a_3 + \frac{1}{24}\pi^2 a_3^3 + \frac{1}{2} a_2 a_3 a_4 + \frac{1}{6}\pi^2 a_2 a_3 a_4 + \frac{1}{8} a_3 a_4^2$$

metalog 4th central moment:

$$\frac{7}{15}\pi^4 a_2^4 + \frac{3}{2}\pi^2 a_2^2 a_3^2 + \frac{7}{30}\pi^4 a_2^2 a_3^2 + \frac{a_3^4}{80} + \frac{1}{24}\pi^2 a_3^4 + \frac{7\pi^4 a_3^4}{1200} + 2\pi^2 a_2^3 a_4 + \frac{1}{2} a_2 a_3^2 a_4 + \frac{2}{3}\pi^2 a_2 a_3^2 a_4 + 2a_2^2 a_4^2 + \frac{1}{6}\pi^2 a_2^2 a_4^2 + \frac{1}{8} a_3^2 a_4^2 + \frac{1}{40}\pi^2 a_3^2 a_4^2 + \frac{1}{3} a_2 a_4^3 + \frac{a_4^4}{80}$$

More generally, the k^{th} central moment of the n-term metalog is simply a k^{th}-order polynomial of the a_i's.

How about simple and flexible semi-bounded or bounded distributions?

Name	Interpretation	CDF (quantile function)	Condition
metalog (unbounded)	generalized logistic distribution	$M_n(y) = a_1 + a_2 \ln\left(\frac{y}{1-y}\right)$ $+ a_3(y-0.5)\ln\left(\frac{y}{1-y}\right) + \ \ldots$	$0 < y < 1$
semi-bounded metalog	$\log(x)$ is metalog distributed	$M_n^{log}(y) = b_l + e^{M_n(y)}$ $= b_l$	$0 < y < 1$ $y = 0$ (given lower bound b_l)
bounded metalog	$\text{logit}(x) = \ln(\frac{x-b_l}{b_u-x})$ is metalog distributed	$M_n^{logit}(y) = \frac{b_l + b_u e^{M_n(y)}}{1 + e^{M_n(y)}}$ $= b_l$ $= b_u$	$0 < y < 1$ $y = 0$ $y = 1$ (given lower and upper bounds b_l and b_u)

Metalog Topics

- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

Types of Continuous Probability Distributions

	Basis of Legitimacy	Criteria	Examples
Type I	Derived from an underlying probability model	Distribution reflects the model	normal exponential …
Type II	Matches *specific* types of empirical data	Distribution matches data	generalized logit-normal (Mead, 1965) skewed generalized t distribution (Theodossiou,1994) … (dozens of others)
Type III	Matches most *any* set of empirical (or assessed) data	Flexibility Simplicity Ease of use	Pearson distributions (1895,1901,1916) Johnson distributions (1949,1982) … Quantile parameterized distributions (Keelin and Powley, 2011) Metalog distributions (this research)

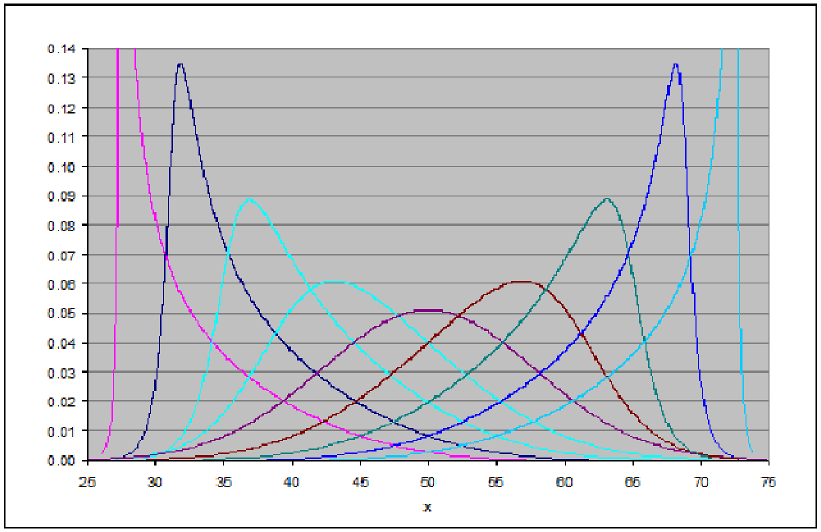
Engineering a new probability distribution – strategy table

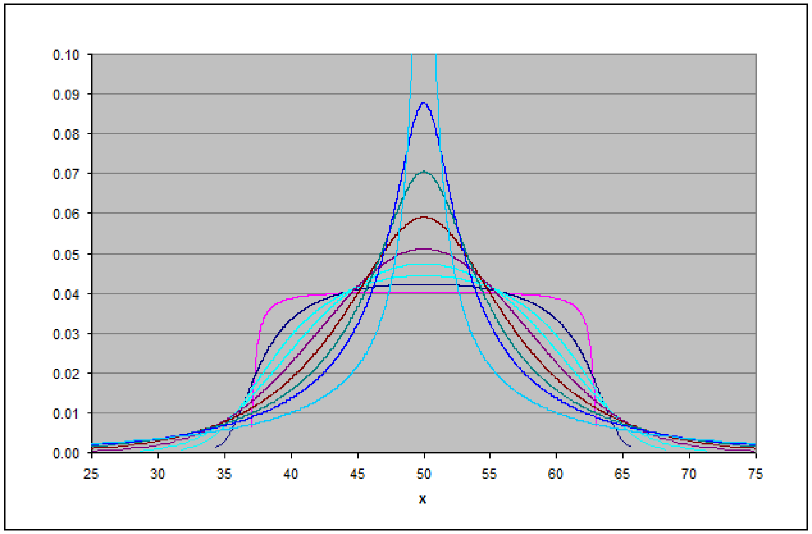
Pearson

metalog

Base Distribution	Form Selected for Modification	Modification Method	Distribution Selection	Parameter Estimation
normal (Edgeworth 1896, 1907; Pearson 1895,1901,1916; Charlier 1928; Johnson 1949)	probability density function (PDF) (Edgeworth 1896, 1907; Pearson 1895,1901,1916; Charlier 1928;)	parameter addition (Mead, 1965; Theodossiou,1994)	ability to match moments (Pearson 1895,1901,1916; Johnson 1949; Tadikamalla and Johnson 1982)	method of moments (Pearson 1895,1901,1916)
logistic (Tadikamalla and Johnson 1982; Balakrishnan, 1992)	cumulative distribution function (CDF) (Burr, 1942)	parameter substitution (Pearson 1895,1901,1916)	match natural bounds	maximum likelihood (Fisher 1932)
student t (McDonald and Newey, 1988; Theodossiou,1994)	quantile function (inverse CDF) (Keelin and Powley, 2011)	transformation (Johnson 1949; Tadikamalla and Johnson 1982; Hadlock and Bickel, tbd)	…	probability-weighted- and L-moments (Greenwood, et. al. 1979; Hosking, 1990)
…	characteristic function (Ord 1972)	series expansion (Edgeworth 1896, 1907; Charlier 1928)		quantile parameterized (Keelin and Powley, 2011. Hadlock and Bickel, tbd.)

The Metalog: A Generalized Logistic Distribution





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

where:

$\mu = a_1 + a_4(y-0.5) + a_5(y-0.5)^2 + a_7(y-0.5)^3 + a_9(y-0.5)^4 + a_{11}(y-0.5)^5 + \ldots$

$s = a_2 + a_3(y-0.5) + a_6(y-0.5)^2 + a_8(y-0.5)^3 + a_{10}(y-0.5)^4 + a_{12}(y-0.5)^5 + \ldots$

$\Big\}$ series expansion

thus: $x \equiv M_n(y) = a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3(y-0.5)\ln\left(\frac{y}{1-y}\right) + a_4(y-0.5) + \ldots$

pdf: $y' \equiv m_n(y) = \left[\frac{a_2}{y(1-y)} + a_3\left(\frac{y-0.5}{y(1-y)} + \ln\left(\frac{y}{1-y}\right)\right) + a_4 + \ldots\right]^{-1}$

where: n = number of series terms in use. a_i's are constants.

Since μ by itself is a power series in $(y-0.5)^k$ with unlimited terms, Taylor's Theorem guarantees that the metalog can *locally* approximate *any* sufficiently smooth distribution *arbitrarily closely*.

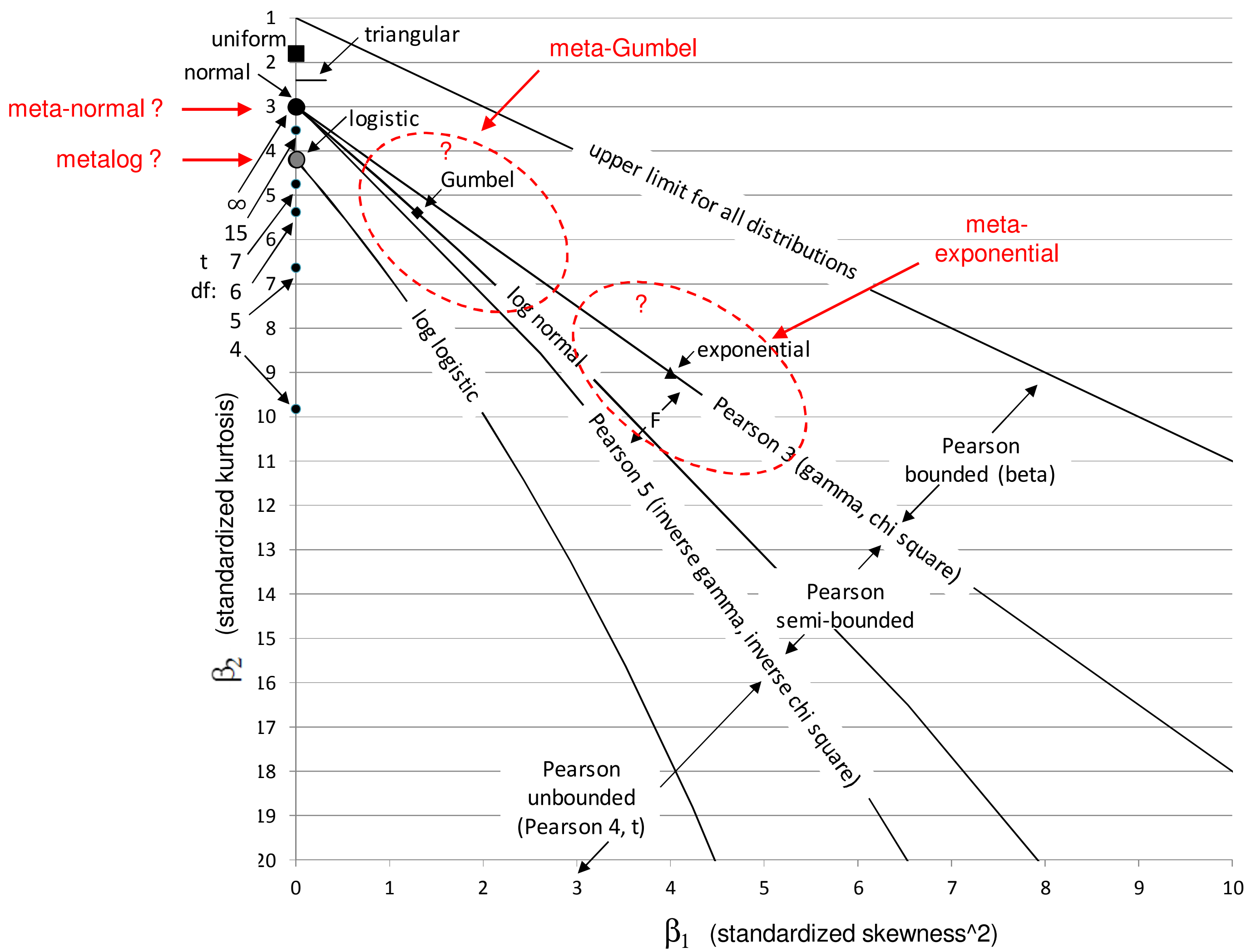
Other practical “meta-distributions” can be formed similarly.

Meta-distribution: a generalization of a base distribution created by substituting for one or more of its parameters an unlimited number of shape parameters

<table>
<tr><th>Name</th><th>Quantile function linear in its parameters</th><th>QPD + easy to simulate</th><th>Flexibility -- (β_1, β_2) plot</th><th>Properties (feasibility, moments, transforms, etc.)</th><th>Advantages relative to other distributions</th><th>Prior research</th></tr>
<tr><td>metalog</td><td>$x = \mu + s\, ln(\frac{y}{1-y})$</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>“The Metalog Distributions”, Keelin, 2016</td></tr>
<tr><td>meta-normal</td><td>$x = \mu + \sigma\, \Phi^{-1}(y)$</td><td>✓</td><td>(unexplored)</td><td>(for 4 terms)</td><td>(for 4 terms)</td><td>“Quantile-Parameterized Distributions”, Keelin and Powley, 2011</td></tr>
<tr><td>meta-exponential</td><td>$x = -\,(1/\lambda)\, ln(1-y)$</td><td>✓</td><td colspan="4" rowspan="4">(unexplored)</td></tr>
<tr><td>meta-Gumbel</td><td>$x = \mu - \beta\, ln(-ln(y))$</td><td>✓</td></tr>
<tr><td>meta-Cauchy</td><td>$x = x_0 - \gamma\, tan\,(\pi(y - 0.5))$</td><td>✓</td></tr>
<tr><td>others … ?</td><td>?</td><td>?</td></tr>
</table>

Unexplored meta-distributions may offer significant additional value.

How much added flexibility does a meta-distribution provide?

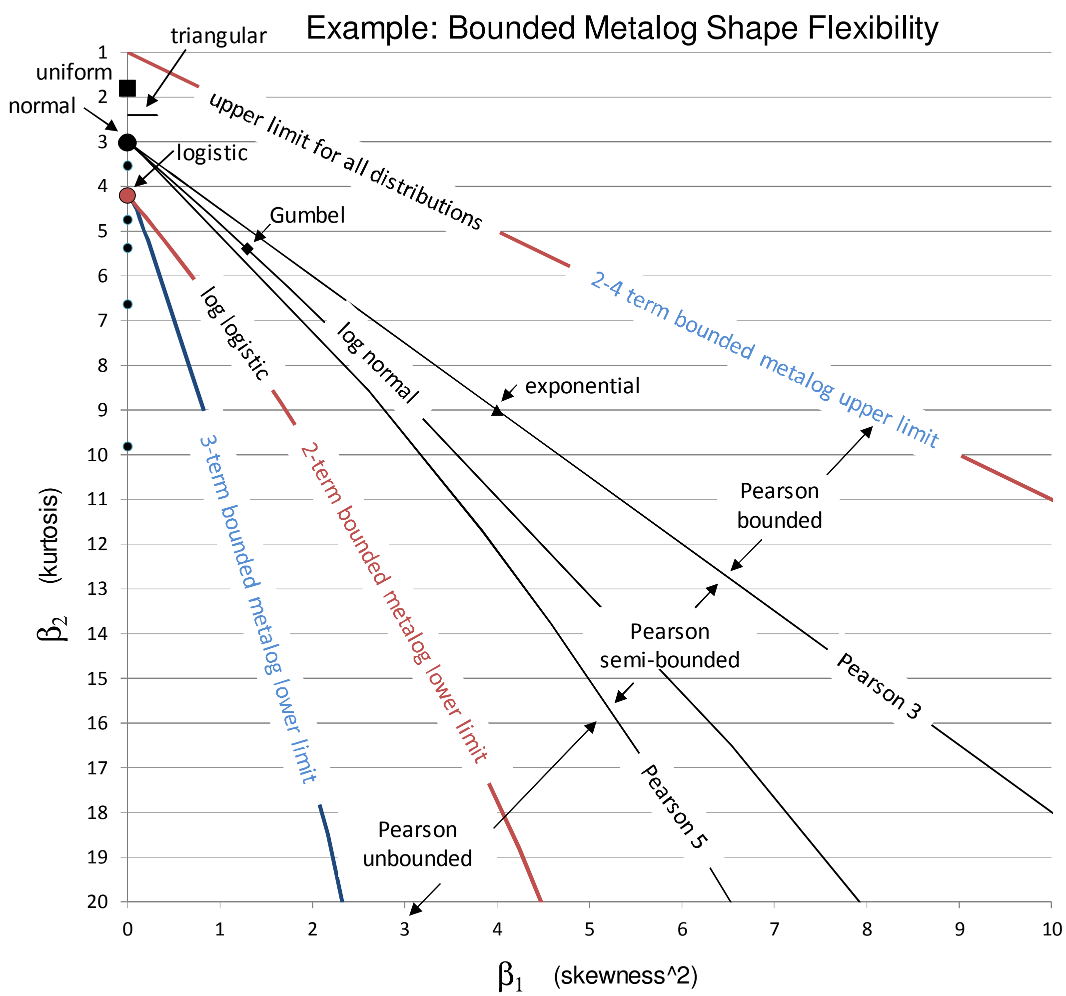


Metalog Topics

- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

Flexibility Comparison: Metalog vs. Pearson Distributions

Flexibility: Metalog flexibility expands with number of terms



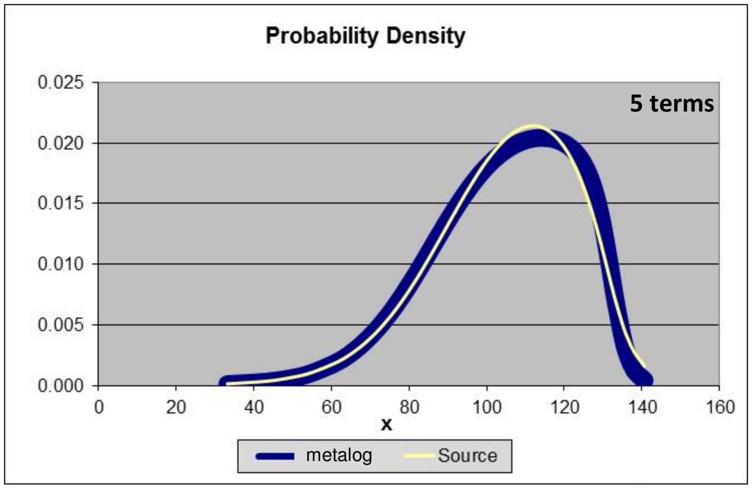
Other Relative Strengths:

- Unlimited shape parameters
- For many areas of (β_1, β_2), the metalog offers
  - ❖ choice of boundedness
  - ❖ ability to match 5th and higher-order moments
- 3 functional forms (one each for unbounded, semi-bounded, and bounded)
- Linear quantile parameterization

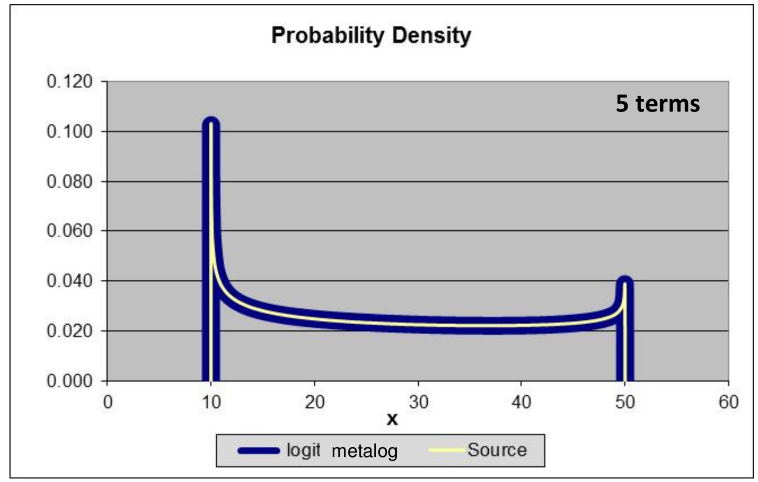
Caveat:

- Certain very extreme distributions (e.g. Cauchy with infinite moments) require transformation in order to enable a good metalog representation.

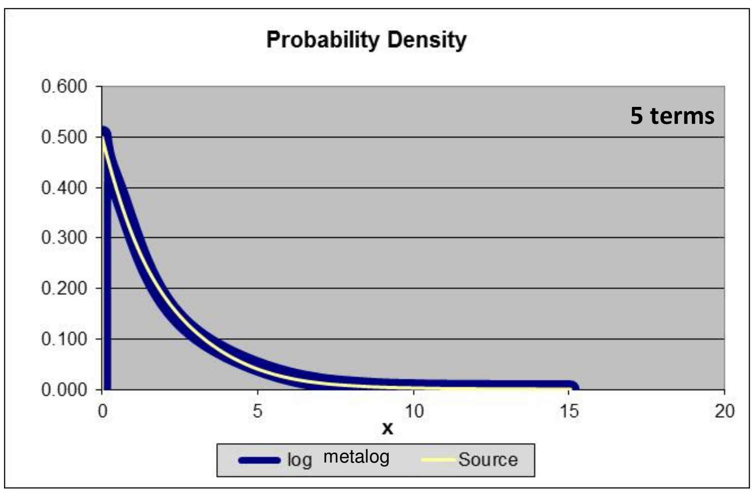
Metalogs can effectively represent a wide range of traditional distributions.



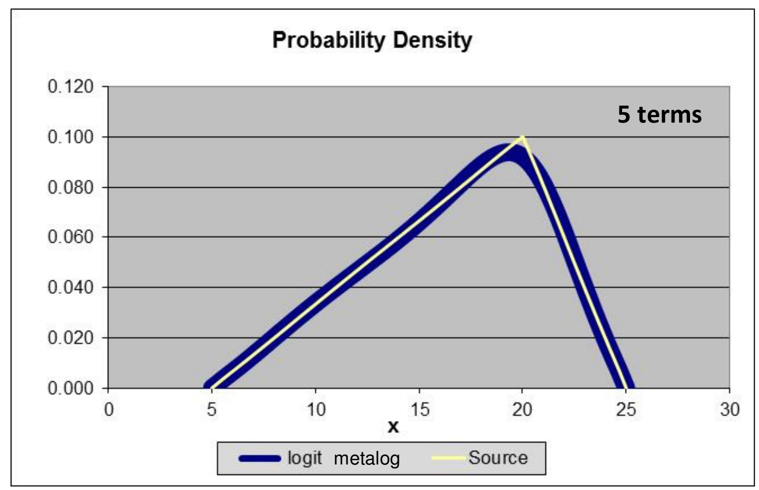
Source: extreme value ($\mu = 100, \sigma = 20, eta = -0.5$)



Source: beta ($\alpha = 0.8$, $\beta = 0.9$, $b_l = 10, bh = 50$)

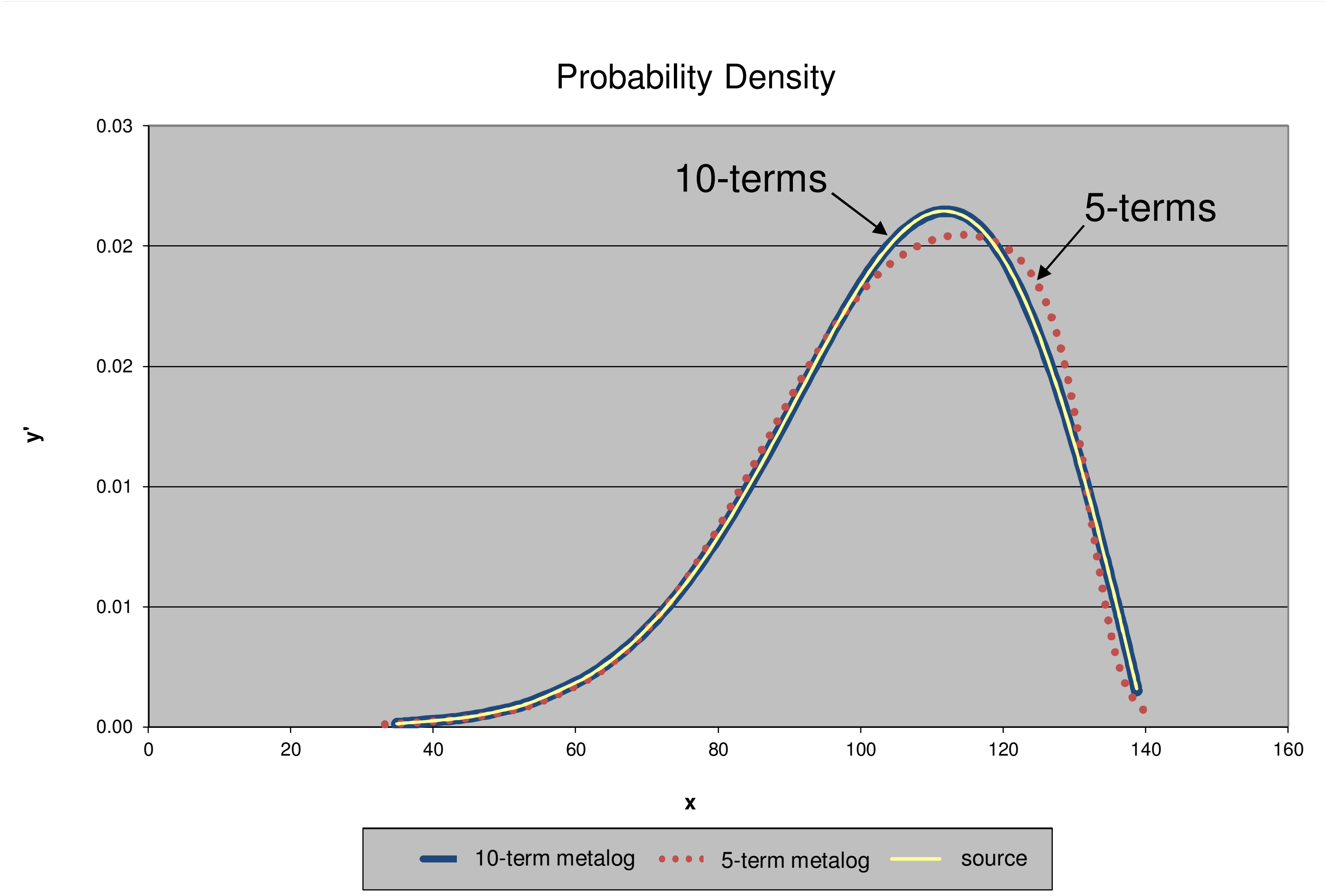


Source: exponential ($\lambda = 0.5$)



Source: triangular ($mode = 20, bl = 10, b_h = 50$)

Metalog representations are increasingly accurate with increased numbers of terms.

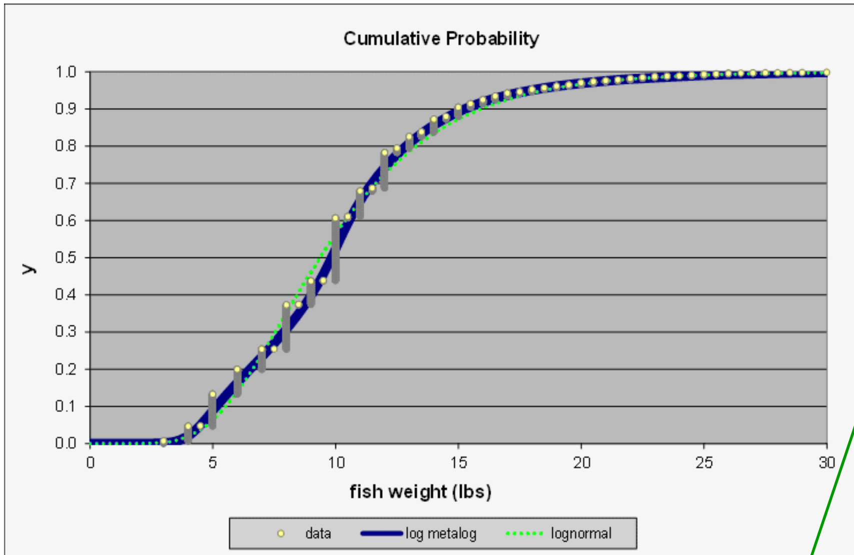


Source: extreme value ($\mu = 100, \sigma = 20, eta = -0.5$)

Metalog Topics

- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

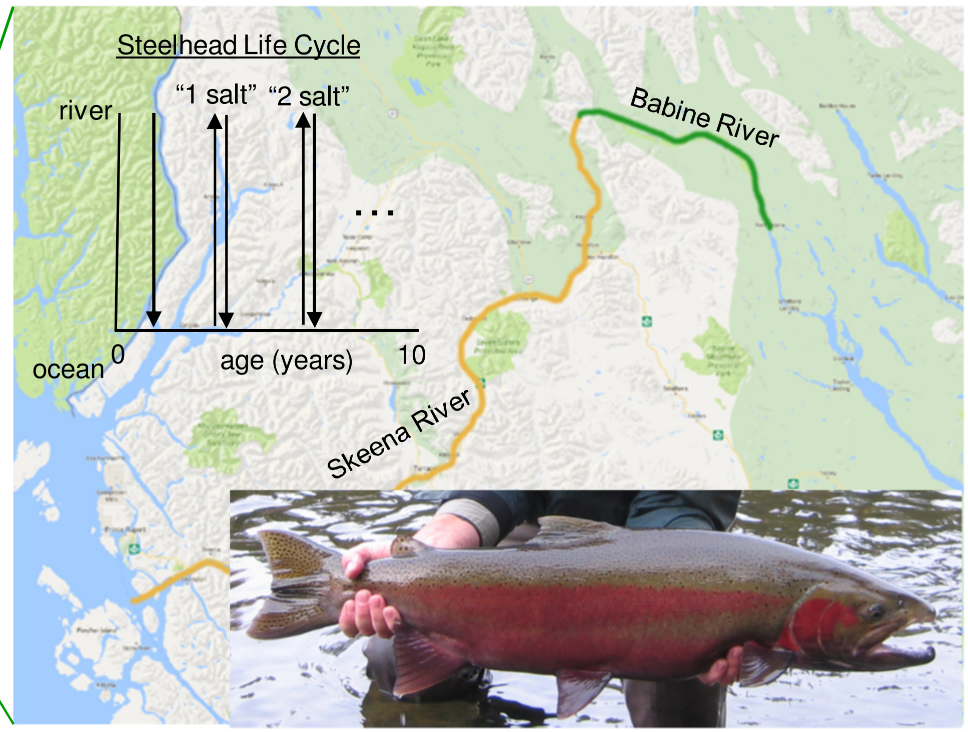
By enabling the data to "speak for itself," metalogs can transform data into knowledge.

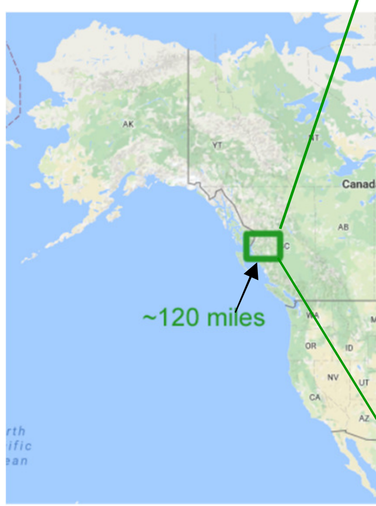


Steelhead Trout Weight (lbs)

3,474 catch-and-release fish records 2010-2014. Babine River, British Columbia.

"1 salt" vs. "2 salt" fish-biology research questions:
fish weights (relative and absolute)?
relative population sizes?



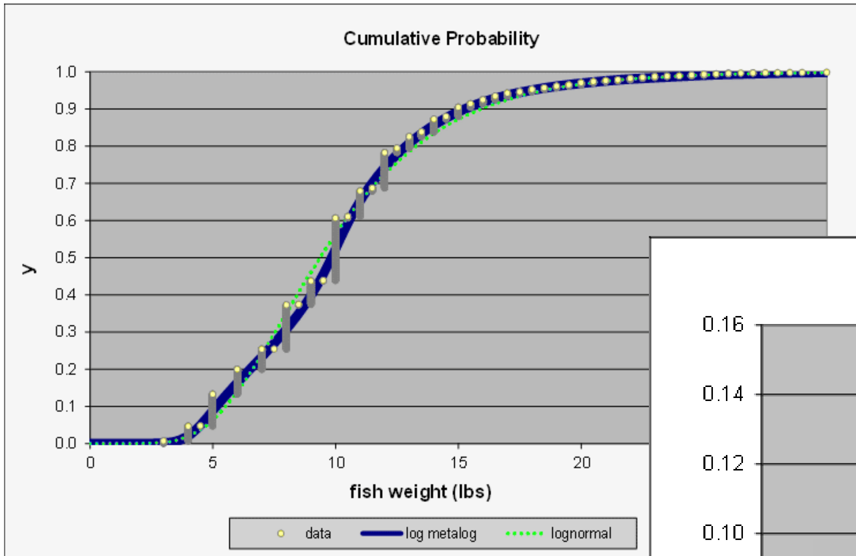


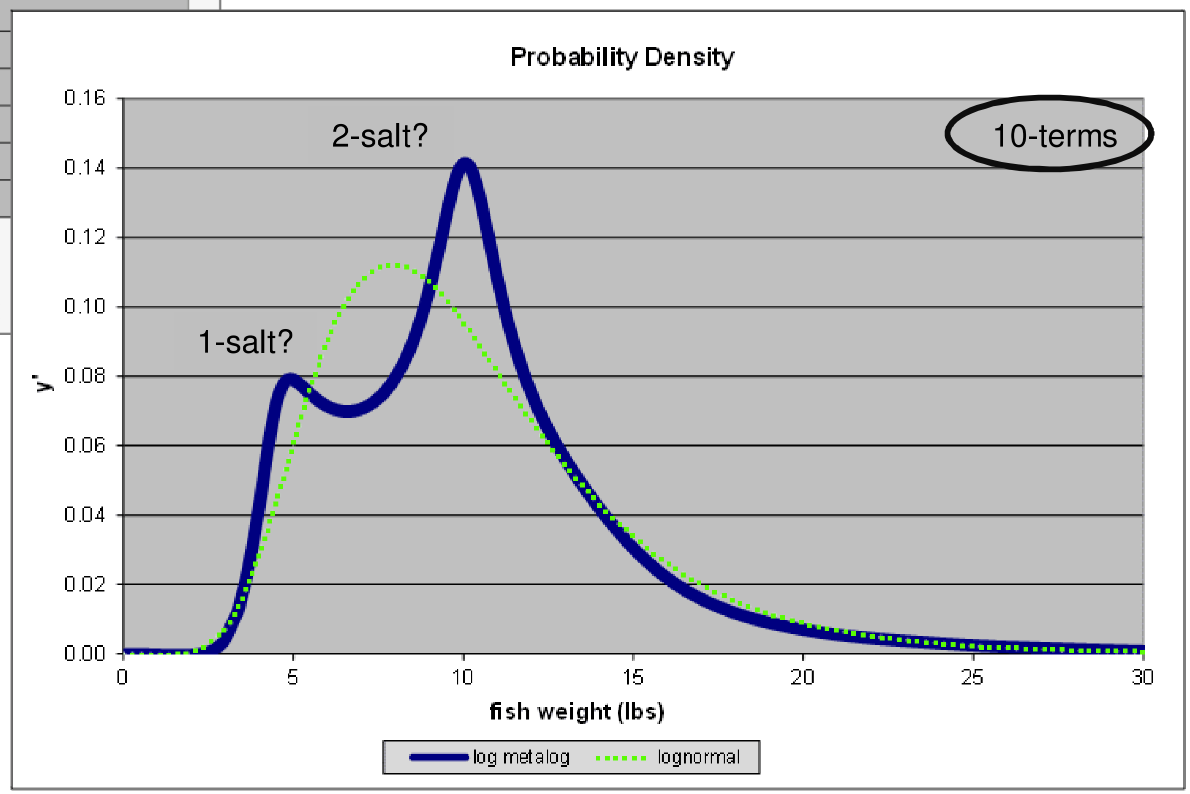
Silver Hilton Steelhead Lodge

By enabling the data to "speak for itself," metalogs can transform data into knowledge.

Steelhead Trout Weight (lbs)

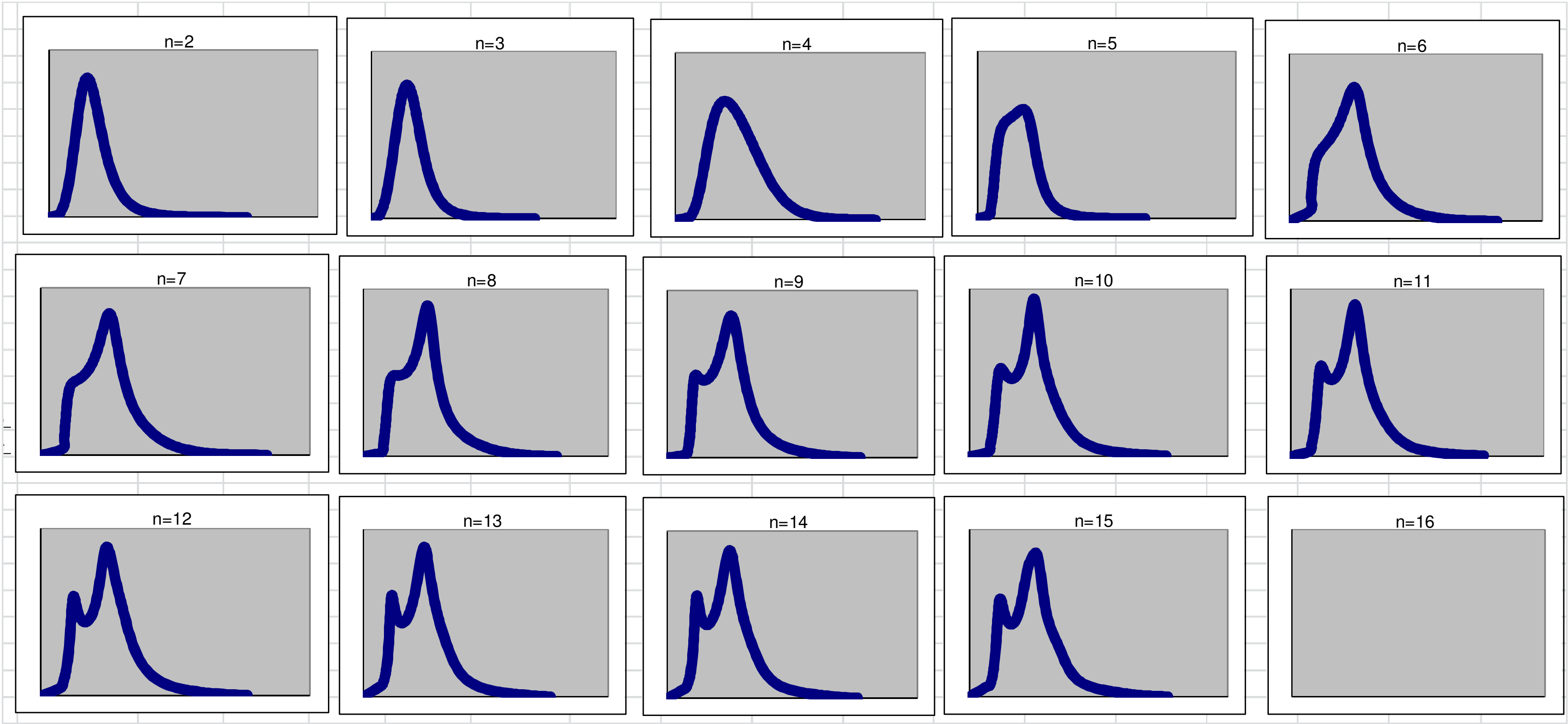
3,474 catch-and-release fish records 2010-2014. Babine River, British Columbia.





Metalog family molds itself to the data -- potentially telling a more nuanced story than previous Type III families.

Metalog Panel for Fish Biology Data (n = 2-16 terms)



Metalog Topics

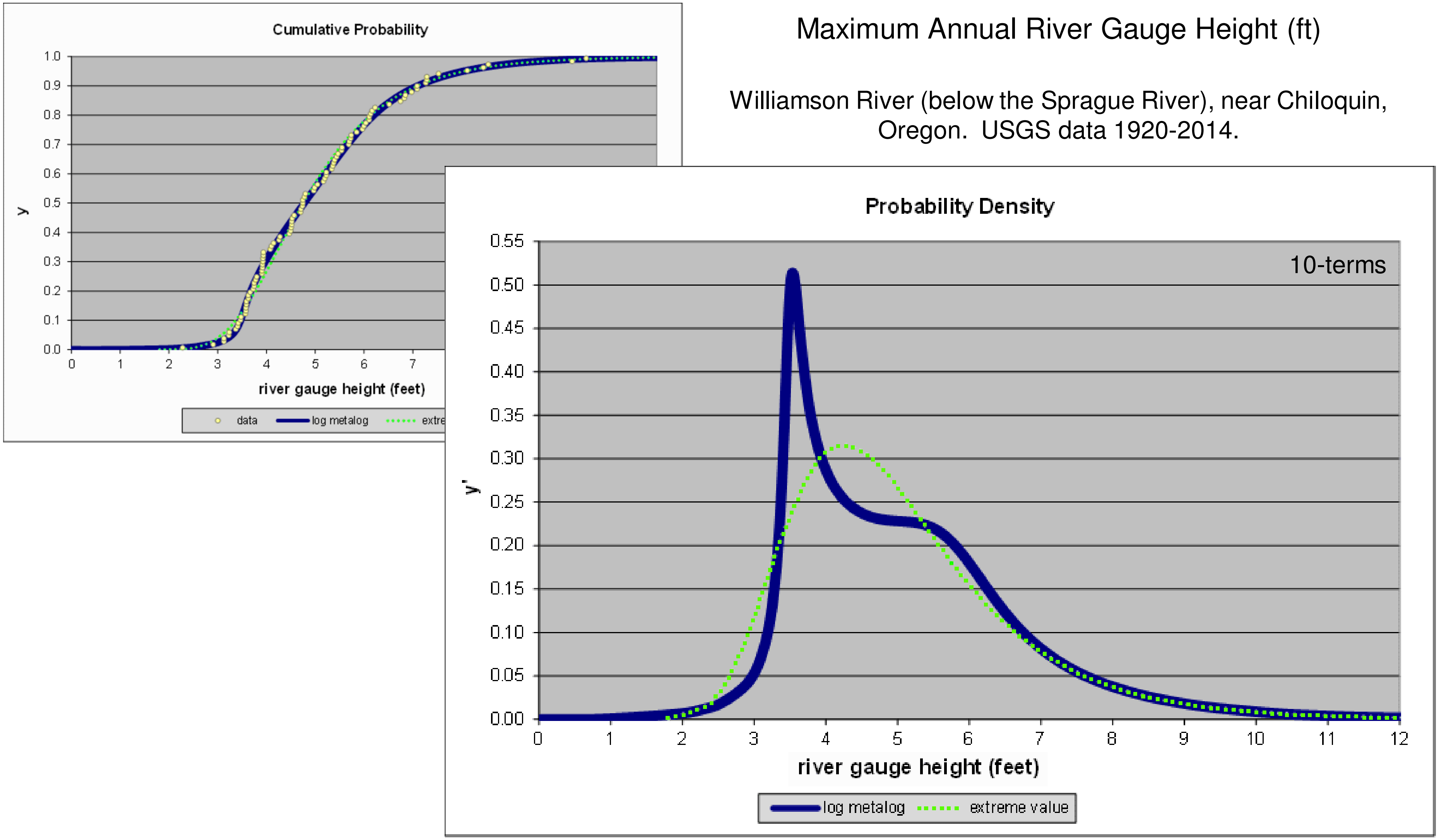
- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

Maximum-annual-river-gauge-height probability distribution?

Metalogs enable examination of whether the “shape of the data” is consistent with a given Type I model.

Maximum Annual River Gauge Height (ft)

Williamson River (below the Sprague River), near Chiloquin, Oregon. USGS data 1920-2014.



Metalog Topics

- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

In 2010, my partners and I faced a major decision: how much to bid in public auction for a pool of 1,456 loans from 16 failed banks.

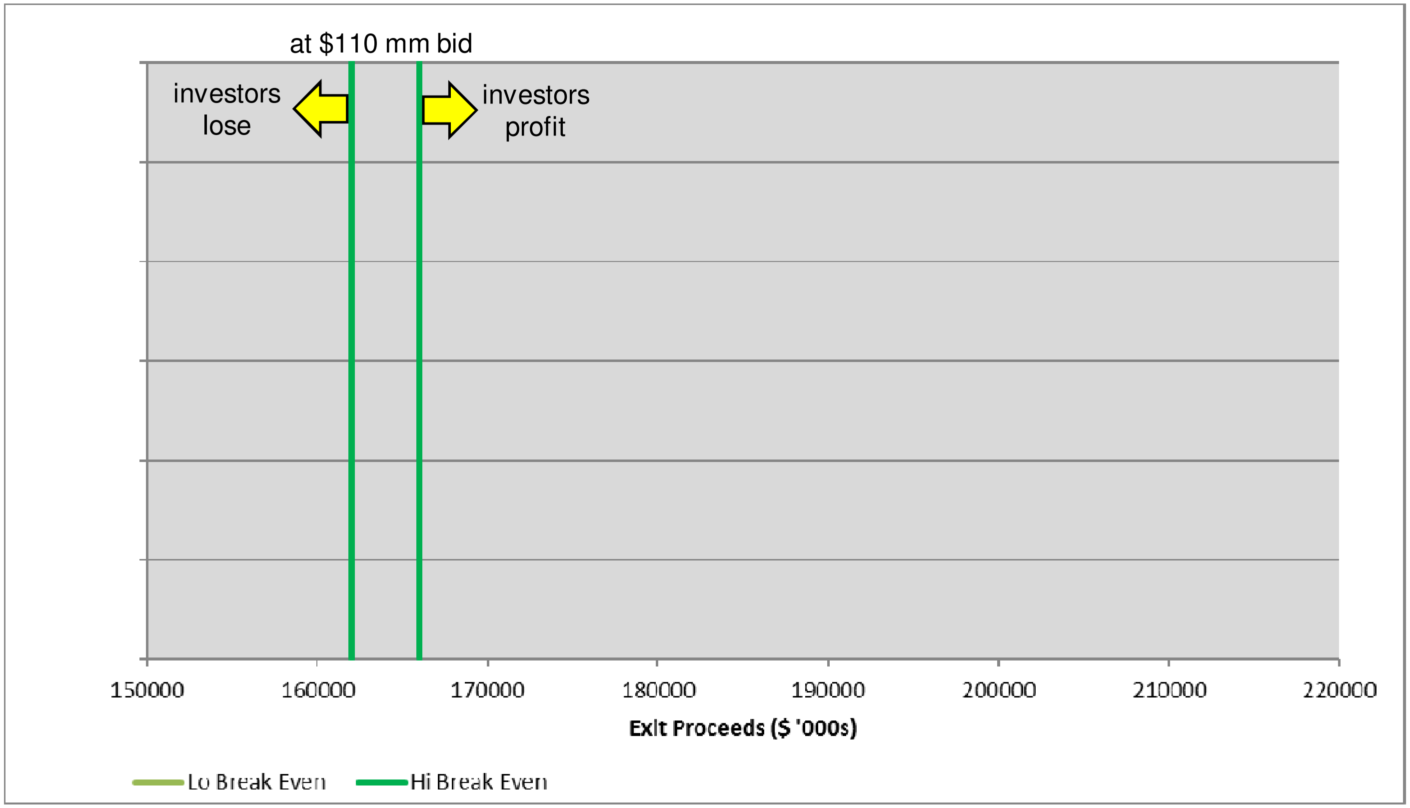
16 Failed Banks

Bank Name	City	State	Closing Date
IndyMac Federal Bank, FSB	Pasadena	CA	July 11, 2008
New Frontier Bank	Greeley	CO	April 10, 2009
American Southern Bank	Kennesaw	GA	April 24, 2009
First Bank of Beverly Hills	Calabasas	CA	April 24, 2009
First Bank of Idaho	Ketchum	ID	April 24, 2009
Michigan Heritage Bank	Farmington Hills	MI	April 24, 2009
America West Bank	Layton	UT	May 1, 2009
Citizens Community Bank	Ridgewood	NJ	May 1, 2009
Silverton Bank, N.A.	Atlanta	GA	May 1, 2009
Westsound Bank	Bremerton	WA	May 8, 2009
Bank of Lincolnwood	Lincolnwood	IL	June 5, 2009
Community Bank of West Georgia	Villa Rica	GA	June 26, 2009
Integrity Bank	Jupiter	FL	July 31, 2009
Community Bank of Nevada	Las Vegas	NV	August 14, 2009
Union Bank, National Association	Gilbert	AZ	August 14, 2009
Corus Bank, N.A.	Chicago	IL	September 11, 2009

FDIC Pool 2010-2

- Total number of loans – 1,456
  - Unpaid Balance (UPB) - $313,848,054
- 1st Liens – 855
  - UPB - $ 261,983,734
  - Performing – 435
  - Non-Performing – 422
- 2nd Liens – 605
  - UPB - $51,864,320
  - Performing – 462
  - Non Performing – 143
- More than 200 cross-collateralized loans

Exit proceeds was the only critical uncertainty, but it was very critical.



Challenge: How to develop the probability distribution over exit proceeds.

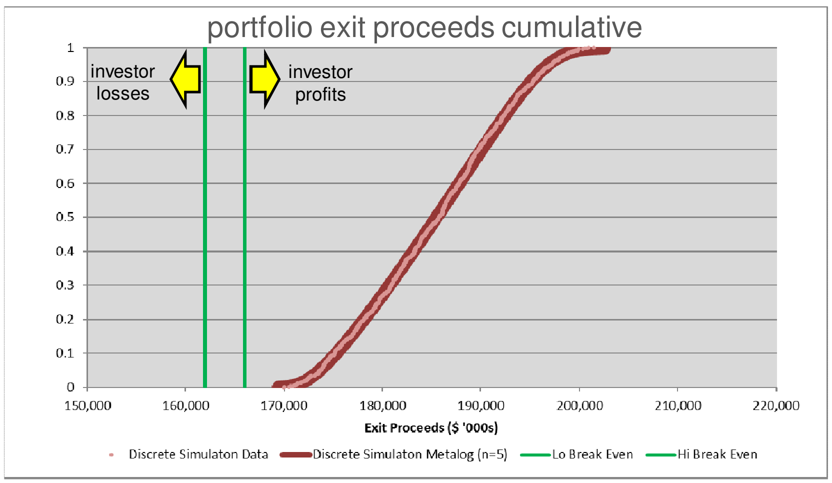
We proceeded much as any good decision analyst would do …

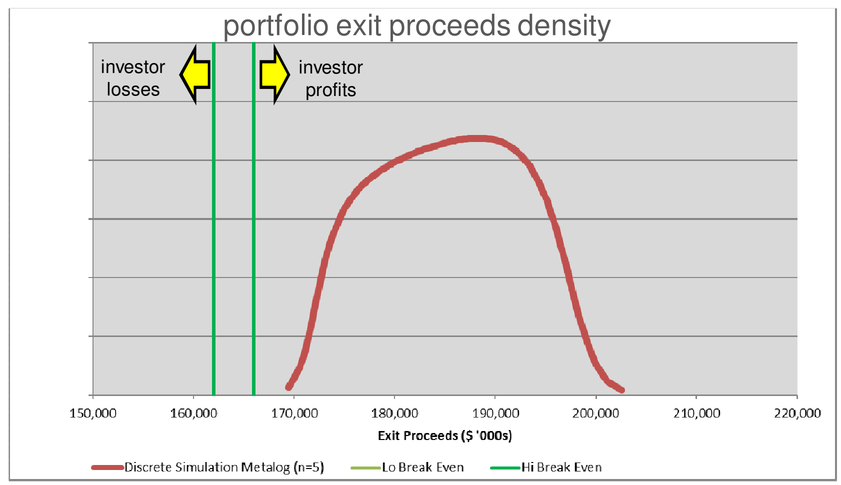
1,456 loans → 259 "asset" assessments

Asset		10%		50%		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		

→

259 discrete uncertainties (correlated with market)





Simulation was the tool of choice.

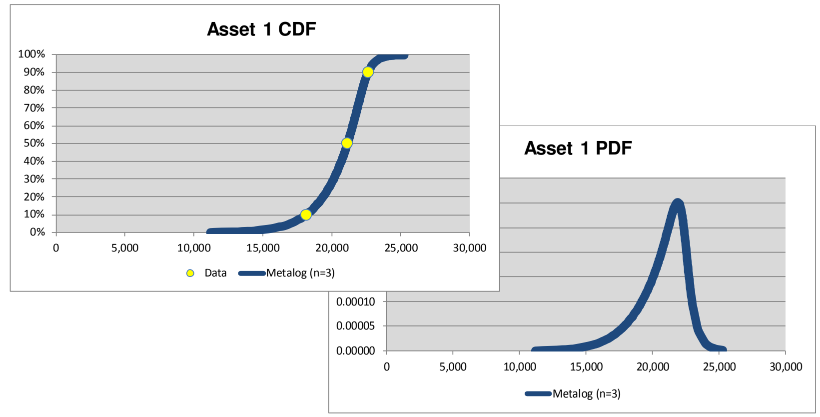
Redoing the same analysis with continuous uncertainties led us to a different and better decision.

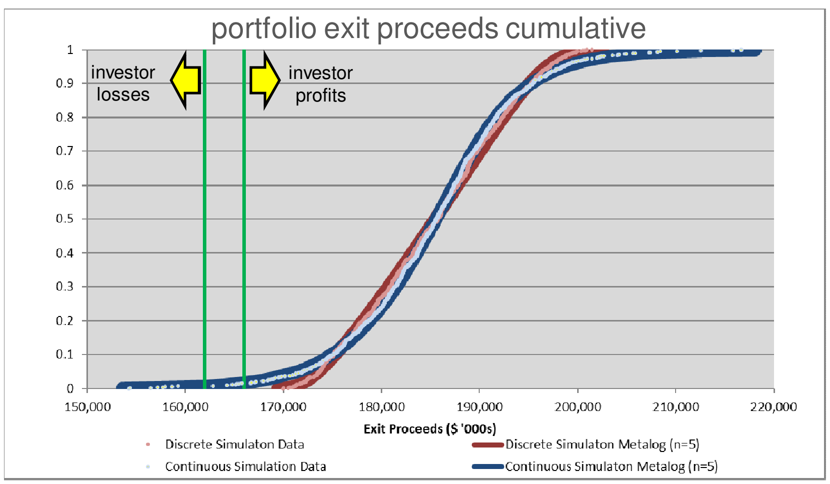
1,456 loans → 259 "asset" assessments

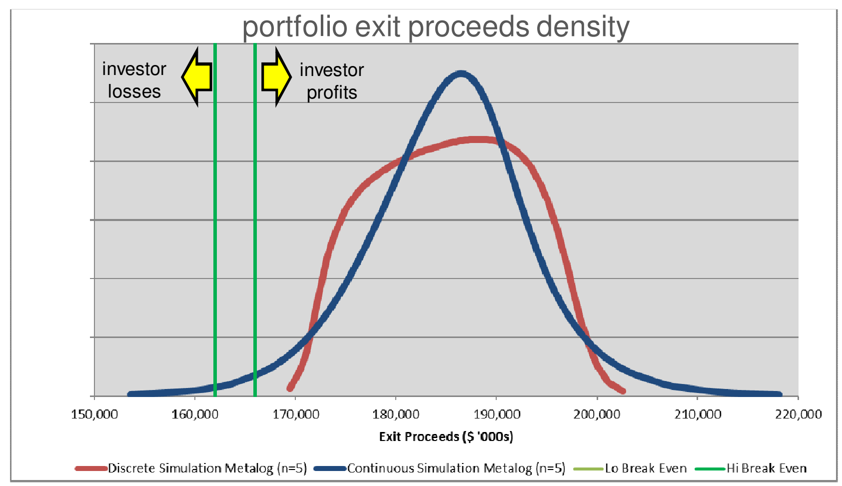
Asset		10%		50%		90%
		probability that realized value is less than ...				
1	$	18,150	$	21,133	$	22,625
2	$	10,465	$	11,362	$	12,408
3	$	15,781	$	16,908	$	18,260
4	$	4,234	$	4,422	$	4,610
5	$	2,629	$	2,979	$	3,295
6	$	13,945	$	14,875	$	16,176
⋮		⋮		⋮		⋮
259	$	3,500	$	4,000	$	4,500
Total			$	185,348		

→

259 continuous uncertainties (correlated with market)



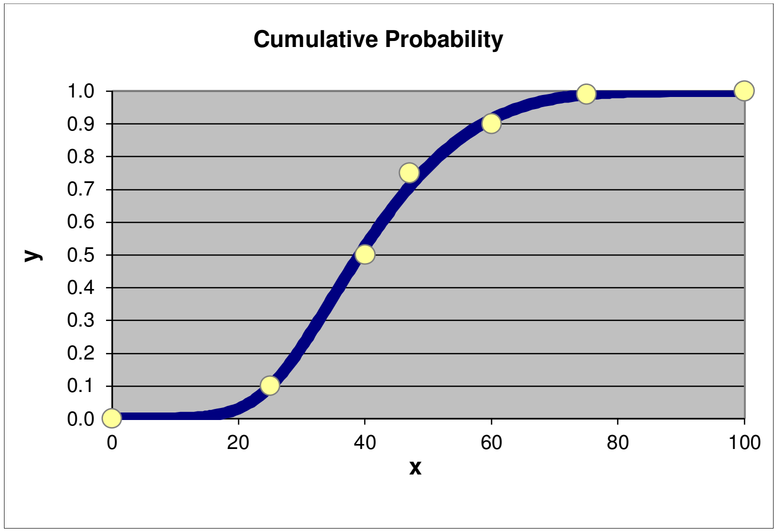


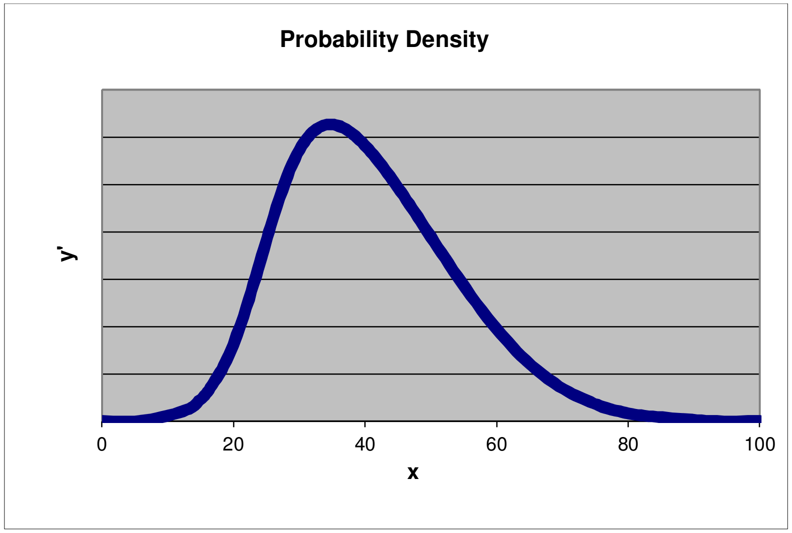


Discrete analysis artificially cut off the tails. If we had believed that analysis, we would have made a wrong decision.

Separately, metalogs can aid expert assessments by providing real-time representations and feedback.

Metalogs enable virtually any shape and can provide real-time feedback as each point is added.





This works for any number of data parameters (including inconsistent ones).

Metalog Topics

- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

Approaches to Characterizing Continuous Multivariate Distributions

A. Decompose with conditional conditional probability

- Easy in concept: {x,z|&} = {z|x,&} {x|&}
- Traditionally difficult in practice:
  - Picking marginal and conditional distributions with sufficient shape and bounds flexibility
  - Conceptualizing how parameters of {z|x,&}, such as standard deviation, skewness, α, β, etc., vary as a function of x for *all* x.

metalogs have practically unlimited shape and bounds flexibility …

… and enable relevance assessments directly in terms of the variable of interest

B. Couple marginal distributions (copulas) with correlation coefficients

- Couple marginal distributions directly (Winkler, et. al, Copulas in Decision Analysis, Decision Analysis, …)
- Simulation of marginal distributions from correlated uniform distributions (correlation accomplished by computing inverse CDF's from the bivariate normal with a given correlation coefficient)
- Difficult in practice: correlation coefficient assessments are
  - A blunt instrument, not clearly interpretable
  - All but impossible for three or more mutually relevant uncertainties

bidding decision example

Let's consider the joint distribution over the future sales prices of a real estate portfolio. (I)

{Shotwell, 24^{th}, Mission, Ashbury, Peters, Minna, Haight | &}

Portfolio manager: "Many decisions (sell now vs. hold vs. exchange) depend critically on the joint distribution over 2023 sales prices of our properties."

Assessment Question	Response
1. How would you think about our range of uncertainty over 2023 selling prices for Shotwell?	It depends on what happens at Shotwell and overall San Francisco market conditions.
2. Assuming your median forecast for 2023 market conditions, what's your 10%, 50%, 90% range for Shotwell selling price?	\$4.9 mm, \$5.3 mm, \$5.8 mm
3. Same question for the other six properties …	…
4. Given median market conditions in 2023, how, if at all, would knowing that one property sold for a high or low price affect your assessments for the other properties.	Not at all.
5. What's your 10%, 50%, 90% range over 2023 market relative to your forecast?	-20%, 0, +20%
6. If you knew that 2023 market would be x%, how, if at all, would you adjust your answers to Question 2 for Shotwell?	I'd multiply all three by k = (1+x%).
7. Would you do this also for the other six properties?	Yes, for all except Haight.
8. What's special about Haight and how would you adjust its assessments for various market outcomes?	Haight has less downside in bad markets. If x% < 0, I'd multiply by k = (1+x%/2) and by k= (1+x%) otherwise.

Let's consider the joint distribution over the future sales prices of a real estate portfolio. (II)

	3-term metalog parameters							
		sale price uncertainty conditional on market						
	market	Shotwell	24th	Mission	Ashbury	Peters	Minna	Haight
bounds	u	u	u	u	u	u	u	u
y_1	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.05
y_2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
y_3	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.95
x_1	-20%	4.9*k	3.9*k	3.9*k	5.0*k	5.0*k	7.2*k	23.0*k
x_2	0%	5.3*k	4.3*k	4.3*k	5.6*k	5.3*k	7.6*k	30.0*k
x_3	20%	5.8*k	4.8*k	4.8*k	6.2*k	5.6*k	8.0*k	35.0*k

where k = (1+market) for all cases except that k = (1+market/2) for Haight when market<0

metalog parameters change with market outcome

y	quantiles	implied quantiles for median market (market = 0%)						
0.01	-42	4.5	3.5	3.5	4.3	4.7	6.8	18.9
0.10	-20	4.9	3.9	3.9	5.0	5.0	7.2	24.9
0.50	0	5.3	4.3	4.3	5.6	5.3	7.6	30.0
0.90	20	5.8	4.8	4.8	6.2	5.6	8.0	33.8
0.99	42	6.4	5.4	5.4	6.9	5.9	8.4	37.7

implied quantiles reviewed and validated

The joint distribution over market and the seven properties is now fully determined.

The joint distribution may be expressed either analytically or as an Outcomes Table.

{market, Shotwell, 24th, Mission, Ashbury, Peters, Minna, Haight | &}

Analytic Expression

$= M_3^{-1}(\text{market} \mid \mathbf{x} = (-20\%,0\%,20\%), \mathbf{y} = (.1,.5,.9),\&)$
$* \prod_{\text{properties}} M_3^{-1}(\text{property} \mid \mathbf{x}_{\text{property}}{*}k, \mathbf{y}_{\text{property}}, \&)$ *joint cumulative*

$= m_3^{-1}(\text{market} \mid \mathbf{x} = (-20\%,0\%,20\%), \mathbf{y} = (.1,.5,.9),\&)$
$* \prod_{\text{properties}} m_3^{-1}(\text{property} \mid \mathbf{x}_{\text{property}}{*}k, \mathbf{y}_{\text{property}}, \&)$ *joint density*

where M_3 and m_3 are well-defined, fully-parameterized, continuous functions in closed form.

Outcomes Table

Aka: realizations array, SLURP (Sam Savage)

simulation number t	market	sales price ($ mm) Shotwell	24th	Mission	Ashbury	Peters	Minna	Haight
1	14%	6.1	4.9	5.0	6.4	6.3	9.1	27.4
2	32%	7.6	5.9	5.7	7.1	7.1	9.8	42.4
3	24%	7.4	5.5	6.1	7.1	6.5	9.1	36.2
4	15%	6.6	4.9	4.7	7.0	5.9	8.8	36.4
5	13%	6.4	5.1	4.6	6.1	5.8	8.9	37.1
6	30%	6.1	5.6	5.4	7.7	7.0	9.9	44.1
7	-3%	5.2	4.0	4.0	5.8	5.4	7.4	24.9
8	-8%	5.2	3.6	3.9	5.2	5.1	6.9	28.8
9	8%	5.9	5.5	4.4	6.9	6.0	9.0	33.9
10	13%	6.5	5.0	5.1	5.6	5.7	9.1	31.7
995	-15%	4.4	3.7	3.8	4.6	4.5	6.0	24.6
996	6%	5.7	4.0	4.7	6.0	5.6	8.3	26.3
997	7%	5.5	4.1	4.5	5.9	5.5	8.3	30.1
998	-5%	5.1	4.4	3.7	4.9	4.6	7.4	28.7
999	12%	6.2	6.5	4.6	6.2	5.4	8.4	35.3
1000	-21%	5.0	3.3	3.5	3.8	4.2	6.4	26.1

Difficulties: many questions of interest are difficult or intractable

- marginals: {Shotwell | &} = ∫ {Shotwell | market, &} {market | &}
- portfolio: sum of mutually relevant sales prices over the properties

$$\int\int\int \dots \int\int\int \textit{joint desnity}$$

- conditionals: portfolio conditional on market
- intuitive continuous representations and closed-forms for all the above

Solution to these difficulties

Outcomes Table
+
additional metalog distributions

How does one form an Outcomes Table?

A. Gather data empirically, or

B. Simulate using uniformly-distributed, mutually irrelevant random numbers y_j:

- Calculate $\text{market}_1 = M_3(y_1, \mathbf{x} = (-20\%, 0\%, 20\%), \mathbf{y} = (.1, .5, .9), \&)$ with the first random number y_1
- Given market_1 outcome, update parameters of M_3 for all properties
- With random numbers $y_2, \ldots, y_8$, calculate sales price outcome = $M_3(y_j \mid \mathbf{x}_{\text{property}}{*}k, \mathbf{y}_{\text{property}}, \&)$ for the seven properties
- Record results and repeat 1-3 with different sets of random numbers enough times (e.g. 1,000) to yield a probabilistic representation that's *equivalent* to the analytic expression.

Marginals are easy to calculate and interpret.

Outcomes Table
(discrete, relevance preserved)

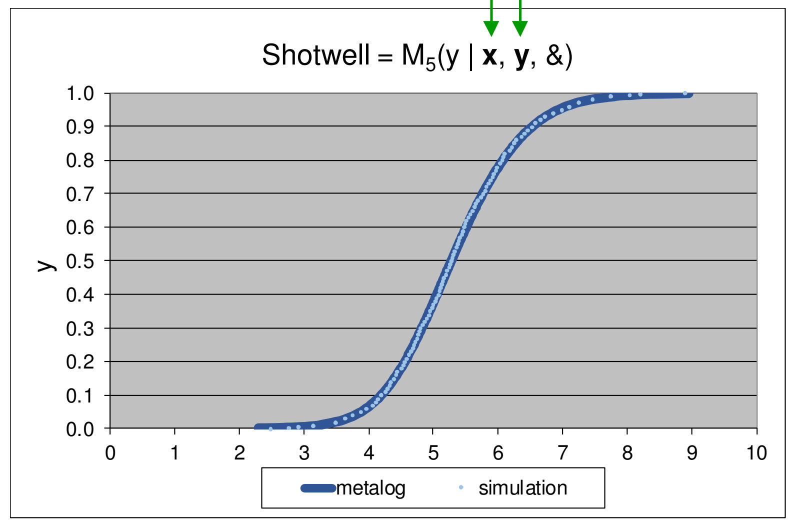
simulation number t	market	sales price ($ mm) Shotwell	24th	Mission	Ashbury	Peters	Minna	Haight
1	14%	6.1	4.9	5.0	6.4	6.3	9.1	27.4
2	32%	7.6	5.9	5.7	7.1	7.1	9.8	42.4
3	24%	7.4	5.5	6.1	7.1	6.5	9.1	36.2
4	15%	6.6	4.9	4.7	7.0	5.9	8.8	36.4
5	13%	6.4	5.1	4.6	6.1	5.8	8.9	37.1
6	30%	6.1	5.6	5.4	7.7	7.0	9.9	44.1
7	-3%	5.2	4.0	4.0	5.8	5.4	7.4	24.9
8	-8%	5.2	3.6	3.9	5.2	5.1	6.9	28.8
9	8%	5.9	5.5	4.4	6.9	6.0	9.0	33.9
10	13%	6.5	5.0	5.1	5.6	5.7	9.1	31.7
995	-15%	4.4	3.7	3.8	4.6	4.5	6.0	24.6
996	6%	5.7	4.0	4.7	6.0	5.6	8.3	26.3
997	7%	5.5	4.1	4.5	5.9	5.5	8.3	30.1
998	-5%	5.1	4.4	3.7	4.9	4.6	7.4	28.7
999	12%	6.2	6.5	4.6	6.2	5.4	8.4	35.3
1000	-21%	5.0	3.3	3.5	3.8	4.2	6.4	26.1

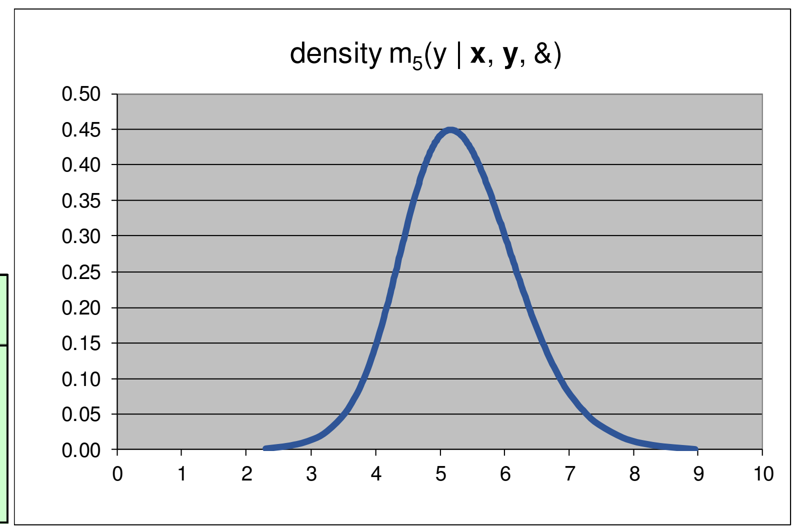
Shotwell y	x
0.0005	2.3
0.0015	2.5
0.0025	2.6
0.0035	2.8
0.0045	2.8
0.0055	2.9
0.0065	2.9
0.0075	2.9
0.0085	3.0
0.0095	3.1
0.9945	8.2
0.9955	8.2
0.9965	8.2
0.9975	8.6
0.9985	8.9
0.9995	9.2

y_t = (t-0.5)/1000 (equally likely)

sort

marginal: {Shotwell | &}





y	Shotwell quantiles median market	unconditional
0.01	4.5	3.2
0.10	4.9	4.2
0.50	5.3	5.3
0.90	5.8	6.5
0.99	6.4	7.8

Any multivariate change of variable is easy to calculate and interpret.

portfolio selling price is sum over properties

Certain equivalent is easy to calculate

risk tolerance
10
e-value of u-value
-0.004
certain equivalent
56.2

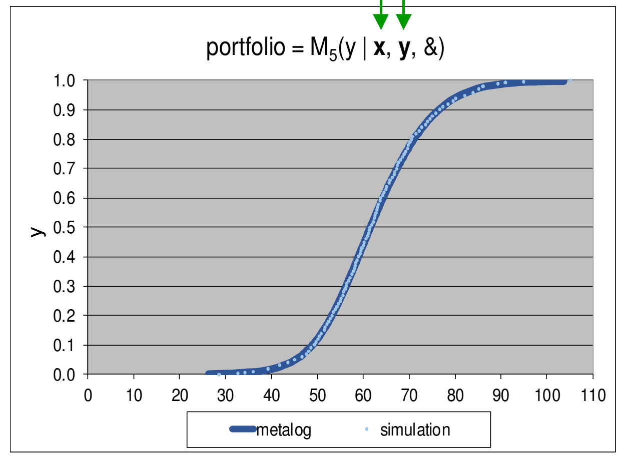
e-value
62.1

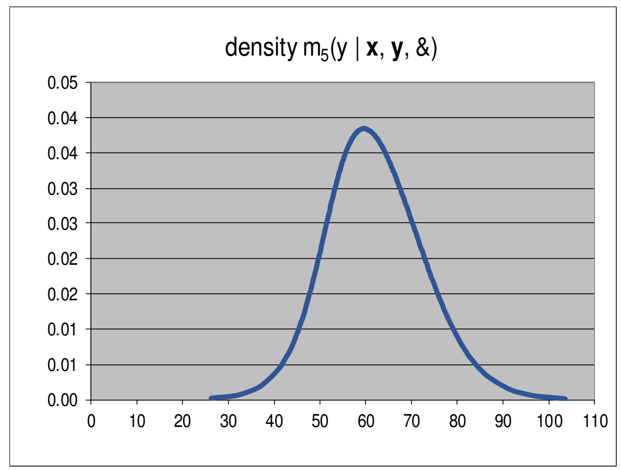
Outcomes Table
(discrete, relevance preserved)

simulation number t	market	sales price ($ mm) Shotwell	24th	Mission	Ashbury	Peters	Minna	Haight	portfolio ($ mm)	portfolio u-value	portfolio y	x
1	14%	6.1	4.9	5.0	6.4	6.3	9.1	27.4	65.1	-0.001	0.0005	27.8
2	32%	7.6	5.9	5.7	7.1	7.1	9.8	42.4	85.7	0.000	0.0015	28.6
3	24%	7.4	5.5	6.1	7.1	6.5	9.1	36.2	77.9	0.000	0.0025	29.2
4	15%	6.6	4.9	4.7	7.0	5.9	8.8	36.4	74.2	-0.001	0.0035	31.2
5	13%	6.4	5.1	4.6	6.1	5.8	8.9	37.1	74.0	-0.001	0.0045	32.8
6	30%	6.1	5.6	5.4	7.7	7.0	9.9	44.1	85.8	0.000	0.0055	33.0
7	-3%	5.2	4.0	4.0	5.8	5.4	7.4	24.9	56.8	-0.003	0.0065	33.5
8	-8%	5.2	3.6	3.9	5.2	5.1	6.9	28.8	58.6	-0.003	0.0075	34.3
9	8%	5.9	5.5	4.4	6.9	6.0	9.0	33.9	71.7	-0.001	0.0085	34.4
10	13%	6.5	5.0	5.1	5.6	5.7	9.1	31.7	68.6	-0.001	0.0095	36.0
995	-15%	4.4	3.7	3.8	4.6	4.5	6.0	24.6	51.6	-0.006	0.9945	94.4
996	6%	5.7	4.0	4.7	6.0	5.6	8.3	26.3	60.6	-0.002	0.9955	94.9
997	7%	5.5	4.1	4.5	5.9	5.5	8.3	30.1	63.8	-0.002	0.9965	97.1
998	-5%	5.1	4.4	3.7	4.9	4.6	7.4	28.7	58.8	-0.003	0.9975	97.3
999	12%	6.2	6.5	4.6	6.2	5.4	8.4	35.3	72.6	-0.001	0.9985	105.0
1000	-21%	5.0	3.3	3.5	3.8	4.2	6.4	26.1	52.1	-0.005	0.9995	110.1

sorted

{portfolio | &}

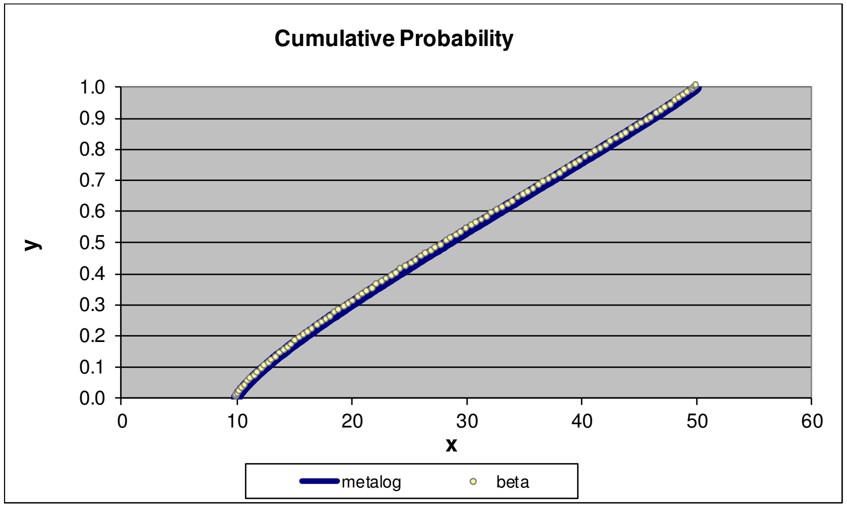


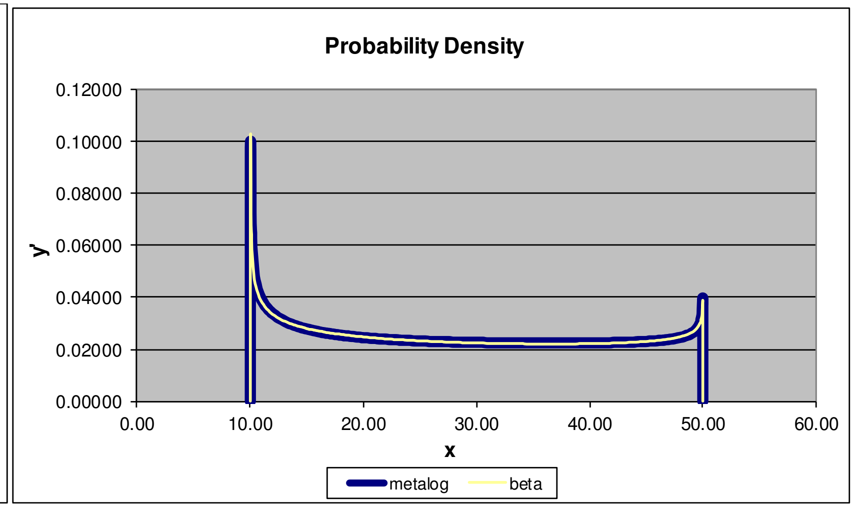


y	portfolio quantiles median market	unconditional
0.01	51.4	37.3
0.10	57.2	49.0
0.50	62.5	61.5
0.90	66.4	76.4
0.99	70.2	90.6

Are these distributions equivalent?

(i.e. can either be legitimately substituted for the other)





beta

y = $b_l + (b_u - b_l)\ I_x(\alpha,\beta)$

(the incomplete beta function)

where $\alpha = 0.8$, $\beta = 0.9$, $b_l = 10$, $b_u = 50$

y' = $\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$

where $\mathrm{B}(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\alpha = 0.8$, $\beta = 0.9$, $b_l = 10$, $b_u = 50$

metalog

$$\mathbf{x} = \frac{b_l + b_u e^{M_n(y)}}{1 + e^{M_n(y)}} \qquad 0 < y < 1$$
$$= b_l \qquad y = 0$$
$$= \qquad y = 1$$

where $b_l = 10$, $b_u = 50$, $M_n(y) = a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3(y-0.5)\ln\left(\frac{y}{1-y}\right) + a_4(y-0.5)$

and

a_1	-0.163
a_2	1.174
a_3	-0.097
a_4	-0.275

$$\mathbf{y'} = m_n(y)\frac{(1+e^{M_n(y)})^2}{(b_u - b_l)e^{M_n(y)}} \quad \text{for } 0 < y < 1,$$

$= 0$ for $y = 0$ or $y = 1$, and where

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

a_1	-0.163
a_2	1.174
a_3	-0.097
a_4	-0.275

Yes -- if the distribution owner declares them to be so.

Process for Developing Multivariate Distributions With Metalogs

1. Define uncertainties z_1, z_2, z_3, … and decompose the joint into a marginal and conditionals convenient for assessment or modeling

   $\{z_1, z_2, z_3, \ldots \mid \&\} = \{z_1 \mid \&\} \{z_2 \mid z_1, \&\} \{z_3 \mid z_2, z_1, \&\} \ldots$

2. Encode the marginal(s) $\{z_1 \mid \&\}$ as a metalog $M_n(z_1; \mathbf{x}_{z1}, \mathbf{y}_{z1})$ – unbounded, semi-bounded, or bounded as appropriate.

3. Select a metalog representation for each conditional uncertainty z_i such that its parameters are expressed as a function of the conditioning variables

   $M_n(z_i; \mathbf{x}_{zi}(z_{i-1}, z_{i-2}, \ldots), \mathbf{y}_{zi}(z_{i-1}, z_{i-2}, \ldots))$

4. Assess or model these parameter functions.

5. Express the implied joint distribution as an Outcomes Table.

6. Explore any desired marginals, conditionals, and/or multivariate changes of variable – using additional metalogs as appropriate to aid interpretation and communication.

7. Fine tune and validate with the distribution owner.

Metalog Topics

- Historical context
- Equations, parameters, and properties
- Theoretical development
- Shape flexibility compared to prior distributions
- Applications
  - Fish biology
  - Hydrology
  - Decision analysis
- Multivariate metalogs
  - Assessment protocol
  - Real estate portfolio
- Conclusions

Metalogs provide simple, flexible, easy-to-use continuous probability distributions to represent CDF data.

- Allow frequency data to "speak for itself" with highly-flexible continuous representations.
- Select among unbounded, semi-bounded, or bounded distributions
- Skip time-consuming parameter estimation
- Facilitate Monte Carlo Simulation by convenient
  - Sampling from input distributions
  - Representing simulation outputs as smooth, continuous distributions
- Use simple, closed-form equations -- easily-programmable-in-Excel
- Apply in both univariate and multivariate contexts

For Excel workbooks, publications, and supporting information, go to www.metalogdistributions.com

References

Abbas AE (2003) Entropy methods for univariate distributions in decision analysis. Williams C, ed. Bayesian Inference and Maximum Entropy Methods Sci. Engrg.: 22nd Internat. Workshop (American Institute of Physics, Melville, NY), 339–349.

Aldrich J (1997) RA Fisher and the making of maximum likelihood 1912–1922. Statist. Sci. 12(3):162–176.

Balakrishnan N (1992) Handbook of the Logistic Distribution (Marcel Dekker, New York).

Bickel JE, Lake LW, Lehman J (2011) Discretization, simulation, and Swanson's (inaccurate) mean. SPE Econom. Management 30(3):128–140.

Burr IW (1942) Cumulative frequency functions. Ann. Math. Statist. 13(2):215–232.

Cheng R (2011 December) Using Pearson type IV and other Cinderella distributions in simulation. Proc. Winter Simulation Conf., Phoenix, AZ, 457–468.

De Moivre A (1756) The Doctrine of Chances: Or, A Method of Calculating the Probabilities of Events in Play, Vol. 1 (Chelsea Publishing Company, London).

Draper NR, Smith H (1998) Applied Regression Analysis, 3rd ed. (Wiley, New York).

Edgeworth FY (1896) XI. The asymmetrical probability-curve. London, Edinburgh, Dublin Philosophical Magazine J. Sci. 41(249): 90–99.

Edgeworth FY (1907) On the representation of statistical frequency by a series. J. Royal Statist. Soc. 70(1):102–106.

Greenwood JA, Landwehr JM, Matalas NC, Wallis JR (1979) Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. Water Resources Res. 15(5):1049–1054.

Hadlock CC, Bickel JE (2016) Johnson quantile-parameterized distributions. Decision Anal. Forthcoming.

Hawkins DM (2004) The problem of overfitting. J. Chemical Inform. Comput. Sci. 44(1):1–12.

Hosking JR (1990) L-moments: Analysis and estimation of distributions using linear combinations of order statistics. J. Roy. Statist. Soc. Series B (Methodological) 52(1):105–124.

Howard RA (1968) The foundations of decision analysis. Systems Sci. Cybernetics, IEEE Trans. 4(3):211–219.

Howard RA, Abbas AE (2015) Foundations of Decision Analysis (Prentice Hall, Upper Saddle River, NJ).

Johnson NL (1949) Systems of frequency curves generated by methods of translation. Biometrika 36(1/2):149–176.

Johnson NL, Kotz S, 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. Comput. Statist. Data Anal. 51(2): 947–959.

Keelin TW, Powley BW (2011) Quantile-parameterized distributions. Decision Anal. 8(3):206–219.

Keelin, T.W., 2016. The Metalog Distributions. Decision Analysis, 13(4), pp.243-277.

Keeney RL, Raiffa H (1993) Decisions with Multiple Objectives: Preferences and Value Trade-Offs (Cambridge University Press, Cambridge, UK).

McDonald JB, Newey WK (1988) Partially adaptive estimation of regression models via the generalized t distribution. Econometric Theory 4(3):428–457.

McGrayne SB (2011) The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy (Yale University Press, New Haven, CT).

Mead R (1965) A generalised logit-normal distribution. Biometrics 21(3):721–732.

Nagahara Y (1999) The PDF and CF of Pearson type IV distributions and the ML estimation of the parameters. Statist. Probab. Lett. 43(3):251–264.

Ord JK (1972) Families of Frequency Distributions, Vol. 30 (Griffin, London).

Pearson K (1895) Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philos. Trans. Roy. Soc. A 186(January):343–414.

Pearson K (1901) Mathematical contributions to the theory of evolution. X. Supplement to a memoir on skew variation. Philos. Trans. Roy. Soc. A 197(January):443–459.

Pearson K (1916) Mathematical contributions to the theory of evolution. XIX. Second supplement to a memoir on skew variation. Philosophical Trans. Royal Soc. London. Series A, Containing Papers Math. Physical Character 216(January):429–457.

Raiffa H (1968) Decision Analysis: Introductory Lectures on Choices Under Uncertainty (Addison-Wesley, Reading, MA).

Spetzler CS, Stael von Holstein CAS (1975) Exceptional paperprobability encoding in decision analysis. Management Sci. 22(3):340–358.

Spetzler C, Winter H, Meyer J (2016) Decision Quality: Value Creation from Better Business Decisions (John Wiley & Sons, New York).

Tadikamalla PR, Johnson NL (1982) Systems of frequency curves generated by transformations of logistic variables. Biometrika 69(2):461–465.

Theodossiou P (1998) Financial data and the skewed generalized t distribution. Management Sci. 44(12-part-1):1650–1661.

Wang M, Rennolls K (2005) Tree diameter distribution modelling: Introducing the logit logistic distribution. Canadian J. Forest Res. 35(6):1305–1313.

Appendices

Three-term unbounded metalog equations.*

cumulative distribution function (cdf):

$$x = M_3(y) = a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3(y-0.5)\ln\left(\frac{y}{1-y}\right)$$

probability density function (pdf):

$$y' \equiv m_n(y) = \left[\frac{a_2}{y\,(1-y)} + a_3\left(\frac{y-0.5}{y(1-y)} + \ln\left(\frac{y}{1-y}\right)\right)\right]^{-1}$$

where constants a_i are derived from quantile assessments

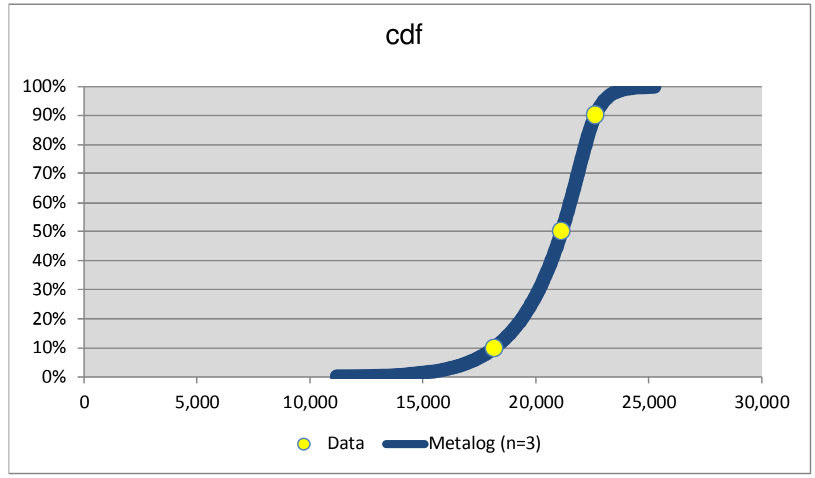
given $\boldsymbol{x} = (q_\alpha, q_{0.5}, q_{1-\alpha})$ then (e.g. for 10-50-90, $\alpha = 0.1$)

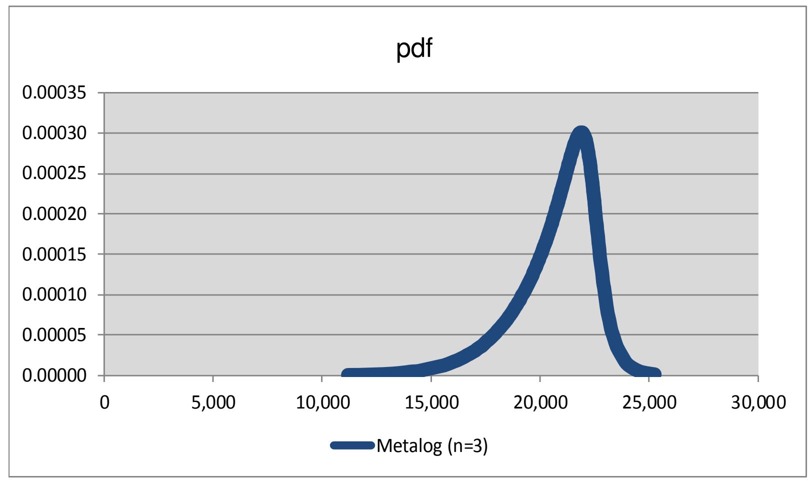
$$a_1 = q_{0.5}$$

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

$$a_3 = \left[(1-2\alpha)\ln\left(\frac{1-\alpha}{\alpha}\right)\right]^{-1}(1-2r)(q_{1-\alpha} - q_\alpha).$$

where $r = \frac{q_{0.5}-q_\alpha}{q_{1-\alpha}-q_\alpha}$.





* For the case where parameters are expressed symmetrically around the median. See definition of SPT (symmetric percentile triplet) in Keelin, 2016.

Personal Journey to a New Family of Probability Distributions (I)

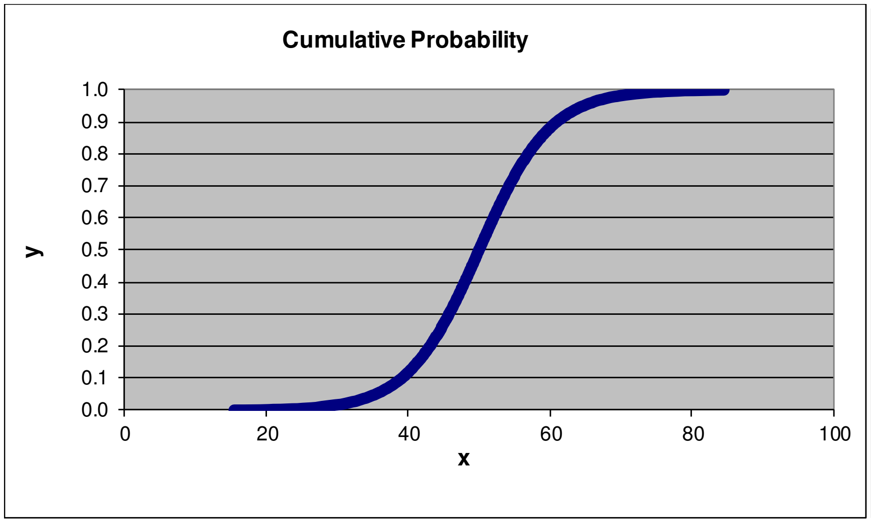
- Early days – learning and experimentation at Stanford
- First 25 years of professional practice (apl, Supertree, Risk Detective, Decision Advisor … reliance largely on others). Continuous distributions were
  - Desirable for smooth representations and density (PDF) displays
  - Impractical (none flexible enough to really "fit" the situation, complex to parameterize, analytically intractable in tree-based tools, no practical way to output PDF displays)
- Founding of KR in 2003 (self reliance for developing DA tools, ended up developing "KR Shell" – using Excel, Crystal Ball, expanded 10-50-90 formats)
  - Simulation solved one problem – making continuous-distribution computations analytically tractable – but did nothing to solve the other problems.
  - 2009 light-bulb moment: why not <u>invent</u> continuous distributions that are practical (simple, flexible, easy and fast to use)?

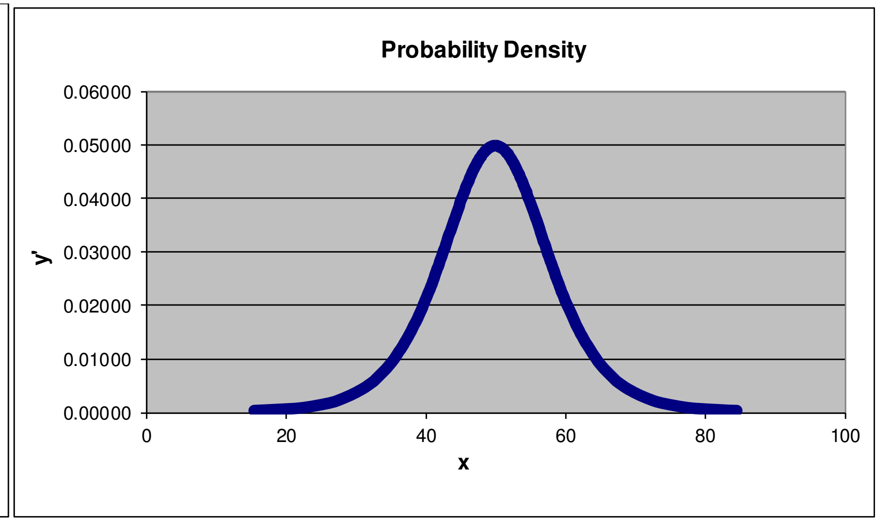
Personal Journey to a New Family of Probability Distributions (II)

- White-board sketches (starting with how to add skewness to the Normal distribution and parameterize it with 10/50/90 assessments) and collaboration with Brad Powley led to
  - “Quantile-Parameterized Distributions” (*Decision Analysis*, Sept 2011)
  - This solved the “difficult-to-parameterize” problem, and -- with the “Simple Q Normal” distribution -- made significant progress toward solving the “lack of flexibility” problem
  - But problems still remained: lack of control over bounds, lack of algebraically-simple closed forms, lack of closed-form moments, need for more flexibility to accurately show PDFs of uncertain inputs <u>and</u> outputs.
- The metalog family of distributions solves all these problems with simplicity, flexibility (unlimited shape parameters), and ease/speed of use (choice of bounds, quantile parameters)
  - A significant improvement over previous families of flexible distributions -- Pearson (1895, 1901, 1916), Johnson (1949), and Tadikamalla and Johnson (1982)
  - Solution to decision problems that tree-based methods can’t solve well … and some other pleasant surprises.

Are these distributions equivalent?

(i.e. can either be legitimately substituted for the other)





$$y = \frac{1}{1+e^{-\frac{x-\mu}{s}}}$$ where $\mu = 50$, $s = 5$

$$y' = \frac{e^{\frac{x-\mu}{s}}}{s\left(1+e^{\frac{x-\mu}{s}}\right)^2}$$ where $\mu = 50$, $s = 5$

Yes, because they are mathematically equivalent.