Proceedings, North American Fuzzy Information Processing Society (NAFIPS 2003), Chicago, pp. 407-412.

Envelopes Around Cumulative Distribution Functions from Interval Parameters of Standard Continuous Distributions

Jianzhong Zhang and Daniel Berleant
Department of Electrical and Computer Engineering
Iowa State University
Ames, Iowa 50014
{zjz, berleant}@iastate.edu

Abstract

A cumulative distribution function (CDF) states the probability that a sample of a random variable will be no greater than a value x, where x is a real value. Closed form expressions for important CDFs have parameters, such as mean and variance. If these parameters are not point values but rather intervals, sharp or fuzzy, then a single CDF is not specified. Instead, a family of CDFs is specified. Sharp intervals lead to sharp boundaries ("envelopes") around the family, while fuzzy intervals lead to fuzzy boundaries. Algorithms exist [12] that compute the family of CDFs possible for some function g(v) where v is a vector of distributions or bounded families of distribution. We investigate the bounds on families of CDFs implied by interval values for their parameters. These bounds can then be used as inputs to algorithms that manipulate distributions and bounded spaces defining families of distributions (sometimes called probability boxes or p-boxes). For example, problems defining inputs this way may be found in [10,12]. In this paper, we present the bounds for the families of a few common CDFs when parameters to those CDFs are intervals.

1. Introduction

Uncertainties are ubiquitous in realistic models. Handling such uncertainty is an important issue in reliable computing. A variety of methods have been developed to deal with this problem [11, 12]. Compared with the traditional method, Monte Carlo, these methods are not subject to noise effects due to randomness that can affect the results obtained from Monte Carlo methods (Ferson 1996 [6]). Such methods offer principled approaches to manipulating uncertain quantities in the presence of 2nd-order uncertainties such as uncertainties in parameters of distributions.

Accurate modeling all too often requires handling the situation that exact distributions are not known, though some information about them is known. To handle this situation, Smith used limited information about distributions to get bounds on the expected value of an arbitrary objective function (1995 [14]). The method is based on moments of distributions. One way to express that information is with interval-valued parameters to standard distributions [10]. Ferson presented some initial results, including examples of envelopes for families of normal distributions defined by interval-valued means and variances, uniform distributions, and Weibull distributions (2003 [7]). The need to formalize and generalize such results helps motivate the present work.

In general, simulation can be adopted to estimate envelopes for distributions with interval parameters. But having CDF envelopes available in closed form can save considerable computation over approximating them when needed using MC simulation. Thus we seek to obtain the left and right envelopes around the family of CDFs for a random variable whose distribution is expressed in closed form with interval parameters.

Then these envelopes can be used to compute envelopes around derived distributions using our Distribution Envelope Determination (DEnv) algorithm or another algorithm [1-5, 8, 12]).

2. Deriving Envelopes Analytically

In order to determine CDF envelopes by analyzing the effect of parameters to the underlying CDF, the core idea is to find the minimum and maximum boundaries, expressed in closed form, for CDFs of random variables when parameter values are specified to be within particular intervals. Then, the curve for the CDF implied by any numerically valued parameters that fall within their respective intervals, will be wholly between those boundaries.

Denote a parameterized CDF with $H(x,\vec{\theta}) = F(x)$ where x is a value of the random variable and $\vec{\theta}$ is a vector of one or more parameters. Assume that each θ_i is not necessarily specified to be a specific numerical value, but instead can be an interval ψ_i. We wish to find the left envelope function

$E_l(x) = \max_{\theta_i \in \psi_i \forall i} H(x,\vec{\theta})$ and the right envelope function $E_r(x) = \min_{\theta_i \in \psi_i \forall i} H(x,\vec{\theta})$.

If $H(x,\vec{\theta})$ is monotonic function about each θ_i, the results are derived as follows. Let $\underline{\theta_i}$ be the minimum value of ψ, and $\overline{\theta_i}$ be the maximum value of ψ. If $H(x,\vec{\theta})$ is non-decreasing, $E_l(x) = H(x,\overline{\theta_1},..., \overline{\theta_I})$ given *I* parameters, and $E_r(x) = H(x,\underline{\theta_1},..., \underline{\theta_I})$. If $H(\theta)$ is non-increasing, $E_l(x) = H(x,\underline{\theta_1},..., \underline{\theta_I})$ and $E_r(x) = H(x,\overline{\theta_1},..., \overline{\theta_I})$.

If $H(x,\vec{\theta})$ is not monotonic, the solution is to partition the domain into regions within which it is monotonic. Different portions of E_l and E_r may derive from different regions and have different functions. In the next section we discuss envelopes which may be derived without partitioning the domain, and in the subsequent section we discuss envelopes for which partitioning is necessary.

3. Envelopes derivable without partitioning

This section gives envelopes for a few common distributions for which the values of the parameters that lead to envelopes whose functions do not depend on the value of the distribution's argument *x*. We first discuss how to get the envelopes for exponential distributions. Then we give the results for uniform and triangular distributions.

3.1 The exponential distribution

The density function of an exponential distribution is $f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$ if $x \geq 0$, parameterized with $\beta > 0$.

From the density function, we can get the cumulative probability function by integrating the density function.

$$F(x) = \int_{-\infty}^{0} f(t)dt = \int_{0}^{x} \frac{1}{\beta} e^{-\frac{t}{\beta}} dt = \int_{0}^{x/\beta} \frac{1}{\beta} e^{-y} d(\beta y)$$

$$= \int_{0}^{x/\beta} e^{-y} dy = -e^{-y} \Big|_{0}^{x/\beta} = -e^{-\frac{x}{\beta}} - (-e^{-0}) = 1 - e^{-\frac{x}{\beta}}$$

if $x \geq 0$.

Next we will show how this parameter affects the probability at given value. Consider the parameterized version of *F*(*x*), which is $G(x,\beta)$. $G(x,\beta) = 1 - e^{-\frac{x}{\beta}} = 1 - \frac{1}{e^{x/\beta}}$, $\beta > 0$. It is clear that $G(x,\beta)$ is a decreasing function of β.

For fixed *x*, if β increases, $G(x,\beta)$ will decrease, so we get a bigger probability if we use a smaller value of β. Assume β belongs to interval [*a*,*b*]. Then $E_l(x) = 1 - e^{-\frac{x}{a}}, x \geq 0$, and $E_r(x) = 1 - e^{-\frac{x}{b}}, x \geq 0$.

For any other β in [*a*,*b*], the CDF $G(x,\beta)$ must lie between envelopes $E_l(x)$ and $E_r(x)$. The following figure shows the case when a=1 and b=3.

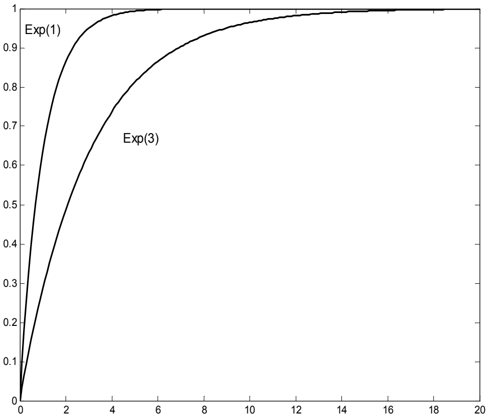

Figure 1. Exponential envelopes $E_l(x)$=Exp(1) **and** $E_r(x)$=Exp(3) **are shown;** $\beta \in [1,3]$.

Now consider another parameter, the location parameter. Since decreasing the location parameter would move the CDF to the left, and increasing it would move it to the right, the left envelope function would use the minimum value of the location parameter and the right envelope function would use its maximum value. Thus if both the location parameter and parameter β were given as intervals, the left envelope would be derived from the low values of both parameters and the right envelope would be derived from their high values.

3.2 Uniform distribution.

If a random variable *X* follows the uniform distribution, 2 parameters may be used to describe it: $X_{\min}$ and $X_{\max}$. $X_{\min}$ is the minimum value and $X_{\max}$ is the maximum value possible for samples of *X*. The relationship between these two parameters is $X_{\min} < X_{\max}$ and the density function is

$$f(x) = \frac{1}{X_{\max} - X_{\min}}, \; X_{\min} \leq x \leq X_{\max}.$$

From the density function, we can get the cumulative distribution function:

$$F(x)=\frac{x-X_{\min}}{X_{\max}-X_{\min}},\ X_{\min}\le x\le X_{\max}.$$

Define a parameterized version of $F(x)$ as $G(x,\ X_{\min},\ X_{\max})$. Since G decreases as $X_{\min}$ and $X_{\max}$ increase, the smaller the parameters the higher the cumulative probability. In general, if we know 2 intervals $[a,b]$ $[c,d]$ for $X_{\min}$ and $X_{\max}$ respectively, then

$$E_l(x)=\begin{cases}\dfrac{x-a}{c-a} & c>x\ge a\\ 1 & x\ge c\end{cases}\text{and } E_r(x)=\begin{cases}\dfrac{x-b}{d-b} & d>x\ge b\\ 1 & x\ge d\end{cases}$$

For any other values of the parameters in those intervals, the CDFs will lie between the envelope CDFs E_l and E_r. The following figure depicts the situation when a=1, b=2, c=5, and d=6.

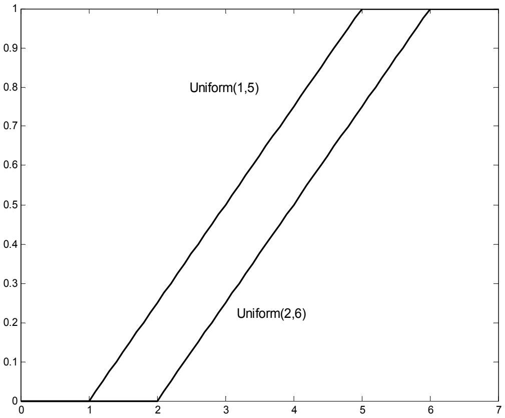

Figure 2. Envelopes based on parameters of the uniform distribution.

3.3 Triangular distribution

Three parameters describe triangular probability density functions. They are $X_{\min}$, X_{mod}, and $X_{\max}$. $X_{\min}$ is the minimum value of X, $X_{\max}$ is the maximum value of X, and X_{mod} is the mode value of X. The relationship between these values is

$$X_{\min}\le X_{\text{mod}}\le X_{\max}\ \text{and}\ X_{\min}<X_{\max}.$$

Its density function is

$$f(x)=\frac{2*(x-X_{\min})}{(X_{\max}-X_{\min})(X_{\text{mod}}-X_{\min})},\ X_{\min}\le x\le X_{\text{mod}}$$

$$f(x)=\frac{2*(X_{\max}-x)}{(X_{\max}-X_{\min})(X_{\max}-X_{\text{mod}})},\ X_{\text{mod}}<x\le X_{\max}$$

From the density function, we can derivate its cumulative probability function.

$$F(x)=\frac{(x-X_{\min})^2}{(X_{\max}-X_{\min})(X_{\text{mod}}-X_{\min})},\ X_{\min}\le x\le X_{\text{mod}}$$

$$F(x)=1-\frac{(X_{\max}-x)^2}{(X_{\max}-X_{\min})(X_{\max}-X_{\text{mod}})},\ X_{\text{mod}}<x\le X_{\max}$$

Based on these CDFs, we can conclude that the smaller the parameter, the higher the cumulative probability F. Let us describe the parameters with three intervals $[a,b]$, $[c,d]$, and $[e,f]$ for $X_{\min}$, X_{mod} and $X_{\max}$ respectively, where $a<b<c<d<e<f$. then $E_l(x)$ and $E_r(x)$ can be written as follows.

$$E_l(x)=\begin{cases}\dfrac{(x-a)^2}{(e-a)(c-a)} & a\le x\le c\\ 1-\dfrac{(e-x)^2}{(e-a)(e-c)} & c<x\le e\\ 1 & x>e\end{cases}$$

and

$$E_r(x)=\begin{cases}\dfrac{(x-b)^2}{(f-b)(d-b)} & b\le x\le d\\ 1-\dfrac{(f-x)^2}{(f-b)(f-d)} & d<x\le f\\ 1 & x>f\end{cases}$$

The space between this pair of envelopes will contain all other CDFs generating from parameters within those intervals. The following figure demonstrates this situation for a=1, b=2, c=3, d=4, e=5, and f=6.

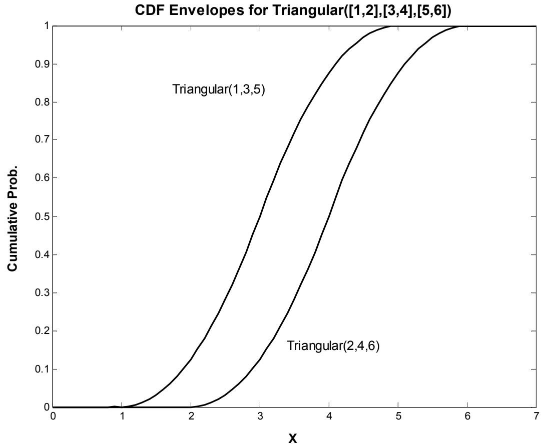

Figure 3. Envelopes around the CDFs of triangular density functions, derived from interval constraints on its parameters.

4. Envelopes requiring partitioning to derive

In this section, we present envelopes for the Cauchy, normal, and lognormal distributions.

4.1 Cauchy distribution

Let us use two parameters to describe the Cauchy distribution, a location parameter μ, and a scale parameter σ. Here $\mu \in R$ and $\sigma > 0$.

The density function of Cauchy distribution is

$$f(x)=\frac{1}{\pi}\frac{\sigma}{\sigma^2+(x-\mu)^2},\ x\in R$$

From the density function, we can get its cumulative probability function by integrating its density function.

$$F(x)=\int_{-\infty}^{x}f(t)dt=\int_{-\infty}^{x}\frac{1}{\pi}\frac{\sigma}{\sigma^2+(t-\mu)^2}dt=\int_{-\infty}^{x}\frac{1}{\pi}\frac{\sigma}{\sigma^2(1+(\frac{x-\mu}{\sigma})^2)}dt$$

$$=\int_{-\infty}^{x}\frac{1}{\pi}\frac{1}{\sigma(1+(\frac{x-\mu}{\sigma})^2)}dy=\int_{-\infty}^{\frac{x-\mu}{\sigma}}\frac{1}{\pi}\frac{1}{\sigma(1+y^2)}d(\mu+\sigma y)=\int_{-\infty}^{\frac{x-\mu}{\sigma}}\frac{1}{\pi}\frac{1}{(1+y^2)}dy$$

$$=\frac{1}{\pi}\tan^{-1}y\Big|_{-\infty}^{\frac{x-\mu}{\sigma}}=\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}-\frac{1}{\pi}\tan^{-1}(-\infty)$$

$$=\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}-\frac{1}{\pi}(-\frac{\pi}{2})$$

$$=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}$$

Let $y=\frac{x-\mu}{\sigma}$ and consider the resulting function $G(y)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}y$. Let us consider the interval for each parameter in turn.

Location parameter μ

$H(x,\mu,\sigma)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}$ is a decreasing function of μ since it is given that $\sigma > 0$. Hence the smaller the value of μ, the higher the value of *H* and hence the higher the cumulative probability for a given value of *x*.

Scale parameter σ

The effect on $y=\frac{x-\mu}{\sigma}$ of changing σ depends on the sign of $x-\mu$. If $x-\mu>0$, then *y* decreases as σ increases, so *G*(*y*) also decreases. So $H(x,\mu,\sigma)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}$ is a decreasing function of σ for $x-\mu>0$. If $x-\mu<0$, then increasing σ increases *y*, so *G*(*y*) also increases. So $H(x,\mu,\sigma)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-\mu}{\sigma}$ is an increasing function of σ for $x-\mu<0$.

Combining the two situations just noted, we have to use different formulas for different regions of an envelope, with the regions meeting at $x=\mu$. Consider intervals [*a*,*b*] and [*c*,*d*] for μ and σ respectively. Then we get the following envelope functions.

$$E_l(x)=\begin{cases}\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-a}{c} & x\geq a\\ \frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-a}{d} & x<a\end{cases}$$

$$E_r(x)=\begin{cases}\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-b}{d} & x\geq b\\ \frac{1}{2}+\frac{1}{\pi}\tan^{-1}\frac{x-b}{c} & x<b\end{cases}$$

For any other values of the parameters consistent with their intervals, the CDF must lie between the region enclosed by the two envelope CDFs. When *a*= –5, *b*=5, *c*=9 and *d*=25, the following figure shows the envelopes.

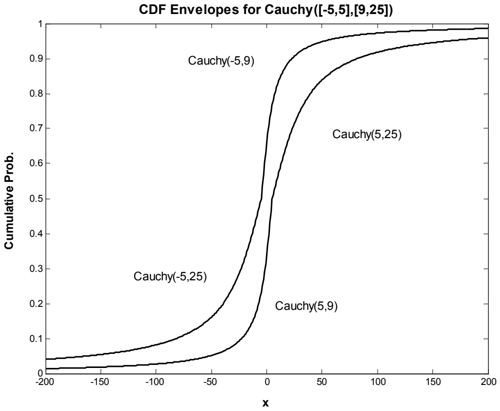

Figure 4. Envelopes around the Cauchy distribution implied by intervals for its two parameters. Each envelope function has two regions which meet at a non-differentiable point, *x*=*a* for E_l **and** *x*=*b* **for** E_r**.**

4.2 Normal distribution

There are two parameters sufficient to describe the normal distribution, the location parameter μ and the scale parameter σ. Possible values for these parameters are $\mu \in R$ and $\sigma > 0$.

The density function of the normal distribution is

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})\ \text{for}\ x\in R.$$

From the density function, we characterize the cumulative function as follows.

$$F(x)=\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma}}\exp(-\frac{(t-\mu)^2}{2\sigma^2})dt=\frac{1}{\sqrt{2\pi\sigma}}\int_{-\infty}^{x}\exp(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2)dt$$

Define $y=\dfrac{t-\mu}{\sigma}$. Then

$$F(x)=\frac{1}{\sqrt{2\pi\sigma}}\int_{y=-\infty}^{y=\frac{x-\mu}{\sigma}}\exp(\frac{-y^2}{2})d(\sigma y+\mu)=\frac{1}{\sqrt{2\pi\sigma}}\int_{-\infty}^{\frac{x-\mu}{\sigma}}\exp(\frac{-y^2}{2})\sigma dy$$

$$=\frac{\sigma}{\sqrt{2\pi\sigma}}\int_{-\infty}^{\frac{x-\mu}{\sigma}}\exp(\frac{-y^2}{2})dy=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-\mu}{\sigma}}e^{\frac{-y^2}{2}}dy=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{w}e^{\frac{-y^2}{2}}dy=H(w)$$

where $w=\dfrac{x-\mu}{\sigma}$. $H(w)$ is an increasing function of w since e to any power is positive. So by considering the direction of change in w caused by changing μ or σ, we can conclude $F(x)$ changes in the same direction.

For w, and so for $H(w)$, the smaller μ is the bigger w and H are. The smaller σ (and therefore σ^2 since σ is positive) is, the bigger w and H are for $x>\mu$, and the smaller w and H are for $x<\mu$.

In general, consider 2 intervals [a,b], [c,d] for μ and σ^2 respectively. $E_l(x)$ and $E_r(x)$ are

$$E_l(x)=\begin{cases} Normal\ \ (a,c) & x\geq a \\ Normal\ \ (a,d) & x<a \end{cases}$$

and

$$E_r(x)=\begin{cases} Normal\ \ (b,d) & x\geq b \\ Normal\ \ (b,c) & x<b \end{cases}$$

where $Normal(\mu,\sigma^2)$ is the CDF of the normal distribution with mean μ and variance σ^2.

For any other values of the parameters in their intervals, the CDF must be within the region enclosed by the two envelope CDFs E_l and E_r. The figure below shows the CDF envelopes for a=1, b=2, c=9 and d=25.

4.3 Lognormal distribution

We parameterize the lognormal distribution as in Siegrist (2002 [13]), one of several alternatives [9]. This parameterization has two parameters, μ and σ. Here $\mu\in R$ and $\sigma>0$. The density function of the lognormal distribution then is

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma x}\exp(\frac{-(\ln x-\mu)^2}{2\sigma^2}),\ x>0.$$

Let $z=\ln x$. Then z is normally distribution. Thus we can apply the results from the case of the normal distribution here. Consequently for z, the smaller the value of μ the higher the cumulative probability, and the lower σ the higher the cumulative probability is if $z\geq\mu$ and the lower the cumulative probability is if $z<\mu$. To derive results for the original argument x from these inequalities for z=ln x, the term ln x may be substituted for z and the inequalities solved for x.

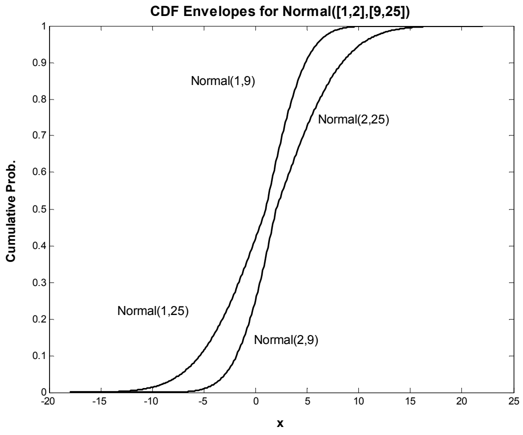

Figure 5. Envelopes around the normal distribution implied by intervals for its location and scale parameters. Each envelope function has two regions which meet at a non-differentiable point, x=a for E_l **and** x=b **for** E_r**.**

Applying those steps yields the following formulation. The smaller μ is, the higher the cumulative probability. The smaller σ is, the higher the cumulative probability is if $x\geq e^{\mu}$ and lower the cumulative probability is if $x<e^{\mu}$. The same rules apply to σ^2 as for σ since $\sigma>0$.

We can now specify intervals for μ and σ^2, the endpoints of which can be used to state the equations of the envelopes E_l and E_r. Let μ and σ^2 be values in [a,b] and [c,d] respectively. Then

$$E_l(x)=\begin{cases} LN\ \ (a,c) & x\geq e^a \\ LN\ \ (a,d) & x<e^a \end{cases}$$

and

$$E_r(x)=\begin{cases} LN\ \ (b,d) & x\geq e^b \\ LN\ \ (b,c) & x<e^b \end{cases}$$

where $LN(\mu,\sigma^2)$ is the CDF of the lognormal distribution with parameters μ and σ.

As an example, let a=3, b=4, c=0.1, and d=0.3. Then the envelopes are shown in the following figure.

5. Discussion: fuzzy interval parameters

The results given may be generalized to the case of parameters described with fuzzy intervals. If one parameter is a fuzzy interval, then each cut set of that interval yields a pair of envelopes. A nested series of envelopes results. A vertical slice through the graph then yields a fuzzy interval for the cumulative probability at a given value on the horizontal axis. A horizontal slice through the graph yields a fuzzy interval for the value on the horizontal axis for which the cumulative probability is a particular value.

6. Conclusion

We analytically derive envelopes for a variety of standard distributions with interval-valued parameters. For some distributions the envelopes have a non-differentiable point. For other distributions, we have not yet been able to derive envelopes analytically. Since there are important distributions which are among those we have not discussed, further work is needed in this direction.

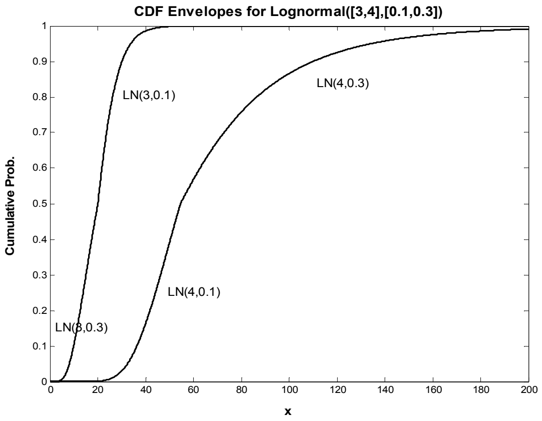

Figure 6. Envelopes around the lognormal distribution implied by intervals for its μ and σ parameters. Values given are for μ and σ^2. Each envelope function has two regions which meet at a non-differentiable point, $x=e^a$ for E_l and $x=e^b$ for E_r.

6. References

[1] Berleant, D. Automatically verified reasoning with both intervals and probability density functions. *Interval Computations* (1993 No. 2), pp. 48-70, http://www.public.iastate.edu/~berleant/.

[2] Berleant, D. and C. Goodman-Strauss. Bounding the results of arithmetic operations on random variables of unknown dependency using intervals. *Reliable Computing* 4(2) (1998), pp. 147-165, http://www.public.iastate.edu/~berleant/.

[3] Berleant, D, L. Xie, and J. Zhang. Statool: a tool for distribution envelope determination (DEnv), an interval-based algorithm for arithmetic on random variables. *Reliable Computing* 9 (2) (2003), pp. 91-108, http://www.public.iastate.edu/~berleant/.

[4] Berleant, D and J. Zhang. Representation and problem solving with the Distribution Envelope Determination (DEnv) method. *Reliability Engineering and System Safety*, **85** (1-3) (2004), pp. 153-168, http://www.public.iastate.edu/~berleant/.

[5] Berleant, D and J. Zhang. Using Pearson correlation to improve envelopes around the distributions of functions. *Reliable Computing*, **10** (2) (2004), pp. 139-161, http://www.public.iastate.edu/~berleant/.

[6] Ferson, S. What Monte Carlo methods cannot do. *Journal of Human and Ecological Risk Assessment* 2 (4)(1996), pp. 990-1007

[7] Ferson, S., V. Kreinovich, L. Ginzburg., D. Myers, and K. Sentz. Constructing Probability Boxes and Dempster-Shafer Structures. *SAND REPORT SAND2002-4015*, Sandia National Laboratories, Jan. 2003.

[8] Neumaier, A. Clouds, fuzzy sets and probability intervals, Reliable Computing **10** (2004), 249-272, http://www.mat.univie.ac.at/~neum/papers.html. See also, On the structure of clouds, submitted, same URL.

[9] *NIST/SEMATECH e-Handbook of Statistical Methods.* Web site http://www.itl.nist.gov/div898/handbook/, as of 2003. Paper at http://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm.

[10] Oberkampf, W., J. Helton, C. Joslyn, S. Wojtkiewicz, and S. Ferson. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safety*, **85** (1-3) (2004).

[11] Regan, H., S. Ferson S, and D. Berleant. Equivalence of five methods for bounding uncertainty. *Journal of Approximate Reasoning*, **36** (2004), pp. 1-30.

[12] Sandia National Laboratory. Epistemic Uncertainty Workshop. August 6-7, 2002, Albuquerque, www.sandia.gov/epistemic/eup_workshop1.htm.

[13] Siegrist, K. Virtual Laboratories in Probability and Statistics. Web site http://www.math.uah.edu/statold/, URL http://www.math.uah.edu/statold/special/special14.html.

[14] Smith, J.E. Generalized Chebychev inequalities: theory and application in decision analysis. *Operations Research* (1995) 43: 807-825.