ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# A Proposal for Construction of P-Box

Palash Dutta*
Dept. of Mathematics
Dibrugarh University,
Dibrugarh-786004, India
palash.dtt@gmail.com

Tazid Ali
Dept. of Mathematics
Dibrugarh University
Dibrugarh-786004, India
tazidali@yahoo.com

***Abstract:*** Risk assessment is an important aid in the decision making process. Generally two types of uncertainties viz., aleatory uncertainty and epistemic uncertainty involve in risk assessment. In order to treat epistemic and aleatory uncertainties in risk assessment probability bounds approach is often used. In this paper we propose a method to construct p-box.

***Keywords:*** Uncertainty, Probability Distribution, Fuzzy set, P-box

## I. INTRODUCTION

Uncertainty comes in many forms in the real world problem and it is an unavoidable component of human being. Risk assessment is an important tool of decision making in the problem of real world. Uncertainty also generally involve in risk assessment. Generally two types of uncertainties: aleatory uncertainty and epistemic uncertainty occur in risk assessment. Aleatory uncertainty arises from heterogeneity or the random character of natural processes while epistemic uncertainty arises from the partial character of our knowledge of the natural world.

In order to treat epistemic and aleatory uncertainties in risk assessment probability bounds approach is often used. Probability bounds analysis combines probability theory and interval arithmetic to produce probability-boxes (p-boxes). In particular, probability bounds analysis provides solution to the problems involving unknown dependencies between variables and uncertainties in the exact nature of distributions. Different authors proposed different methods to compute probability bounds. Chebyshev (1874) [4] described bounds on a distribution only when mean and standard deviation of the variable are known. Markov (1886) [1] found similar bounds for a positive variable when only its mean is known. Frechet (1935) [2] proposed how to compute bounds on probabilities of logical conjunctions and disjunctions without making independence assumptions. Yager (1986) [7] describes the elementary procedures by which bounds on convolution can be computed an assumption of independence. At the same time, Frank et al. (1987) [3] solve a question posted by A. N. Kolmogorov about how to finds bounds on distributions of sum of random variables when no information about their interdependency was available.

Extending the approach of frank et al (1987) [3], Williamson and Downs (1990) [6] develop a semi analytical approach that computes rigorous bounds on the cumulative distribution functions of convolution without necessarily assuming independents between the operands. Ferson et al (2003, 2004) [8, 9] gave method to compute probability bounds. Tucker et al (2003) [12] studied probability bounds analysis in environmental risk assessment. In this paper we have proposed a method to construct probability bounds when parameters of probability distribution viz., mean and standard deviation (variance) are available in the form of interval or fuzzy number.

## II. BASIC CONCEPT OF PROBABILITY THEORY

Probability theory [10] frequently used in uncertainty analysis. If parameters used in prescribed models are random in nature and follow well defined distributions, then probabilistic methods are most suitable and well accepted approach for risk assessment.

A random variable is a variable in a study in which subjects are randomly selected. Let X be a discrete random variable.

A probability mass function is a function such that

(i) $P_X(x) = P(X = x)$

(ii) $P_X(x) \geq 0$, (iii) $\sum_{i=1}^{n} P_X(x_i) = 1$,

The cumulative distribution function of a discrete random variable X, denoted as F(x) is

$$F_X(x) = P(X \leq x) = \sum_{x \leq x_i} P_X(x_i)$$

Let X be a continuous random variable. A probability density function of X is a non-negative function *f*, which satisfies

$$P(X \in B) = \int_B f(x)dx$$

for every subset B of the real line.
As X must assume some value, f must satisfy

$$P(X \in (-\infty, \infty)) = \int_{-\infty}^{\infty} f(x)dx = 1$$

This means the entire area under the graph of the PDF must be equal to unit.
In particular, the probability that the value of X falls within an interval [a, b] is

$$p(a \leq X \leq b) = \int_a^b f(x)dx$$

The CDF of a continuous random variable X is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

### *A. Most Frequently and Commonly Used Probability Distribution in Uncertainty Analysis in Risk Assessment:*

Normal distribution, lognormal distribution, triangular distribution, uniform distribution are the probability distributions most frequently and commonly occur in risk assessment in the analysis of uncertainty.

### *a. Normal Probability Distribution:*

A random variable X is said to be normally distributed with mean $\mu$ and variance $\sigma^2$, if its probability distribution function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty$$

The cumulative distribution function of a normally distributed random number is

$$F(x) = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right], -\infty < x < \infty$$

### *b. Lognormal Probability Distribution:*

The lognormal distribution results when the logarithm of the random variable is described by a normal distribution. That is, if X is log normally distributed, then Y=lnX is normally distributed. The probability density function of the lognormal distribution is given by

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

The cumulative distribution function of a log normally distributed random number is

$$F(x) = \frac{1}{2} + \frac{1}{2} erf\left[\left(\frac{\ln x - \mu}{2\sigma^2}\right)\right], x > 0$$

### *c. Triangular Probability Distribution:*

A random variable X is said to be triangularly distributed with lower limit a, upper limit c and mode b such that $a < b < c$, if its probability density function is given by

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)}, a \le x \le b \\ \dfrac{2(c-x)}{(c-b)(c-a)}, b \le x \le c \\ 0, otherwisw \end{cases}$$

The cumulative distribution function of a triangular probability distribution is given by

$$F(x) = \begin{cases} 0, x < a \\ \dfrac{(x-a)^2}{(b-a)(c-a)}, a \le x \le b \\ 1 - \dfrac{(c-x)^2}{(c-a)(c-b)}, b \le x \le c \\ 1, b < x \end{cases}$$

It should be noted that instead of having lower limit, mode and upper limit if we have mean, variance and mode then also we can have probability density function and cumulative distribution function of triangular probability distribution.

### *d. Uniform Probability Distribution:*

A random variable X is said to be uniformly distributed over the interval [a,b] f its probability density function is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a}, a \le x \le b \\ 0, otherwise \end{cases}$$

The cumulative distribution function of a uniform probability distribution is given by

$$F(x) = \begin{cases} 0, x \le a \\ \dfrac{x-a}{b-a}, x \in [a,b] \\ 1, x \ge b \end{cases}$$

Unlike triangular probability distribution, instead of having the interval [*a, b*] one can have PDF and CDF of uniform distribution from mean and variance.

It is observed from experiment or further studies that two same distribution (Normal distribution, lognormal distribution, triangular distribution, uniform distribution) having same mean but different standard deviations the CDFs of the two distributions meet at mean and 0.5.

## III. BASIC CONCEPT OF FUZZY SET THEORY

Environmental/human health risk assessment is an important element in any decision-making process in order to minimize the effects of human activities on the environment. Unfortunately, often environmental data tends to be vague and imprecise, so uncertainty is associated with any study related with these kinds of data. Fuzzy set theory provides a way to characterize the imprecisely defined variables, define relationships between variables based on expert human knowledge and use them to compute results.
In this section, some necessary backgrounds and notions of fuzzy set theory [5], [11] are reviewed.

***Definition 1)*** Let *X* be a universal set. Then the fuzzy subset *A* of *X* is defined by its membership function

$$\mu_A : X \to [0,1]$$

Which assign a real number $\mu_A(x)$ in the interval [0, 1], to each element $x \in A$, where the value of $\mu_A(x)$ at *x* shows the grade of membership of *x* in *A*.

***Definition 2)*** Given a fuzzy set *A* in *X* and any real number $\alpha \in [0,1]$. Then the $\alpha$ -cut or $\alpha$ -level or cut worthy set of *A*, denoted by $^{\alpha}A$ is the crisp set

$$^{\alpha}A = \{x \in X : \mu_A(x) \ge \alpha\}$$

The strong *a* cut, denoted by $^{\alpha+}A$ is the crisp set

$$^{\alpha}A = \{x \in X : \mu_A(x) > \alpha\}$$

For example, let A be a fuzzy set whose membership function is given as

$$\mu_A(x) = \begin{cases} \dfrac{x-a}{b-a}, a \le x \le b \\ \dfrac{c-x}{c-b}, b \le x \le c \end{cases}$$

To find the α-cut of A, we first set α □ [0,1] to both left and right reference functions of A.

That is, $\alpha = \dfrac{x-a}{b-a}$ and $\alpha = \dfrac{c-x}{c-b}$.

Expressing x in terms of α we have

$$x = (b-a)\alpha + a \text{ and } x = c - (c-b)\alpha .$$

which gives the α-cut of A is

$$^{\alpha}A = [(b-a)\alpha + a, c - (c-b)\alpha]$$

0-cut of the fuzzy number A is the support of the fuzzy number *A* that is, [a, c].

***Definition 3)*** The support of a fuzzy set *A* defined on *X* is a crisp set defined as

$$\text{Supp}(A) = \{x \in X : \mu_A(x) > 0\}$$

***Definition 4)*** The height of a fuzzy set *A*, denoted by *h*(*A*) is the largest membership grade obtain by any element in the set.

$$h(A) = \sup_{x \in X} \mu_A(x)$$

***Definition 5)*** **A** fuzzy number is a convex normalized fuzzy set of the real line *R* whose membership function is piecewise continuous.

***Definition 6)*** A triangular fuzzy number A can be defined as a triplet (*a, b, c*). Its membership function is defined as:

$$\mu_A(x) = \begin{cases} \dfrac{x-a}{b-a}, a \le x \le b \\ \dfrac{c-x}{c-b}, b \le x \le c \end{cases}$$

***Definition 7)*** A trapezoidal fuzzy number *A* can be expressed as (*a,b,c,d*) and its membership fuzzy number is defined as:

$$\mu_A(x) = \begin{cases} \dfrac{x-a}{b-a}, a \le x \le b \\ 1, b \le x \le c \\ \dfrac{d-x}{d-c}, c \le x \le d \end{cases}$$

## IV. PROPOSAL FOR CONSTRUCTION OF P-BOX

P-box can be constructed when parameters of probability distributions (Normal distribution, lognormal distribution, triangular distribution, uniform distribution) are not precisely known. Here, we consider parameters of probability distributions are available as closed intervals or fuzzy numbers. If parameters are fuzzy numbers then we first find the support of the fuzzy number using 0-cut.

Suppose *prodist* indicates one of the probability distribution i.e., normal distribution, lognormal distribution, triangular distribution, uniform distribution. Let A be a *prodist* whose parameters are available in the form of intervals, mean $[a,b]$ and standard deviation (variance) $[c,d]$. The p-box for A has to be calculated by taking all the combination such as $(a,c),(a,d),(b,c)$ and $(b,d)$. Then the envelope over four distributions $prodist(a,c), prodist(a,d), prodist(b,c)$ and $prodist(b,d)$ gives the resulting p-box for

### *A. Algorithm for p-box:*

**Step 1:** Generate *N* number of uniformly distributed random numbers in between 0 and 0.5 and *N* numbers of uniformly distributed random numbers in between 0.5 and 1.

i.e., say $r = unifrnd(0,0.5,1,N)$ and

$$s = unifrnd(0.5,1,1,N)$$

**Step 2:** Take the inverse of the cumulative distributions function. That is,

$$x1 = prodistinv(r,a,d), x2 = prodistinv(s,a,c)$$

$$y1 = prodistinv(r,b,c), y2 = prodistinv(s,b,d)$$

**Step 3:** Consider $x = [x1, x2]$ and $y = [y1, y2]$

**Step 4:** Plot cumulative distribution function (cdf) for *x* and *y*. Which will give the resulting p-box for *A*.

## V. NUMERICAL EXAMPLE

To demonstrate and make use of the algorithm we illustrated four examples for all the distributions most frequently and commonly occur in risk assessment in the analysis of uncertainty.

***Example 1)*** Consider A be a normal probability distribution whose parameters i.e., mean and standard deviation are available in the form of interval or fuzzy number, say mean [20,30] or fuzzy number [20,25,30] and standard deviation [2,6] or fuzzy number [2,4,6]. We need to construct p-box using the algorithm. If parameters are available in the form of fuzzy number we first calculate 0-cut to obtain the complete support. That is, 0-cut of the fuzzy number [20,25,30] and [2,4,6] are [20,30] and [2,6] respectively. Now envelope of the four combination normal(20,2), normal(20,6), normal(30,2) and normal(30,6) will give the p-box for A.

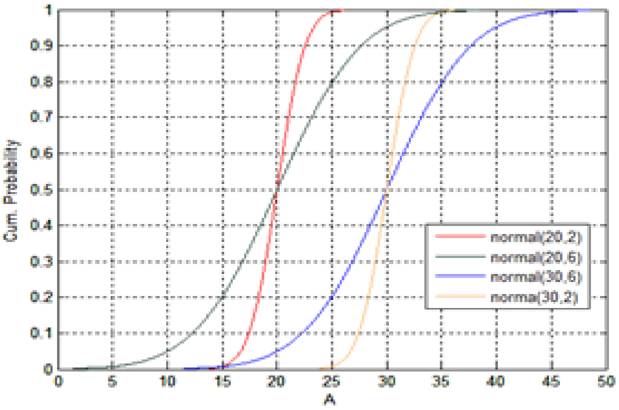

Figure 1: Combination of distributions-A

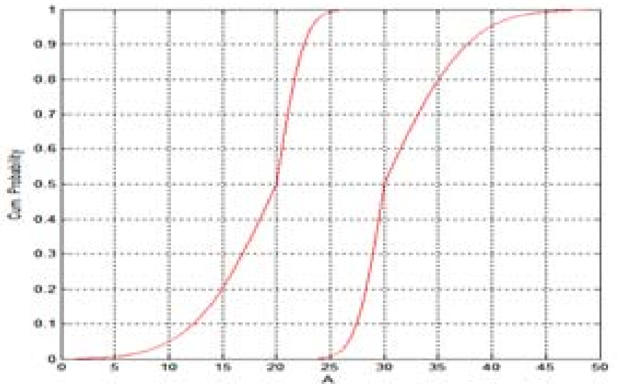

Figure 2: p-box for A

***Example 2)*** Consider B be a lognormal probability distribution whose parameters are available in the form of fuzzy number, say, mean [10,15,20] and standard deviation [2,2.5,3]. 0-cut of [10,15,20] and [2,2.5,3] are [10,20] and [2,3]. Envelope of the four combination lognormal(10,2), lognormal(10,3), lognormal(20,2) and lognormal(20,3) give the p-box for B.

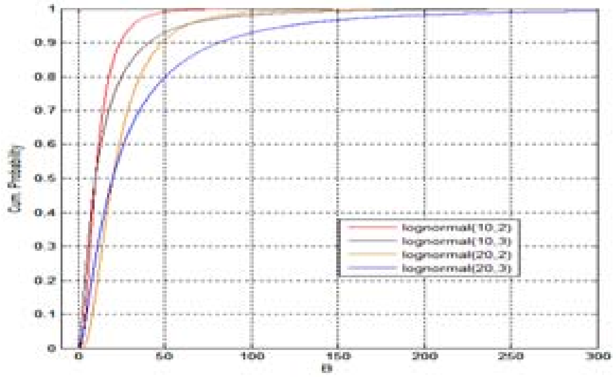

Figure 3: Combination of distributions-B

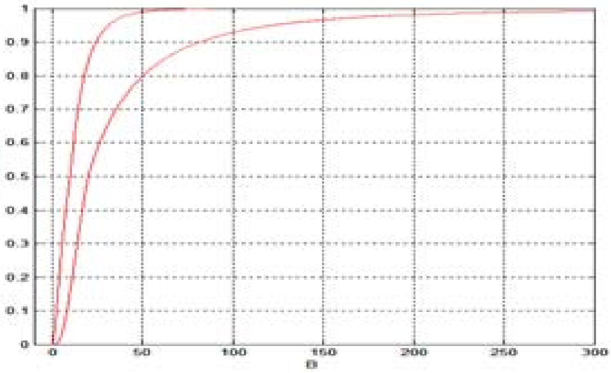

Figure 4: p-box for B

***Example 3)*** Consider C be a symmetric triangular probability distribution with mean [20,18,30] and variance[5,10]. 0-cut of [20,18,30] is [20,30]. Now, the four combination triangular(20,5), triangular(20,10), triangular(35,5) and triangular(35,10) gives the p-box for C.

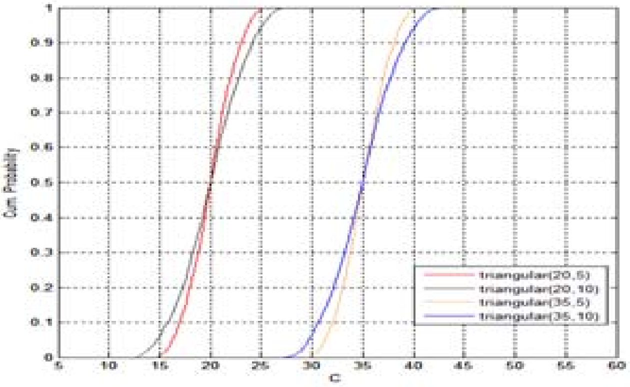

Figure 5: Combination of distributions-C

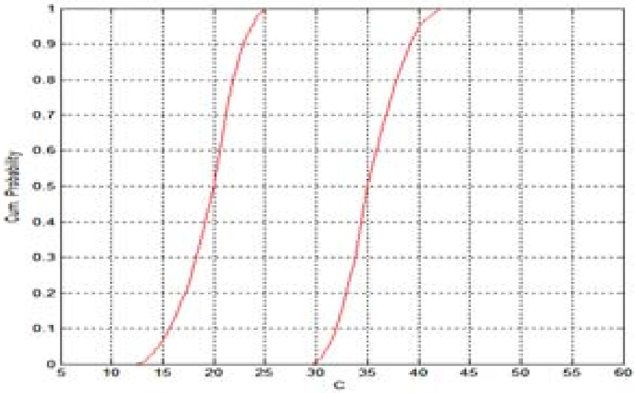

Figure 6: p-box for C

***Example 4)*** Consider D be a uniform probability distribution with mean [20,25] and variance [8,10,15]. From the four combination uniform(20,8), uniform(20,15), uniform(25,8) and uniform(25,15) we obtain the p-box for D.

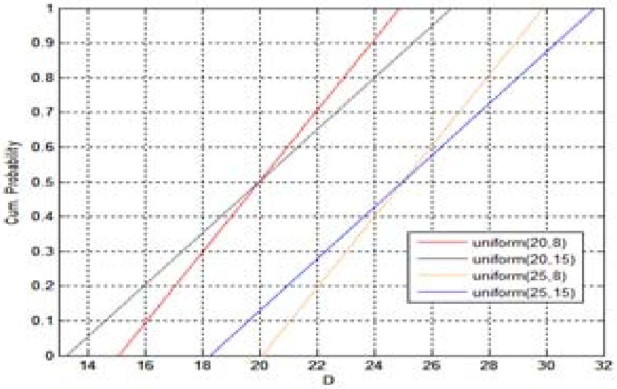

Figure 7: Combination of distributions-D

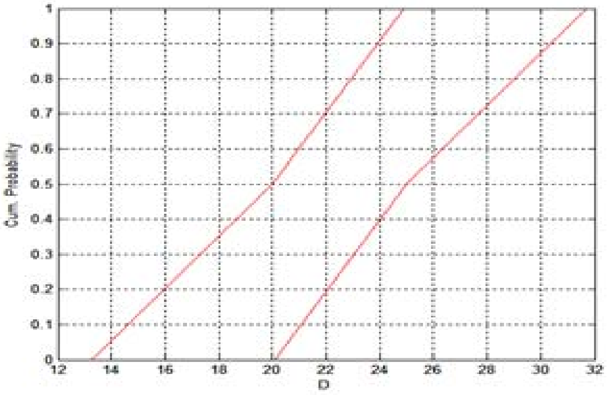

Figure 8: p-box for D

## VI. CONCLUSION

Risk assessment is an important aid in the decision making process. Generally two types of uncertainties viz., aleatory uncertainty and epistemic uncertainty involve in risk assessment. Aleatory uncertainty arises from heterogeneity or the random character of natural processes while epistemic uncertainty arises from the partial character of our knowledge of the natural world. To deal with epistemic and aleatory uncertainties in risk assessment probability bounds approach is often used. Probability bounds analysis combines probability theory and interval arithmetic to produce probability-boxes (p-boxes). P-box can be constructed when parameters of probability

© 2010, IJARCS All Rights Reserved

distributions (normal distribution, lognormal distribution, triangular distribution, uniform distribution etc.) are not precisely known. In this paper we have proposed a method to construct probability bounds when parameters of probability distribution viz., mean and standard deviation (variance) are not precisely known i.e., they are available in the form of interval or fuzzy number. To demonstrate and make use of the proposed method we illustrated four examples for all the distributions most frequently and commonly occur in risk assessment in the analysis of uncertainty. P-box also can be constructed when parameters of probability distributions are also available in the form of probability distributions. In this case bounds can be obtained using confidence interval.

## VII. ACKNOWLEDGMENT

The work done in this paper is under a research project funded by Board of Research in Nuclear Sciences, Department of Atomic Energy, Govt. of India.

## VIII. REFERENCES

[1]. Markov [Markoff] “Sur une question de maximum et de minimum propose´e par M Tchebycheff”. Acta Math, 1886, Vol. 9 pp. 57–70. G.

[2]. M. Fre´chet “Ge´ne´ralisations du the´ore`me des probabilite´s totals”. Fundam Math, 1935, Vol. 25, pp. 379–387.

[3]. M. J. Frank, R. B. Nelsen, and B. Schweizer , “Best-possible bounds for the distribution of a sum—a problem of Kolmogorov”. Probab. Theory Related Fields 1987, Vol. 74(2), 199–211.

[4]. P. Chebyshev [Tchebichef] “Sur les valeurs limites des integrals”. J Math Pure Appl, 1874, Vol. 19(2), pp.157–160.

[5]. P. Dutta, H. Boruah and T. Ali “Fuzzy arithmetic with and without using $\alpha$-cut method: a comparative study”, International Journal of Latest Trends in Computing, 2011, vol-2, pp : 99-108.

[6]. R.C. Williamson and T. Downs “Probabilistic Arithmetic I: numerical methods for calculating convolutions and dependency bounds”. International journal of approximate reasoning, 1990, Vol. 4, pp. 89-158.

[7]. R. R. Yager, “Arithmetic and other operations on Dempster-Shafer structures”. International Journal of Man-machine Studies 1986, Vol. 25: 357-366.

[8]. S. Ferson and J.G. Hajagos “Arithmetic with uncertain numbers: rigorous and (often) best possible answers”, Reliability Engineering and System safety, 2004 , Vol. 85, pp. 135-152.

[9]. S. Ferson, V. Kreinovich, L. Ginzburg, D.S. Myers and K. Sentz “Constructing Probability boxes and Dempster-Shafer Structure”, 2003, SAND2002-4015. Sandia national Laboratories, Albuquerque, NM.

[10]. S. M. Ross “Probability and statistics for engineers and scientists”, Academic press, third edition, 2005.

[11]. T. Ross “Fuzzy logic with engineering application”, John Wiley & Sons Ltd., Second edition, 2004.

[12]. W. T Tucker and S. Ferson “Probability bounds analysis in environmental risk assessment”, Applied Biomathematics, 2003.