Fitting Absolute Distributions to Limited Data

ISPA / SCEA Conference
Orlando, Florida
June 2012

Blake Boswell

Table Of Contents

Introduction: Choosing Distributions with Limited Data

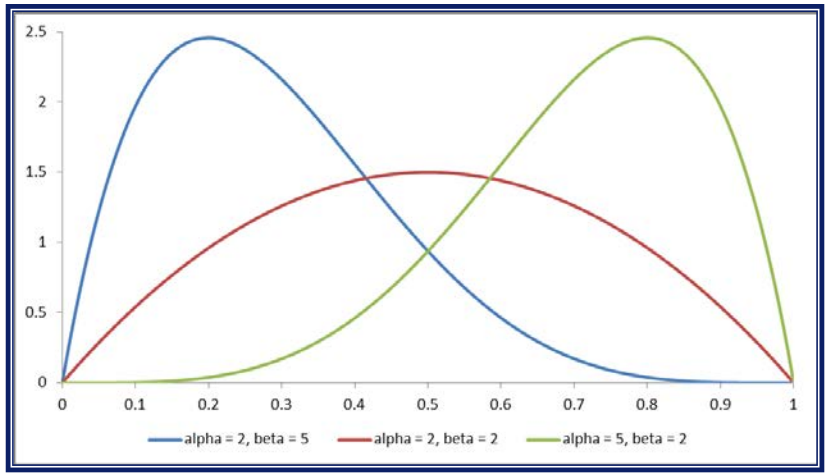
- The choice of probability distributions is a critical component for cost risk and uncertainty modeling
  - When data is available, distribution fitting techniques such as Goodness of Fit (GoF) tests and Information Criteria (IC) can be applied to determine probability distribution functions (PDFs) that accurately describe potential cost realizations
  - With limited data, GoF tests and IC based methods provide little or no insight into the best PDF choice
- When data is limited, it is standard practice in cost risk and uncertainty analysis to choose a PDF based on expert opinion
  - A popular approach to constructing PDFs relies on three point estimates, with the points representing high and low extremes and a measure of central tendency (5th, 50th, 95th percentiles)
  - Three point estimates are usually mapped to bounded distributions such as Triangular and Beta
- This paper investigates the use of a new approach to distribution fitting, called **Decision on Belief** (DoB)[1], to guide the choice of distributions in cost risk and uncertainty models when limited data is available
  - DoB was developed for statistical estimation of processes incurring high cost and risk, such as testing new drugs, prototyping industrial products, experimenting with nuclear material and launching missiles
  - The goal of incorporating DoB with three point estimation is to choose a best fit PDF from a candidate set of both bounded and absolute distributions based on limited information

Three Point Estimation and Bounded Distributions

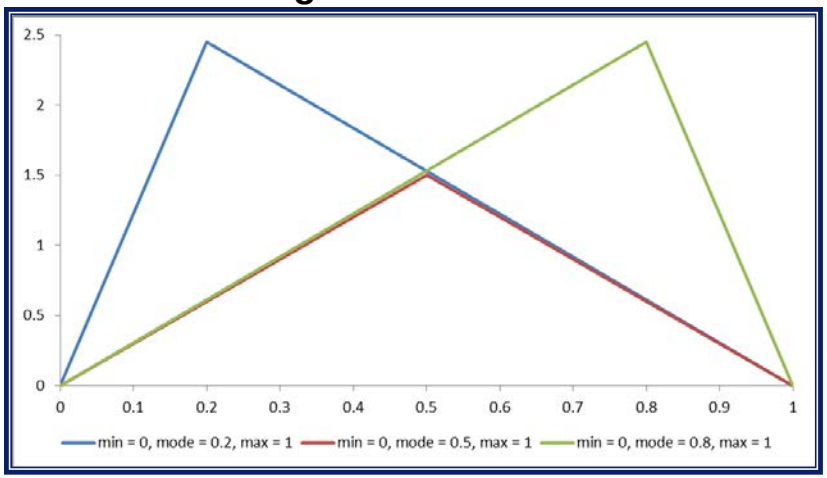
- Many uncertain quantities can be conceptualised in terms of a continuous PDF with the majority of mass concentrated around a most likely value
  - Therefore, this study focuses on bell & triangular shaped distributions as opposed to U, or J shaped distributions

- Uncertain variables often have estimable min and max values, therefore a distribution with finite limits is intuitively plausible for many scenarios in risk and uncertainty modeling
  - The beta and triangular distributions are seen as suitable models in this context as they provides a wide variety of distribution shapes over a finite interval

Beta Distribution



Triangular Distribution



Three Point Estimates for the Beta Distribution

- Common approaches to Three Point Estimation of the first two moments of the Beta Distribution
  - The "PERT" approach to defines the first two moments of the Beta distribution as follows

$$\mu = \frac{a+b+4m}{6}; \qquad \sigma = \frac{(b-a)}{6}$$

where a = best case, m = most likely, and b = worse case. Knowing μ and σ, and given a, and b, one can estimate the beta distribution parameters (α, β) via the following:

$$\alpha = \frac{(1-\mu_s)\mu_s^2}{\sigma_s^2} - \mu_s; \qquad \beta = \frac{\alpha(1-\mu_s)}{\mu_s}$$

where μ_s and σ_s refer to the standard beta distribution, and are given by:

$$\mu_s = \frac{\mu-a}{b-a}; \qquad \sigma_s = \frac{\sigma}{b-a}$$

  - Another approach the Pearson-Tukey formula:

$$\mu = .185a_{.05} + .63m + .185b_{.95}; \qquad \sigma = \frac{b_{.95}-a_{.05}}{3.25}$$

where $a_{.05}$ = 5th percentile and $b_{.95}$ = 95th percentile

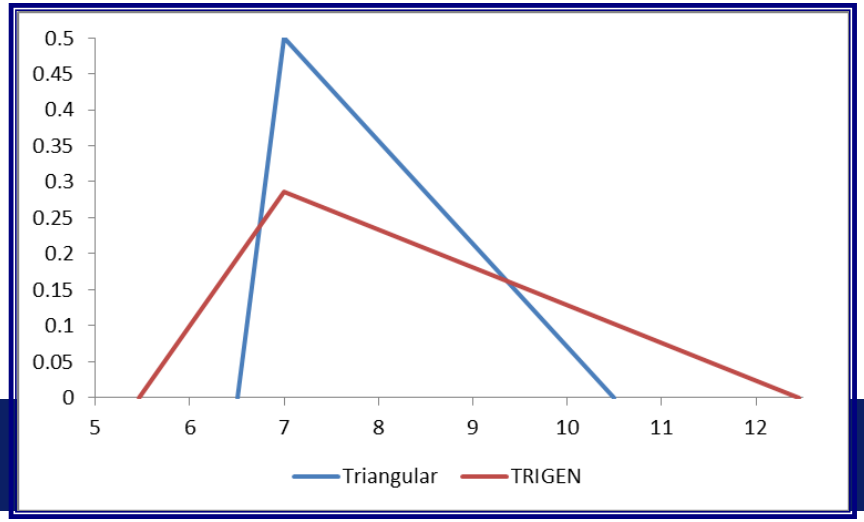
Three Point Estimates for the Triangular Distribution

- Three Point Estimation of the first two moments of the Triangular Distribution
  - The standard formula for defining the first two moments of a triangular distribution over the interval (a, b) with mode m, such that $a \leq m \leq b$ (denoted as Triangular in this study)

$$\mu = \frac{1}{3}(a + m + b); \qquad \sigma^2 = \frac{1}{18}\left((b-a)^2 - (m-a)(b-m)\right)$$

  - The first two moments can also be estimated via the TRIGEN method by using percentiles, (a_p, b_r), to determine the interval (a, b) such that $a < a_p \leq m \leq b_r < b$ by numerically solving the following:

$$q = \frac{\left((m-a_p)\left(1-\sqrt{\frac{1-r}{1-q}}\right)\right)}{(b_r-m)(1-\sqrt{\frac{p}{q}}+(m-a_p)\left(1-\sqrt{\frac{1-r}{1-q}}\right)} \text{ then setting } a = \frac{a_p - m\sqrt{\frac{p}{q}}}{1-\sqrt{\frac{p}{q}}} \text{ and } b = \frac{b_r - m\sqrt{\frac{1-r}{1-q}}}{1-\sqrt{\frac{1-r}{1-q}}}$$

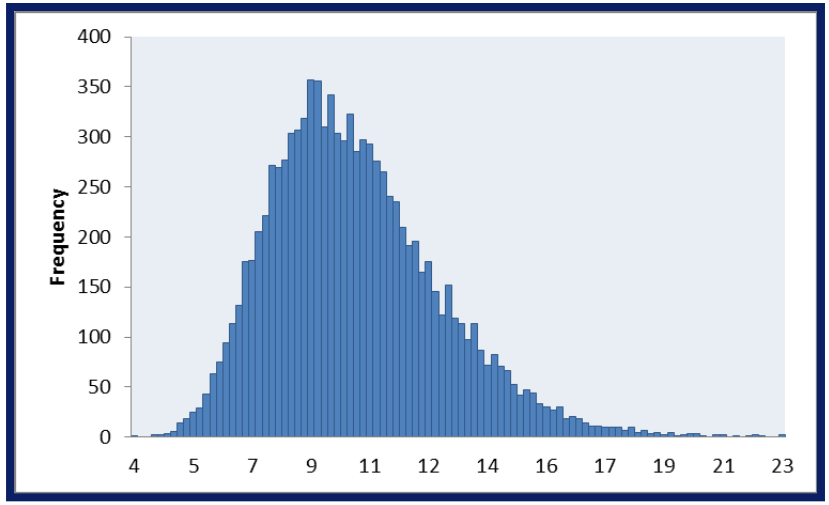


The figure to the left displays a Triangular distribution with parameters $a = 6.5$, $m = 7$, and $b = 10.5$ and an analogous adjusted TRIGEN distribution where $a_{.10} = 6.5$ and $b_{.90} = 10.5$

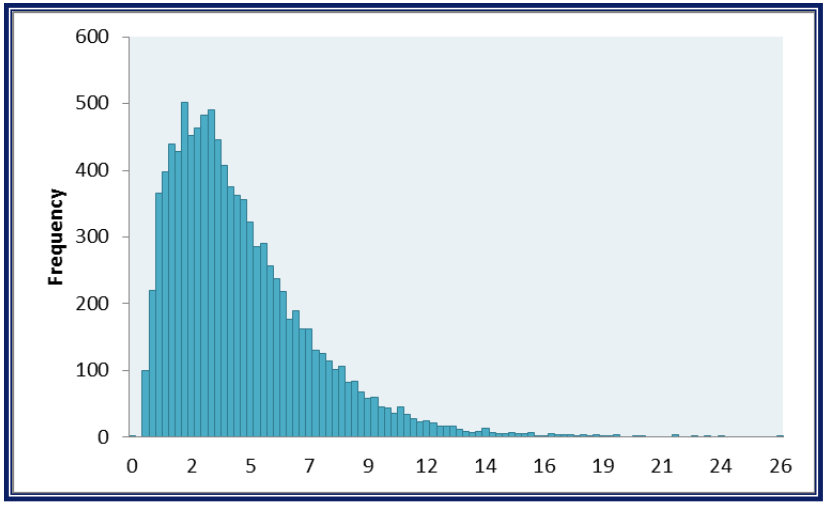
Limitations of Bounded Distributions

- Situations arise when it is necessary to represent an uncertain variable having a highly skewed distribution that cannot be adequately modeled by a bounded distribution
  - The skewness of a bell-shaped beta distribution cannot exceed 2
  - The skewness of a triangular distribution it is less than $\frac{2\sqrt{2}}{5} \approx .56$
- Absolute distributions that are often seen as plausible alternatives to beta and triangular are the gamma and lognormal distributions

Gamma Distribution



Lognormal Distribution



Approximation with Bounded Distributions

- Consider the following example:
  - An expert gives an estimate as $a = 64$ (best case), $m = 97$ (typical), $b = 145$ (worse case)
  - Using this information, and the formulas in the previous slides, we can construct the triangular, TRIGEN, PERT, and beta distributions
  - Assume that the actual PDF underlying the process described is a lognormal distribution with $a_{.05} = 64$, $m = 97$, and $b_{.95} = 145$; the below table represents the approximation performance of the absolute distributions

	Mean	Delta	StDev	Delta	CV	Delta	Skew	Delta
Lognormal	100.2936		24.99		0.249		0.717	
Triangular	101.9843	-2%	16.25	35%	0.159	36%	0.202	72%
TRIGEN	104.1584	-4%	23.1	8%	0.222	11%	0.196	73%
PERT	99.88513	0%	14.72	41%	0.147	41%	0.172	76%
Beta	100.2994	0%	19.49	22%	0.194	22%	0.192	73%

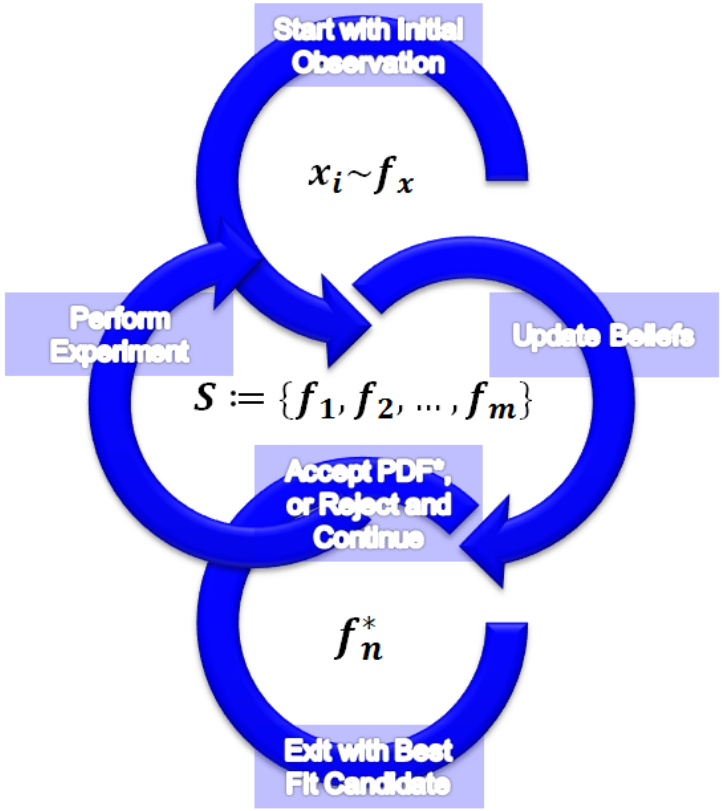
Minimizing the Error Resulting from Distribution Selection

- In the previous example the three point estimate was derived from a lognormal distribution to assess the resulting approximations by absolute distributions
- The absolute distributions performed very well with regards to estimating the mean of the underlying distribution, however, they consistently underestimated the StDev and drastically underestimated the skew
  - Underestimation of variance compounds as misspecified lower level distributions are rolled up
- In the previous example the underlying distribution was **known**
  - This is never the case in practice
- This study will now focus on developing a method for minimizing the error resulting from distribution selection
  - First a family of candidate distributions is created by three point estimation
  - Then a “best-fit” distribution is chosen as being the average representation of the family of distributions
  - The selected distribution will represent the family’s first two moments, and CDF with the least error from other candidate distributions in the family
- To select this distribution with maximum confidence, we consider the method of DoB

Statistical Estimation by Decision on Belief

- DoB is a new approach to distribution fitting developed by Eshragh and Modarres[1]
- The estimation method can be stated in the following way:
  - Given a random variable, X, having unknown PDF, f_X, the distribution best describing X can be selected from a set of candidate PDFs, $S := \{f_1, f_2, \cdots, f_m\}$, via a multi-stage algorithm formulated as a special case of Optimal Stopping Problem
  - At each stage an experiment is conducted to generate a random observation from f_X and then a decision is made, either to select a candidate in S, or to move on to a new experiment
  - It is assumed that a cost, C, is incurred for each new observation, and that there is a maximum number of observations, N, that can be obtained

Selection by Optimal Stopping Problem



Background: Related Concepts and Terminology

- The theory of **optimal stopping** is concerned with the problem of choosing a time to take a particular action, in order to maximize an expected reward or minimize an expected cost
  - You are observing a sequence of random variables, and at each step, you can choose to either stop observing or continue
  - If you stop observing at a step, you will receive a reward (or not incur further penalties)
  - You want to choose a stopping rule to maximize your expected reward
- A **belief** is a real number representing the probability of the event $\{f_X \equiv f_i\}$, where f_X is an unknown PDF and f_i is a reasonable candidate distribution.
  - Beliefs for f_i are based on a vector of observations from f_X denoted by $O_k = (x_1, x_2, \cdots, x_k)$
  - The belief of f_i at observation k is denoted as $B_i(x_k, O_{k-1})$
- The explanation of the DoB algorithm discussed herein is adapted from a systematic, three phased approach proposed by Monfared and Ranaiefar[2]
  - For a mathematical proof based presentation see (1)

Phase I: Constructing Beliefs of f_X on S

1. Define the set of candidate PDFs, $S := \{f_1, f_2, \cdots, f_m\}$, where all m functions are considered "reasonable" fits for f_X, the unknown PDF

2. Initialize belief, $B_i(\cdot) = \frac{1}{m}$, for each $f_i \in S$. Also set α as the discount rate, $V(N)$ as the maximum probability of correct selection and N as the maximum number of observations which can be generated

3. Start with observation x_k from f_X and estimate the posterior belief values, $B_i(\cdot)$, for $i = 1,2,\cdots,m$ by using the following formula derived from Bayes Theorem:

$$B_i(O_k) = B_i(x_k, O_{k-1}) = \frac{B_i(O_{k-1}) \cdot f_i(x_k)}{\sum_{i=1}^{m} B_i(O_{k-1}) \cdot f_j(x_k)}$$

4. Order the beliefs in ascending order from 1 to m and set $B_{gr}(\cdot) = B_m$, and $B_{sm}(\cdot) = B_{m-1}$

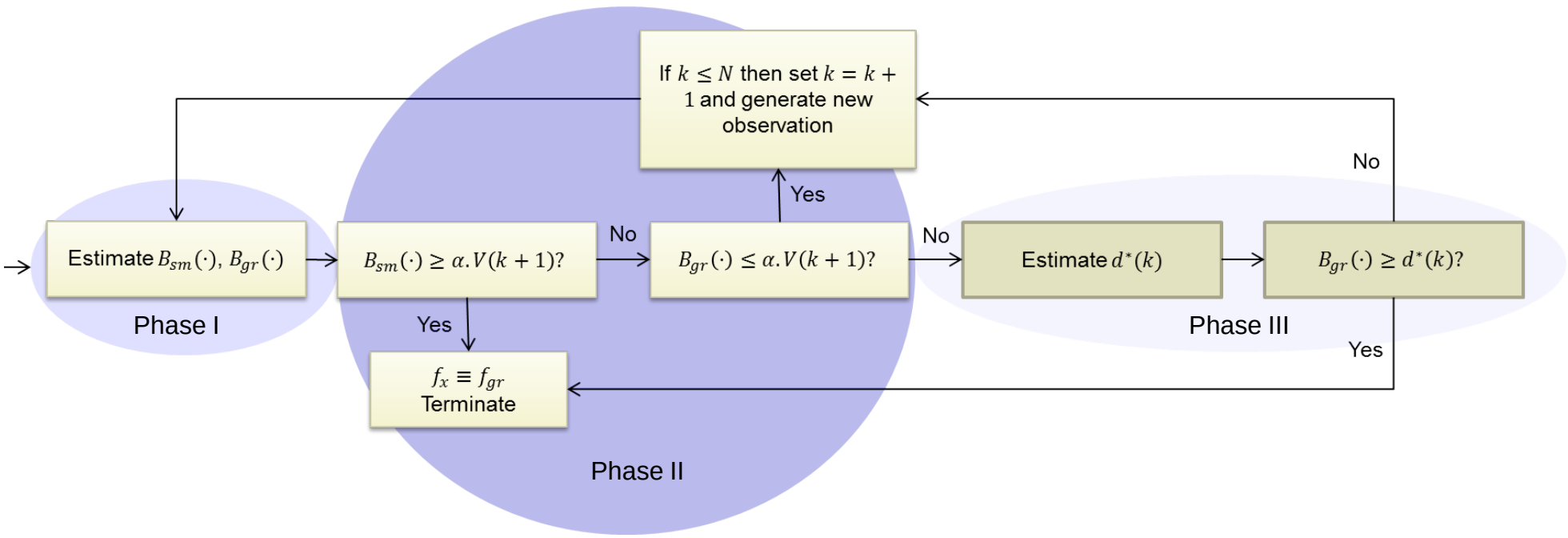
5. Normalize $B_{sm}(\cdot)$ and $B_{gr}(\cdot)$ using the following

$$B_{sm}(O_k) = \frac{B_{sm}(O_k)}{B_{sm}(O_k)+B_{gr}(O_k)}, \text{ and } \quad B_{gr}(O_k) = \frac{B_{gr}(O_k)}{B_{sm}(O_k)+B_{gr}(o_k)}$$

Phase II: Selection

6. If $B_{gr} > \alpha.V(k+1)$, where $\alpha.V(k+1) = \alpha^{N-k} \cdot V(N)$, then $f_X \equiv f_{gr}$, hence, f_{gr} is the best fit candidate PDF and the algorithm should terminate

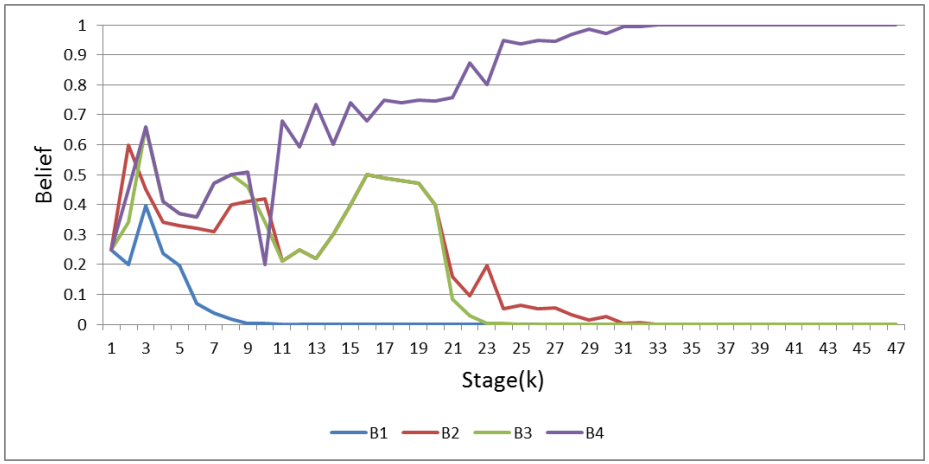
7. If $B_{gr} < \alpha.V(k+1)$, then $f_X \neq f_{gr}$, and a new observation is required. If $k \leq N$, set $k = k+1$, then go to Step 3. If $k > N$ then f_{gr} is the best fit distribution out of the candidate set achievable given the limited observation set



Convergence of Phase I & II

- A best fit PDF from S can be determined utilizing only phases I and II of the DoB algorithm
  - A large number of observations may be needed, diminishing the effectiveness of DoB to handle processes with limited observations
  - The belief formula for determining $B_i(O_k)$ in Phase I is proved to be convergent in (1), i.e. after getting enough observations and updating the beliefs with probability one, the belief from which the observations came converges to one and the other beliefs converge to zero
- Phase III proposes a novel method for formulating decision criteria allowing DoB to select the best PDF in S while requiring the

Belief at each Stage for 4 Candidate PDFs



Consider $S := \{f_1, f_2, f_3, f_4\}$, where $f_1 = \Gamma(0,1)$, $f_2 = \Gamma(4, \sqrt{3})$, $f_3 = \Gamma(16, \sqrt{3})$ and $f_4 = \Gamma(3,4)$. Random numbers from $\Gamma(3,4)$ were generated in Excel to serve as observations. The above chart shows B_4 converging to 1 at around k = 31 revealing that f_4 is the best candidate.

Phase III: Efficient Selection by Decision Criteria $d^*(k)$

- $d^*(k)$ is a criteria that is applied to determine the best fit PDF efficiently. The procedure for determining $d^*(k)$ utilizes non-linear dynamic programming to maximize the probability of accurate PDF selection.

- For a full mathematical treatment of Phase III see (1) & (2)

The goal is to define $B^*_{x_{k+1}}$ as the most plausible belief on f_{gr} while in the kth stage. Because x_{k+1} is unknown, the sequential nature of DoB is leveraged to assume that the "next" observation is the best possible observation one can expect to have at the present stage to select f_{gr} as f_x. Under this assumption one considers estimating the highest plausible belief one can get on the present best fit function, f_{gr}.

The underlying idea here is that, if the next forthcoming observation were considered to be the best possible one, would it be possible to terminate the process and make a decision on a best fit function or not? This idea will help to minimize the need for additional experiments

- Phase III is best illustrated through a numerical example that ties together all phases for

DoB Using SME Input: Excel Demo

Contact

Blake Boswell
Senior Consultant

Booz | Allen | Hamilton

Booz Allen Hamilton Inc.
Tel (202) 412-7516
Boswell_James@bah.com

Excel demo file available by email.

References

1. Eshragh Jahromi, A., Modarres Yadzi, M., “A New Approach to Distribution Fitting: Decision on Beliefs”, Journal of Industrial and Systems Engineering, Vol 3, No.1, pp 56 – 71, 2009

2. Monfared M.A.S, Ranaeifar F., “Further Analysis and Developments of the Eshragh-Modarres (E-M) Algorithm and Statistical Estimation”, Scientia Iranica, Vol. 14, No. 5, pp 425 -434, 2007