Progress Report on "Convergence of Risk Estimates Obtained from Highly Skewed Distributions," a Problem Posed by Jim Sinclair of AEA Technology

David Draper* Callum McKail* David Allwright†

12 September 1997

1 Introduction and problem statement

Safe disposal of radioactive wastes is an important problem for nuclear power industries worldwide. Deep underground disposal in special-purpose chambers is a leading possible solution to this problem. Assessments of the likely risk to individuals in the future if such a disposal system were implemented have been mandated by government agencies in the UK and elsewhere.

One leading approach to such risk evaluation is a method called *Probabilistic Safety Assessment* (PSA). To paraphrase Jim Sinclair of AEA Technology, in his memo laying out the problem studied here,

> With this approach uncertainty about the disposal system, in its initial state and in its long-term evolution, is assessed by Monte Carlo simulation methods, in which a (preferably large) number n of randomly sampled possible realizations of the system are generated. For each realization, performance measures such as radiological dose experienced by a hypothetical exposed group are calculated as a function of time, using complex mathematical models with hundreds of parametric and scenario inputs. Uncertainty about the system is embodied in uncertainty about these model inputs, which is quantified through probability density functions (PDFs) meant to express expert opinion about the likelihood of each possible value of the inputs.
>
> Risk to an individual from radiological exposure is defined as the probability of a serious health effect as a result of the exposure. It is a universally-used assumption that, for small doses, the risk from a given dose over a period of time, such as a year, is proportional to the size of the dose. When the dose received is uncertain, this leads to annual expected risk being proportional to the expected value of annual dose. As a risk performance measure, then, the aim of PSA calculations is to estimate the expected value of dose as a function of time. Because the system is complex, there can be no reliable expert opinion about the PDF of dose itself. Instead, the mathematical model is used as an "inference engine," to turn justifiable opinions about the model inputs into probabilistic assessments of likely dose.
>
> Since an infinite (at least in practice) number of input-value combinations is possible, the Monte Carlo calculations deliver only a finite sample from an extremely large population. With equal-probability sampling, the sample mean provides an unbiased estimate of the desired population mean μ; with appropriately and deliberately

*Statistics Group, Department of Mathematical Sciences, University of Bath
†Department of Mathematics, University of Oxford

biased sampling, a weighted sample mean can achieve the same goal with smaller uncertainty about μ. In either case, along with an estimate $\hat{\mu}$, it is necessary to provide an assessment of the uncertainty of $\hat{\mu}$.

The problem is that with highly skewed and long-tailed dose distributions F, of the type regularly encountered in this form of risk assessment, the mean $\mu = \mu(F)$ is a functional of F that is extremely sensitive to assumptions made about F, even with fairly large values of n. Computing constraints, arising from the complexity of the models being simulated, place severe downward pressure on achievable values of n in practice; often sample sizes in the low to middle hundreds are all that is possible.

Question: Can a usable and defensible upper bound to μ be created with modest n, even when F is at least partially unknown?

One possible answer to this question is to attempt to redefine the problem, by asserting that μ is the wrong functional—because of its sensitivity to small tail perturbations in F—and concentrating instead on a more robust summary of F such as the median. Unfortunately, this approach is not available: the current regulatory process demands the focus on μ.

2 Progress toward a solution

The working group assembled for this problem included David Allwright, Ryan Cheal, David Draper, David Epstein, Chris Jennison, Robert Ketzscher, Callum McKail, Martin O'Gorman, and Vladic Putyatin. We pursued two avenues of attack in attempting to move toward an answer to the above question: an approach, based on a Chebyshev bound, developed by David Allwright, and a Bayesian nonparametric approach developed by David Draper and Callum McKail.

2.1 A Chebyshev bound approach

AEA Technology needs reliable estimates of the health risk at time t from a given quantity of radioactive waste that is buried and sealed at time 0. The risks at time t depend on many unknown parameters, e.g., the adsorption coefficients of the isotopes on the sealing materials (such as cement and concrete) and on the rocks, the rock porosities, and so on. Given the parameter vector $\mathbf{x}$, AEA have a deterministic model that calculates the risk $r(\mathbf{x}, t)$, and this is relatively expensive to calculate. The accepted practice is to gloss over the ignorance of the *true* value of $\mathbf{x}$ by regarding it as a *random* vector with probability density f, determined by expert opinion concerning the likely values and ranges of the physical, chemical, geological, radiological, biological and other parameters involved. Then it is necessary to evaluate the expectation of $r(\mathbf{x}, t)$ where $\mathbf{x}$ has the probability density f—in other words, we need to find

$$R(t) = \mathbf{E}(r(\mathbf{x}, t)) = \int r(\mathbf{x}, t) f(\mathbf{x}) d\mathbf{x}, \tag{1}$$

where $\mathbf{x}$ is the vector of parameters, $r(\mathbf{x}, t)$ is the risk at time t given $\mathbf{x}$, and $f(\mathbf{x})$ is the probability density function of $\mathbf{x}$. Part of AEA's method has been to draw sample vectors $\mathbf{x}$ from the distribution f, and calculate r for those $\mathbf{x}$. So, let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be independent random variables each with density f, and let $r_i(t) = r(\mathbf{x}_i, t)$. The $r_i(t)$ are independent, with mean $R(t)$ and variance $\sigma(t)^2$. We shall be thinking of a fixed t, and so we shall suppress the explicit mention of t as a variable. Clearly the average,

$$R_e = \frac{r_1 + \cdots + r_N}{N}, \tag{2}$$

is an unbiased estimator of R. The questions posed are:

- How do we get a confidence interval for R? And in particular,
- How do we bound the probability that R_e may seriously underestimate the true value of R, because $\mathbf{x}_1, \ldots, \mathbf{x}_N$ may happen to miss all the rare-but-bad events?

We recall that a 95% confidence interval on R means a random interval $[L, U]$ depending on $r_1, \ldots, r_N$ such that

$$\mathbf{P}(L \leq R \leq U) = 0.95. \tag{3}$$

For a symmetric confidence interval we wish U to have

$$\mathbf{P}(U < R) \leq 0.025, \tag{4}$$

i.e., the probability that U *under*estimates R is at most 1 in 40. (The health consequences of *over*estimating the risk are of course not a concern, so we concentrate on finding U rather than L.) One specific point raised by AEA was to find how the question is affected if there is a known upper bound on the values of r, so

$$0 \leq r(\mathbf{x}, t) \leq r_X(t) \tag{5}$$

for all $\mathbf{x}$. Again suppressing the t-dependence from the notation, the random variables r_i lie in

$$0 \leq r_i \leq r_X. \tag{6}$$

So

$$\sigma^2 = Var(r_i) \leq R(r_X - R), \tag{7}$$

from which

$$Var(R_e) = \frac{\sigma^2}{N} \leq \frac{R(r_X - R)}{N}. \tag{8}$$

Then by the one-sided Chebyshev bound,

$$\mathbf{P}(R_e \leq R - \sqrt{cVar(R_e)}) \leq \frac{1}{1+c}. \tag{9}$$

Hence

$$\mathbf{P}(R_e \leq R - \sqrt{\frac{cR(r_X - R)}{N}}) \leq \frac{1}{1+c}. \tag{10}$$

This implies that if we now let

$$U = \frac{2NR_e + cr_X + \sqrt{c^2 r_X^2 + 4NcR_e(r_X - R_e)}}{2(N+c)} \tag{11}$$

then we shall have

$$\mathbf{P}(U < R) \leq \frac{1}{1+c} \tag{12}$$

as required.

Thus if we take $c = 39$ then this U gives us the required upper end of a confidence interval, using the upper bound r_X as the only additional information. The importance of this kind of result is that it is entirely independent of the detailed distribution of the r_i, and it does not depend on assuming that R_e is approximately normally distributed. It is not a best-possible distribution-free result, because the inequalities used cannot all be sharp simultaneously. But it illustrates the kind of problems that arise, and the kind of approach that can be used, when it is the expected risk R that is the critical quantity.

2.1.1 Upper confidence limits on quantities other than the risk

Suppose instead of R that we wished to estimate a certain percentage point c_p of the risk distribution, i.e., we wish to estimate the value c_p such that

$$\mathbf{P}(r_i \leq c_p) = p. \tag{13}$$

Then it follows trivially from this that

$$\mathbf{P}(\max(r_i) \leq c_p) = p^N. \tag{14}$$

If we wish to obtain $p^N < \frac{1}{40}$ where p is close to 1, then we just let $N \approx \frac{3.7}{1-p}$ and $\max(r_1, \ldots, r_N)$ will be an upper confidence limit on c_p. This illustrates how much easier it is to obtain confidence intervals on quantities other than the integrated risk of the distribution.

2.2 A Bayesian nonparametric approach

Let $y = (y_1, \ldots, y_n)$ be the sampled doses resulting from the Monte Carlo runs for a given time horizon T. If the model inputs are all chosen randomly from their PDFs, the y_i will constitute an *IID* (independent, identically distributed) sample from the underlying cumulative distribution function (CDF) F. The quantity of interest is $\mu = \mu(F) = \int y \, dF(y)$, and the desired goal is inference about μ in the presence of substantial uncertainty about F. This is typically referred to as a *nonparametric* problem.

The standard frequentist nonparametric method in this situation is the *bootstrap*, which attempts to construct interval estimates for μ via IID sampling from the *empirical CDF* $\hat{F}_n$ (the distribution that places mass $\frac{1}{n}$ at each of the observed y_i). The problem with the bootstrap is that it has no way of learning anything about the part of F beyond the largest of the y_i, and it is precisely this region that has the largest influence on functionals such as μ. The result is interval estimates of μ that are too narrow and centered too far to the left with skewed and long-tailed data.

A natural way out of this dilemma is to incorporate prior information about the tail-weight of F in a Bayesian fashion. Probabilistically, the frequentist model $(y_i|F) \stackrel{\text{IID}}{\sim} F$ above is expanded by including a prior distribution on F:

$$\begin{aligned} F &\sim p(F) \\ (y_i|F) &\stackrel{\text{IID}}{\sim} F. \end{aligned} \tag{15}$$

This is a Bayesian nonparametric model, and the chief difficulties in working with it are of three kinds:

- Specifying $p(F)$ in a rich enough way to realistically encompass all practical possibilities is mathematically troublesome, because $p(F)$ is a distribution on function space rather than $\Re^k$;

- Choosing a $p(F)$ that directly incorporates whatever "soft" knowledge is available, even if it is of several types (e.g., I am 90% sure that F is unimodal, skewed, with a very long right tail; the mean of F is somewhere between μ_{lo} and μ_{hi}; and the variance of F is somewhere between σ^2_{lo} and σ^2_{hi}) is problematic; and

- Computing in this model is not easy, e.g., with a Monte Carlo approach a strategy for simulating from $p(F|y)$, the posterior distribution for F given the data, is needed.

We have been exploring a solution to a more difficult problem—that of nonparametrically capturing uncertainty about the *variance* of F, rather than its mean—which employs the family of *Dirichlet process* priors for $p(F)$. The bootstrap performs abysmally in this problem: even the currently best available bootstrap technology for constructing nonparametric confidence intervals (CIs), the BC_a method (Efron and Tibshirani, 1993), yields only 63% coverage of nominal 90% CIs with samples of size $n = 50$ from the lognormal distribution, and the coverage is still less than 80% with $n = 500$.

Walker et al. (1997) define the Dirichlet process as follows: Let Ω be a space and $\mathcal{B}$ a σ-field of subsets, and let α be a finite non-null measure $(\Omega, \mathcal{B})$. Then a stochastic process indexed by elements A of $\mathcal{B}$ is said to be a Dirichlet process on $(\Omega, \mathcal{B})$ with parameter α if, for any measurable partition $(A_1, \ldots, A_k)$ of Ω, the random vector $(F(A_1), \ldots, F(A_k))$ has a Dirichlet distribution[1] with parameter $(\alpha(A_1), \ldots, \alpha(A_k))$.

The standard notation is $F \sim \mathcal{D}(\alpha)$ and $\alpha(\cdot) = cF_0(\cdot)$ where $c > 0$ and F_0, a probability distribution on Ω, "centers" the process in the sense that $E(F) = F_0$. We have been using the lognormal distribution for F_0, as a prototypical highly skewed and long-tailed distribution. The scale parameter c is used as a tuning constant: it acts like a *prior sample size*—using the heuristic of a prior distribution being equivalent to a data set that is in effect merged with the y_i—and c can be chosen for good coverage behavior of interval estimates.

Sethuraman and Tiwari (1982) detail how to simulate from a Dirichlet process. If $F \sim \mathcal{D}(cF_0)$, we can generate an F via

$$F = \sum_{j=1}^{\infty} V_j \delta_{\theta_j}, \tag{16}$$

where $V_1 = W_1, V_j = W_j(1 - W_{j-1}) \ldots (1 - W_1), \quad j = 2, 3, \ldots,$ $W_1, W_2 \ldots$ are IID Beta$(1, c)$ and $\theta_1, \theta_2, \ldots$ are IID from F_0.

In practice, the infinite sum $\sum_{j=1}^{\infty} V_j \delta_{\theta_j}$ can be replaced by $\sum_{j=1}^{M} V_j \delta_{\theta_j}$, with M a suitably chosen value so that, for $N > M$, V_N is effectively zero. Figure 1 illustrates the "envelope" around the centering distribution F_0 that is produced by sampling from a Dirichlet process prior.

It turns out that as c increases, the envelope around F_0 becomes tighter. This is because the variance of draws from $\mathcal{D}(cF_0)$ is given by

$$V[F(A)] = \frac{F_0(A)(1 - F_0(A))}{c + 1}, \tag{17}$$

for any set A, and so increasing c decreases the variability about F_0.

[1] The Dirichlet distribution is defined as follows: if $\theta \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_k)$ the density function is $p(\theta) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1}; \theta_1, \ldots, \theta_k \geq 0; \sum_{j=1}^{k} \theta_j = 1$.

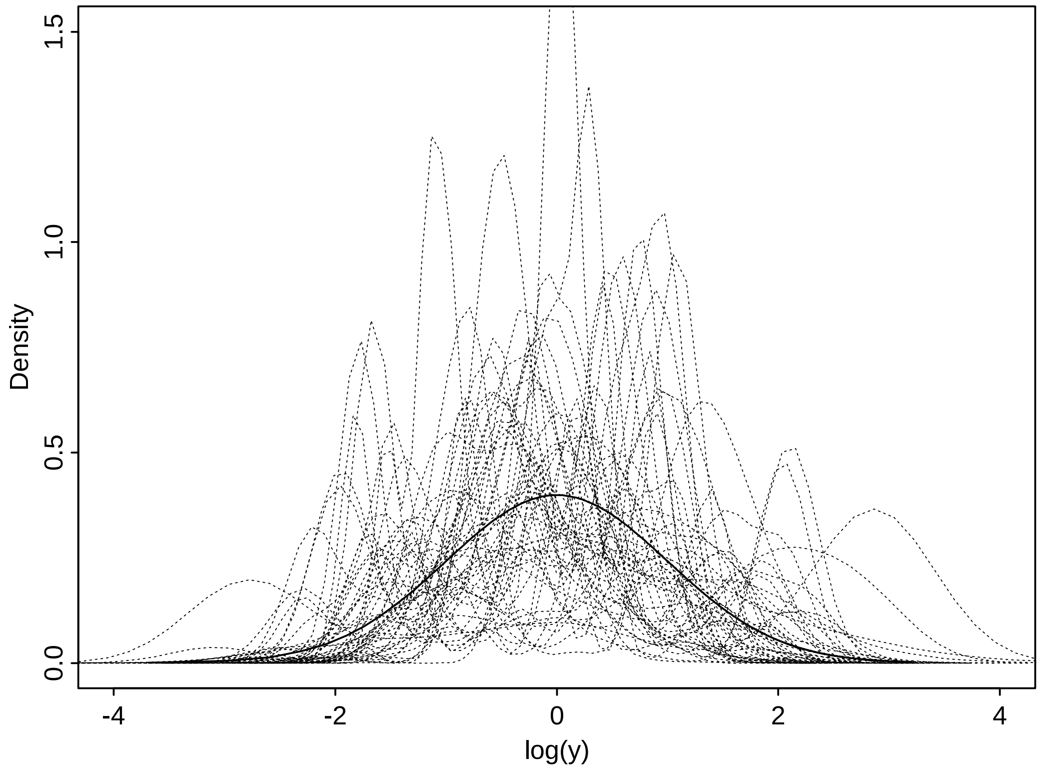

Figure 1. *Standard normal density (solid curve) and smoothed density traces of 50 draws (plotted on the log scale) from the Dirichlet process prior* $\mathcal{D}(cF_0)$ *with* $c = 5$ *and* $F_0 =$ *the standard lognormal distribution (dotted curves).*

The obvious drawback of sampling a distribution F from $\mathcal{D}(cF_0)$ is that the resulting F is discrete. This is somewhat unsatisfactory given that in general we are thinking of F as continuous, but—in trying to improve on the bootstrap—it is in keeping with the bootstrap's use of the discrete $\hat{F}_n$ as an approximation to F.

By means of simulation studies we have shown (McKail, 1997) that this approach holds significant promise to remedy the defects of the bootstrap. With suitable c we are able to achieve coverage essentially at nominal values for the variance when the data-generating mechanism ranges from lognormal to gamma to Gaussian (Table 1), and it is possible to find a value of c that acts as a good compromise between these three (and other) distributions. We have not yet tried this methodology on the problem of the mean of F, but its performance with the variance provides a lower bound on its effectiveness with the mean.

Table 1. *Actual coverage of nominal 90% intervals for the population variance with* $n = 50$ *and using a Dirichlet process prior centered at the lognormal distribution.*

Distribution	c	Actual Coverage
Gaussian	4.8	90.4
Gamma	14.1	89.8
Lognormal	20.2	90.4

3 Future work

- We look forward to working with Jim Sinclair to test (and improve as necessary) the Bayesian nonparametric method of Section 2.2 with his data.

- *Sensitivity analysis* should also prove useful: the idea would be to use importance sampling to adaptively oversample on (expanded versions of) regions in input space that lead to very large doses. In other words, an initial Monte Carlo IID sample of (say) $n_1 = 250$ model runs could be made and analyzed for input configurations that lead to large dose values; and then, in a second round of sampling, neighborhoods in input space could be defined around the promising configurations from round 1, with oversampling from these neighborhoods in round 2 using an additional (say) $n_2 = 250$ model runs likely to produce a more valuable set of y_i than obtainable with a purely IID sample of 500 runs.

References

Efron B, Tibshirani RJ (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

McKail C (1997). Fixing the broken bootstrap: Inference with skewed and long-tailed data. MSc dissertation, School of Mathematical Sciences, University of Bath.

Sethuraman J, Tiwari R (1982). Convergence of Dirichlet measures and the interpretation of their parameters. In *Proceedings of the Third Purdue Symposium on Statistical Decision Theory and Related Topics*, SS Gupta, JO Berger (eds.), New York: Academic Press.

Walker SG, Damien P, Laud PW, Smith AFM (1997). Bayesian nonparametric inference for random distributions and related functions. Technical report, Department of Mathematics, Imperial College, London.