A Bayesian approach to the assessment of contaminant spread Part I. Discrete case[1]

Tommy Norberg
Department of Mathematical Statistics
Chalmers University of Technology
and Göteborg University

21 February, 2002

Abstract

The paper provides a Bayesian method for estimating totals such as the polluted area or cost of remediation of a possibly contaminated hazardous waste site. After specifying a prior distribution on the total polluted area (here expert opinion may be taken into account), its focus is on how to chose the number of measurements to make in order to achieve a specified accuracy goal. This paper treats the discrete case in which the site is partitioned into a finite number of remediation units (cells). An accompanying paper [4] treats the continuous case.

Key words: *beta-binomial distribution, hypergeometric measurement model, pre-posterior analysis, cost distribution, value at risk, expected shortfall.*

[1] Preprint no 2002:09, `http://www.math.chalmers.se/Stat/Research/Preprints/`

1 The problem

In the very early stages of an investigation of a possibly contaminated hazardous waste site, there is often no interest in detailed maps contouring the contaminant spread with high precision. Instead focus is on estimating quantities such as the total polluted area A_c or the total remediation cost C. Often this kind of problem is handled by Kriging methods, the accuracy of which depends crucially on the quality of the variogram estimate (as noticed e g by Myers [3, p 379]). However, if one only is interested in totals, a Bayesian approach inspired by classical survey sampling theory may be both simpler to use and more appropriate since it does not rely on the variogram. Also, Kriging methods may provide bad answers if the range of the variogram (i e, spatial correlation length) is long compared to the size of the site under investigation or if the intrinsic hypothesis (see e g Wackernagel [6, p 36]) is questionable. In the former case, assessing the variogram is difficult, and both cases may very well turn out true for many hazardous man made waste sites.

It is supposed in this paper that the site is divided into N cells (remediation units), of which an unknown number T are contaminated, and that measurements will be made in n randomly selected cells in order to determine their contamination status. The paper addresses the problem of choosing the appropriate number n of cells to examine and its purpose is to provide some tools for making this choice. An accompanying paper [4] addresses the case when there is no natural division of the site into cells.

The approach of the paper is Bayesian and consists of specifying a mathematically convenient prior distribution of the number T of contaminated cells. This is done Section 3. We further in this section specify the measurement model, which conditionally on the value t of T is hypergeometric, and calculate the posterior distribution of T and the pre-posterior (or predictive) distribution of the actual result, X, of the measurements. In the section we also derive an explicit formula for the mean of the pre-posterior variance $\mathrm{V}[T|X]$, and show how to calculate other relevant pre-posterior means such as $E[c_p(X)]$, where $c_p(x)$ is the posterior pth quantile of a cost distribution. These quantities are of major importance when it comes to making decisions on n. In Section 2 some plausible cost functions C are mentioned and studied to some depth. In Section 4 we provide some faked case studies, the purposes of which are to illustrate the results of this paper and how they may be used in the process of determining a suitable number n of cells to investigate. Finally, in Section 5, we summarise and make some general comments.

Whenever the number T of contaminated cells is provided with a probability distribution, the expectation $E[T]$ and variance $\mathrm{V}[T]$ may be calculated w r t this distribution. These quantities, however are not the only ones that provide valuable insight into our problem. We thus define the pth quantile

$$t_p = \min\{t : P(T \leq t) \geq p\}$$

and write $\tilde{t}$ instead of $t_{0.5}$. Clearly, t_p is the unique whole number satisfying $P(T \leq t_p) \geq p$ and $P(T < t_p) < p$. We further let

$$\hat{t} = \arg\max_t P(T = t)$$

be the most likely value of T.

A to this paper central distribution is the beta-binomial, the definition and some facts of which are given in Appendix A. Included is also a short appendix quoting some results for the gamma distribution.

2 Some cost functions

Let C be the total cost of remediating the T contaminated cells.

2.1 $C = \mu T$

The most obvious way in which C could be specified is to think of a unit cost μ per contaminated cell, making $C = \mu T$. Then

$$E[C] = \mu E[T]$$

and

$$\mathrm{V}[C] = \mu^2 \mathrm{V}[T]$$

Moreover, any quantile in the distribution of C is just μ times the corresponding quantile in the distribution of T.

2.2 $C|T \sim \mathrm{N}(\mu T, \sigma^2 T)$

The cost of remediating one cell, however, need not be deterministic. It may very well be random with some mean μ and standard deviation σ. If the total number T of cells to remediate is large then, by the central limit theorem, the total cost C is (approximately) normal with mean μT and variance $\sigma^2 T$. If the central limit theorem is not applicable, there is a need for a specific model for the cost of remediating one cell.

An example of the latter is the gamma distribution. Suppose that the cost of remediating one cell is gamma distributed with parameters $p > 0$, $\lambda > 0$. (Refer to Appendix B for the definition and some facts of the gamma distribution.) If each such cost is independent of all others, then, by the addition property of the gamma distribution, the total cost C is gamma distributed with parameters pT and λ. Below it will be assumed that C is normal with mean μT and variance $\sigma^2 T$. Results for the gamma case are obtained similarly.

Assume, however, first only that $E[C|T] = \mu T$ and $\mathrm{V}[C|T] = \sigma^2 T$. (If C is gamma distributed with parameters pT and λ, then $\mu = p/\lambda$ and $\sigma^2 = p/\lambda^2$.) Then, by the double expectation formula,

$$E[C] = E[E[C|T]] = \mu E[T]$$

as in Section 2.1, and

$$\begin{aligned} \mathrm{V}[C] &= E[\mathrm{V}[C|T]] + \mathrm{V}[E[C|T]] \\ &= \sigma^2 E[T] + \mu^2 \mathrm{V}[T] \end{aligned}$$

Notice the extra variance component compared to the case studied in Section 2.1. Notice also that by letting $\sigma = 0$, we are back in the case studied in that section.

Next, let C, given $T = t$, be normal with mean μt and variance $\sigma^2 t$. Then

$$\begin{aligned} F(c) &= P(C \leq c) \\ &= \sum_t P(C \leq c|T = t)P(T = t) \\ &= \sum_t \Phi\left(\frac{c - \mu t}{\sigma\sqrt{t}}\right) P(T = t) \\ &= E\left[\Phi\left(\frac{c - \mu T}{\sigma\sqrt{T}}\right)\right] \end{aligned}$$

where Φ is the standard normal distribution function, given by

$$\Phi(z) = \int_{-\infty}^{z} \varphi(t)\, dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

The density $f(c)$ of C is obtained by derivation of $F(c)$. Interchanging the order of derivation and expectation is permitted, since the support of T is finite. Thus,

$$\begin{aligned} f(c) &= \frac{d}{dc} F(c) \\ &= E\left[\varphi\left(\frac{c-\mu T}{\sigma\sqrt{T}}\right)\frac{1}{\sigma\sqrt{T}}\right] \\ &= E\left[\frac{1}{\sqrt{2\pi\,\sigma^2 T}}\, e^{-(c-\mu T)^2/(2\sigma^2 T)}\right] \end{aligned}$$

Notice finally that the pth quantile c_p in the distribution of C solves $F(c) = p$, since F is continuous. Thus the risk of getting a cost bigger than c_p is $1 - p$. One may therefore refer to any quantile c_p as a *value at risk* (VaR) $1-p$ if the latter is small. Another quantity, which is of great interest besides the VaR, is its associated *expected shortfall*, which by definition is the conditional expectation of C given that $C \geq c_p$, to be denoted e_p. Thus,

$$e_p = E[C|C \geq c_p] = \frac{1}{1-p}\int_{c_p}^{\infty} c f(c)\, dc$$

3 Bayesian analysis

3.1 Prior distribution of T

Recall that the site of interest is divided into N cells, of which T are contaminated. As prior distribution of T we take the beta-binomial distribution with parameters N and $\alpha, \beta > 0$, to be referred to as $\mathrm{BB}(N, \alpha, \beta)$ (refer to Appendix A for the definition and some facts for the beta-binomial distribution). The prior belief is conveyed in the parameters α, β. The interpretation is that there are $\alpha + \beta - 2$ extra hypothetical known cells, of which $\alpha - 1$ are contaminated and the remaining $\beta - 1$ are not. In particular $\alpha = \beta = 1$ corresponds to no prior knowledge and in this case T is uniform on $\{0, 1, \dots, N\}$.

Our reasons for choosing this prior distribution are intrinsically mathematical. It will be seen below that if the measurement model is hypergeometric, then also the posterior distribution of T is beta-binomial. Surprisingly also the pre-posterior (or predictive) distribution of X, the number of contaminated cells found out of n examined, is beta-binomial.

Write $p(t) = P(T = t)$ for the prior probability mass function of T. Then

$$p(t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{N!}{\Gamma(\alpha+\beta+N)} \frac{\Gamma(\alpha+t)}{t!} \frac{\Gamma(\beta+N-t)}{(N-t)!}$$

for $t = 0, 1, \ldots, N$. Moreover, the prior mean and variance are

$$E[T] = N \frac{\alpha}{\alpha+\beta}$$

and

$$\mathrm{V}[T] = N \frac{\alpha}{\alpha+\beta} \frac{\beta}{\alpha+\beta} \frac{\alpha+\beta+N}{\alpha+\beta+1}$$

respectively. Thus, $p_C = \alpha/(\alpha+\beta)$ is our prior guess of the proportion of contaminated cells and the sum $\alpha+\beta$ measures our belief in it. A large value of $\alpha+\beta$ corresponds to a small prior variance. Moreover, as $\alpha+\beta \uparrow \infty$ holding $p_C = \alpha/(\alpha+\beta)$ fixed, $\mathrm{V}[T] \downarrow N p_C(1-p_C)$, and T converges to a binomial random variable.

3.2 The measurement model

Let X be the number of contaminated cells found by examining n randomly selected cells. Write $p(x|t) = P(X = x|T = t)$. Then

$$p(x|t) = \frac{\binom{t}{x}\binom{N-t}{n-x}}{\binom{N}{n}}$$

for $\max[0, n-(N-t)] \leq x \leq \min(n, t)$. This is the simplest possible measurement model. We emphasise that it is crucial for the method of this paper to be applicable that the n cells are randomly sampled. Notice also that it is assumed that it is possible to determine with certainty the contamination status of each cell. Notice finally that the conditional mean and variance of X are

$$E[X|t] = n \frac{t}{N}$$

and

$$\mathrm{V}[X|t] = n \frac{t}{N} \frac{N-t}{N} \frac{N-n}{N-1}$$

respectively.

3.3 The posterior distribution of T

Let $p(t|x) = P(T = t|X = x)$. An appropriate use of Bayes' rule yields

$$p(t|x) \propto \frac{\Gamma(\alpha+t)}{(t-x)!} \frac{\Gamma(\beta+N-t)}{(N-n+x-t)!}$$

for $x \le t \le N-n+x$. Hence, conditionally on $X = x$,

$$T - x \sim \mathrm{BB}(N-n, \alpha+x, \beta+n-x)$$

Therefore,

$$p(t|x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \frac{(N-n)!}{\Gamma(\alpha+\beta+N)} \frac{\Gamma(\alpha+t)}{(t-x)!} \frac{\Gamma(\beta+N-t)}{(N-n+x-t)!}$$

for $x \le t \le N-n+x$, and the posterior mean and variance are

$$E[T|x] = x + (N-n)\,\frac{\alpha+x}{\alpha+\beta+n}$$

and

$$\mathrm{V}[T|x] = (N-n)\,\frac{\alpha+x}{\alpha+\beta+n}\,\frac{\beta+n-x}{\alpha+\beta+n}\,\frac{\alpha+\beta+N}{\alpha+\beta+n+1}$$

respectively. Other posterior quantities of interest are the pth quantile

$$t_p(x) = \min\{t : P(T \le t|X = x) \ge p\}$$

(e g the median $\tilde{t}(x) = t_{0.5}(x)$) and the most likely value of T,

$$\hat{t}(x) = \arg\max_t p(t|x)$$

Next, let C be the total remediation cost. Assume first that $E[C|T] = \mu T$ and $\mathrm{V}[C|T] = \sigma^2 T$ (cf Section 2.2). Then the posterior expectation and variance of C are

$$E[C|x] = \mu E[T|x]$$

and

$$\mathrm{V}[C|x] = \sigma^2 E[T|x] + \mu^2 \mathrm{V}[T|x]$$

respectively.

Finally, assume that C, conditionally on $T = t$, is normal with mean μt and variance $\sigma^2 t$ as in Section 2.2. Write $F(c|x)$ and $f(c|x)$ for the distribution function and density of the posterior cost distribution (i e, the distribution of C w r t the posterior distribution of T). Then,

$$F(c|x) = E\left[\Phi\left(\frac{c-\mu T}{\sigma\sqrt{T}}\right)\middle| X = x\right]$$

and

$$f(c|x) = E\left[\varphi\left(\frac{c-\mu T}{\sigma\sqrt{T}}\right)\frac{1}{\sigma\sqrt{T}}\middle| X = x\right]$$

Some cost quantities of interest in addition to the posterior mean $E[C|x]$ and variance $\mathrm{V}[C|x]$ mentioned already, are the posterior VaR $1-p$, $c_p(x)$, and the associated expected shortfall

$$e_p(x) = \frac{1}{1-p}\int_{c_p(x)}^{\infty} cf(c|x)\,dc$$

3.4 The pre-posterior (or predictive) distribution of X

Pre-posterior to doing the actual measurements, their outcome, X, is random with probability mass function $p(x) = P(X = x)$. Since $p(x)p(t|x) = p(t)p(x|t)$,

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\frac{n!}{\Gamma(\alpha+\beta+n)}\,\frac{\Gamma(\alpha+x)}{x!}\,\frac{\Gamma(\beta+n-x)}{(n-x)!}$$

for $0 \leq x \leq n$. Thus $X \sim \mathrm{BB}(n, \alpha, \beta)$, so the pre-posterior mean and variance of X are

$$E[X] = n\,\frac{\alpha}{\alpha+\beta}$$

and

$$\mathrm{V}[X] = n\,\frac{\alpha}{\alpha+\beta}\,\frac{\beta}{\alpha+\beta}\,\frac{\alpha+\beta+n}{\alpha+\beta+1}$$

respectively.

3.5 Some further pre-posterior means

A purpose of this section is to derive an explicit formula for the mean $E[\mathrm{V}[T|X]]$ of the pre-posterior variance $\mathrm{V}[T|X]$. Notice, however, first that

$$E[E[T|X]] = E[T]$$

by the double expectation formula. Thus, in practise by finding x contaminated cells when examining n randomly sampled, a new (posterior) mean of T is obtained. However, if the prior distribution is reasonably well chosen, this updated mean can not be expected to be different from $E[T]$.

Contrary to this,

$$E[\mathrm{V}[T|X]] < \mathrm{V}[T]$$

since

$$\mathrm{V}[T] = E[\mathrm{V}[T|X]] + \mathrm{V}[E[T|X]]$$

and

$$E[T|X] = X + (N-n)\,\frac{\alpha + X}{\alpha+\beta+n}$$

making

$$\mathrm{V}[E[T|X]] > 0$$

Indeed, $E[\mathrm{V}[T|X]]$ is a strictly decreasing function of n that equals 0 when $n = N$, as we now shall see.

First notice,

$$\mathrm{V}[T|X] = (N-n)\,\frac{\alpha+X}{\alpha+\beta+n}\,\frac{\beta+n-X}{\alpha+\beta+n}\,\frac{\alpha+\beta+N}{\alpha+\beta+n+1}$$

Cf Section 3.3.

Lemma 1 *Let $X \sim \mathrm{BB}(n, \alpha, \beta)$. Then*

$$E[(\alpha+X)(\beta+n-X)] = \frac{\alpha\beta(\alpha+\beta+n)(\alpha+\beta+n+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

Proof Notice

$$\begin{aligned}E[(\alpha+X)(\beta+n-X)] &= E[(\alpha+X)(\alpha+\beta+n-(\alpha+X))]\\ &= (\alpha+\beta+n)E[\alpha+X] - E[(\alpha+X)^2]\\ &= (\alpha+\beta+n)E[\alpha+X] - E[\alpha+X]^2 - \mathrm{V}[X]\end{aligned}$$

The remaining part of the proof consists of straightforward algebraic manipulations. □

Hence, by Lemma 1,

$$E[\mathrm{V}[T|X]] = \frac{N-n}{N}\frac{\alpha+\beta}{\alpha+\beta+n}\mathrm{V}[T]$$

Notice also that in spite of the fact that we have no closed form expression for the pre-posterior mean of the pth quantile, $t_p(X)$, it can readily be calculated numerically by means of the formula

$$E[t_p(X)] = \sum_x t_p(x)p(x)$$

where $p(x)$ is the probability mass function of $X \sim \mathrm{BB}(n, \alpha, \beta)$.

Next, let C be the total remediation cost. If $E[C|T] = \mu T$ and $\mathrm{V}[C|T] = \sigma^2 T$ as in Section 2.2, then the pre-posterior variance of C is

$$\mathrm{V}[C|X] = \sigma^2 E[T|X] + \mu^2 \mathrm{V}[T|X]$$

from which we conclude,

$$\begin{aligned} E[\mathrm{V}[C|X]] &= \sigma^2 E[E[T|X]] + \mu^2 E[\mathrm{V}[T|X]] \\ &= \sigma^2 E[T] + \mu^2 \frac{N-n}{N}\frac{\alpha+\beta}{\alpha+\beta+n}\mathrm{V}[T] \end{aligned}$$

Notice that it is only the variance component that depends on the uncertainty in T, that can be reduced by means of taking further samples. Notice also that if

$$n \approx \frac{N(\alpha+\beta)(\mu^2\mathrm{V}[T] - \sigma^2 E[T])}{\mu^2(\alpha+\beta)\mathrm{V}[T] + N\sigma^2 E[T]}$$

then both variance components are approximately equal, and

$$E[\mathrm{V}[C|X]] \approx 2\sigma^2 E[T]$$

Finally, in the case when C is normal with mean μT and variance $\sigma^2 T$, we may calculate the posterior VaR $1-p$, $c_p(x)$ and the associated expected shortfall $e_p(x)$. The pre-posterior means of these quantities may readily be calculated by means of the formula

$$E[h(X)] = \sum_x h(x)p(x)$$

where h is either c_p or e_p.

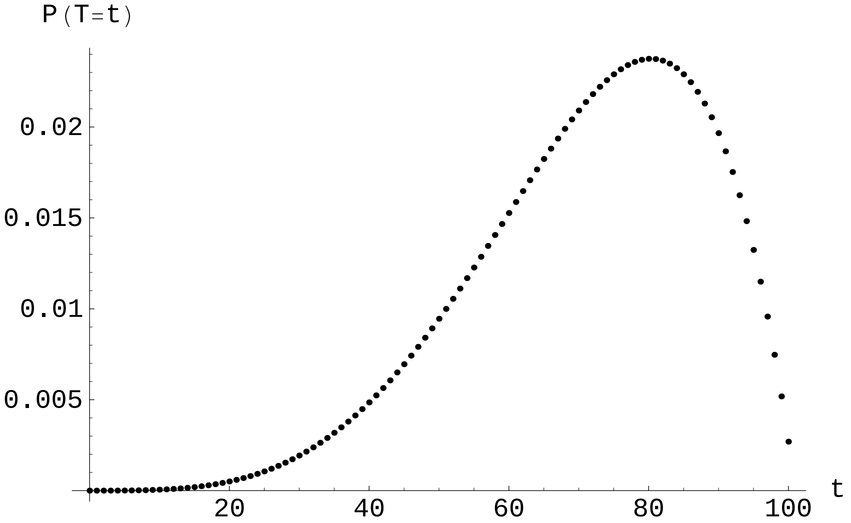

Figure 1: Prior probability mass function of $T \sim \text{BB}(100, 5, 2)$ in Case study A ($E[T] = 71.4$, $\text{V}[T] = 273.0 = 16.5^2$, $\tilde{t} = 74$, $\hat{t} = 80$, $t_{0.99} = 98$, $t_{0.95} = 94$, $t_{0.9} = 91$, $t_{0.75} = 84$).

4 Some faked case studies

4.1 Case study A

In Case study A, the site of interest is divided into $N = 100$ equally sized cells. The aim was to provide a sufficiently good prediction of T. Local experts were of the opinion that $E[T/N] \approx 0.7$ and that it is very likely that $T/N \geq 0.5$. Thus the probability $q = P(T/N \geq 0.5)$ was considered to be large.

The first of these subjective facts, says that $3\alpha \approx 7\beta$. Notice then that a value close to one of the probability q corresponds to a large value of $\alpha + \beta$ and to a strong belief in the opinion of the experts. In order to put not too much nor too less faith in the experts opinion, the compromise $q \approx 0.9$, which is obtained for $\alpha + \beta \approx 7$, was accepted. Therefore $\alpha = 5, \beta = 2$ seems to be a reasonable specification of the prior information and as prior distribution of T we thus take $\text{BB}(100, 5, 2)$, which is plotted in Figure 1. Its mean and variance are $E[T] = 71.4$ and $\text{V}[T] = 273.0 = 16.5^2$ respectively. Moreover, $\tilde{t} = 74$, $\hat{t} = 80$ and $t_{0.99} = 98$, $t_{0.95} = 94$, $t_{0.9} = 91$ and $t_{0.75} = 84$.

It was decided that a reasonable goal of the measurements is to reduce

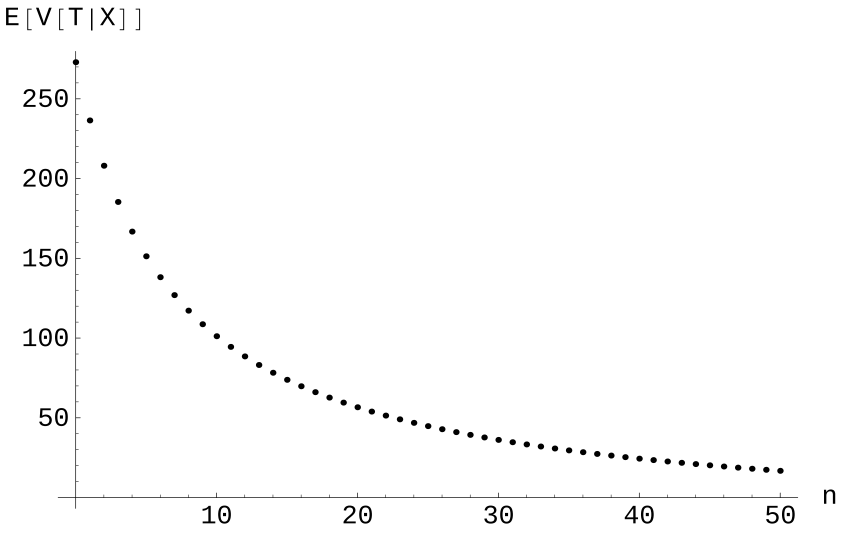

Figure 2: Plot of $E[\mathrm{V}[T|X]]$ vs n in Case study A. For $n = 20$, $E[\mathrm{V}[T|X]] = 56.6$.

the variance of T by a factor of four to below 70. In order to determine how many cells that need to be examined, $E[\mathrm{V}[T|X]]$ was plotted as a function of n. See Figure 2. The plot reveals that $E[\mathrm{V}[T|X]] < 70$ for $n \geq 16$.

Figure 3 furthermore shows a plot of $E[t_{0.95}(X)]$ vs n.

Next the risk of getting a posterior variance bigger than 70 was calculated for some selected values of n. Table 1 shows the range $x_1 \leq x \leq x_2$ of outcomes x of X for which $\mathrm{V}[T|x] > 70$ and the probability $P(\mathrm{V}[T|X] > 70)$ for $13 \leq n \leq 22$. For $n \geq 22$, $P(\mathrm{V}[T|X] > 70) = 0$.

n	x_1	x_2	risk
13	0	11	0.778
14	0	11	0.657
15	1	11	0.550
16	1	12	0.580
17	2	12	0.489

n	x_1	x_2	risk
18	3	12	0.411
19	4	12	0.344
20	5	12	0.286
21	7	11	0.163
22	-	-	0.000

Table 1: Tabulation of the range $x_1 \leq x \leq x_2$ of outcomes x of X for which $\mathrm{V}[T|x] > 70$ and the risk $P(x_1 \leq X \leq x_2)$ of obtaining $\mathrm{V}[T|X] > 70$ for $13 \leq n \leq 22$ in Case study A. For $n \geq 22$, $P(\mathrm{V}[T|X] > 70) = 0$.

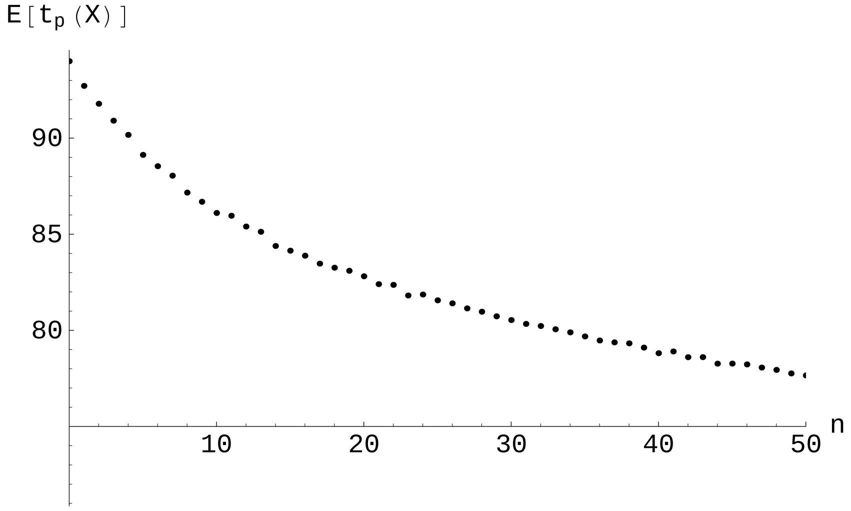

Figure 3: Plot of $E[t_{0.95}(X)]$ vs n in Case study A. For $n = 20$, $E[t_{0.95}(X)] = 82.8$.

The final decision was to examine $n = 20$ cells, and that if $5 \leq x \leq 12$, in which case the posterior variance of T still is bigger than 70, perhaps examine some more cells until the posterior variance is smaller than 70.

The result of the measurements was $x = 15$, so there was no need to do further examinations. The posterior distribution of T is $15 + \mathrm{BB}(80, 20, 7)$, which is plotted in Figure 4. The posterior mean and variance of T are $E[T|X = 15] = 15 + 59.3 = 74.3$ and $\mathrm{V}[T|X = 15] = 58.7 = 7.7^2$, respectively. Moreover, the posterior median, mode and some quantiles are $\tilde{t}(15) = 75$, $\hat{t}(15) = 76$ and $t_{0.99}(15) = 89$, $t_{0.95}(15) = 86$, $t_{0.90}(15) = 84$ and $t_{0.75}(15) = 80$.

4.2 Case study B

A small site is divided into 48 cells of which 4 'hopefully' randomly selected already are known to be contaminated and 3 not. Prior to invoking this knowledge, assume $T \sim \mathrm{BB}(48, 1, 1)$ (this is a uniform distribution on $0 \leq t \leq 48$). Then, after invoking it, $T - 4 \sim \mathrm{BB}(41, 5, 4)$.

Next, after doing measurements in n randomly selected cells, $T - 4 - x \sim$

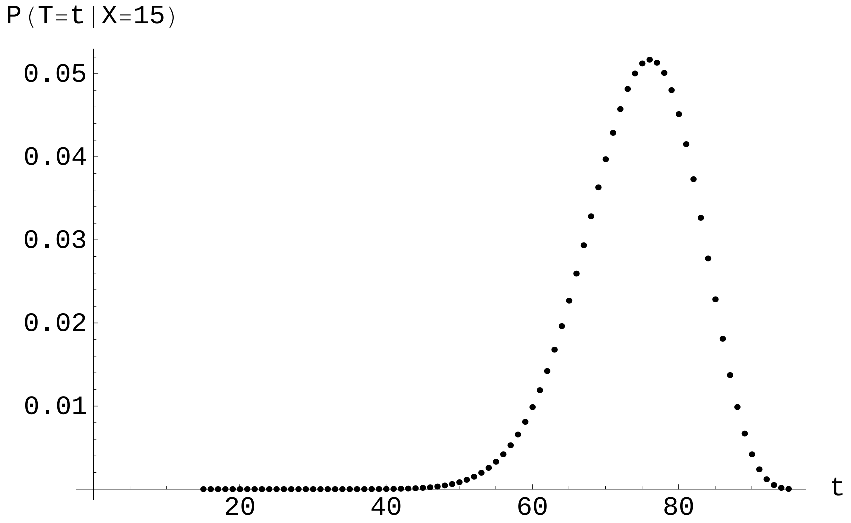

Figure 4: Posterior probability mass function of $T \sim 15 + \mathrm{BB}(80, 20, 7)$ after observing $x = 15$ contaminated cells out of $n = 20$ randomly selected in Case study A ($E[T|X = 15] = 74.3$, $\mathrm{V}[T|X = 15] = 58.7 = 7.7^2$, $\tilde{t}(15) = 75$, $\hat{t}(15) = 76$, $t_{0.99}(15) = 89$, $t_{0.95}(15) = 86$, $t_{0.90}(15) = 84$, $t_{0.75}(15) = 80$).

$\mathrm{BB}(41 - n, 5 + x, 4 + n - x)$. The pre-posterior mean of $\mathrm{V}[T|X]$ is

$$E[\mathrm{V}[T|X]] = \frac{41 - n}{41} \frac{9}{9 + n} \mathrm{V}[T]$$

where $\mathrm{V}[T] = 50.6$.

4.3 Case study C

Assume that the site of interest is divided into, say $N = 50$, cells and that the remediation of a cell is quite expensive and variable. Say that the mean cost is $\mu = 10$ with a standard deviation of $\sigma = 10/3$ in some monetary unit. Assume further that there is no prior information, so $\alpha = \beta = 1$ making, prior to any measurement, $T \sim \mathrm{BB}(50, 1, 1)$.

Then $E[T] = 25$ and $\mathrm{V}[T] = 216.7 = 14.7^2$ so that, $E[C] = 250$ and $\mathrm{V}[C] = 21944.4 = 148.1^2$ (cf Section 2.2). If $\sigma = 0$, $\mathrm{V}[C] = 21666.7 = (147.2)^2$. Thus, in spite of the fact that σ seems to be relatively large compared to μ, the randomness in the cost of remediating one cell does not contribute much to the prior calculation of the uncertainty in the total cost C.

Referring to Sections 2.2 and 3.5, we next see that the means of $E[C|X]$ and $\mathrm{V}[C|X]$ are

$$E[E[C|X]] = E[C] = 250$$

and

$$E[\mathrm{V}[C|X]] = 277.8 + 21666.7\,\frac{50-n}{50}\,\frac{2}{2+n}$$

Notice that making measurements does not decrease the uncertainty component that is due to the fact that $\sigma > 0$. Notice also that the two variance components are approximately equal and that

$$E[\mathrm{V}[C|X]] = 566.7 \approx 2 \cdot 277.8$$

if $n = 37$.

After discussions similar to those indicated in Case study A (cf Section 4.1), it is decided to examine $n = 7$ cells. For this n, $E[\mathrm{V}[C|X]] = 4418.5 = 66.5^2$. Thus the fact that $\sigma > 0$ contributes considerably more to the pre-posterior mean of $\mathrm{V}[C|X]$ than to $\mathrm{V}[C]$.

4.4 Case study D

Here $N = 800$ and the experts have agreed on using $\mathrm{BB}(800, 4.5, 1.5)$ as prior distribution for T. Their argument was that then $\alpha/(\alpha+\beta) = 0.75$ and $P(T \geq 400) = 0.912$. It implies $E[T] = 600$ and $\mathrm{V}[T] = 17271.4 = 131.4^2$.

It is assumed that the total cost of remediating T cells is normal with mean μT and variance $\sigma^2 T$ as in Section 2.2. The aim of this case study is to illustrate how the parameters of the prior and posterior cost distributions vary as a function of σ/μ. The results are stated in units of μ or μ^2. It is assumed that $n = 80$ randomly selected cells are examined and that $x = 60$ are found to be contaminated, so that posterior $T \sim 60 + \mathrm{BB}(720, 64.5, 21.5)$. Notice, $E[T|X = 60] = 600$ and $\mathrm{V}[T|X = 60] = 1250.7 = 35.6^2$.

Figure 5 show plots of the prior and posterior distributions of T. Figure 6 show plots of the prior and posterior distributions of the cost C in the case when $\sigma/\mu = 1/5$. In both figures, the posterior distribution is concentrated around the value 600.

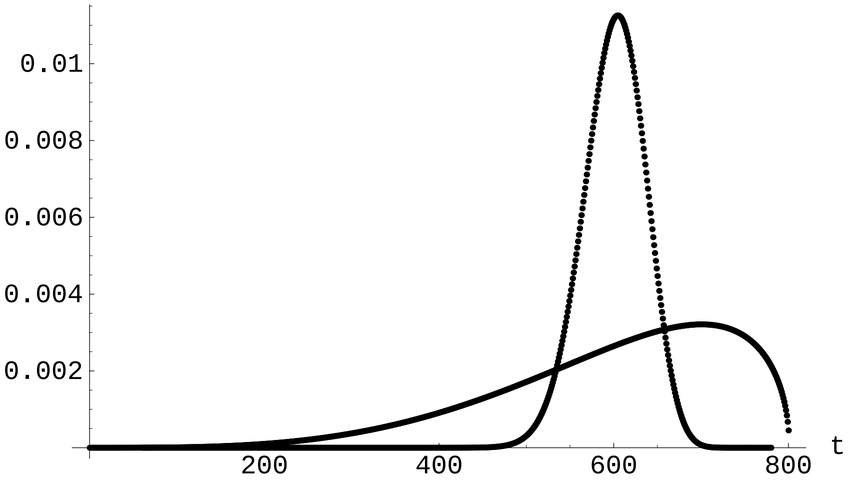

Figure 5: Plots of the prior (BB(800, 4.5, 1.5)) and posterior (60 + BB(720, 64.5, 21.5)) distributions of T in Case study D.

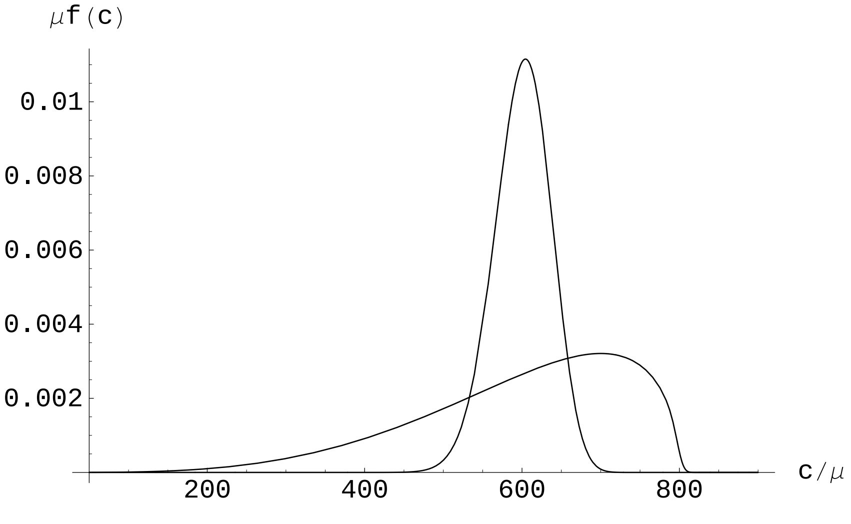

Figure 6: Plots of the prior and posterior cost distributions for the case $\sigma = \mu/5$ in Case study D.

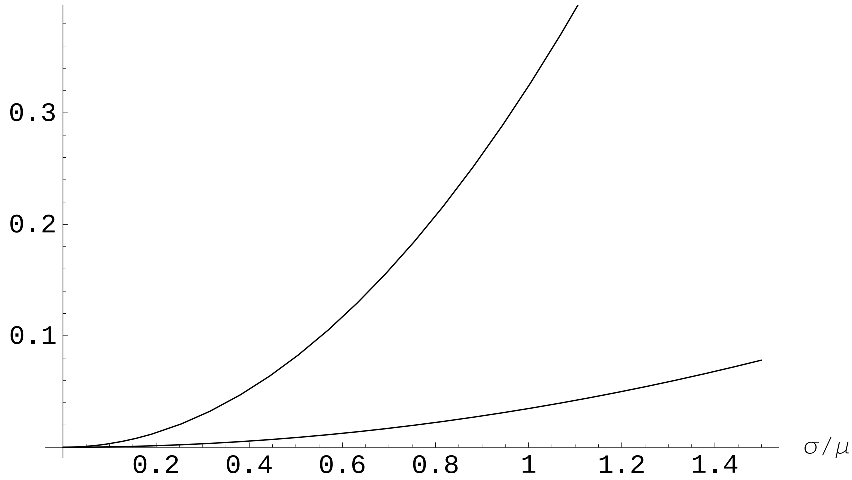

Figure 7: Plot of $E[T](\sigma/\mu)^2/\mathrm{V}[T]$ as a function of σ/μ in Case study D. The upper curve is w r t the posterior distribution of T, given that $X = 60$, while the lower is w r t the prior distribution.

Next notice,

$$\mathrm{V}[C] = \left(600\left(\frac{\sigma}{\mu}\right)^2 + 17271.4\right)\mu^2$$

$$\mathrm{V}[C|X = 60] = \left(600\left(\frac{\sigma}{\mu}\right)^2 + 1250.7\right)\mu^2$$

Refer to Figure 7 for plots of the ratios $600(\sigma/\mu)^2/17271.4$ and $600(\sigma/\mu)^2/1250.7$ as functions of σ/μ. Table 2 further tabulates some quantiles in the prior and posterior distributions of C for various values of σ/μ. Table 3 finally tabulates the VaR 0.05 (i e, $c_{0.95}$) and the corresponding expected shortfall $e_{0.95}$ w r t the prior and the posterior distribution of C, for various values of σ/μ.

5 Summary, some conclusions and extensions

In this paper we have shown that if the prior distribution on the number T of contaminated cells is $\mathrm{BB}(N, \alpha, \beta)$ (see Appendix A), where N is the

σ/μ	$c_{0.999}$	$c_{0.99}$	$c_{0.95}$	$c_{0.90}$	$c_{0.75}$	$\tilde{c}$	$\hat{c}$
1	844.4	812.7	778.3	755.3	704.0	621.8	696.6
1/2	816.6	797.4	772.8	753.2	704.2	622.9	698.8
1/3	808.6	793.9	772.0	752.9	704.2	623.0	699.9
1/4	805.0	792.6	771.8	752.8	704.2	623.1	700.2
1/5	803.1	792.0	771.7	752.8	704.3	623.1	700.4
1/10	797.1	791.3	771.6	752.7	704.3	623.2	700.7
	799	791	771	753	704	623	700

σ/μ	$c_{0.999}$	$c_{0.99}$	$c_{0.95}$	$c_{0.90}$	$c_{0.75}$	$\tilde{c}$	$\hat{c}$
1	727.0	697.0	669.6	654.7	629.4	600.7	602.0
1/2	704.4	681.4	659.5	647.2	626.0	601.2	603.8
1/3	699.4	678.0	657.3	645.7	625.3	601.4	604.3
1/4	697.5	676.8	656.6	645.1	625.1	601.4	604.3
1/5	696.6	676.2	656.2	644.9	625.0	601.4	604.4
1/10	695.4	675.4	655.7	644.5	624.8	601.5	604.5
	695	675	656	644	625	601	605

Table 2: Tabulation of some quantiles (values at risk) for various values of the ratio σ/μ in units of μ in Case study D. The last line shows the corresponding quantiles in the distribution of T. The upper table is w r t the prior distribution, while the lower is w r t the posterior distribution of C.

	Prior		Posterior	
σ/μ	$c_{0.99}$	$e_{0.99}$	$c_{0.99}$	$e_{0.99}$
1	778.3	799.4	669.6	686.4
1/2	772.8	788.0	659.5	672.9
1/3	772.0	785.5	657.3	670.0
1/4	771.8	784.6	656.6	668.9
1/5	771.7	784.2	656.2	668.4
1/10	771.5	783.6	655.7	667.8

Table 3: Tabulation of the VaR 0.05 (i e, $c_{0.95}$) and the corresponding expected shortfall $e_{0.95}$ w r t the prior and the posterior distribution of C, for various values of the ratio σ/μ in units of μ in Case study D.

total number of cells and $\alpha, \beta > 0$, and x contaminated cells are found by examining n randomly selected cells, then the posterior distribution of the remaining number $T' = T - x$ of contaminated cells out of now totally $N' = N - n$ unknown cells is $\mathrm{BB}(N-n, \alpha+x, \beta+n-x)$.

We next showed that the pre-posterior (or predictive) distribution of the number X of contaminated cells to be found in an investigation of n randomly selected cells, is $\mathrm{BB}(n, \alpha, \beta)$.

We were then able to calculate explicitly the mean of the pre-posterior variance $\mathrm{V}[T|X]$ w r t this distribution,

$$E[\mathrm{V}[T|X]] = \frac{N-n}{N} \frac{\alpha+\beta}{\alpha+\beta+n} \mathrm{V}[T]$$

Another pre-posterior quantity of interest is the pth quantile $t_p(X)$, the mean of which may easily be calculated according to the formula

$$E[t_p(X)] = \sum_x t_p(x) p(x)$$

where $p(x)$ is the probability mass function of $X \sim \mathrm{BB}(n, \alpha, \beta)$.

We also showed how to calculate the prior and posterior distribution of the total cost C in the case when C is normal with mean μT and variance $\sigma^2 T$. We gave explicit formulae for how to calculate $\mathrm{V}[C]$ w r t the prior distribution,

$$\mathrm{V}[C] = \sigma^2 E[T] + \mu^2 \mathrm{V}[T]$$

and for how to calculate the mean of the pre-posterior variance $\mathrm{V}[C|X]$ in terms of the prior quantities $E[T]$ and $\mathrm{V}[T]$,

$$E[\mathrm{V}[C|X]] = \sigma^2 E[T] + \mu^2 \frac{N-n}{N} \frac{\alpha+\beta}{\alpha+\beta+n} \mathrm{V}[T]$$

Notice that these two formulae hold true if only $E[C|T] = \mu T$ and $\mathrm{V}[C|T] = \sigma^2 T$. Other pre-posterior means of interesting posterior quantities like the VaR $1-p$, $c_p(x)$, and the associated expected shortfall $e_p(x)$ may readily be calculated by means of the formula

$$E[h(X)] = \sum_x h(x) p(x)$$

where $p(x)$ is the probability mass function of $X \sim \mathrm{BB}(n, \alpha, \beta)$ and h is either c_p or e_p.

These results were illustrated in 4 faked case studies. The purpose of Case study A was to indicate how expert knowledge can be summarised in the parameters α and β, and to show how pre-posterior means and probabilities can assisst in determining the value of the number n of randomly sampled cells to examine.

In Case study B, notice that if the 7 known cells are not randomly sampled, then the prior mean of T may be biased. In such a case, it may still be a good idea to let the knowledge from the known cells carefully influence the choice of prior distribution. However, one must be aware of the direction of the biasedness. Notice also that in such a case it is important that the number n of randomly sampled cells is not too small, since the influence of the (perhaps biased) prior decreases with n. Moreover, the basis of posterior (economic) descisions may be in doubt if n is small.

Case study C next demonstrates the fact that initially the almost only contributor to the randomness in the total cost C is the uncertainty due to T and that if one wants to reduce this uncertainty to the same size as the uncertainty in C that is due to the fact that $\sigma > 0$, than one needs to make very many measurements.

Case study D finally studies how some important parameters of the prior and posterior distributions of the total cost C vary with the ratio σ/μ, where σ is the standard deviation and μ is the mean of the cost of remediating a cell. Figure 7 demonstrates the fact that while almost negligible w r t the prior distribution, the component of the posterior variance of C that is due to the fact that $\sigma > 0$ (i e, $\sigma^2 E[T]$) needs to be taken into account even when the ratio σ/μ is not too large. Notice next that the prior and posterior distributions of T and C are very alike when σ/μ is small. This is demonstrated in Figures 5 and 6 and in Table 2. Table 3 finally is an example of the fact that the expected shortfall e_p may be considerably larger than its lower bound c_p.

There are two natural extensions of the results of this paper, one of which is how to treat remediation sites with more than one contaminant of concern and another is concerned with the case when the cells are partitioned into two or more strata depending on what is known about the type of activities that has been going on in the area. We plan to return to both extensions in forthcoming publications. Recall that the method of this paper pressumes that the status of each cell can be determined with certainty. We therefore plan to invoke detection errors in forthcoming work.

Acknowledgement

I am grateful to Lars Rosén for many valuable conversations regarding remediation of contaminated sites.

A The beta-binomial distribution

Let n be a strictly positive integer, and let $\alpha, \beta > 0$. The beta-binomial Distribution with parameters n, α, β (referred to as $\mathrm{BB}(n, \alpha, \beta)$) has probability mass function

$$g(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{n!}{\Gamma(\alpha+\beta+n)} \frac{\Gamma(\alpha+x)}{x!} \frac{\Gamma(\beta+n-x)}{(n-x)!}$$

for $x = 0, 1, \ldots, n$, where Γ is the Euler gamma function defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, dt$$

Notice, $\Gamma(1) = 1$. Moreover, $\Gamma(x) = (x-1)\Gamma(x-1)$ for $x > 1$ as can be seen by a straightforward integration by parts argument. In particular, $\Gamma(k) = (k-1)!$ for positive integers k.

The beta-binomial distribution is sometimes called the negative hypergeometric or compound binomial distribution. Our earliest reference to this distribution is Skellam [5], who in addition to calculating moments and the maximum likelihood estimates also used it in a study of the secondary association of chromosomes in Brassia. The beta-binomial distribution have also been used in many toxicological studies (an example of this is Williams [7]), in analysing consumer purchasing behaviour (see Chatfield and Goodhardt [1]) and in analysing point quadrant data (Kemp and Kemp [2].

Plots of beta-binomial distributions can be seen in Figures 1, 4 and 5.

The mean and variance of a $\mathrm{BB}(n, \alpha, \beta)$-variable are

$$\mu = n \frac{\alpha}{\alpha+\beta}$$

and

$$\sigma^2 = n \frac{\alpha}{\alpha+\beta} \frac{\beta}{\alpha+\beta} \frac{\alpha+\beta+n}{\alpha+\beta+1}$$

respectively.

To see that $g(x)$ is a probability mass function for a discrete random variable X taking on the values $0, 1, \ldots, n$, such that $E[X] = \mu$ and $\mathrm{V}[X] = \sigma^2$, let

$$g(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}\, (1-\theta)^{\beta-1}, \quad 0 < \theta < 1$$

(this is a beta distribution with parameters $\alpha, \beta > 0$) and let X conditionally on θ be binomial with parameters n and θ, so that

$$g(x|\theta) = \binom{n}{x}\, \theta^{x}\, (1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n$$

Then

$$\begin{aligned} P(X = x) &= \int_0^1 g(x|\theta) g(\theta)\, d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{x} \int_0^1 \theta^{\alpha+x-1}\, (1-\theta)^{\beta+n-x-1}\, d\theta \\ &= g(x) \end{aligned}$$

Moreover,

$$E[X] = E[E[X\,|\,\theta]] = E[n\theta] = \mu$$

and

$$\begin{aligned} \mathrm{V}[X] &= E[\mathrm{V}[X\,|\,\theta]] + \mathrm{V}[E[X\,|\,\theta]] \\ &= E[n\theta(1-\theta)] + \mathrm{V}[n\theta] \\ &= \sigma^2 \end{aligned}$$

by well known facts for the beta distribution.

B The gamma distribution

The density of any gamma distributed random variable is

$$g(x) = \frac{\lambda^p}{\Gamma(p)}\, x^{p-1}\, e^{-\lambda x}$$

for $x > 0$. Here, $p > 0$ and $\lambda > 0$ are parameters. Any random variable with this density will be referred to as a $\Gamma(p, \lambda)$-variable in this appendix. The gamma function Γ is defined in Appendix A.

The mean and variance of a $\Gamma(p, \lambda)$-variable are

$$\mu = \frac{p}{\lambda}$$

and

$$\sigma^2 = \frac{p}{\lambda^2}$$

respectively. Moreover, the sum of n independent $\Gamma(p_i, \lambda)$-variables is a $\Gamma(\sum_i p_i, \lambda)$-variable. Notice finally that $g(x)$ reduces to the exponential density if $p = 1$, and to the $\chi^2(k)$-density if $p = k/2$ and $\lambda = 1/2$.

References

[1] Chatfield, C. and Goodhardt, G. J.: The beta-binomial model for consumer purchasing behaviour. *Appl. Statist.* **19**, 240-250, 1970.

[2] Kemp, C. D. and Kemp, A. W.: The analysis of point quadrant data. *Aust. J. Botany* **4**, 167-174, 1956.

[3] Myers, Jeffrey C.: *Geostatistical Error Management. Quantifying uncertainty for environmental sampling and mapping.* Van Nostrand Reinhold, 1997.

[4] Norberg, Tommy: A Bayesian approach to the assessment of contaminant spread. Part II. Continuous case. Preprint no 2002:10. Department of Mathematical Statistics, Chalmers University of Technology, 2002.

[5] Skellam, J. G.: A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Society, Series B* **10**, 257-261, 1948.

[6] Wackernagel, Hans: *Multivariate Geostatistics.* Springer, 1995.

[7] Williams, D. A.: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**, 949-952, 1975.