# ON THE WORTH OF GROUNDWATER CONTAMINANT SAMPLING

## 1. Non-spatial case

**Lars Rosén**
**Department of Geology**
**Chalmers University of Technology**
**SE-412 96 Gothenburg, SWEDEN**
**E-mail:** rosen@geo.chalmers.se

**Tommy Norberg**
**Department of Mathematical Statistics**
**Chalmers University of Technology**
**SE-412 96 Gothenburg, SWEDEN**
**E-mail:** tommy@math.chalmers.se

## *Abstract*

This is the first in a series of three papers studying the economical worth of groundwater contaminant sampling in applications of increased complexity. This paper presents the theoretical basis for Bayesian data-worth analysis and discusses the non-spatial case with one hydraulically isolated cell of uniform and isotropic conditions. The two subsequent papers will discuss data worth in two one-dimensional models and in a two-dimensional spatial setting. The main contribution of this paper is that it provides a new theory for estimating the worth of simultaneous sampling in more than one location, which is an especially important aspect in cases of large sampling errors. Data worth depends directly on the prior probability of contamination and we compare a case in which this probability is completely known with a case in which it is known with given uncertainty. In the latter case, we show that the data worth is sensitive with respect to the selected design criteria.

## *Introduction*

Contaminated soil and groundwater is a problem of growing concern in our society and is today a major issue in land use planning and management, real estate assessment, and property selling. Investigation and remediation of contaminated areas are often associated with high costs. As an example, the Swedish Environmental Protection Agency (EPA) provides more than the equivalent of $50 million annually of governmental resources to the Swedish county authorities for investigation and remediation of contaminated areas where no responsible part can be found. Still, the bulk of necessary investments for investigation and remediation come from responsible landowners and operators. The Swedish EPA (1999) estimates that there are 22 000 contaminated sites in Sweden, of which approximately 4000 are in need of remediation. Similar, or even more severe, situations are present throughout Europe and North America.

The Polluter Pays Principle, which is the regulatory philosophy in the European Union, puts large financial pressure on responsible companies and authorities. The regulatory framework is therefore a strong incentive for both the public and private companies for applying cost-efficient investigation and remediation strategies. In addition, the environmental legislation in many countries, e.g. in Sweden, state that the environmental value of remediation must be in proportion to the remediation costs in order to provide proper prioritization of resources and to achieve sustainable use of land and water.

Due to a combination of complex geological, hydrogeological, and geochemical conditions at contaminated sites and high investigation costs, it is usually not possible to obtain complete information or characterize a site with a high degree of certainty. Investigations of contaminated areas are therefore typically associated with large uncertainties regarding e.g. type and extent of contamination and possible future contaminant spreading. These uncertainties transform into substantial economical risks in the remediation process. To handle these risks properly and to achieve a cost-efficient management of contaminated sites, it is therefore of primary importance to design investigation strategies considering both the sampling costs *and* the economical risks associated with making incorrect decisions regarding remediation.

The number and type of new data in contamination studies are often estimated with respect to the potential for reduction of uncertainty. However, in order to be cost-effective the worth of new data should be evaluated with respect to the reduction of the risk-cost of making incorrect decisions, rather than uncertainty. In a cost-benefit perspective, the worth of sampling depends on the sampling cost and the risk-cost reduction it provides with respect to failure of successfully solving the specific problem at hand. The risk-cost must then incorporate all expected costs for the decision-maker related to improper evaluation of the situation, e.g. change of remediation strategy, prolonged clean-up times, restrictions on land-use, and loss of environmental values. A tool for achieving cost-effective investigation is Bayesian data-worth analysis, evolving from an economical decision analysis framework. Bayesian data-worth analysis in hydrogeology has been described by e.g. Freeze et al (1990; 1992), James & Freeze (1993) and James & Gorelick (1994). The basic philosophy behind this approach in remediation projects is to integrate the cost of sampling, the reduction of the risk-cost of failure, and the cost and efficiency of the remedial design in order to find the best sampling strategy among a given set of alternative strategies. The data-worth analysis is made iteratively according to Figure 1 until no further sampling is considered justifiable.

Our work will be presented in a series of three papers of which this is the first. The main purpose is two-fold: (1) to analyze and describe how economical factors, sampling uncertainty, and prior assumptions regarding contamination affect cost-efficiency of the sampling strategy and (2) to present a spatial statistical model for Bayesian data-worth analysis. We focus on the non-spatial aspects of data-worth in the first paper, 1-dimensional aspects in the second paper, and 2-dimensional aspects in the third paper. The fist two papers in this series are directed at analyzing economical and statistical implications of data-worth analysis in a single cell[1] and along a line of several cells. We present conclusions and give recommendations regarding a number of strategic decisions to be taken when applying Bayesian data-worth analysis. In the third paper we present a 2D spatial statistical data-worth model, based on Monte Carlo Markov Chain (MCMC) analysis. To fully understand and appreciate the capabilities and limitations of Bayesian data-worth analysis, implications on lower dimensionalities need to be studied and evaluated before full 2D and/or 3D applications can be made correctly. We believe that the work presented in these three papers provide important knowledge for an increased cost-efficiency in projects on investigation and remediation of contaminated soil and groundwater.

## *General description of Bayesian data-worth analysis*

We state that the worth of sampling data can be best evaluated by the use of Bayesian decision analysis, which has been applied for contaminated soil and groundwater by several authors, e.g. Massmann & Freeze (1987), Marin et al (1989), Freeze et al (1990, 1992), Massmann et al (1991) James & Freeze (1993), James & Gorelick (1994), James et al (1996), Rosén & Wladis (1998), Wladis et al (1999), Norrman (2001), Back (2001) and Rosén (2001). A special application of Bayesian decision analysis is referred to as Bayesian data-worth analysis, which is described below, primarily based on the works by Freeze et al (1992) and James & Freeze (1993). We describe the methodology, assuming a simple case involving a decision of whether to remediate or not, that this decision depends on the contaminant levels with respect to a defined action level, and that failure occurs if no action is taken when contamination occurs.

---

[1] A cell is a homogeneous part of the contaminated site

Let $C$ denote the event of contamination above a specific action level in a hypothetical cell, which is hydraulically isolated, and let $C^c$ denote the complementary event that contamination is not present. The probability Pr($C$) is then the prior probability of contamination for this cell. Typically, Pr($C$) is not known with certainty. In early stages, an estimate may therefore be based on professional judgment using existing site information and experiences from other sites. Given that remediation has to be made if contamination occurs, the prior risk-cost is defined as:

$$R_0 = k_F \Pr(C) \tag{1}$$

where $k_F$ denotes the monetary consequences of failure. Thus, $R_0$ is the expected cost of no action. Let $k_R$ denote the remediation cost. Another alternative is to take a sample to see if any contaminant is present and to act accordingly. Let $D$ and $D^c$ denote the events that the contaminant is detected and not detected, respectively, given that the cell is sampled. Bayes' theorem gives us the pre-posterior probability of failure to remediate, given that no contaminant is detected:

$$\Pr(C \mid D^C) = \frac{\Pr(C)\Pr(D^C \mid C)}{\Pr(C)\Pr(D^C \mid C) + \Pr(C^C)\Pr(D^C \mid C^C)} \tag{2}$$

and the pre-posterior risk-cost is:

$$R_1 = k_F \Pr(C \mid D^c) \tag{3}$$

Let $k_M$ denote the sampling cost. The expected risk-cost of taking a sample is:

$$k_R \Pr(D) + \min[k_R, R_1] \Pr(D^c) \quad (4)$$

since the cell is remediated if contamination is detected, and if contamination is not detected the cell is remediated only if the remediation cost is less than the pre-posterior risk-cost. Following Freeze et al (1992) we define the data worth as:

$$W = \min[k_R, R_0] - \{k_R \Pr(D) + \min[k_R, R_1] \Pr(D^c)\} \quad (5)$$

Since the quantities involved in the calculation of the worth $W$ typically are known only approximately, one may require that $W$ be significantly larger than the sampling cost in order to justify sampling. This analysis applies to all cells of the region of interest and the most elementary and naive way to proceed would be to say that the cells that should be analyzed further are those with $W$ sufficiently larger than the sampling cost. However, such an analysis does not take into account the obvious spatial dependence that exists in any remediation site. The effects of spatial dependence are further described in the second and third papers in this series. Below, we present an analysis of economical and statistical implications on the data-worth without spatial considerations.

## *The worth of taking additional samples*

Now consider a cell to be investigated with respect to groundwater contamination. Prior to analyzing the value of additional samples, we have two alternative options: (1) to remediate the cell to the cost of $k_R$; (2) to "do nothing" with a remaining risk-cost of $R_0=k_F\Pr(C)$. The decision rule is to do nothing if $\Pr(C) \leq k_R/k_F$ and remediate if $\Pr(C) > k_R/k_F$. A third alternative is to take new samples and act according to the outcome of these (see also Figure 1). An objective function, $\Phi(n)$, is defined for each sampling alternative in which the number of new samples is $n$=1,2…. The objective function of the optimal action is *a priori*, before considering any new samples:

$$\Phi(0) = \min(k_R, k_F \Pr(C)) \qquad (6)$$

We will now consider the possibilities of taking *one or several* imperfect samples with the conditional detection probabilities $\Pr(D|C)$ (which may be much less than 1) and $\Pr(D|C^c)$ (which may be considerably larger than 0). Let $k_M$ denote the cost of one new sample and $k_M(n)$ denote the cost of simultaneously taking $n$ new samples. A sample cost function that we have used in our work is:

$$k_M(n) = k_M n r^{n-1}, \quad n = 1, 2, ..... \qquad (7)$$

where $r \leq 1$ is a discount factor. The objective function of taking $n$ new samples is:

$$\Phi(n) = k_M(n) + \sum_{k=0}^{n} \min(k_R, k_F \Pr(C \mid D_n = k)) \Pr(D_n = k)$$
$$= k_M(n) + E[\min(k_R, k_F \Pr(C \mid D_n))] \qquad (8)$$

where $D_n$ ($n$=1, 2, …) is the total number of contaminant detections in $n$ independent samples. The conditional probabilities $\Pr(C|D_n = k)$ can be shown, by means of Bayes' rule, to be:

$$\Pr(C \mid D_n = k) = \frac{\Pr(C) \Pr(D_n = k \mid C)}{\Pr(D_n = k)} \qquad (9a)$$

where

$$\Pr(D_n = k) = \Pr(C) \Pr(D_n = k \mid C) + \Pr(C^c) \Pr(D_n = k \mid C^c) \qquad (9b)$$

$\Pr(D_n = k|C)$ is the binomial probability of getting $k$ successes in $n$ independent trials with success probability $\Pr(D|C)$ (an analogous statement holds for $\Pr(D_n = k|C^c)$).

The optimal number of samples is:

$$n_{opt} = \arg\min_{n \geq 0} \Phi(n) \qquad (10)$$

Note that $\Phi_{opt} = \Phi(n_{opt})$ is the expected cost of the optimal action, which is "do nothing" if $n_{opt} = 0$ and $\Pr(C) < k_R/k_F$, "take $n_{opt}$ samples and act accordingly" if $n_{opt} \geq 1$, and "remediate" if $n_{opt} = 0$ and $\Pr(C) \geq k_R/k_F$.

The “Expected Net Value” (ENV) of taking $n$ samples is:

$$\mathrm{ENV}(n) = \Phi(0) - \Phi(n), \quad n = 1, 2, .... \tag{11}$$

This is related to the “Expected Value of Sample Information” (EVSI) discussed in Bruce & Freeze (1993) and Bruce & Gorelick (1994). It is necessary to include the sample cost in ENV, since it depends on the number $n$ of samples, while it is not included in the EVSI, which therefore is positive for all values of Pr($C$). Another difference of our approach to evaluating the worth of additional samples is that Bruce & Freeze (1993) and Bruce & Gorelick (1994) discuss the value of taking one additional sample, whereas we are open to the possibility that it may be cost-effective to simultaneously take several additional samples.

Obviously, it is cost-effective to take samples only if:

$$\mathrm{ENV} = \max_{n \geq 1} \mathrm{ENV}(n) > 0$$

## *Examples – ignoring uncertainty in Pr(C) estimates*

Having described the general theory of Bayesian data-worth analysis, we now present two hypothetical cases on using the described methodology for evaluation of the data-worth of additional samples. In this first presentation of cases, the purpose is to illustrate the application of the data-worth assessment approach described in this paper. In the subsequent section, we introduce uncertainty assessment of input variables and investigate the impact of this on decisions regarding the design of the sampling program.

### *Case 1*

Consider a hydraulically isolated cell of uniform and isotropic conditions. The cost functions and the conditional detection probabilities, which reflect the errors of the chosen sampling technique, are shown in Table 1. As seen from the table, we assume relatively large sampling errors. We further assume that the sampling cost is small compared to the remediation cost, and that the failure cost is significantly larger than the remediation cost. Note that we used no discount factor in the sample cost function, $k_M$, in this example. Using the procedure for calculating the optimal number of samples, $n_{opt}$, described above, the optimum number of samples for different $\Pr(C)$ was calculated and is shown in Figure 2. Figure 3 shows the prior expected cost ($\Phi(0)$) and the optimal cost ($\Phi_{opt}$) for different $\Pr(C)$. Figure 4 shows the ENV for different values of $\Pr(C)$. Considering the results shown in these figures, we take $\Pr(C) = k_R/k_F = 0.4$ prior to any samples, since this represents the “maximum decision uncertainty”. As can be seen from Figure 2, the optimal number of samples for $\Pr(C) = 0.4$ is 3. The four possible outcomes of taking three samples, $D_3$, are shown in Table 2. Given the information about the cell and the data-worth analysis, we thus conclude that three additional samples represent the most cost-effective sample alternative. The example also illustrates that even if contamination was to be detected in one sample ($D_3 = 1$), the decision would still be to “do nothing”, due to the large sampling error associated with the sample technique that we have chosen. In real world applications the sampling error is often ignored.

### *Case 2*

The conditions of the second hypothetical case, also assuming a cell of uniform and isotropic conditions, are shown in Table 3. Also here, we let Pr($C$) = $k_R/k_F$ = 0.4 prior to any samples. In this case we estimate the sampling error to be smaller than in the first example and we use a discount factor for the $k_M$ cost function. Figure 5 shows the prior expected cost ($\Phi(0)$) and the optimal cost ($\Phi_{opt}$) for different values of Pr($C$). Figure 6 shows the ENV for different Pr($C$). Calculating the optimum number of additional samples yields $n_{opt}$ = 1. The two possible outcomes of taking one sample, $D_1$, are shown in Table 4. If the single new sample is made the decision would be "do nothing" if contamination was not detected and "remediate" if contamination was detected. This example illustrates the situation of an expensive, but also reasonably accurate sampling technique, which makes remediation cost-effective when contamination is detected in a single sample.

## *Examples – including uncertainty in P(C) estimates*

In the two hypothetical cases just studied, Pr($C$) is taken to equal $k_R/k_F$ since for this value, the expected net value (ENV) is maximal. However, it is typical that one has some idea about the true Pr($C$), but that the estimate is associated with some uncertainty. An evaluator may therefore want to express the estimation as Pr($C$) ≈ 0.65 and 0.5 ≤ Pr($C$) ≤ 0.8. This information may be turned into a prior distribution of Pr($C$). The most obvious prior is the triangular distribution with mode 0.65 and support (0.5, 0.8). This prior distribution gives mass 1 to the interval (0.5, 0.8). Thus, it does not account for the fact that the expert specification 0.5 ≤ Pr($C$) ≤ 0.8 might be wrong.

A more realistic prior is therefore the beta distribution, having density function:

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{12}$$

for 0 < $\theta$ < 1 (in order to avoid a cumbersome notation, we now write $\theta$ instead of Pr($C$)). The parameters $\alpha$ and $\beta$ concretize the expert opinion and $\Gamma$ denotes the Euler gamma function. A beta prior allows for wrong expert estimates. For example, note that Pr{$\theta$< 0.5} = Pr{$\theta$> 0.8} = 0.05 if $\alpha$ = 16.4, $\beta$ =8.6. This beta distribution is plotted in Figure 7.

This is one strategy to concretize and quantify the reliability of the experts. Another strategy would be to specify the uncertainty using an estimate of the most likely value, $\theta$, and a number $n$ of equivalent samples representing the reliability of this estimate. The appropriate beta distribution then is the one with $\alpha = n\theta + 1, \beta = n(1-\theta) + 1$. This follows from the well-known fact that the beta distribution is conjugate to the binomial (Bickel & Doksum, 1977).

Providing $\theta$ (=Pr($C$)) with a beta prior forces us to redefine $n_{opt}$. Firstly, we specify a design criterion such as the expected value or a suitable percentile in the distribution of $\Phi(n)$ for various number of samples $n$. Secondly, if the expected value is taken, then

$$n_{opt} = \arg\min_{n \geq 0} E[\Phi(n)]$$

Similarly, if the 100$p^{\text{th}}$ percentile is chosen, then

$$n_{opt} = \arg\min_{n \geq 0} G_n^{-1}(p)$$

where $G_n^{-1}(p)$ denotes the percentile function of the distribution function $G_n(x) = \Pr\{\Phi(n) \leq x\}$. We now return to the two hypothetical cases.

*Case 1*

The conditions for this case are identical to the first case in the previous section (see Table 1), except that we now introduce uncertainty in the $\theta$ (=Pr($C$)) estimate. Let us assume that the experts agree on the following: $\theta \approx 0.45$ and the uncertainty (or reliability) of this statement is equivalent to the one achieved after taking 8 independent samples. An appropriate parameterization of the prior beta distribution is then $\alpha$ = 4.6 and $\beta$ = 5.4. This beta distribution is plotted in Figure 8. From Figure 2 it is seen that the optimal number of samples is $n_{\text{opt}} = 3$ if $\theta = 0.45$. However, if the uncertainty is taken into account, the optimal number of samples is 2 if the design criterion is the expected value, i.e. $n_{opt} = \arg\min_{n \geq 0} E[\Phi(n)]$. In Figure 9 the density of $\Phi(2)$ is plotted. Notice that it has a discrete component: $\Pr\{\Phi(2) < 11\} = 0.982$ and $\Pr\{\Phi(2) = 11\} = 0.018$. Note also that if the uncertainty in $\theta$ instead corresponds to a larger number of independent samples, e.g. 50, then $n_{opt}$ =3 if the design criterion is the expected value.

*Case 2*

In the second hypothetical case, $n_{opt}$ = 1 for 0.157 ≤ Pr(C) ≤ 0.627, 2 for 0.628 ≤ Pr(C) ≤ 0.657 and 0 otherwise. This means that if it is estimated that $\theta \approx 0.4$, then $n_{opt}$ = 1 also if any reasonable uncertainty is taken into account. The situation changes, however, and becomes similar to what we observed in *Case 1*, if $\theta \approx 0.62$.

Assume that the expert opinion is $0.50 \leq \theta \leq 0.75$. Assume also that the experts are regarded as highly qualified and reliable by the responsible decision-maker. A quantification of the reliability of the expert opinions may therefore be expressed as: $\Pr\{\theta < 0.5\} = \Pr\{\theta > 0.75\} = 0.05$. This holds for $\alpha = 24.82$ and $\beta = 14.65$. In this case $\min_n E[\Phi(n)]$ is obtained for both $n = 1$ and $n = 2$ if the design criterion is expectation. This holds true also if the median is chosen. However, if the design criterion is the 75$^{th}$ percentile, $n_{opt} = 2$.

The outcome would differ if the decision-maker would be less willing to depend on the expert opinion. If the reliability of the expert opinion was instead expressed as: $\Pr\{\theta < 0.5\} = \Pr\{\theta > 0.75\} = 0.3$, this would correspond to $\alpha = 2.70$ and $\beta = 1.71$. In this case, the optimal number of samples would be 1 if the design criterion is expectation, 2 if the median is the design criterion and 1 if the criterion is the 75$^{th}$ percentile.

## *Conclusions*

The following conclusions were made with respect to the non-spatial case of Bayesian data-worth analysis:

1. Given the expected costs for sampling, remediation and failure, the probability of contamination, and the sample error, the optimal number of samples to be taken simultaneously in the next investigation phase can be calculated.

2. Bayesian data-worth analysis is directed at minimizing the sum of investment and risk cost. The sampling cost is compared to the reduction of the economical risk of failing to meet set up management goals of the site, implying that the risk reduction should exceed the sampling cost in order to make sampling cost-effective. This is in contrast to more commonly applied variance reduction approaches, which are directed at reaching a level of acceptable uncertainty, rather than looking at the cost-efficiency of the sampling program.
3. Bayesian data-worth analysis provides a possibility for formal incorporation of the sampling error in the valuation of the cost-efficiency of the sampling program. As shown in the 1st hypothetical case, that even if contamination is detected remediation may not be economically motivated if the sampling error is large.
4. Different results can be obtained, depending on whether uncertainty in the prior estimate of the probability of contamination is included or not.
5. Different results can be obtained, depending on the strategy for handling the uncertainty of the prior estimation of the probability of contamination. If the uncertainty of the prior estimate is taken into account, the decision is sensitive to the selected design criteria, e.g. expected value, median or another percentile. We present an approach to model this type of uncertainty using a beta-distribution. Also remediation costs and failure costs may be associated with substantial uncertainties, which may have large impact on the final decision. We do not present specific approaches for handling these uncertainties, but recommend that they are included in the same manner as the contamination uncertainty.

Finally, the Bayesian data-worth analysis approach presented in this paper incorporates the possibility of taking several new samples simultaneously, and therefore differs to some extent to the previous works by e.g. James & Freeze (1993) and James & Gorelick (1994). We also illustrate the importance for handling uncertainties in the estimates of the key variables of the data-worth analysis. We do this without the complexities of spatial statistical models. This will be further explored in the 2nd and 3rd papers in this series.

## *Acknowledgments*

## *References*

Back, P-E, 2001. Sampling strategies and data-worth analysis for contaminated land – a literature review. Department of Geology, Chalmers University of Technology. Publ B 486 and SGI Varia 500.

Bickel, P.J. & K.A. Doksum, 1977. Mathematical Statistics. Basic Ideas and Selected Topics. Holden-Day, Inc.

Freeze, R. A., Massmann, J., Smith, J. L., Sperling, T. & B. R. James, 1990. Hydrogeological decision analysis: 1. A Framework. *Ground Water*. Vol. 28, no. 5, pp 738-766.

Freeze, R. A., Massmann, J., Smith, J. L., Sperling, T. & B. R. James, 1992. Hydrogeological decision analysis: 4. The concept of data worth and its use in the development of site investigation strategies. *Ground Water*. Vol. 28, no. 5, pp 738-766.

James, B. R. & S. M. Gorelick, 1994. When enough is enough: The worth of monitoring data in aquifer remediation design. *Water Resources Research*. Vol. 30, no. 12, pp. 3499-3513.

James, B. R., Huff, D. D., Trabalka, J. R., Ketelle, R. H. & C. T. Rightmire, 1996. Allocation of environmental funds using economic risk-cost-benefit analysis: A case study. *Ground Water Remedaition and Monitoring*. Fall 1996, pp 95-105.

James, B.R. & Freeze, R.A., 1993: The worth of data in predicting aquitard continuity in hydrogeological design. Water Resources Research, Vol 29, No. 7, pp 1721-1730.

Marin, C. M., Medina, M. A. & J. B. Butcher, 1989. Monte Carlo analysis and Bayesian decision theory for assessing the effects of waste sites on groundwater, I: Theory. *Journal of Contaminant Hydrology*. No. 5, pp 1-13.

Massmann, J. & R. A. Freeze, 1987. Groundwater contamination from waste sites: The interaction between risk-based engineering design and regulatory policy, 1: Methodology. *Water Resources Research*. Vol. 23, no. 2, pp 351-367.

Massmann, J., Freeze, R. A., Smith, L., Sperling, T. & B. James, 1991. Hydrogeological Decision Analysis: 2.Applications to Ground-Water Contamination. *Ground Water*. Vol. 29, no. 4, pp. 536-548.

Norrman, J, 2001. Decision Analysis under Risk and Uncertainty at Contaminated Sites A literature review. Department of Geology, Chalmers University of Technology. Publ B 486 and SGI Varia 500.

Rosén, L. and Wladis, D., 1998: Hydrogeological Decision Analysis for Monetary Evaluation of Nitrate Reduction Measures in Europe. *Proceedings of the International Association of Hydrogeologists XXVII Congress*.

Wladis, D., Rosén, L. and Kros, J., 1999: Risk-Based Decision Analysis of Emission Alternatives to Reduce Groundwater Acidification on the European Scale. *Ground Water*, Vol. 37, No. 6, pp 818-826.

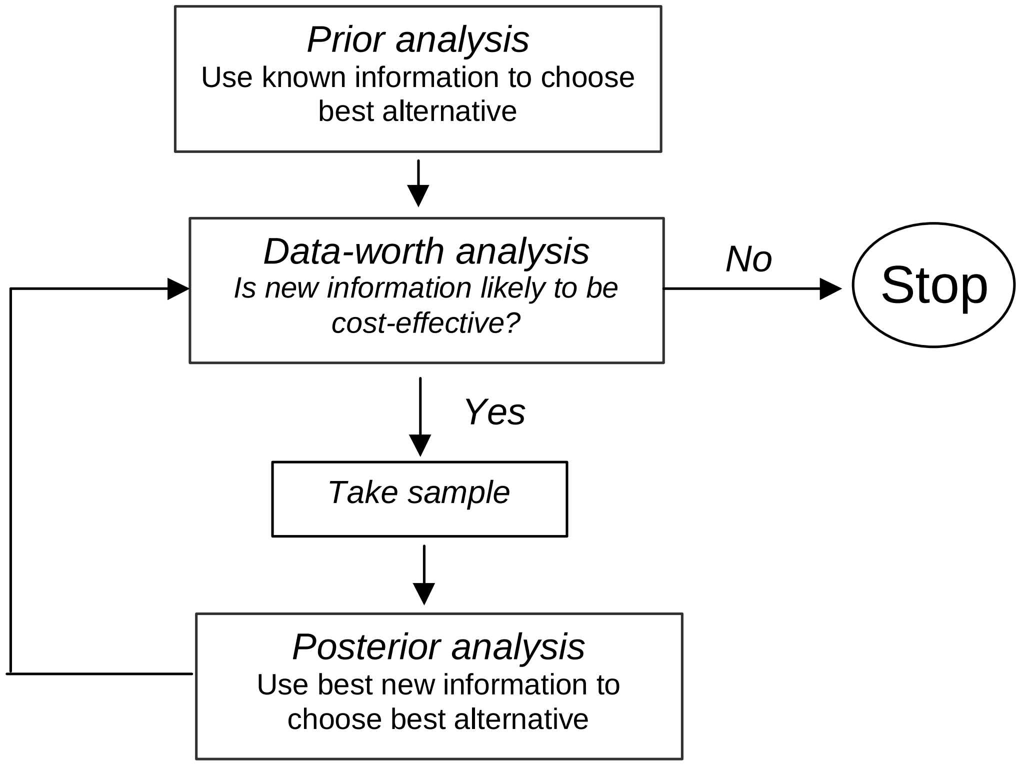


**Figure 1.** Schematic description of Bayesian data-worth analysis (after James et al, 1996).

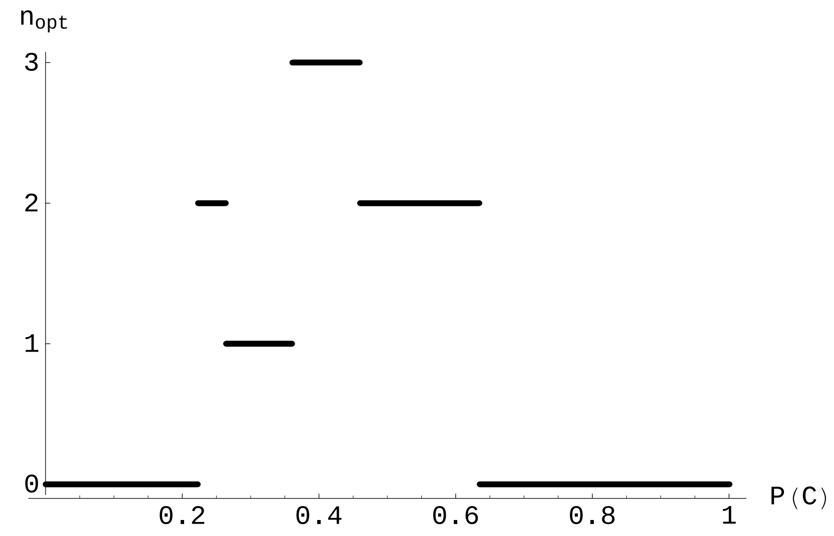


***Figure 2***. Plot of $n_{opt}$ vs Pr($C$) for the 1st hypothetical case.

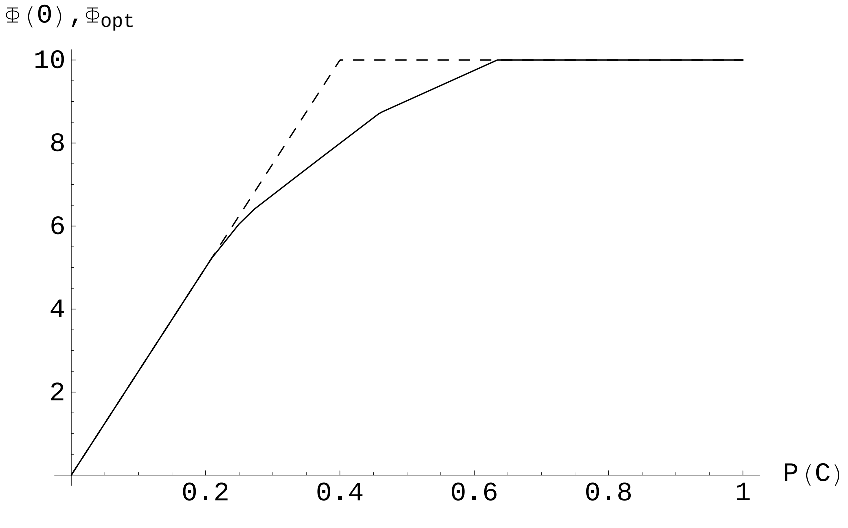


***Figure 3.*** Plot of Φ(0) (dashed) and $\Phi_{opt}$ vs Pr($C$) for the 1st hypothetical case.

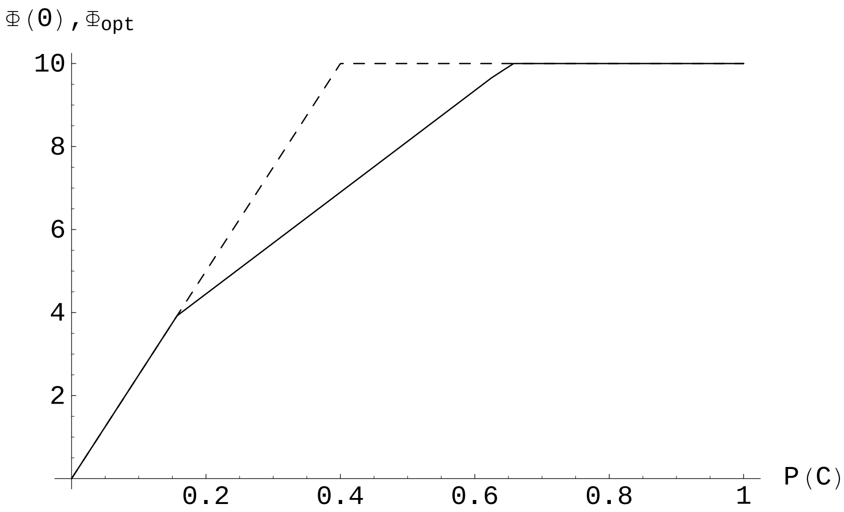


***Figure 4.*** Plot of Φ(0) (dashed) and $\Phi_{opt}$ vs Pr(*C*) for the 2nd hypothetical case.

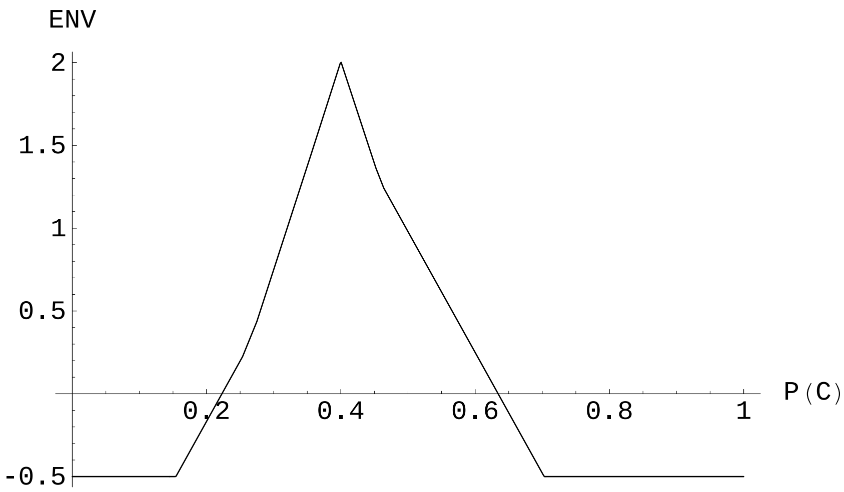


***Figure 5.*** Plot of ENV vs Pr(*C*) for the 1st hypothetical case.

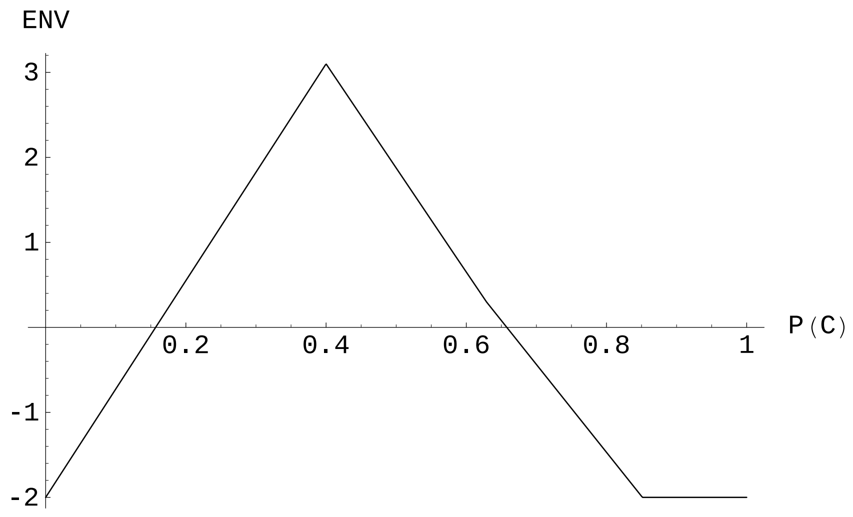


***Figure 6.*** Plot of ENV vs Pr($C$) for the $2^{nd}$ hypothetical case.

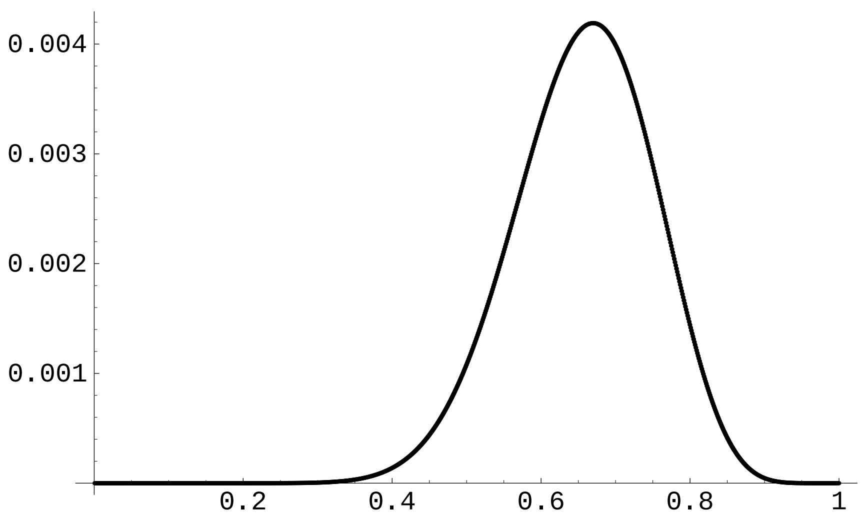


***Figure 7***. The beta(16.4, 8.6) density function.

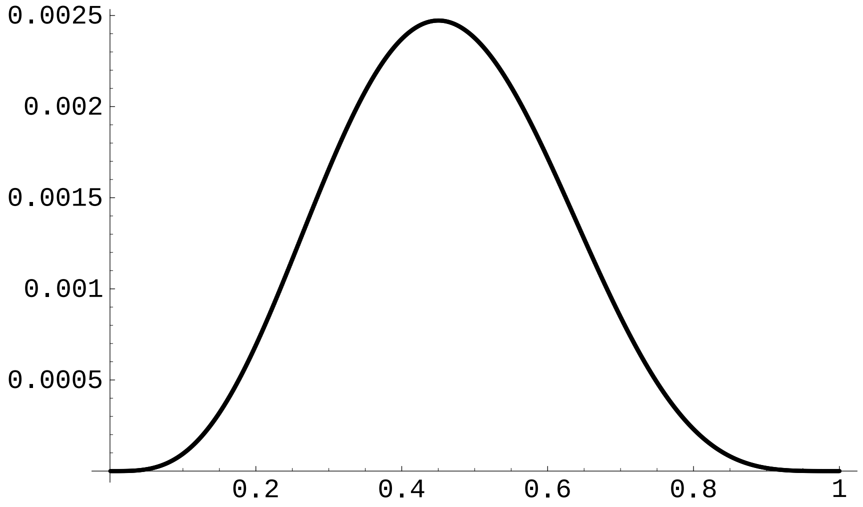


***Figure 8.*** The beta(4.6, 5.4) density function.

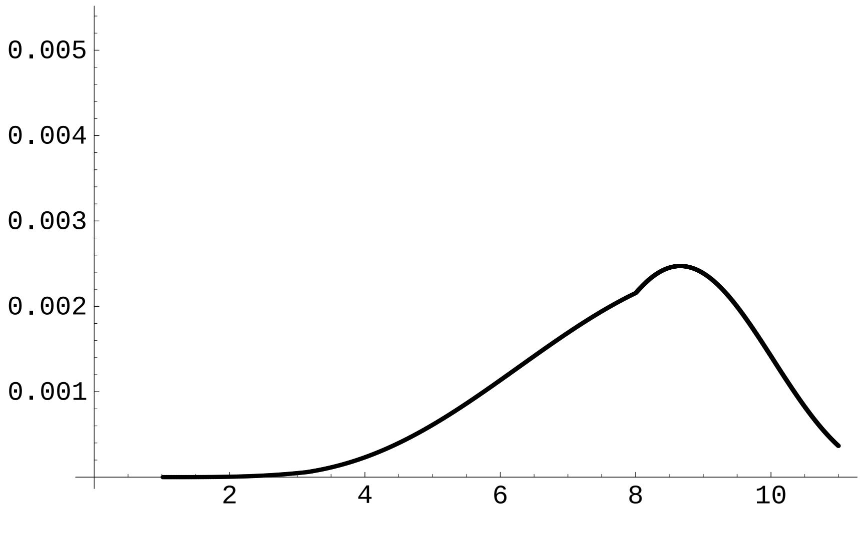


***Figure 9.*** The distribution of $\Phi(2)$ for the 1st hypothetical case for a beta(4.6, 5.4) prior on $\theta =$ Pr($C$).

***Table 1.*** *Costs and probabilities for 1st hypothetical case.*

| $k_M(n)$ | $k_R$ | $k_F$ | $\Pr(D\|C)$ | $\Pr(D\|C^c)$ |
|---|---|---|---|---|
| $0.5n$ | 10 | 25 | 0.67 | 0.25 |

***Table 2.*** *The four possible outcomes of taking three additional samples, $D_3$.*

| $k$ | $\Pr(D_3 = k)$ | $\Pr(C\|D_3 = k)$ | $k_F\Pr(C\|D_3 = k)$ | Decision |
|---|---|---|---|---|
| 0 | 0.27 | 0.06 | 1.4 | “do nothing” |
| 1 | 0.34 | 0.26 | 6.5 | “do nothing” |
| 2 | 0.26 | 0.68 | 17.0 | “remediate” |
| 3 | 0.13 | 0.93 | 23.2 | “remediate” |

***Table 3.*** *Costs and probabilities for 2nd hypothetical case.*

| $k_M(n)$ | $k_R$ | $k_F$ | $\Pr(D\|C)$ | $\Pr(D\|C^c)$ |
|---|---|---|---|---|
| $2n0.8^{n-1}$ | 10 | 25 | 0.85 | 0 |

***Table 4.*** *The two possible outcomes of taking one additional sample, $D_1$.*

| $k$ | $\Pr(D_1 = k)$ | $\Pr(C\|D_1 = k)$ | $k_F\Pr(C\|D_1 = k)$ | Decision |
|---|---|---|---|---|
| 0 | 0.66 | 0.09 | 2.3 | “do nothing” |
| 1 | 0.34 | 1.00 | 25.0 | “remediate” |