Appendix A3.6: Estimation of Distribution Parameters from Data with Observations Below Detection Limit: An example from South Nahanni River area, District of Mackenzie, Canada

Chang-Jo F. Chung[1] and Wendy A. Spirito[2]

Chung, C.F. and Spirito, W.A., Estimation of distribution parameters from data with observations below detection limit: an example from South Nahanni River area, District of Mackenzie, Canada; in Statistical Applications in the Earth Sciences; ed. F.P. Agterberg and G.F. Bonham-Carter; Geological Survey of Canada, Paper 89-9, p.233-242, 1989.

Abstract

Gold and tungsten stream sediment geochemical data obtained for a resource assessment of South Nahanni River area are used to illustrate maximum likelihood estimators (MLE) when applied to data sets with undetected values. Ideally, the data should be complete and the distribution function should be known before statistics are computed. However, as data are commonly not complete, the maximum likelihood estimation method allows statistical analysis of censored data with the inherent assumption that their distribution is normal or lognormal. Although data with a high proportion of undetected values may not be reliable, in a mineral resource assessment these may be the only data available. Comparison of curves generated from data with ad-hoc substitutions versus curves generated by maximum likelihood estimation, shows that the MLE method provides more realistic generalizations of the sample mean and variance.

1 : Geological Survey of Canada, Ottawa, Ontario K1A 0E8
2 : University of Western Ontario, London, Ontario N6A 5B7

Introduction

Stream sediments from proposed extensions to Nahanni National Park Reserve, N.W.T. were analyzed by neutron activation for gold, tungsten and other elements as part of a non-renewable resource inventory study (Jefferson et al., 1989). An initial statistical analysis of the data was presented in Spirito et al. (1988). This paper selects the gold and tungsten data to illustrate the different results obtained by using conventional ad-hoc substitution methods versus the maximum likelihood estimator method outlined by Chung (1989a, 1989b).

Commonly, geochemical data are incomplete because of the difficulty in determining rare elements present in extremely small amounts (ie below the detection limit). The detection limit is dependent upon the characteristics of the specific element, the analytical technique and the sample itself (quantities of other elements present). Highly variable detection limits for gold and tungsten in the Nahanni data can be attributed to interference from varying abundances of radioactive elements present in the stream sediment samples.

Neither standard computer packages for statistical analyses nor statistical techniques in textbooks can properly handle censored data. To analyze such incomplete data, ad-hoc "substitution methods" are commonly used: observations below the detection limit are replaced by a certain percentage of that limit (cf. Spirito et al., 1988). For example, if a sample contains Au values less than 2 ppb, then the Au value of the sample is set to 1.2 (= 2 x 0.6) ppb or 1 (= 2 x 0.5) ppb. After the substitution, the data are assumed to be complete and a statistical analysis is performed.

If a small portion of samples is below the detection limit or the detection limits are relatively low, then the results may be reasonable and most geological interpretations or implications are probably valid. However, if a large part of the data are below the detection limit, then, for example, statistics to calculate

background values, produce distribution curves and detect geochemical anomalies, may be meaningless. In addition, where the detection limits are high, methods that automatically substitute some arbitrary value (eg. 0.5, 0.6) for the elevated "less than" values may create artificial geochemical anomalies. We propose the maximum likelihood estimation method (Chung, 1989a,b) to estimate parameters in the distribution functions of Au and W in the South Nahanni River area as a means of enhancing assessment of the mineral potential of the area.

Geological Background

Nahanni National Park Reserve is located east of the Yukon border and north of the British Columbia border in the Northwest Territories. It covers an area of approximately 4800 km^2 which transects the southern Mackenzie Mountains fold and thrust belt. Surficial deposits from alpine and continental glaciations are found throughout the study area. The data for this project were collected from proposed park extension areas at the western and eastern ends of the park (Spirito et al., 1988).

The western extension area, known as the Ragged Ranges is located near Tungsten, NWT. It is characterized by Paleozoic shelf margin carbonates and basinal shales intruded by Cretaceous plutons. Three main types of mineral deposits are known in this study area (Scoates et al., 1986): 1. tin-tungsten associated with granitic plutons similar to those mined at Tungsten, NWT; 2. shale-hosted lead-zinc similar to that found at MacMillan Pass and Howards Pass; 3. precious metal-bearing veins.

The bedrock geology is grouped into eight main units (after Spirito et al., 1988) which are simplified from numerous sources cited by Scoates et al. (1986):

1,2) Late Proterozoic: glaciomarine conglomerate, iron formation, argillite, shale, quartz arenite and carbonate of the Windermere Supergroup.

3,4) Early Paleozoic: platformal and carbonate strata (Rock Type 4) on the NE side and shales to shaly carbonates (Rock Type 3) on the SW side of the facies boundary. Analyses of heavy mineral concentrates taken from stream gravel derived from these two rock types were selected for discussion in this paper.

5,6,7) Late Devonian and Younger: basinal shale and porcellanite; Carboniferous shallow marine carbonates and coal-bearing continental sandstones overlain by Permian to Triassic basinal cherts and mudstones.

8) Granitoid Rock Types: mainly Early and Mid-Cretaceous quartz monzonites.

9) Quaternary Deposits: in valleys, bedrock is deeply covered by talus, glacial till and alluvial deposits. This "rock type" contributes to the geochemical signature of the samples.

Sample Collection and Preparation

In 1985, an orientation survey tested the sampling method and identified potential problems. Known mineralized zones were sampled at a 1:50,000 scale at Lened (W-Mo-Cu) and Prairie Creek (Ag-Pb-Zn). The following summer a reconnaissance survey at 1:250,000 covered all large drainage basins in the study regions. During the summer of 1987, more detailed sampling investigated geochemical anomalies that were detected in the 1986 samples.

The sample sites were chosen on the basis of rock type, basin size and in rare cases, accessibility. The density of sampling was limited by funding. Samples representing all rock types and 244 drainage basins were taken. At each site a stream silt and gravel were collected. Data from silts are not used in this paper as all of the Au and W determinations are below the detection limit.

The gravels were sieved from -841μ to +63μ. In 1985, heavy liquids (SG >3.2) were used to separate the heavy minerals. This method was not efficient for the large number and size of samples collected in 1986 and 1987. These samples were sieved and the heavy minerals were separated using a concentrating table. The magnetic fraction was removed from the heavy mineral concentrate and the concentrate was analyzed by neutron activation. Anomalous values for W, Au and Zn were published in Spirito et al. (1988). The complete list of W and Au values from Rock Types 3 and 4 are found in Tables 1A and 1B.

Table 1A: W and Au values for rock type 3 in Ragged Ranges

Sample	W (ppm)	Au (ppb)
6041	35	< 19
6042	95	< 12
6043	231	< 14
6050	200	< 18
6052	< 13	98
6053	< 20	140
6057	496	110
6062	9	< 15
6067	< 6	< 13
6069	655	20
6070	< 8	< 20
6071	28	< 12
6073	44	< 25
6074	66	< 37
6075	86	26
6076	37	< 36
6078	< 7	< 11
6089	< 6	< 11
6090	6	37
6091	< 2	< 5
6092	< 4	< 5
6096	< 2	< 5
6099	< 2	< 5
6116	255	< 9
6117	32	< 21
6120	15	< 26
6123	601	1300
6124	< 8	< 5
6130	351	81
6131	52	< 19
6132	180	< 11
6134	< 5	< 5
6137	< 8	< 17
6142	< 19	< 5
6143	< 8	< 19
6144	12	21
6280	< 37	< 100
7001	86	< 23
7003	88	< 18
7006	282	< 25
7008	327	< 15
7016	< 2500*	100
7017	< 3500*	< 120
7021	130	< 21
7027	257	< 22
7034	39	< 26
7035	< 10	< 5
7036	< 3000*	< 180
7039	< 4900*	< 170
*calculations done with and without these values		

Table 1B: W and Au values for rock type 4 in Ragged Ranges

Sample	W (ppm)	Au (ppb)
6040	1130	< 18
6044	< 6	< 13
6046	2370	32
6049	< 5	< 5
6051	74	< 46
6054	< 8	69
6055	1130	869
6056	1630	160
5058	< 11	< 22
6059	< 6	< 10
6060	< 2	< 5
6061	< 5	< 5
6063	9	< 12
6064	4	< 5
6065	5	< 5
6109	< 8	< 5
6110	< 8	< 11
6111	< 2	< 5
6112	< 7	< 17
6113	8	< 11
6114	9	< 12
6115	< 13	< 23
6126	6	< 5
6138	10	25
6139	25	< 22
6140	< 5	28
6141	10	< 17
6159	< 69	< 110
6160	< 20	53
6162	< 23	< 49

6072	< 8	< 16
6077	214	< 35
6079	4	< 5
6081	< 2	< 5
6082	< 2	< 5
6083	< 2	< 5
6085	1920	44
6086	140	< 12
6093	4	9
6095	9	< 5
6097	< 2	< 5
6098	< 8	< 5
6100	< 2	< 5
6101	7	19
6102	< 6	< 5
6103	< 8	< 10
6104	< 8	< 11
6106	< 10	< 17
6108	9	< 5

6163	< 21	60
6164	< 11	41
6282	< 12	< 34
6284	29	< 27
7004	< 3600*	< 160
7005	< 3400*	448
7018	< 17	31
7019	< 26	44
7020	< 50	410
7025	13200*	79
7026	< 3600*	< 120
7029	< 7	< 5
7032	37	290
7033	< 41	180
7037	306	< 49
7038	110	< 11
7042	< 2900*	< 180
*calculations done with and without these values		

Multiple Censored Data

Consider n observations $X_1, X_2, \ldots, X_n$ from a population with the continuous distribution function $F(x : u_k, k=1,...m) = P\{X_i < x\}$ where u_k are the population parameters such as the mean (the location parameter) and the variance (the scale parameter). Suppose that the first h observations are censored, but that $X_i < \alpha$ for i=1, ... , h where α is a known constant. That is, instead of $X_1, \ldots, X_n$, the observations are $<\alpha, <\alpha, \ldots, <\alpha, X_{h+1}, X_{h+2}, \ldots, X_n$ where $<\alpha$ denotes that the value is less than α. This is called a single left censored data set. A geochemical data set with some observations below a single detection limit α is a typical example of single left censored data.

For a data set with multiple censoring, the observations are $<\alpha_1, \ldots, <\alpha_h, X_{h+1}, \ldots, X_{h+k}, >\beta_1, \ldots, >\beta_g$ instead of $X_1, \ldots, X_n$, where n = h+k+g and $>\beta_j$ indicates that the value of the sample is greater than a constant β_j. The first h samples are called multiple left censored data and the last g samples are referred as multiple right censored data. The tungsten and gold values in Table 1A and B are two examples of multiple left censored data.

Maximum likelihood estimation

$F(x : u_k, k=1,...,m)$ implies that the distribution function F is completely characterized by m parameters $u_1, ..., u_m$. The statistical problem consists of how to estimate these m parameters from the n observed samples $X_1, ..., X_n$. Let $f(x : u_k, k=1,...,m)$ be the corresponding density distribution function of F. Then the maximum likelihood estimators (MLE) of u_k, k=1, ... , m from n multiple censored observations, $<\alpha_1, \ldots, <\alpha_h, X_{h+1}, \ldots, X_{h+k}, >\beta_1, \ldots, >\beta_g$, where n = h+k+g, are obtained by determining u_k, k=1, ... , m which maximize the log-likelihood function:

$$(1) \quad L(u_k,k=1,...,m) = \sum_{j=1}^{h} \log (F(\alpha_j:u_k,k=1,...,m)) + \sum^{k} \log (f(X_{h+i}:u_k,k=1,...,m))$$

$$i=1$$

$$+ \sum_{v=1}^{g} \log(1 - F(\beta_v : u_k, k=1,...,m))$$

The maximum likelihood (ML) estimators are dependent upon not only the observations but also the distribution F. Even for the most commonly used distribution functions, such as normal, log-normal, exponential or gamma, the analytical solutions of the ML estimators from multiple-censored observations cannot be obtained unless an iterative numerical procedure is applied.

There are several iterative algorithms to obtain the ML estimators maximizing $L(u_k,k=1,...,m)$. Three commonly used techniques are the scoring method (Rao, 1975), the EM-algorithm (Dempster et al., 1977) and the conjugate gradients method (Stoer and Bulirsch, 1980).

Although the ML estimators of the parameters can be obtained from any distributional assumption on F, only the normal and lognormal distributions will be discussed here. The scoring method, assuming that F is a normal distribution function with two parameters, the mean μ and the variance σ^2, is illustrated in Appendix A.

Properties of Maximum Likelihood Estimators

In geoscience applications, the sample mean and variance (or the sample logarithmic mean and variance) are computed. Where the data are complete (no observations below detection) and normally distributed, the ML estimators of the mean μ and variance σ^2 are simply the sample mean and variance. If the data contain multiple censored observations, the proposed ML estimators are not as easy to obtain. However, they are the only proper generalization of the sample mean and variance. If the normality assumption is violated (the observations did not come from a normal population), then the ML estimators have no meaning regardless of whether or not the observations are complete.

Suppose that an element has a relatively high detection limit and therefore the value cannot be determined. For example, W in sample #7039 is less than a detection limit of 4900 ppm. This sample contains almost no information (only that the value is between 0 and 4900 ppm) and it should be removed from any further analysis. The next question is how high must the detection limit be before the sample is disregarded. This question is particularly relevant if the substitution method is used. A value of 2940 ppm (0.6x4900 ppm) substituted for <4900 ppm will distort the estimators. However, if the ML estimators are used, then it can be shown that this kind of sample has almost no effect on the estimators. The reason is that, for example, $\log(F(4900:u_k,k=1,...,m))$ will be near 0 regardless of u_k, k=1,...,m, and thus, in maximizing $L(u_k,k=1,...,m)$ in (1), this sample (<4900 ppm) will not have any influence on the ML estimators. This is illustrated in Table 2 where the presence or absence of four samples with high detection limits has very little effect on the ML estimator while it has a noticeable effect on the substitution method means (Table 3A). It should also be noted that the means and standard deviations are log values and cannot be applied to the data set directly.

Table 2: Maximum likelihood estimates for means and standard deviations for W and Au from rock type 3

	$\hat{\mu}$	$\hat{\mu}^*$	$\hat{\sigma}$	$\hat{\sigma}^*$
W (ppm)	2.89	2.90	2.39	2.41
Au (ppb)	0.15		3.17	
* estimates with <2500, <3500, <3000, <4900 removed				

Table 3A: Sample means and standard deviations for W from rock type 3 using substitution method

W (ppm)	$\hat{\mu}$	$\hat{\mu}^*$	$\hat{\sigma}$	$\hat{\sigma}^*$
Sub (0.4)	3.09	3.07	1.97	2.05
Sub (0.5)	3.22	3.15	1.91	1.98
Sub (0.6)	3.36	3.22	1.88	1.83
Sub (0.7)	3.49	3.30	3.62	3.37
Sub (0.8)	3.62	3.37	1.90	1.77
* estimates with <2500, <3500, <3000, <4900 removed				

Distribution of Au and W in the Ragged Ranges, South Nahanni River area

The most common distribution functions in the geosciences are the two parameter lognormal distribution functions. The two population parameters are the log-mean and the log-variance denoted, by μ and σ^2, respectively.

For the distribution of W and Au in Rock Type 3 in the Ragged Ranges area, 49 samples were collected. Among these, 21 samples have W values less than detection limits varying from 2 ppm to 4900 ppm; 39 samples have Au values less than detection limits varying from 5 ppb to 180 ppb as shown in Table 1A. Because of the multi-level detection limits, not even simple statistics such as the sample mean, median or percentiles are easily calculated. The substitution method would not provide any reasonable statistics related to the population because there is a large portion of data below the detection limit and these limits are commonly high (e.g. the detection limit of sample #7039 is 4900 pm). This is illustrated in Table 3A where five different values are substituted for the observations below detection in the W data of Rock Type 3. The five estimated means are distinct, and selecting one of them as an estimator would be difficult. Table 3A also contains the five substitution method estimates for W where four samples with high detection limits (#7016, #7017, #7036 and #7039) are deleted. The removal of these four samples has a much greater effect on the substitution method than the ML method, especially where the commonly substituted values of 0.5 and 0.6 are used. From Table 2, the MLE means are similar regardless of whether or not the four samples are used. This is important because it illustrates that it is not necessary to subjectively remove values from the data set before proceeding with statistical analyses; the ML method objectively recognizes that such samples contribute little to the knowledge of the distribution of the data. Table 3B shows the sample means using the substitution method for the Au data of Rock Type 3. Once again, the sample means produced by each substitution are unique.

Suppose that W and Au in Rock Type 3 are distributed as lognormal distributions with unknown parameters μ_w, σ_w and μ_{Au}, σ_{Au} respectively. Using the observations from 49 samples including values below detection, μ_w, σ_w and μ_{Au}, σ_{Au} are to be estimated. Estimates by MLE with and without the four samples for W (samples #7016, #7017, #7036 and #7039) are shown in Table 2. As noted in the previous section, the two sets of ML estimates for W - one

with and the other without the four samples, are virtually identical, since the four samples provide little information and do not influence the MLE's. However, this is not the case for the substitution method as shown in Table 3A where the sample means differ.

In particular, the ML estimates of μ_{Au} and σ_w in Table 2 are distinctly different from those in Table 3B because close to 80% of the observations are below the detection limit. Although the ML estimators are appropriate, the lognormality assumption is very important, and if violated, the estimators are meaningless. In Figure 1, the distribution function generated by the ML method is shown for Au in Rock Type 3. This curve is compared to three distribution curves generated by the substitution method using 0.4, 0.6 and 0.8 of the detection limit. The effect of substituting arbitrary values is seen by the shift to the right of the substitution-method curves. In all cases, the Au values appear to be higher than they probably are. This is an extreme example of the misleading effect of the substitution method because 41 of 48 samples are below the detection limit. However, it illustrates that the ML method can produce more meaningful and realistic results.

Table 3B: Sample means and standard deviations for Au from rock type 3 using substitution method

Au (ppb)	$\hat{\mu}$	$\hat{\sigma}$
Sub (0.4)	1.84	1.50
Sub (0.5)	2.07	1.42
Sub (0.6)	2.29	1.35
Sub (0.7)	2.52	1.29
Sub (0.8)	2.75	1.24

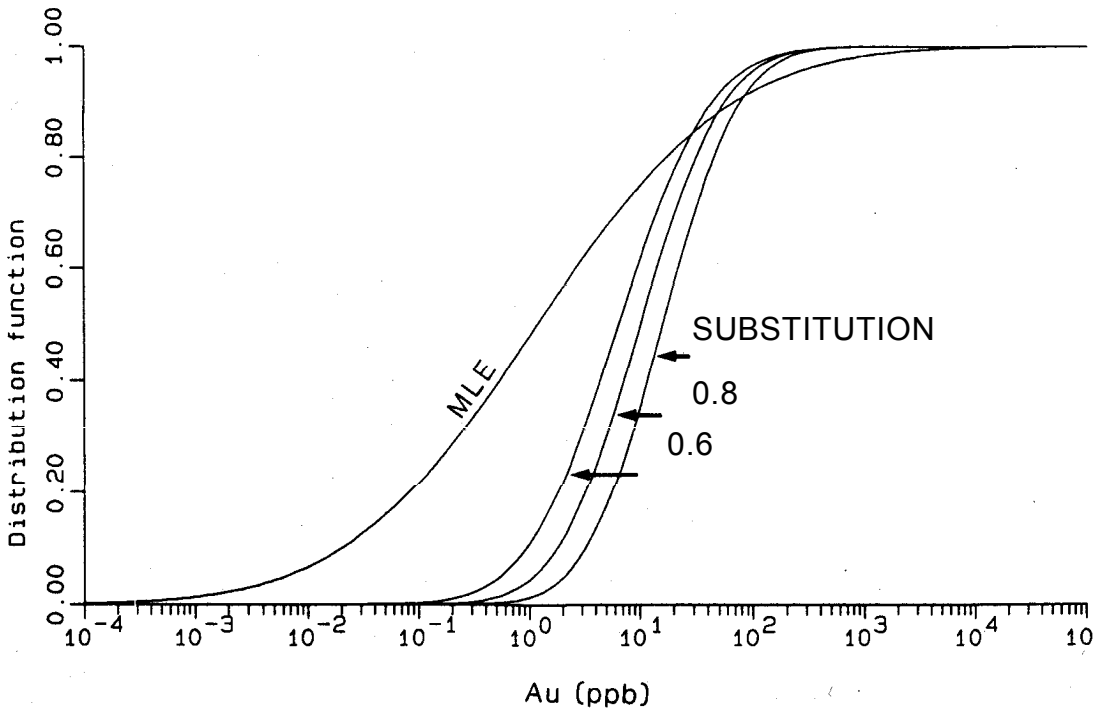

Figure 1: Four lognormal distribution functions for Au in rock type 3 estimated by the ML method (data from Table 2) and the substitution method (0.4, 0.6, 0.8) (data from Table 3B).

In Table 1B, W and Au values of 66 samples from Rock Type 4 are listed. Among them, 39 samples for W and 47 samples for Au

are below the detection limit. Similar to Tables 2A, 2B and 3, Table 4 includes the estimates for u_w, s_w, u_{Au} and s_{Au} of Rock Type 4 using the ML and substitution methods. In order to compare these two types of estimators, the confidence bands for the distribution function of W in Rock Type In Table 1B, W and Au values of 66 samples from Rock Type 4 are listed. Among them, 39 samples for W and 47 samples for Au are below the detection limit. Similar to Tables 2A, 2B and 3, Table 4 includes the estimates for u_w, s_w, u_{Au} and s_{Au} of Rock Type 4 using the ML and substitution methods. In order to compare these two types of estimators, the confidence bands for the distribution function of W in Rock Type 4 are constructed (Chung, 1987; Csorgo & Horvath, 1985). These are compared with two lognormal distribution functions for W which were estimated by the ML and the substitution methods (Fig. 2). The lognormal distribution function for W, estimated by the substitution method, falls outside of the confidence band. Therefore, the hypothesis that the 66 samples came from the lognormal distribution estimated by the substitution method is rejected. However, the hypothesis that the samples came from the lognormal distribution estimated by the ML method may be accepted, because the distribution function is constrained by the confidence bands. An empirical distribution curve (a modified product limit estimator), constructed using the observations for W in Rock Type 4 is shown between the confidence bands. This curve estimates the distribution of the data without the assumption of normality. Even the ML estimator does not fit well with this modified product-limit estimator for the distribution (cf. Chung, 1987) suggesting that the assumption of lognormality may be inaccurate.

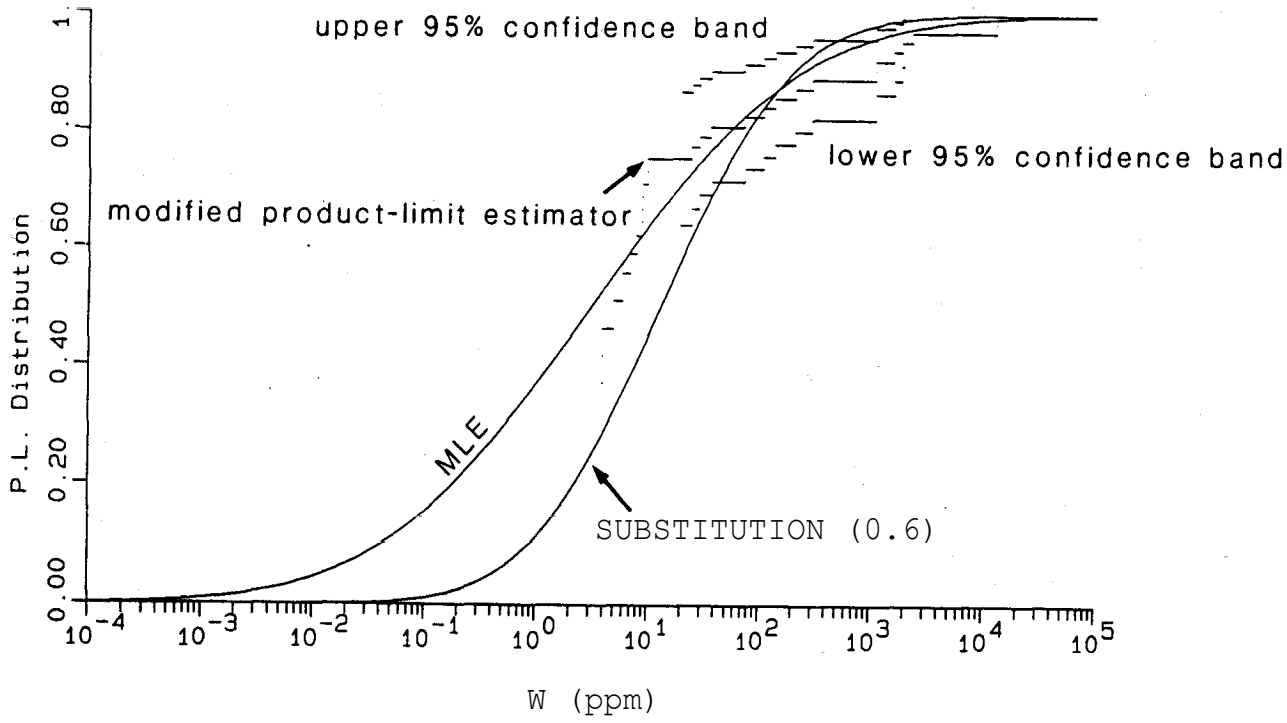

Figure 2: Csőrgo-Horváth 95% confidence band and a modified product-limit estimator for the distribution of W in rock type 4. Two lognormal distributions estimated by the ML and substitution methods are also shown..

Table 4: Estimators for means and standard deviations for W and Au from rock type 4 using maximum likelihood estimation and substitution methods

W (ppm)	$\hat{\mu}$	$\hat{\mu}^*$	$\hat{\sigma}$	$\hat{\sigma}^*$
MLE	1.13	1.13	3.37	3.39
Sub (0.4)	2.25	2.19	2.16	2.22
Sub (0.5)	2.41	2.31	2.13	2.16
Sub (0.6)	2.57	2.42	2.12	2.10
Sub (0.7)	2.73	2.54	2.13	2.05
Sub (0.8)	2.89	2.66	2.15	2.01
* values with <3600, <3400, <3600, <2900 removed				
Au (ppb)	$\hat{\mu}$		$\hat{\sigma}$	
MLE	0.82		2.92	
Sub (0.4)	1.96		1.64	
Sub (0.5)	2.14		1.56	
Sub (0.6)	2.32		1.49	
Sub (0.7)	2.51		1.43	
Sub (0.8)	2.69		1.38	

To compare the distribution functions of W in Rock Types 3 and 4, two lognormal distribution functions estimated by the ML method are illustrated in Figures 3A and B. Figure 3A shows the distribution functions of W in Rock Types 3 and 4 in probability density function form. The same distribution functions are shown in the cumulative distribution function form in Figure 3B. The mean for W in Rock Type 3 (shales) is greater than the mean for Rock Type 4 (platformal carbonates). The variance is much greater in Rock Type 4 and 59% of the data (vs 43% in Rock Type 3) are below the detection limit.

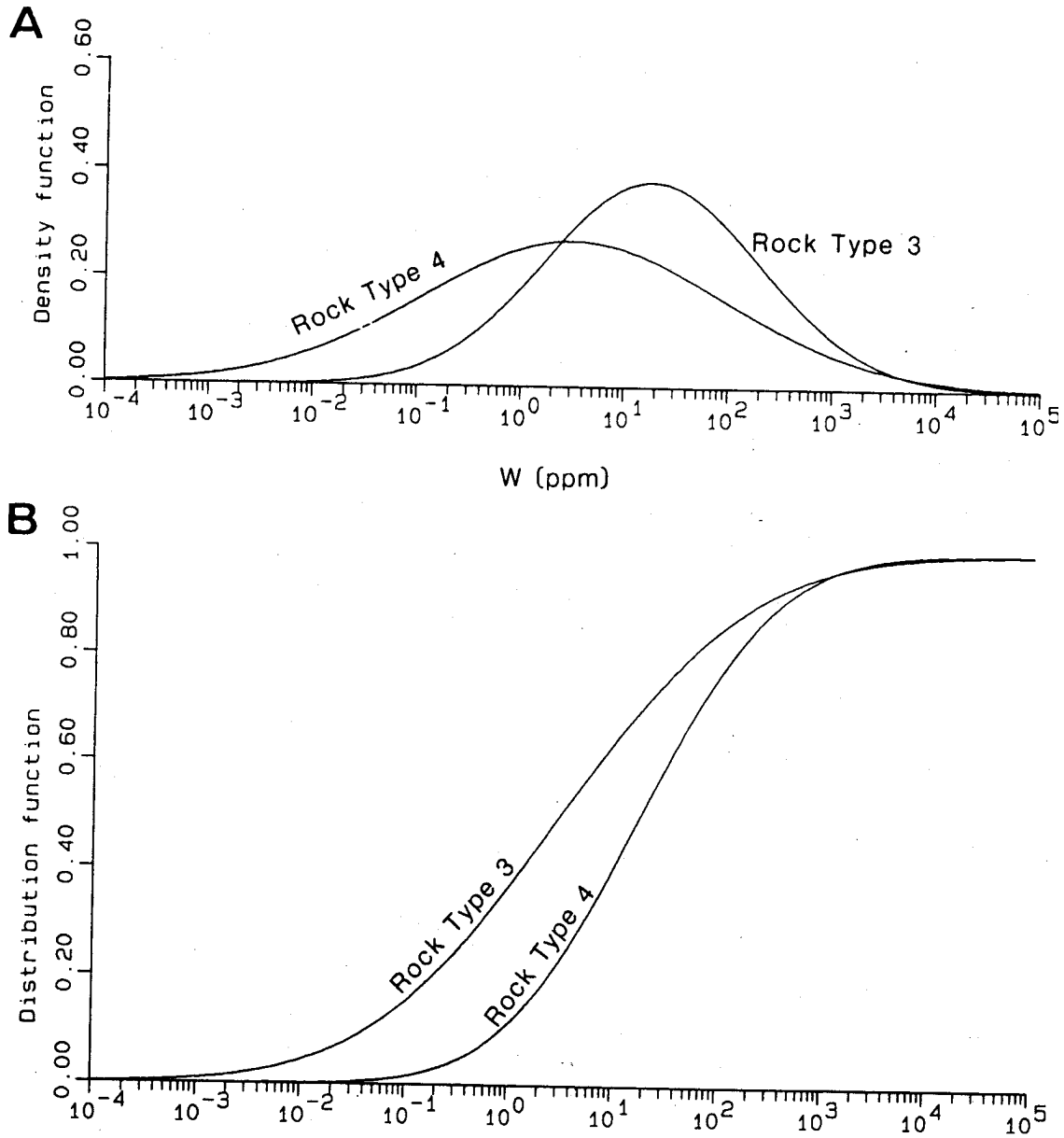

Figure 3A: Two lognormal density functions for W in rock types 3 and 4 estimated by the ML method (data from Tables 2 and 4), **3B:** Cumulative distribution function form of the distribution functions shown in Figure 3A.

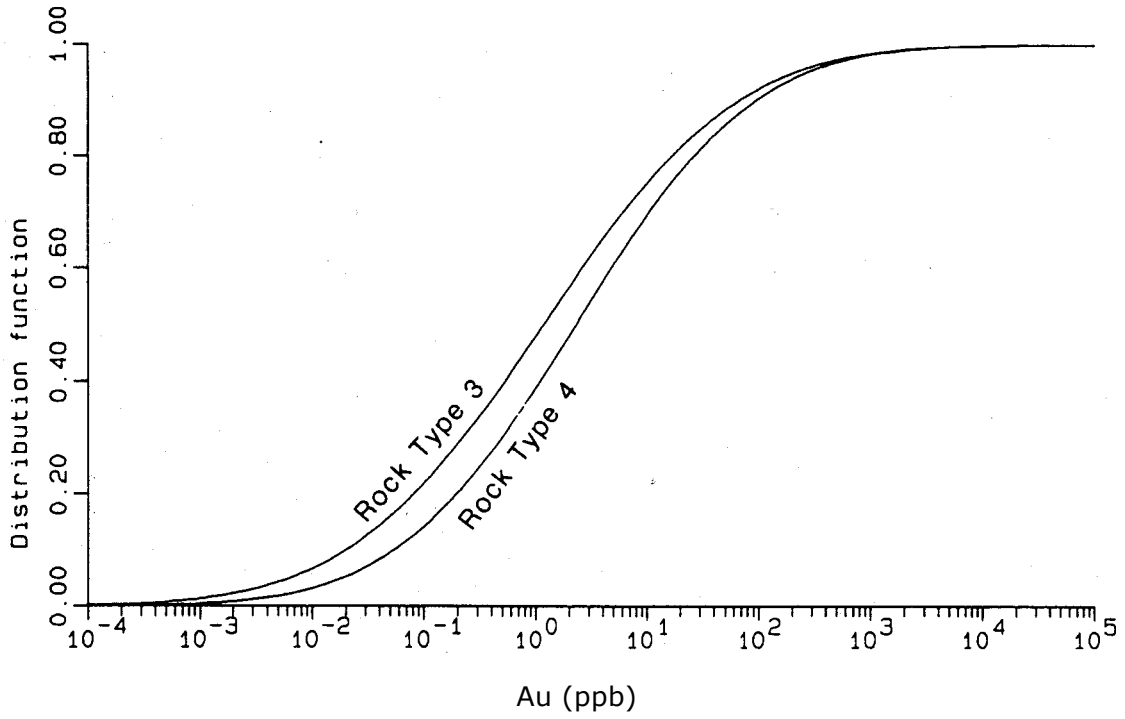

Figure 4: Two lognormal distribution functions for Au in rock types 3 and 4 estimated by the ML method (data from Tables 2 and 4)

Two lognormal distribution functions for Au in Rock Types 3 and 4 estimated by the ML method are illustrated in Figure 4. The two distribution functions for Au have similar shapes by the distribution of Rock Type 4 is shifted to the right because the mean is greated than in Rock Type 3. The variance of Au in these two rock types is similar.

Selection of Anomalously High Samples

The number of samples above a certain probability level (i.e. anomalous) is different for the ML and substitution methods. Tables 5A and 5B show critical values for the 98th, 95th and 90th percentiles based on means calculated by MLE and by substitution (0.6 of detection). The number of samples above that critical value is also listed. In all cases, the number of samples above a certain probability level is less for MLE than substitution. This means that the MLE method is more discriminating than the substitution method and, depending on other parameters used in the resource assessment, requires fewer samples to be re-checked in a follow-up survey.

Table 5A: Comparison of the number of samples above critical values for rock type 3 using MLE and substitution methods (critical value/# of samples above that value) d.l. = detection limit

		$\hat{\mu}$	98th	95th	90th	n	# < d.l.
W	MLE	2.89	2440/0	917/0	383/3	49	21
	SUB	3.36	1371/4	634/5	319/10	49	21
Au	MLE	0.15	784/1	214/1	67/7	49	41
	SUB	2.29	158/2	91/8	56/12	49	41

Table 5B: Comparison of the number of samples above critical values for rock type 4 using MLE and substitution methods (critical value/# of samples above that value) d.l. = detection limit

		$\hat{\mu}$	98th	95th	90th	n	# < d.l.
W	MLE	1.13	1736/2	491/5	160/7	66	39
	SUB	2.57	1019/11	427/11	197/13	66	39
Au	MLE	0.82	916/0	277/5	95/7	66	47
	SUB	2.32	217/5	118/9	69/12	66	47

Figure 5 uses data from Table 5B to plot distribution curves based on the MLE and substitution methods for Au in Rock Type 4. In addition, the three probability levels are plotted. Where these lines intersect the distribution curve is the value for each probability level. For example, the 95th level line intersects the ML curve at 277 and intersects the substitution curve at 118. In all three cases, the value at the point of intersection is less for the ML distribution curve.

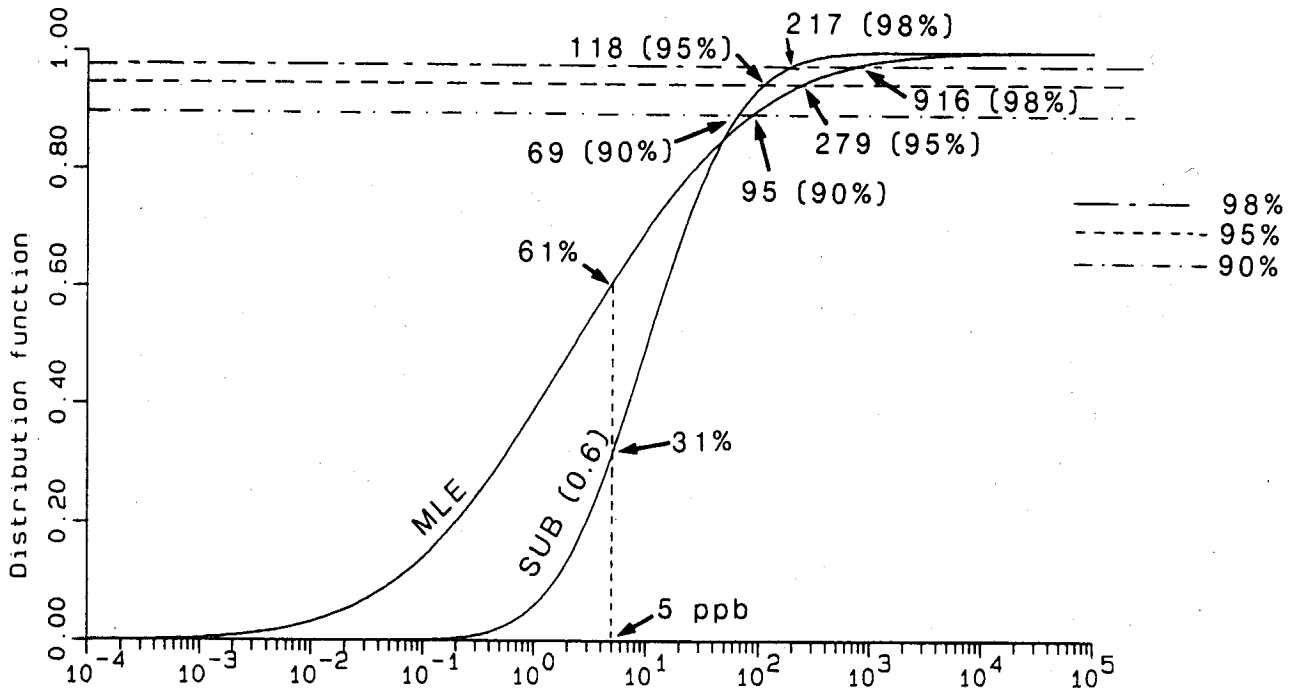

Figure 5: Two lognormal distributions for au in rock type 4 estimated by the ML and substitution methods. The three horizontal lines indicated three probability levels as discussed in text and Table 5.

Inferred Distribution of Low Values

The line representing 5 ppb (the laboratory's detection level for Au) intersects the MLE curve at 61% and the substitution curve at 31%. For the substitution curve, this means that 31% or 20 of the 66 Au values are expected to be <5. In the raw data, 19 samples are known to be <5 so that only 1 of the remaining 28 undetected values is expected to be <5. The MLE curve intersects the 5 ppb line at 61% so that 40 of 66

samples are expected to be <5. Again, because 19 of the samples are known to be <5, then 21 of the remaining undetected samples are expected to be <5. Because the laboratory's detection level is <5, and the presence of radioactive elements is common for this area, it is reasonable to assume that more than just one of the remaining 28 undetected values is actually <5, suggesting that the curve generated by the MLE method more accurately reflects the distribution of Au in Rock Type 4.

Concluding remarks

Statistical analysis of incomplete geochemical data is facilitated by using maximum likelihood estimators. For data sets which are distributed normally and are complete, the maximum likelihood estimators are simply the sample mean and variance. In the case of incomplete censored data, the maximum likelihood estimators (MLE) obtained by an iterative procedure, are the most appropriate estimates of the population mean and variance. If the assumption that the data are normal or lognormal is violated, then the estimators are meaningless, even when the data set is complete.

The gold and tungsten values of heavy mineral concentrates from the Ragged Ranges contain a large proportion of undetected values. If the data are to be used in a geological assessment, they should ideally be re-analyzed to reduce the size and variability of the detection limit. If this is not possible, the data must be used as they are. The MLE method can handle such data, provides more reasonable results than the substitution method, and is more discriminating for the comparison of different geologic environments.

Acknowledgements

We wish to thank Dr. C.W. Jefferson of the Geological Survey of Canada for providing not only the data, which are part of a set that was acquired for a resource assessment of South Nahanni River area, but also for useful comments on our earlier draft of this paper. We also wish to acknowledge the referees whose suggestions improved the manuscript. The resource assessment was jointly funded by Environment Canada (Parks), Indian and Northern Affairs Canada and the Geological Survey of Canada. Polar Continental Shelf Project provided additional logistical support.

References

Chung, C.F

1987: Confidence bands for the distribution and quantile functions for truncated and randomly censored data; in Quantitative Analysis of Mineral and Energy Resources; ed. C.F. Chung et al., Reidel, Dordrecht, p. 433-457.

1989a: Regression analysis of geochemical data with observations below detection limits. in Computer Applications in Resource Estimation: Prediction and Assessment for Minerals and Petroleum, ed. G. Gaal and D. Merriam, Pergamm, Oxford (in press).

1989b: Estimating mean and variance for geochemical data with multi-level detection limits (in preparation).

Csorgo, S. and Horvath, L.

1986: Confidence bands from censored samples; Canadian Journal of Statistics, V.14, p.131-144.

Dempster, A.P., Laird, N.M. and Rubin, D.B.

1977: Maximum likelihood from incomplete data via the EM algorithm; Journal of Royal Statistical Society, Series B, v. 39, p. 1-22.

Jefferson, C.W., Spirito, W.A., Hamilton, S.M., Michel, F.A. and Pare, D.

1989: Geochemistry of stream sediments, bedrock and spring waters in resource assessment of the South Nahanni River area, Yukon and N.W.T.; in Program and Abstracts, Contributions of the Geological Survey of Canada, Cordilleran Geology and Exploration Roundup, Geological Survey of Canada Miscellaneous Publication, February 7-10, 1989, p. 18-21.

Rao, C.R.

1975: Linear Statistical Inference and its Applications, 2nd ed., John Wiley and Sons: New York, 625p.

Scoates, R.F.J., Jefferson, C.W. and Findlay, D.C.

1986: Northern Canada mineral resource assessment; in Mineral Resource Assessment on Public Lands: Proceedings of the Leesburg Workshop, ed. S.M. Cargill and S.B. Green. U.S. Geological Survey Circular 980, p. 111-139.

Spirito, W.A., Jefferson, C.W. and Pare, D.

1988: Comparison of gold, tungsten and zinc in stream silts and heavy mineral concentrates, South Nahanni resource assessment area, District of Mackenzie; in Current Research Part E, Geological Survey of Canada Paper 88-1E, p. 117-126.

Stoer, J. and Bulirsch, R.

1980: Introduction to Numerical Analysis; Springer-Verlag, Berlin.

APPENDIX A. SCORING METHOD.

Assuming that F in (1) is the normal distribution function with the mean μ and variance σ^2, the log-likelihood function $L(\mu, \sigma)$ is written as:

$$L(\mu, \sigma) = \sum_{i=1}^{k} \log \phi(X_{h+i}: \mu, \sigma) + \sum_{j=1}^{h} \log \Phi(\alpha_j: \mu, \sigma) + \sum_{v=1}^{g} \log (1 - \Phi(\beta_v: \mu, \sigma)), \qquad \text{(A.1)}$$

where $\Phi(y: \mu, \sigma)$ and $\phi(x: \mu, \sigma)$ denote the normal distribution and density functions, respectively, with the mean μ and variance σ^2. The ML estimates μ and σ are obtained such that the log-likelihood function L in (A.1) is maximized.

The scoring method (Rao, 1975) based on the Taylor series expansion is an iterative procedure as follows:

$$\begin{pmatrix} \mu_{i+1} \\ \sigma_{i+1} \end{pmatrix} = \begin{pmatrix} \mu_i \\ \sigma_i \end{pmatrix} - \begin{pmatrix} \frac{\partial^2 \log L}{\partial \mu^2} & \frac{\partial^2 \log L}{\partial \sigma \partial \mu} \\ \frac{\partial^2 \log L}{\partial \mu \partial \sigma} & \frac{\partial^2 \log L}{\partial \sigma^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial \log L}{\partial \mu} \\ \frac{\partial \log L}{\partial \sigma} \end{pmatrix}_{\substack{\mu = \mu_i \\ \sigma = \sigma_i}} \qquad \text{(A.2)}$$

where μ_1 and σ_1 are initial estimates for μ and σ, and

$$\frac{\partial \log L}{\partial \mu} = \sigma^{-1} \left(\sum_{i=1}^{k} y_{h+i} - \sum_{j=1}^{h} \eta_j + \sum_{v=1}^{g} \tau_v \right)$$

$$\frac{\partial \log L}{\partial \sigma} = \sigma^{-1} \left(\sum_{i=1}^{k} y_{h+i}^2 - \sum_{j=1}^{h} \eta_j a_j + \sum_{v=1}^{g} \tau_v b_v - k \right)$$

$$\frac{\partial^2 \log L}{\partial \mu \partial \mu} = -\sigma^{-2} \left(k + \sum_{j=1}^{h} (a_j + \eta_j)\eta_j - \sum_{v=1}^{g} (b_v - \tau_v)\tau_v \right)$$

$$\frac{\partial^2 \log L}{\partial \sigma \partial \mu} = -\sigma^{-2} \left(2 \sum_{i=1}^{k} y_{h+i} + \sum_{j=1}^{h} ((a_j + \eta_j)a_j - 1)\eta_j - \sum_{v=1}^{g} ((b_v - \tau_v)b_v - 1)\tau_v \right)$$

$$\frac{\partial^2 \log L}{\partial \sigma^2} = -\sigma^{-2} \left(3 \sum_{i=1}^{k} y_{h+i}^2 + \sum_{j=1}^{h} ((a_j + \eta_j)a_j - 2)\eta_j a_j - \sum_{v=1}^{g} ((b_v - \tau_v)b_v - 2)\tau_v b_v \right)$$

$$y_{h+i} = \frac{Y_i - \mu}{\sigma}, \; a_j = \frac{\alpha_j - \mu}{\sigma}, \; b_v = \frac{\beta_v - \mu}{\sigma}$$

$$\eta_j = \frac{\phi(a_j)}{\Phi(a_j)}, \; \tau_v = \frac{\phi(b_v)}{1 - \Phi(b_v)}.$$

The iteration in (A.2) is continued until the differences of two successive estimates are less than a specified value.