USING EXPLORATION HISTORY
TO ESTIMATE UNDISCOVERED RESOURCES

by

DAN RABINOWITZ
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 131
OCTOBER 1989

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH (SIMS)

PREPARED UNDER A GRANT TO SIMS FROM
ENVIRONMENTAL PROTECTION AGENCY (EPA)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
DEPARTMENT OF APPLIED EARTH SCIENCES
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
USING EXPLORATION HISTORY 
TO ESTIMATE UNDISCOVERED RESOURCES 

Daniel Rabinowitz 
Stanford University 

Introduction and Summary

Methods of inference about deposits left to be discovered in a partially explored oil field that use the exploration history, the sequence of discovered deposits and dry holes, have typically made two kinds of assumptions. The first of these is assumptions about a "super-population" from which the deposits in the field are supposedly drawn. The second is assumptions about randomness in the exploration procedure.

Several models for the super-population have been proposed, and difficulties in testing the validity of such models have been described. Kauffman (1974) contains a review. Kauffman (1989) described how unbiased estimates of various characteristics of undiscovered deposits may be determined, without any assumptions about a super-population, after the field is observed on a random set of lines.

The purpose of this paper is to discuss inference when no assumptions about a super-population are made, and when the exploration procedure is supposed to follow the Arps-Roberts (1958) model. The Arps-Roberts model postulates that each well location is chosen at random from previously unexplored areas. It is hoped that such a model is sufficiently rich that the methods suggested here can be straightforwardly generalized to more realistic scenarios.

In what follows, the model is detailed, Kauffman's unbiased estimator is generalized to the model, and it is shown how confidence intervals based on the estimator may be found, a computational formula for UMVUE's is derived, the rudiments of a likelihood based approach are explored, some open questions are posed, and the methods are applied to a contrived data set.

The Model

Throughout, it is supposed that $N$ wells are sequentially drilled in an area, $A$, containing an unknown number, $K$, of deposits with unknown areas, $U = \{u_1, u_2, \ldots, u_k\}$. It is supposed that a deposit is observed if a well is contained in that deposit, and that each dry well is supposed to
allow an area $\epsilon$, to be observed as containing no deposit. The crux of the model is that the location of each well is assumed to be chosen at random from $A$ minus all previously discovered deposits and areas previously observed to contain no deposit.

Let

$$(u(1), u(2), \cdots, u(N))$$

denote the sequence of areas of discovered wells (with $u(i) = 0$ if the $i$th well is a dry hole). It follows that the probability of a sequence of observations is

$$\prod_{i=1}^{N} \left( \frac{A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}}{A - \sum_{j=i+1}^{N} u(j) - \epsilon \sum_{j=i+1}^{N} 1\{u(j) = 0\}} \right) \frac{u(i)}{A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}}$$

if $u(i) = 0$

and

$$\prod_{i=1}^{N} \left( \frac{A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}}{A - \sum_{j=i+1}^{N} u(j) - \epsilon \sum_{j=i+1}^{N} 1\{u(j) = 0\}} \right) \frac{u(i)}{A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}}$$

if $u(i) \neq 0$.

(1)

Let $F(t)$ denote the sum of the areas of deposits with areas less than $t$. Then, many quantities of interest may be expressed in the form $\int dF(t) \phi(t)$. For example, the total surface area of deposits is given by $\phi(t) \equiv 1$, the number of deposits by $\phi(t) = \frac{1}{t}$, and letting $\phi_i$ be the volume of the $i$th deposit, the total volume is given by

$$\phi(t) = \sum_{i=1}^{k} \phi_i 1\{u_i = t\} / t.$$

When the asymptotic behavior of the methods described here are considered, it is supposed that $A$, $N$ and $F(\infty)$ tend to infinity at the same rate, as this is the scenario in which a nontrivial proportion of the deposits are discovered. The following technical conditions are sufficient to justify the claims made in the proofs: $u_i$ and $\phi(u_i)$ are uniformly, as $N \to \infty$, bounded, and there exist constants $a$ and $b$, $0 < a < b < 1$, such that for all $N$ and all $i$,

$$A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}$$

and

$$A - \sum_{j=1}^{k} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}$$

are bounded between $aN$ and $bN$. Finally, it will be convenient to denote the algebra generated by the first $i$ observations by $F_i$.

**The Unbiased Estimator**

It follows from (1) that the probability that the $i$th well will discover a deposit with area less than or equal to $t$, given the previous exploration history is

$$\frac{F(t) - \sum_{j=1}^{i-1} u(j) 1\{u(j) \leq t\}}{A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}}.$$

2
and thus that

\[ \hat{F}_U(t) = \frac{1}{N} \sum_{i=1}^{N} \left[ \left( A - \sum_{j=1}^{i-1} u_{(j)} - \epsilon \sum_{j=1}^{i-1} 1_{\{u_{(j)} = 0\}} \right) 1_{\{0 < u_{(i)} \leq t\}} + \sum_{j=1}^{i-1} u_{(j)} 1_{\{0 < u_{(j)} \leq t\}} \right] \]

is unbiased for \( F(t) \), so that

\[ \int d\hat{F}_U(t) \phi(t) = \frac{1}{N} \sum_{i=1}^{N} \left[ \left( A - \sum_{j=1}^{i-1} u_{(j)} - \epsilon \sum_{j=1}^{i-1} 1_{\{u_{(j)} = 0\}} \right) 1_{\{u_{(i)} \neq 0\}} \phi(u_{(i)}) + \sum_{j=1}^{i-1} u_{(j)} \phi(u_{(j)}) \right] \]  

is unbiased for \( \int dF(t) \phi(t) \).

The expression, (2), suggests using the fact that the quantities

\[ \left[ \left( A - \sum_{j=1}^{i-1} u_{(j)} - \epsilon \sum_{j=1}^{i-1} 1_{\{u_{(j)} = 0\}} \right) 1_{\{u_{(i)} \neq 0\}} \phi(u_{(i)}) \right] / N + \sum_{j=1}^{i-1} u_{(j)} \phi(u_{(j)}) - \int dF(t) \phi(t) \]

are martingale differences with respect to \( \{\mathcal{F}_i\}_{i=1}^N \), so that

\[ \hat{\sigma}_U^2 = \sum_{i=1}^{N} \left[ \left( A - \sum_{j=1}^{i-1} u_{(j)} - \epsilon \sum_{j=1}^{i-1} 1_{\{u_{(j)} = 0\}} \right) 1_{\{u_{(i)} \neq 0\}} \phi(u_{(i)}) + \sum_{j=1}^{i-1} u_{(j)} \phi(u_{(j)}) - \int dF(t) \phi(t) \right]^2 / N^2 \]

is unbiased for the variance of \( \int d\hat{F}_u(t) \phi(t) \). It is shown in the appendix that

\[ \left[ \int d\hat{F}(t) \phi(t) - \int dF(t) \phi(t) \right] / \hat{\sigma}_U \]

is asymptotically normal with mean 0 and variance 1. It follows that \( \int d\hat{F}(t) \phi(t) \) may be substituted for \( \int dF(t) \phi(t) \) in the definition of \( \hat{\sigma}_U \) when using (3) as a pivot for confidence intervals.

The UMVUE

Let \( \{u_{(1)}, u_{(2)}, \ldots, u_{(N)}\} \) denote the collection of discovered deposits and dry holes. It follows from (1) that the collection is sufficient for \( \mathcal{U} \). It is shown in the appendix that it is complete as well. Since \( A1_{\{0 < u_{(1)} \leq t\}} \) is unbiased for \( F(t) \), it follows that

\[ \hat{F}_{\text{UMVUE}}(t) = E\{A1_{\{0 < u_{(1)} \leq t\}} \mid \{u_{(1)}, u_{(2)}, \ldots, u_{(N)}\}\} \]

is the UMVUE for \( F(t) \). (See Lehmann (1983, p. 75).) Similarly, \( E\{A\phi(u_{(1)}) \mid \{u_{(1)}, u_{(2)}, \ldots, u_{(N)}\}\} \) is the UMVUE for \( \int dF(t) \phi(t) \). It is shown in the appendix that

\[ P\{u_{(1)} = u \mid \{u_{(1)}, u_{(2)}, \ldots, u_{(N)}\}\} \]

may be expressed as
\[
\frac{1}{A} \int_0^\infty e^{-(A - \sum_{j=1}^N u(j) - \epsilon \sum_{j=1}^N 1\{u(j) = 0\})z} \prod_{\substack{j=1 \\ u(j) \neq u \neq 0}}^{N} (1 - e^{-u(j)x}) \, dz
\]
\[
\int_0^\infty e^{-(A - \sum_{j=1}^N u(j) - \epsilon \sum_{j=1}^N 1\{u(j) = 0\})z} (1 - e^{-\epsilon z}) \sum_{j=1}^{i-1} 1\{u(j) = 0\} \prod_{\substack{j=1 \\ u(j) \neq 0}}^{N} (1 - e^{-u(j)x}) \, dz
\]
\[4\]

This may be computed by standard numerical integration routines. Expanding the products in the integrals and integrating term by term provides sums which may either be computed exactly or by Monte Carlo. When \( \epsilon = 0 \), the limit as \( \epsilon \to 0 \) of this expression (given by substituting \( z \sum_{j=1}^N 1\{u(j) = 0\} \) for \( (1 - e^{-\epsilon z}) \sum_{j=1}^N 1\{u(j) = 0\} \)) is appropriate.

The Likelihood Approach

The probability, (1), may be rewritten as
\[
\prod_{i=1}^N \left\{ \frac{\frac{F(\infty) - \sum_{j=1}^{i-1} u(j)}{A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}}}{\frac{A - F(\infty) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}}{A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}}} \right\}^{1\{u(i) \neq 0\}} \prod_{i=1}^N \left( \frac{u(i)}{F(t) - \sum_{j=1}^{i-1} u(j)} \right)^{1\{u(i) \neq 0\}}.
\]

The first product is the partial likelihood of the indicators of the events \( \{u(i) \neq 0\} \) given the previous \( i - 1 \) wells. Let \( \hat{F}_{PL}(\infty) \) be that value of \( F(\infty) \) which maximizes the partial likelihood. The second derivative of the log partial likelihood is
\[
\hat{\sigma}_{PL}^2 = \sum_{i=1}^N \left\{ \frac{(F(\infty) - \sum_{j=1}^{i-1} u(j))^{-2}}{(A - F(\infty) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\})^{-2}} \right\} \text{ if } u(i) \neq 0
\]
\[
\left\{ \frac{(A - F(\infty) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\})^{-2}}{} \right\} \text{ if } u(i) = 0.
\]

It is shown in the appendix that
\[
[\hat{F}_{PL}(\infty) - F(\infty)]/\hat{\sigma}_{PL}
\]
\[5\]
is asymptotically normal with mean 0 and variance 1. In using (5) as a pivot, \( \hat{F}_{PL}(\infty) \) may be substituted for \( F(\infty) \) in the definition of \( \hat{\sigma}_{PL}^2 \). Note that
\[
E[\hat{\sigma}_{PL}^2] = E \sum_{i=1}^N (A - F(\infty) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\})^{-1} (F(\infty) - \sum_{j=1}^{i-1} u(j))^{-1}
\]
while for \( \phi = 1 \),
\[
E[\hat{\sigma}_{PL}^2] = E \sum_{i=1}^N (A - F(\infty) - \epsilon \sum_{j=1}^{i-1} 1\{u(j) = 0\}) (F(\infty) - \sum_{j=1}^{i-1} u(j)).
\]
It follows by Jensen’s inequality that, when estimating \( F(\infty) \) the estimator based on the likelihood approach is asymptotically superior to the unbiased estimator.

Open Questions

The likelihood approach may be extended to estimating \( F \) for various values of \( t \). One possible partial likelihood is

\[
\ell_t^{PL} = \prod_{i=1}^{N} \left\{ \frac{F(t) - \sum_{j=1}^{i-1} u(i) 1_{\{u(j) \leq t\}}}{A - \sum_{j=1}^{i-1} u(i) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j) > t\}} - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j) = 0\}}} \right\} \text{ if } 0 < u(i) \leq t
\]

\[
\frac{A - F(t) - \sum_{j=1}^{i-1} u(i) 1_{\{u(j) > t\}} - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j) = 0\}}}{A - \sum_{j=1}^{i-1} u(i) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j) = 0\}}} \text{ if } u(i) = 0 \text{ or } u(i) > t.
\]

It would be of interest to compare the variance of \( \int d\hat{F}_{PL}(t)\phi(t) \) (where \( \hat{F}_{PL}(t) \) maximizes \( \ell_t^{PL} \)) to the variance of \( \int d\hat{F}_{U}(t)\phi(t) \). It would also be of interest to determine to what extent optimal choices for \( \ell_t^{PL} \) may be made and analogously to explore the question of what other indicators (other than \( 1_{\{0 < u(i) \leq t\}} \)) might be used to construct unbiased estimates.

Example

In this section, the methods are applied to a very small contrived data set in order to illustrate the calculations. Suppose that the area \( A \) is 10, that the number of wells drilled \( N \) is 3, that the observed sequence of deposit areas is \( (0, 1, 2) \) and that \( \epsilon \) is 1. Consider the problem of measuring the total surface area of deposits \( (\phi(t) \equiv 1) \).

Table 1a details the calculations involved in using the unbiased estimator. From the third column

\[
\int d\hat{F}_n(t)\phi(t) = 26/3.
\]

From the fourth,

\[
N^2 \hat{\sigma}_U^2 = 144.6.
\]

An “asymptotic” \( 1 - \alpha \) confidence interval is

\[
\frac{26}{3} \pm Z_{1-\alpha/2}\sqrt{16.07}.
\]

Of course, the interval should be truncated to exclude values outside of \((0, A)\).

Table 1b details the calculation if the sequence had been \((2, 1, 0)\). In this case, the confidence interval is \( \frac{23}{3} \pm Z_{1-\alpha/2}\sqrt{\frac{98}{9}} \). Table 2 details the calculations for the UMVUE. Note that

\[
\sum_{i=1}^{N} u(i) = 0 + 1 + 2 = 3
\]
and that the denominator in (4) is

\[
\int_0^\infty e^{-\left(A - \sum_{j=1}^N u(j) - \epsilon \sum_{j=1}^N 1_{(u(j) = 0)}\right)z} \left(1 - e^{-\epsilon z}\right) \sum_{j=1}^{i-1} 1_{(u(j) = 0)} \prod_{\substack{j = 1 \\ u(j) \neq 0}} (1 - e^{-u(j)z}) \, dz
\]

\[
= \int_0^\infty e^{-6z}(1-e^{-z})^2(1-e^{-2z}) \, dz
\]

\[
= .003174595.
\]

From column 4, the UMVUE is 6.5.

Table 3 details the calculation of the likelihood and the formula for \( \hat{\sigma}_{\hat{\alpha}}^2 \). The likelihood is maximized at \( F(\alpha) = 0.68 \). Substituting into the variance formula provides \( \hat{\sigma}_{\hat{\alpha}}^2 = 11.93 \) so the confidence interval is \( 6.8 \pm Z_{1-\alpha/2} \times 2.59 \). Table 3b details the calculations for the likelihood method when the observations are \( 2, 1, 0 \), the MLE is 5.07, and the estimated standard error is 2.42.
Appendix

Asymptotic Normality of the Unbiased Estimator

The following notation will be used in this section:

\[
\xi_i = \left(A - \sum_{j=1}^{i-1} u_{(j)} \epsilon \sum_{j=1}^{i-1} 1_{\{u_{(j)} = 0\}} \phi(u_{(i)}) + \sum_{j=1}^{i-1} u_{(j)} \phi(u_{(j)}) - \int dF(t) \phi(t)\right) \cdot 1_{\{u_{(i)} \neq 0\}} \phi(u_{(i)}) + \sum_{j=1}^{i-1} u_{(j)} \phi(u_{(j)}) - \int dF(t) \phi(t)
\]

\[
v_i = E\{\xi_i^2 | F_{i-1}\}
\]

\[
\sigma^2 = E \sum_{i=1}^{N} \xi_i^2.
\]

In order to show that

\[
\left(\int d\hat{F}_U(t) \phi(t) - \int dF(t) \phi(t)\right) / \hat{\sigma}_U \to N(0,1)
\]

in distribution it suffices to show that

\[
\sum_{i=1}^{N} \xi_i / \sigma \to N(0,1)
\]

in distribution, and that

\[
\hat{\sigma}_U^2 / \sigma^2 \to 1
\]

in probability. The result then follows from Slutsky’s theorem.

That \(\hat{\sigma}_U^2 / \sigma^2 \to 1\) is the content of Lemma 1.

In order to show the asymptotic normality of \(\sum_{i=1}^{N} \xi_i / \sigma\), it is sufficient, by Theorem 1, Chapter VIII of Pollard (1984), to show that

\[
\sum_{i=1}^{N} v_i / \sigma^2 \to 1
\]

in probability, and that, for every \(\epsilon > 0\),

\[
\sum_{i=1}^{N} E\{\xi_i^2 / \sigma^2 1_{\{\xi_i^2 > \epsilon \sigma^2\}} | F_{i-1}\} \to 0.
\]

The second of these conditions follows from the fact that the \(\xi_i^2\) are uniformly, (in \(i\) and \(N\)) \(O(N^2)\), while \(\sigma^2\) is uniformly (in \(N\)) \(O(N^3)\). That \(\sum v_i / \sigma^2 \to 1\) in probability is the content of Lemma 2.
Lemma 1.

\[
\left[ \sum_{i=1}^{N} \frac{\xi_i^2 - \sigma^2}{\sigma^2} \right]/\sigma^2 \to 0
\]

in probability.

Proof. Note that

\[
\frac{\sum_{i=1}^{N} \xi_i^2 - \sigma^2}{\sigma^2} = \frac{\sum_{i=1}^{N} (\xi_i^2 - v_i)}{\sigma^2} + \frac{\sum_{i=1}^{N} (v_i - Ev_i)}{\sigma^2}.
\]

Now the \( \frac{\xi_i - v_i}{\sigma^2} \) are mean 0 uncorrelated random variables with variance on the order of \( N^{-2} \), so by Chebychev's inequality, the first sum on the right hand side tends to 0 in probability.

The convergence of the second sum follows by Lemma 2.

Lemma 2.

\[\forall \, \epsilon > 0, \quad P \left\{ \left| \sum_{i=1}^{N} v_i - \sum_{i=1}^{N} Ev_i \right| > \epsilon \sigma^2 \right\} \to 0.\]

Proof. Since

\[P \left\{ \left| \sum_{i=1}^{N} v_i - \sum_{i=1}^{N} Ev_i \right| > \epsilon \sigma^2 \right\} < \sum_{i=1}^{N} P \{|v_i - Ev_i| > \epsilon \sigma^2 / N\}\]

and since \( \sigma^2 = O(N^3) \), it suffices to show that \( \forall \, \epsilon > 0, \forall i \)

\[NP\{|v_i - Ev_i|/N^2 > \epsilon\} \to 0.\]  \hspace{1cm} (5)

Now,

\[v_i = \left( A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j)=0\}} \right)^2 \text{var}(\phi(u(i))1_{\{u(i)=0\}}|F_{i-1})\]

and \(\text{var}(\phi(u(i))1_{\{u(i)=0\}}|F_{i-1})\) is bounded, so it suffices to show that

\[NP\left\{ \left[ \left( A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j)=0\}} \right)^2 - E \left( A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j)=0\}} \right)^2 \right]/N^2 > \epsilon \right\} \to 0\]

and, since for \( i \neq j \),

\[\text{cov}(u(j) + \epsilon 1_{\{u(j)=0\}}, u(i) + \epsilon 1_{\{u(i)=0\}}) = 0 \left( \frac{1}{N} \right)\]

and

\[\text{var}(u(j) + \epsilon 1_{\{u(j)=0\}}) = O(1)\]
it follows that
\[ \text{var} \left( \left( A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j) = 0\}} \right) / N \right) = 0 \left( \frac{1}{N} \right) . \]

So, by Chebychev’s inequality,
\[ NP \left\{ \left[ \left( A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j) = 0\}} \right) - E \left( A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j) = 0\}} \right) \right] / N > \epsilon \right\} \rightarrow 0 . \]

The result thus follows by algebra together with the supposition that
\[ A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j) = 0\}} \]
is \( O(N) \).

**Completeness of the Sufficient Statistic and Derivation of the Formula for the UMVUE**

To show that \( \{u(1), u(2), \ldots, u(N)\} \) is complete, it suffices to show that if there exists a function \( \delta \) with the property that for any \( \mathcal{U} \), \( E\delta(\{u(1), u(2), \ldots, u(N)\}) = 0 \), then \( \delta \) must be identically 0. This can be shown by mathematical induction on the number of nonzero elements of \( \{u(1), u(2), \ldots, u(N)\} \).

(i) Suppose that \( \mathcal{U} \) is the empty set, then
\[ E\delta(\{u(1), u(2), \ldots, u(N)\}) = 0 \]
\[ \Rightarrow E\delta(\{0, 0, \ldots, 0\}) = 0 \]
\[ \Rightarrow \delta(\{0, 0, \ldots, 0\}) = 0 . \]

This establishes the base case.

(ii) Suppose that \( \forall n < m, \forall x \in S_n, \delta(x) = 0 \), where \( S_n \) is the class of \( \{u(1), u(2), \ldots, u(N)\} \)'s with \( n \) or less nonzero elements. Let \( y \in S_m \) and let \( \mathcal{U} \) be the nonzero elements of \( y \). Then
\[ E\delta(\{u(1), u(2), \ldots, u(N)\}) = 0 \]
\[ \Rightarrow E[\delta(\{u(1), u(2), \ldots, u(N)\})1_{\{\{u(1), u(2), \ldots, u(N)\} \in S_n\}}] + \delta(\{u(1), u(2), \ldots, u(N)\})1_{\{\{u(1), u(2), \ldots, u(N)\} = y\}} = 0 \]
\[ \Rightarrow 0 + \delta(y)P\{\{u(1), u(2), \ldots, u(N)\} = y\} = 0 \]
\[ \Rightarrow \delta(y) = 0 . \]

This is the induction step, and so the result is proved.

The basis for the computational formula for the UMVUE lies in the following lemma which describes an alternative probabilistic mechanism for generating the \( \{u(1), u(2), \ldots, u(N)\} \)'s.
Lemma 3. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_k\}$ and suppose that $A - \sum_{i=1}^{k} u_i = r\epsilon$ where $r$ is an integer. Let $X_1, X_2, \ldots, X_k$ be independent exponential random variables with parameters $u_1, u_2, \ldots, u_k$ and let $X_{k+1}, X_{k+2}, \ldots, X_{k+r}$ be independent exponential random variables with parameter $\epsilon$, independent of the $X_1, X_2, \ldots, X_k$.

Let $X(1), X(2), \ldots, X(k+r)$ denote the order statistics of the $X$'s. Finally, define

$$\tilde{u}_i = \begin{cases} u_j & \text{if } X(i) = X_j \text{ for } j \leq k \\ 0 & \text{if } X(i) = X_j \text{ for } j > k. \end{cases}$$

Then, the distribution of the sequence

$$(\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_N)$$

is the same as the distribution of

$$(u_1, u_2, \ldots, u_N)$$

under $\mathcal{U}$.

Proof. It suffices to show that the conditional distribution of $\tilde{u}_i$ given $\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_{i-1}\}$ is the same as the conditional distribution of $u(i)$ given $\{u(1), u(2), \ldots, u(i-1)\}$.

There is some notational convenience in restricting attention to the following comparisons: For arbitrary $p$, between

$$P\{u_i = u(i) \mid \{u_1, u_2, \ldots, u_{i-1}\} = \underbrace{0, 0, \ldots, 0}_{p \text{ zeros}}, u_1, u_2, \ldots, u_{i-p-1}\}$$

and

$$P\{\tilde{u}_i = u_i \mid \{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_{i-1}\} = \underbrace{0, 0, \ldots, 0}_{p \text{ zeros}}, u_1, u_2, \ldots, u_{i-p-1}\}.$$
(where $S_p$ is now the class of subsets of \( \{X_{k+1}, X_{k+2}, \ldots, X_{k+r} \} \) of size $p$)

\[
\sum_{S \in S_p} \int dF_{\text{max}}(S \cup \{X_1, X_2, \ldots, X_{i-p-1}\}) \\
\quad \times P\{X_i = \min(\{X_1, X_2, \ldots, X_{k+r}\} \setminus (S \cup \{X_1, X_2, \ldots, X_{i-p-1}\})) \} \\
\quad \times P\{\min(\{X_1, X_2, \ldots, X_{k+r}\} \setminus (S \cup \{X_1, X_2, \ldots, X_{i-p-1}\})) > z \} \\
\quad \max(S \cup \{X_1, X_2, \ldots, X_{i-p-1}\}) = z \}
\]

\[
\sum_{S \in S_p} P\{\max(S \cup \{X_1, X_2, \ldots, X_{i-p-1}\}) < \min(\{X_1, X_2, \ldots, X_{k+r}\} \setminus (S \cup \{X_1, X_2, \ldots, X_{i-p-1}\}))\}
\]

By independence and memorylessness, this equals

\[
\sum_{S \in S_p} \int dF_{\text{max}}(S \cup \{X_1, X_2, \ldots, X_{i-p-1}\})(z) \\
\quad P\{X_i = \min(\{X_1, X_2, \ldots, X_{k+r}\} \setminus (S \cup \{X_1, X_2, \ldots, X_{k-p-1}\}))\} \\
\quad \times P\{\min(\{X_1, X_2, \ldots, X_{k+r}\} \setminus (S \cup \{X_1, X_2, \ldots, X_{i-p-1}\})) > z\} \\
\sum_{S \in S_p} P\{\max(S \cup \{X_1, X_2, \ldots, X_{i-p-1}\}) < \min(\{X_1, X_2, \ldots, X_{k+r}\} \setminus (S \cup \{X_1, X_2, \ldots, X_{i-p-1}\}))\}
\]

Cancelling and integrating obtains

\[
P\{X_i = \min(\{X_1, X_2, \ldots, X_{k+r}\} \setminus (S \cup \{X_1, X_2, \ldots, X_{k-p-1}\}))\} \\
= u_i/ \left( A - \sum_{j=1}^{i-p-1} u_i - \rho \right).
\]

The computational formula for the UMVUE is derived by calculating

\[
E\{1_{\{u(1) = u\}}(\{u(1), u(2), \ldots, u(N)\}) = \{u_1^*, u_2^*, \ldots, u_N^*\}\}
\]

in terms of the exponential random variables. First note that, by sufficiency, the conditional expectation may be computed for any $U$ containing the nonzero elements of $\{u_1^*, u_2^*, \ldots, u_N^*\}$. It will be convenient to take $k$ equal to 1 plus the number of nonzero elements, and $U$ to be those nonzero elements together with an element of size

\[
A - \sum_{j=1}^{N} u_j^* - \rho \sum_{i=1}^{N} 1_{\{u_i^* = 0\}}.
\]

Let the number of nonzero elements be denoted by $n$ and denote the exponential random variables corresponding to those elements by $X_1, X_2, \ldots, X_N$. Let $X_{n+1}$ denote the exponential random variable corresponding to the additional element of $U$. Finally, denote the $N - n$ variables corresponding to the zero elements of $\{u_1^*, u_2^*, \ldots, u_N^*\}$ by $X_{n+2}, X_{n+3}, \ldots, X_{N+1}$. 

11
Thus we have

\[ P\{u(1) = u_j \mid \{u(1), u(2), \ldots, u(N)\}\} = P\{X(1) = X_j \mid X_{n+1} = \max\{X_1, X_2, \ldots, X_{N+1}\}\}. \]

By the memoryless property of exponentials and the independence assumption this probability may be computed as

\[
\frac{P\{X(1) = X_j; X_{n+1} = \max\{X_1, X_2, \ldots, X_{N+1}\}\}}{P\{X_{n+1} = \max\{X_1, X_2, \ldots, X_{N+1}\}\}} = \frac{P\{X(1) = X_j\}P\{X_{n+1} = \max\{X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_{N+1}\}\}}{P\{X_{n+1} = \max\{X_1, X_2, \ldots, X_{N+1}\}\}}
\]

\[
\frac{\int_0^\infty e^{-u_{n+1}z} \prod_{i=1 \atop i \neq j}^n (1 - e^{-u_i z})(1 - e^{-\epsilon z})^{N-n} \, dz}{\int_0^\infty e^{-u_{n+1}z} \prod_{i=1}^N (1 - e^{-u_i z})(1 - e^{-\epsilon z})^{N-n} \, dz}.
\]

This technique may be used to approximate other quantities of interest. As an example, consider the asymptotic distribution of the number of dry holes as a function of \(U\). For simplicity, take \(\epsilon = 0\). Then the observations

\(\{u(1), u(2), \ldots, u(N)\}\)

may be generated by observing a Poisson process \(N_t\) with parameter \((A - F(\infty))\) and \(k\) independent exponentials, \(z_1, z_2, \ldots, z_k\) with parameters \(u_1, u_2, \ldots, u_k\) until

\[ N_t + \sum_{i=1}^k 1_{\{Z_i < t\}} = N, \]

letting \(u_{(i)}\) be 0 if the \(i\)th event (either a jump in \(N_t\) or a jump in one of the indicators \(1_{\{Z_i < t\}}\)) is a jump in \(N\), and letting \(u_{(i)}\) be \(u_j\) if the \(i\)th event is a jump in \(1_{\{Z_j < t\}}\).

Let

\[ \tau_r = \inf\{t : N_t = r\}, \]

then \(P\{N \leq r\}\) may be expressed as

\[
\int_0^\infty dF_{\tau_r}(t) P\left\{\sum_{i=1}^k 1_{\{Z_i < t\}} \geq N - r \right\}.
\]

The distribution \(F_{\tau_r}\) is gamma with parameters \(r\) and \((A - F(\infty))\), so as \(r \to \infty\) and \(A \to \infty\), the integral converges with \(P\{\sum_{i=1}^k 1_{\{Z_i < t\}} \geq N - r\}\). The Liapunov central limit theorem may be used to approximate this last probability by

\[
1 - \Phi\left( \frac{N - r - \sum_{i=1}^k (1 - e^{-u_i \lambda - F(\infty)} \sqrt{k} \sum_{i=1}^k (1 - e^{-u_i \lambda - F(\infty)} e^{-u_i \lambda - \tau_r})}{\sqrt{k} \sum_{i=1}^k (1 - e^{-u_i \lambda - F(\infty)}) e^{-u_i \lambda - \tau_r}} \right).
\]

12
Asymptotic Normality of the Partial Likelihood Estimator

The following notation will be convenient

\[
\hat{\ell}(\theta) = \sum_{i=1}^{N} \frac{-\mathbb{1}_{\{u_{(i)}=0\}}}{A - \theta - \epsilon \sum_{j=1}^{i-1} \mathbb{1}_{\{u_{(j)}=0\}}} + \sum_{i=1}^{N} \frac{\mathbb{1}_{\{u_{(i)} \neq 0\}}}{\theta - \sum_{j=1}^{i-1} u_{(j)}}
\]

\[
\tilde{\ell} = \sum_{i=1}^{N} \frac{-\mathbb{1}_{\{u_{(i)}=0\}}}{A - \theta - \epsilon \sum_{j=1}^{i-1} \mathbb{1}_{\{u_{(j)}=0\}}} + \sum_{i=1}^{N} \frac{-\mathbb{1}_{\{u_{(i)}=0\}}}{\theta - \sum_{j=1}^{i-1} u_{(j)}}
\]

\[
\tilde{\ell}(\theta) = \sum_{i=1}^{N} \frac{-\mathbb{1}_{\{u_{(i)} \neq 0\}}}{\left(A - \theta - \epsilon \sum_{j=1}^{i-1} \mathbb{1}_{\{u_{(j)}=0\}}\right)^{3}}.
\]

Note that \(\hat{\sigma}_{PL} = [\hat{\ell}(F(\infty))]^{-1/2}\). By Taylor's theorem,

\[
[\hat{F}_{PL}(\infty) - F(\infty)]/\hat{\sigma}_{PL} = [\hat{\ell}(F(\infty))/\hat{\ell}^{1/2}(F(\infty))]/(1 + (\hat{F}_{PL}(\infty) - F(\infty)) \tilde{\ell}(\theta)/\hat{\ell}(F(\infty)))
\]

where \(\theta\) is between \(F(\infty)\) and \(\hat{F}_{PL}(\infty)\). It follows then, that in order to prove asymptotic normality, it suffices to show that

(i) \(\hat{\ell}(F(\infty))/\hat{\sigma}_{PL} \rightarrow N(0, 1)\) in distribution, and
(ii) \((\hat{F}(\infty) - F(\infty)) \tilde{\ell}(\theta)/(\hat{\ell}(F(\infty))) \rightarrow 0\) in probability.

For (ii), note that \(\tilde{\ell}(\theta) = O(N^{-2})\) and \(\hat{\ell}(F(\infty)) = O(N^{-1})\), so it suffices to show that \(\hat{F}_{PL}(\infty) - F(\infty)\) is \(O(N)\). By the convexity of the partial likelihood, it suffices to show that

\[
\forall \epsilon > 0, \quad P\{\hat{\ell}(F(\infty) + \epsilon N) > 0\} \xrightarrow{N} 1 \quad \text{and} \quad P\{\hat{\ell}(F(\infty) - \epsilon N) < 0\} \xrightarrow{N} 1.
\]

Well,

\[
\hat{\ell}(F(\infty) + \epsilon N) = \hat{\ell}(F(\infty)) + \epsilon N \hat{\ell}'\theta
\]

where \(\theta'\) is between \(F(\infty)\) and \(F(\infty) + \epsilon N\), and since \(\epsilon N \hat{\ell}'\theta\) is \(O(\epsilon)\), \(E\hat{\ell}(F(\infty)) = 0\) and \(\text{var}(\hat{\ell}(F(\infty))) = 0(N^{-1})\). The first convergence follows. The \(F(\infty) - \epsilon N\) case is similar.

The martingale central limit theorem referenced in the first section of the appendix may be used to establish (i). The two conditions which must be checked follow from the fact that the increments in \(\hat{\ell}_{\theta}\) are \(O(\frac{1}{N^2})\) while \(\text{var}(\hat{\ell}(F(\infty))) = O(\frac{1}{N})\) and the fact that \(\text{var}(\hat{\ell}(F(\infty)))/\text{var}(\hat{\ell}) = O(\frac{1}{N})\).
References


Table 1a

<table>
<thead>
<tr>
<th>i</th>
<th>u(i)</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>((\frac{26}{3})^2)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>((9 - \frac{26}{3})^2)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>17</td>
<td>((17 - \frac{26}{3})^2)</td>
</tr>
</tbody>
</table>

26

144.6

a) \(\sum_{j=1}^{i-1} u(j) + \epsilon \sum_{j=1}^{i-1} 1\{u(j)=0\}\)

b) \(\sum_{j=1}^{i-1} u(j)\phi(u(j))\)

c) \(\left( A - \sum_{j=1}^{i-1} \epsilon \sum_{j=1}^{i-1} 1\{u(j)=0\} \right) 1\{0 \neq u(i)\} \phi(u(i)) + \sum_{j=1}^{i-1} u(j)\phi(u(j))\)

d) \[\left( A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1\{u(j)=0\} \right) + \sum_{j=1}^{i-1} u(j)\phi(u(j)) - \int d\tilde{F}(t)\phi(t)\]^2

Table 2

<table>
<thead>
<tr>
<th>u</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/10 (\int_0^\infty e^{-6z}(1-e^{-z})(1-e^{-2z})dz = 0.00992063)</td>
<td>1/10 (\times 0.00992063/0.03174595 = 0.3125007)</td>
</tr>
<tr>
<td>2</td>
<td>1/10 (\int_0^\infty e^{-6z}(1-e^{-z})^2dz = 0.00595237)</td>
<td>1/10 (\times 0.00595237/0.03174595 = 0.3375002)</td>
</tr>
</tbody>
</table>

a) \(\int_0^\infty e^{(A-\sum_{j=1}^{N} u(j)-\epsilon \sum_{j=1}^{N} 1\{u(j)=0\})} (1 - e^{-\epsilon z})^{\sum_{j=1}^{i-1} 1\{u(j)=0\}} \prod_{u(j) \neq u_{(1)}}^{N} (1 - e^{-u(j)z})dz\)

b) \(P\{u_{(1)} = u | \{u_{(1)}, \cdots, u_{(N)}\}\})

Table 3a

<table>
<thead>
<tr>
<th>i</th>
<th>(\sum_{j=1}^{i-1} u(j))</th>
<th>(\sum_{j=1}^{i-1} 1{u(j)=0})</th>
<th>multiplicand for likelihood</th>
<th>addend for variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>10 - (F(\infty))</td>
<td>(\frac{1}{(10-F(\infty))^2})</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>(F(\infty))</td>
<td>(\frac{1}{(F(\infty))^2})</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>(F(\infty) - 1)</td>
<td>(\frac{1}{(F(\infty)-1)^2})</td>
</tr>
</tbody>
</table>
Table 1b

<table>
<thead>
<tr>
<th>$i$</th>
<th>$u(i)$</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>$(10 - \frac{22}{3})^2$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>$(10 - \frac{22}{3})^2$</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>$(3 - \frac{23}{3})^2$</td>
</tr>
</tbody>
</table>

23 98

a) $\sum_{j=1}^{i-1} u(j) + \epsilon \sum_{j=1}^{i-1} 1_{\{u(j)=0\}}$

b) $\sum_{j=1}^{i-1} u(j) \phi(u(j))$

c) $\left( A - \sum_{j=1}^{i-1} - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j)=0\}} \right) 1_{\{0 \neq u(i)\}} \phi(u(i)) + \sum_{j=1}^{i-1} u(j) \phi(u(j))$

d) $\left[ (A - \sum_{j=1}^{i-1} u(j) - \epsilon \sum_{j=1}^{i-1} 1_{\{u(j)=0\}}) + \sum_{j=1}^{i-1} u(j) \phi(u(j)) - \int dF(t) \phi(t) \right]^2$

Table 3b

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\sum_{j=1}^{i-1} u(j)$</th>
<th>$\sum_{j=1}^{i-1} 1_{{u(j)=0}}$</th>
<th>multiplicand for</th>
<th>addend for</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>likelihood</td>
<td>variance</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$F(\infty)$</td>
<td>$\frac{1}{F(\infty)^2}$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>$F(\infty) - 2$</td>
<td>$\frac{1}{(F(\infty) - 2)^2}$</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>$10 - (F(\infty) - 3)$</td>
<td>$\frac{1}{(10 - F(\infty) - 3)^2}$</td>
</tr>
</tbody>
</table>