SUCCESSIVE SAMPLING AND SOFTWARE RELIABILITY

by

Gordon M. Kaufman*

MIT Sloan School Working Paper 3316
July 1991

*Supported by AFOSR Contract #AFOSR-890371. I wish to thank Nancy Choi and Tom Wright for valuable programming assistance.

SUCCESSIVE SAMPLING AND SOFTWARE RELIABILITY

by

Gordon M. Kaufman*

1. Introduction

A software system is tested and times between failures are observed. How many faults remain in the system? What is the waiting time to the next failure? To the occurrence of the next n failures? Conditional on the observed history of the test process, knowledge of properties of the time on test necessary to discover the next n faults is very useful for making test design decisions.

Times between failures models of software reliability are designed to answer such questions. Many versions of such models appear in the literature on software reliability and most such models rely on the assumption that the failure rate is proportional to the number of remaining faults or to some single valued function of the number of remaining faults. Goel (1985) observes that this is a reasonable assumption if the experimental design of the test assures equal probability of executing all portions of the code – – a design seldom achieved in practice. The character of testing usually varies with the test phase: requirements, unit, system or operational. The impact of such considerations have been recognized by some authors: Littlewood's criticism of the Jelinski–Moranda assumption that software failure rate at any point in time is directly proportional to the residual number of faults in the software is cited by Langberg and Singpurwalla (1985) in an excellent overview paper. Only recently have some researchers come to grips with the implications of replacing this assumption. In terms of counts of failures, it may be labelled an "equal bug size" postulate (Scholz (1985). Littlewood (1981) and Singpurwalla and Langberg (1985) do incorporate the assumption that different bugs may have different failure rates, but the empirical Bayes (superpopulation) approach adopted by Littlewood and the Bayesian approach adopted by Singpurwalla and Langberg "averages out" the effects of this assumption. According to Scholz "...it was not recognized by some proponents of reliability growth models that relaxing the equal bug size assumption also entails some complications concerning the independence and exponentiality [of waiting times between failures]". He and Miller (1986) are the first to investigate systematically (in the absence of a super-

* Supported by AFOSR Contract #AFOSR–89–0371. I wish to thank Nancy Choi and Tom Wright for valuable programming assistance.

population process or of a Bayesian prior for failure rates) the implications of assuming that given an observational history, each of the remaining bugs in a software system may possess different probabilities of detection at a given point in time. In contrast to most times between failures models, for successive sampling - EOS models, times between failures are **not** independent. As Goel [1985] points out, independence would be acceptable if "...successive test cases were chosen randomly. However, testing especially functional testing, is not based on independent test cases, so that the test process is not likely to be random".

Scholz presents a multinomial model for software reliability that is identical to Rosén's characterization of successive sampling stopping times. (Rosén, 1972) The connection seems to have gone unnoticed. The "continuous" model based on independent, non–identically distributed exponential random variables suggested by Scholz as an approximation to multinomial waiting times is in fact in exact correspondence with a representation of successive sampling in terms of non–identically distributed but independent exponential order statistics. Scholz's approximation is in fact Ross's (1985) exponential order statistics model which Ross treats Bayesianly. Gordon (1983) was among the first to observe that successive sampling is representable in this fashion. Miller's study of such order statistics is focused on similarities and differences between types of models derivable from this particular paradigm.

Joe (1989) provides an asymptotic (large sample) maximum likelihood theory for parametric order statistics models and non-homegenous Poisson models of fault occurrence that, when the parameter is of fixed dimension, yields asymptotic confidence intervals. He states that for the general exponential order statistics model, one cannot expect any estimate [of the conditional failure rate] to be good because the ratio of parameters to random variables is too big".

Successive sampling as described in the next section has been successfully used as a model for the evolution of magnitudes of oil and gas field discovery and has its roots in the sample survey literature. (Hájek (1981), for example.) In this application magnitudes of fields in order of discovery are observed and used to make predictions of the empirical frequencies of magnitudes of undiscovered fields. Logically tight theories of maximum likelihood, moment type and unbiased estimation for this class of problems have been developed by Bickel, Nair and Wang, (1989), Gordon, (1989) and Andreatta and Kaufman, (1986). The problem of estimation of software reliability based on observation of times between failures of a software system may be viewed as the dual to the problem of inference when only magnitudes of population elements are observed. The principal purpose of this paper is to establish connections between these two disparate lines of research and to lay out

possibilities for applying methods of estimation developed for successive sampling schemes to successive sampling as a model for software reliability. Our attention is restricted to successive sampling of elements of a finite population of software faults in a software system; that is,

(1) Individual faults may possess distinct failure rates that depend on covariates particular to the stage of testing and on other features of the software environment. For a given fault, that fault's failure rate as a function of such covariates is called the fault **magnitude.**

(2) Faults are sampled (a) without replacement and (b) proportional to magnitude.

Some recent studies of successive sampling schemes have assumed the existence of a superpopulation process generating finite population magnitudes. We shall not.

The accuracy of model structure as a depiction of the physics of software fault occurrence depends in part on the validity of choice of definition for the magnitude of a fault. Different definitions of magnitude may be required for different environments. Here we shall assume that the appropriate definition of the magnitude of a fault for the particular application considered has been resolved. It is NOT easy to resolve and considerable effort must be devoted to defining operationally meaningful definitions of fault magnitudes. For example, the effort in programmer time required to fix a fault could be adopted. (Programmer effort expended to correct faults is measured and recorded for six software projects studied by the Software Engineering Laboratory at NASA-Goddard). Empirical work on fault prediction by Basili and Perricone (1982), and Basili and Patniak (1986) provides an excellent starting point. It is a subject for a different paper.

Following a formal description of successive sampling properties of successive sampling schemes needed in the sequel, two distinct sampling (observational) schemes are examined in section three. The first is a scheme in which both the time from start of testing to time of occurrence **and** the magnitude of each fault in a sample of n faults are jointly observed. With this scheme we can order faults observed from first to last and assign a "waiting time" to each fault. In the second scheme magnitudes of faults in a sample of n faults are observed along with the waiting time to occurrence of the last fault in the sample; waiting times to occurrences of individual faults are **not** observed. The order in which faults occurred is then lost.

Section 4 is devoted to properties of unbiased estimators of unobserved finite population parameters for each of the two aftermentioned sampling schemes. The connection between maximum likelihood estimation **(MLE)** and unbiased estimation established by Bickel, Nair and Wang (1989) for a successive sampling scheme in which magnitudes alone

are observed is developed for a scheme in which both waiting times to failures and magnitudes are observed. The results of a Monte Carlo study of properties of both types of estimators presented in Section 5.

Section 6 returns a principal interest of the software manager: conditional on observing the history of the process up to and including the m^{th} failure, what is the waiting time to the occurrence of the next $n - m$ failures? Successive sampling theory suggests a simple point estimator of this waiting time, dependent on the waiting time $z_{(m)}$ to occurrence of the first m faults and on the unordered set $\{y_1, \ldots, y_m\}$ of magnitudes of faults observed in $(0, z_{(m)})$. A Monte Carlo study of its behavior suggests that this class of estimators of returns to test effort measured in faults/unit time on test is worth further study.

2 Successive Sampling

Let $U = \{1, 2, \ldots, N\}$ denote the set of labels of N finite population elements and associate with each $k \epsilon U$ a **magnitude** $a_k > 0$. A successive sampling scheme is characterized by a permutation distribution of labels induced by sampling of $\mathcal{A}_N = \{a_1, \ldots, a_N\}$ proportional to magnitude and with replacement. The **parameter** $\underline{\mathcal{A}}_N \equiv (a_1, \ldots, a_N)$ of a finite population with label set U takes on values in the parameter space $(0, \infty)^N$.

Define for $1 \leq n \leq N$, $\underline{s}_n = (i_1, \ldots, i_n)$, $i_1, \ldots, i_n \epsilon U$ to be an **ordered** sample $\underline{s}_n$ of size n with i^th component i_j and $s_n = \{i_1, \ldots, i_n\} \subseteq U$ to be an **unordered** sample of size n. We distinguish a random variable from a value assumed by it with capital letter; e.g., the random variable $\underline{S}_n$ assumes a values $\underline{s}_n$. With $r_N = \sum_{k=1}^{N} a_k$, a successive sampling scheme is implemented via

$$P\{\underline{S}_n = \underline{s}_n \mid \underline{\mathcal{A}}_N\} \equiv P\{\underline{s}_n \mid \underline{\mathcal{A}}_N\} = \prod_{j=1}^{n} a_{i_j} / [r_N - a_{i_j} - \cdots - a_{i_{j-1}}] \tag{2.1}$$

with $a_{i_0} \equiv 0$; equivalently with $p_j = a_j / r_N$, $\sum_{j=1}^{N} p_j = 1$ and

$$P\{\underline{s}_n \mid \underline{\mathcal{A}}_N\} = \prod_{j=1}^{n} p_{i_j} / [1 - p_{i_1} - \cdots - p_{i_{j-1}}]. \tag{2.2}$$

The permutation distribution (2.1) of labels induces a distribution of magnitudes: setting $y_j = a_{i_j}, y_j$ is the magnitude of the j^{th} observation and with $\underline{y}_n = (y_1, \ldots, y_n)$ for $n = 1, 2, \ldots, N$,

$$P\{\underline{Y}_n = \underline{y}_n | \underline{\mathcal{A}}_N\} = P\{\underline{s}_n | \underline{\mathcal{A}}_N\} \tag{2.3}$$

is the probability of observing magnitudes in the order y_1 first, y_2 second, etc.

An alternative representation of (2.1) in terms of exponential order statistics (Gordon (1983)) is as follows: let $X_1, \ldots, X_n$ be miid exponential random variables with means one. Then

$$P\{\underline{s}_n \mid \underline{\mathcal{A}}_N\} = P\left\{ \frac{X_{i_1}}{a_{i_1}} < \frac{X_{i_2}}{a_{i_2}} < \ldots < \frac{X_{i_N}}{a_{i_N}} \right\}. \tag{2.4}$$

Defining $Z_j = X_j / a_j$, $Z_{(j)}$ as the j^{th} smallest of $Z_1, \ldots, Z_N$ and $U/s_n = \{k \mid k \epsilon U$ and $k \not\epsilon s_n\}$, the probability that an unordered sample s_n is observed is

$$\begin{aligned} P\{s_n \mid \underline{\mathcal{A}}_N\} &= P\{\max\{Z_j \mid j \epsilon s_n\} \leq \min\{Z_k \mid k \epsilon U/s_n\}\} \\ &= \sum_{k \epsilon U/s_n} P\{\max\{Z_j \mid j \epsilon s_n\} \leq Z_k < \min\{Z_\ell \mid \ell \epsilon U/s_n \text{ and } \ell \neq k\}\}. \end{aligned} \tag{2.5}$$

An integral representation of $P\{s_n \mid \underline{\mathcal{A}}_N\}$ suggested by (2.5) is, with $\alpha \equiv r_N - \sum_{k \epsilon s_n} a_k$,

$$P\{s_n \mid \underline{\mathcal{A}}_N\} \equiv P(s_n \mid \alpha) = \int_0^\infty e^{-\alpha x} d\left[\prod_{j \epsilon s_n} (1 - e^{-a_j x})\right]$$

$$= \alpha \int_0^\infty e^{-\alpha x} \prod_{j \epsilon s_n} (1 - e^{-a_j x}) dx. \tag{2.6}$$

We shall need derivatives of $P(S_n \mid \alpha)$ with respect to α and in order to maintain consistency of notation will call the rightmost integral in (2.6) $P_{n+1}(s_n \mid \alpha)$ as the density of $Z_{(n+1)}$ conditional on $S_n = s_n$ is just the normalized integrand of this integral:

$$D_{n+1}(z \mid s_n; \alpha) = \alpha e^{-\alpha z} \prod_{j \epsilon s_n} (1 - e^{-a_j z}) / P_{n+1}(s_n \mid \alpha) \tag{2.7}$$

for $z \epsilon (0, \infty)$ and zero otherwise. In a similar vein, the middle integral in (2.6) will be called $P_n(s_n \mid \alpha)$ as the density of $Z_{(n)}$ conditional on $S_n = s_n$ is

$$D_n(z \mid s_n; \alpha) = [e^{-\alpha z} \sum_{j \epsilon s_n} a_j e^{-z a_j} \prod_{\substack{\ell \epsilon s_n \\ \ell \neq j}} (1 - e^{-z a_\ell})] / P_n(s_n \mid \alpha). \tag{2.8}$$

From (2.6) we have

$$\frac{d}{d\alpha} \log\ P_{n+1}(s_n \mid \alpha) = \frac{d}{d\alpha} \log\ P_n(s_n \mid \alpha)$$

which is equivalent to

$$E(Z_{(n+1)} \mid s_n; \alpha) = \frac{1}{\alpha} + E(Z_{(n)} \mid s_n; \alpha). \tag{2.9}$$

The next two propositions record properties of marginal expectations of $Y_1, \ldots, Y_N$. The first documents the intuitively obvious notion that when elements of A are distinct from one another, the marginal expectation $E(Y_k)$ of Y_k strictly decreases with increasing k.

Proposition 2.1: When elements of $\underline{\mathcal{A}}_N$ are distinct $E(Y_1) > E(Y_2) > \ldots > E(Y_N)$.

Expectations $E(Y_k)$, $k = 1, 2, \ldots, N$ may be usefully expressed in terms of inclusion probabilities $\pi_j(n)$ as follows: let $\pi_j(n) = P\{j \epsilon S_n | \underline{\mathcal{A}}_N\}$.

Proposition 2.2: For sample size is $n = 1, 2, \ldots, N$, with $\pi_j(0) \equiv 0$,

$$E(Y_n) = \sum_{j=1}^{N} [\pi_j(n) - \pi_j(n-1)] a_j. \tag{2.10}$$

We shall adopt Hájek bounds on the $a_j s$ and examine large N behavior in the following uniform asymptotic regime:

(A_1): For a positive constant ϵ independent of N, $\epsilon < a_k < 1/\epsilon, \quad k \epsilon U_N$.

(A_2): As $N \to \infty$, $n/N = f_n \to f \ \epsilon \ (0,1)$ with f bounded away from zero and one.

In order to simplify notation we eschew double subscripting of sequences. Henceforth, $N \to \infty$ will serve as shorthand for (A_2). Gordon (1989) has established that under more general conditions than (A_1), (A_2) implies $Z_{(n)} \xrightarrow{P} z(f)$ where $z(f_n)$ is the solution to

$$\frac{1}{N} \sum_{j=1}^{N} e^{-a_j z(f_n)} = 1 - f_n, \tag{2.11}$$

and $\lim_{N \to \infty} z(f_n) \to z(f)$. An immediate consequence of (A_1), (A_2) and (2.11) is that for large N, $z(f_n) = 0(1)$ because $-\epsilon \ log \ (1 - f_n) < z(f_n) < - \ log \ (1 - f_n)/\epsilon$ and $-log \ (1 - f_n) = 0(1)$.

3. Likelihood Functions for Ordered and Unordered Samples

Consider a sample of size $n \leq N$ generated according to a successive sampling scheme as defined by (2.1) and (2.2) and let $\underline{z}_{(n)} = (z_{(1)}, \ldots, z_{(n)})$. When both Z_n and Y_n are observed,

$$Prob\{\underline{Z}_{(n)} \epsilon\, d\underline{z}_{(n)} \text{ and } \underline{Y}_n = \underline{y}_n \mid \underline{\mathcal{A}}_N\} = \text{Prob}\{\underline{Z}_{(n)} \epsilon\, d\underline{z}_{(n)} \text{ and } \underline{S}_n = \underline{s}_n \mid \underline{\mathcal{A}}_N\}$$

$$= e^{-\alpha(s_n) z_{(n)}} \prod_{j=1}^{n} y_j e^{-z_{(j)} y_j}\, dz_{(j)}, \qquad z_{(1)} < z_{(2)} < \cdots < z_{(n)}. \tag{3.1}$$

Equivalently, with $b_j = y_j + \cdots + y_n$, $t_j = z_{(j)} - z_{(j-1)}$, and $\underline{t}_n = (t_1, \ldots, t_n)$,

$$Prob\{\underline{T}_n \epsilon\, d\underline{t}_n \text{ and } \underline{Y}_n = \underline{y}_n \mid \underline{\mathcal{A}}_N\}$$

$$= \prod_{j=1}^{n} y_j e^{-(b_j + \alpha(s_n)) t_j}\, dt_j, \qquad t_1, \ldots, t_n > 0. \tag{3.2}$$

Likelihood functions formed from (3.1) or (3.2) for purposes of inference about unobserved elements U/s_n of U (or about $\alpha(s_n) \equiv r_N - \sum_{j \epsilon s_n} a_j$) are not regular (see Kaufman (1989)) and so are not useful for maximum likelihood estimation. For example, (3.1) as a function of $\alpha(s_n) \equiv \alpha$ is proportional to $\exp\{-\alpha z_{(n)}\}$ and a maximizer of this function over $\alpha \epsilon (0, \infty)$ occurs at $\alpha = 0$, irrespective of the observed sample.

If we condition on either the ordered sample or on the unordered sample, it is possible to deduce a regular likelihood function with a corresponding efficient score function whose expectation is zero.

3.1 Ordered Samples

To this end with $\underline{T}_n = (T_1, T_2, \ldots, T_n)$ consider first

$$Prob\{\underline{T}_n \epsilon\, d\underline{t}_n \mid \underline{Y}_n = \underline{y}_n; \underline{\mathcal{A}}_N\} = \prod_{j=1}^{n} (b_j + \alpha(s_n)) e^{-(b_j + \alpha(s_n)) t_j}\, dt_j. \tag{3.3}$$

Conditional on $\underline{Y}_n = \underline{y}_n$ or equivalently on $\underline{S}_n = \underline{s}_n$, elements of U/s_n are fixed and the closure of successive sampling under conditioning leads to a likelihood function

$$\log \mathcal{L}(\alpha \mid \underline{t}_n, \underline{y}_n) = \sum_{j=1}^{n} -t_j[b_j + \alpha] + \log[b_j + \alpha] \tag{3.4}$$

and corresponding efficient score function

$$\frac{d}{d\alpha}\log\mathcal{L}(\alpha \mid \underline{t}_n, \underline{y}_n) = -z_{(n)} + \sum_{j=1}^{n} \frac{1}{b_j + \alpha}. \tag{3.5}$$

Conditional on $\underline{S}_n = \underline{s}_n$, $E(Z_{(n)} \mid \underline{s}_n; \alpha) = \sum_{j=1}^{n} [b_j + \alpha]^{-1}$, so (3.5) is representable as

$$\frac{d}{d\alpha}\log\mathcal{L}(\alpha \mid z_{(n)}; \underline{s}_n) = -z_{(n)} + E(Z_{(n)} \mid \underline{s}_n; \alpha) \tag{3.6}$$

and a conditional MLE (CMLE) of $\alpha(s_n)$ must be a solution of

$$z_{(n)} = E(Z_{(n)} \mid \underline{s}_n; \alpha). \tag{3.7}$$

The likelihood function (3.4) is regular as

$$\frac{d}{d\alpha}E_{Z_{(n)}|s_n}\log\mathcal{L}(\alpha \mid Z_{(n)}; \underline{s}_n) = E_{Z_{(n)}|s_n}\frac{d}{d\alpha}\log\mathcal{L}(\alpha \mid Z_{(n)}, \underline{s}_n) = 0. \tag{3.8}$$

In addition

$$\frac{d^2}{d\alpha^2}\log\mathcal{L}(\alpha \mid z_{(n)}; \underline{s}_n) = -\sum_{j=1}^{n} \frac{1}{[b_j + \alpha]^2} = -\mathrm{Var}(Z_{(n)} \mid \underline{s}_n; \alpha) < 0 \tag{3.9}$$

for all $\alpha > 0$, so $\log\mathcal{L}$ is concave on $(0, \infty)$.

It is instructive to examine the special case $a_j = a > 0$, $j = 1, 2, \ldots, N$. Then (3.5) becomes $-z_{(n)} + \frac{1}{a}\sum_{j=1}^{n}[N - j + 1]^{-1}$, a familiar equation appearing in many studies of software reliability via the assumption that all faults are of equal magnitude.

For finite samples, the score function (3.7) may not possess a zero in $(0, \infty)$. As $E(Z_{(n)} \mid \underline{s}_n; \alpha)$ is monotone decreasing from $\sum_{j=1}^{n} b_j^{-1}$ at $\alpha = 0$ to zero as $\alpha \to \infty$, $\sum_{j=1}^{n} b_j^{-1} > z_{(n)}$ is both necessary and sufficient for existence of a unique $\alpha(z_{(n)}; \underline{s}_n) \epsilon (0, \infty)$ such that $z_{(n)} - E(Z_{(n)} \mid \underline{s}_n; \alpha(z_{(n)}; \underline{s}_n)) = 0$. Namely,

Proposition 3.1: A unique zero in $(0, \infty)$ of (3.6) exists iff $\sum_{j=1}^{n} b_j^{-1} > z_{(n)}$.

While use of (3.6) to define an estimator of $\alpha(s_n)$ may produce desultory results for small examples (unacceptably large probability of no solution), in the uniform asymptotic regime dictated by (A_2) a unique zero of (3.7) exists with probability one.

A first consequence of (A_1) and (A_2) is that $\alpha(s_n) = O(N)$. A second is that

$$-\epsilon \log(1 - f_n) < \sum_{j=1}^{Nf_n} \frac{1}{\alpha(s_n) + b_j} < -\frac{1}{\epsilon} \log(1 - f_n), \tag{3.10}$$

and for some positive $\theta_N = 0(1)$,

$$Var(Z_{(n)}|\underline{s}_n) = \sum_{j=1}^{Nf_n} \frac{1}{(\alpha(s_n) + b_j)^2} = \frac{f_n}{N\theta_N(1 - f_n)} = 0(N^{-1}). \tag{3.11}$$

Consequently, by Chebychev's inequality, for each possible sequence of realizations $\underline{s}_n$ of $\underline{S}_n, n = 1, 2, \ldots$, the rv $Z_{(n)} \mid \underline{s}_n$ converges in probability to $\lim_{N\to\infty} \sum_{j=1}^{Nf_n} [\alpha(s_n)+b_j]^{-1} = 0(1)$. We shall examine the limiting behavior of $E(Z_{(n)} \mid \underline{s}_n; \alpha(s_n))$ in more detail later, but for now the facts that $\alpha = O(N)$, $E(Z_{(n)} \mid \underline{s}_n; \alpha(s_n)) = 0(1)$, and $Var(Z_{(n)}|\underline{s}_n) = 0(N^{-1})$ for each possible $\underline{S}_n = \underline{s}_n$ suffice to guarantee existence of a unique solution to (3.7) with probability one as $N \to \infty$.

Since $\alpha(s_n) = 0(N)$ by assumption, it is convenient to define $\xi_N = \alpha(s_n)/N$ and a corresponding estimator $\hat{\xi}(Z_{(n)}; \underline{S}_n)$ of ξ_N.

Proposition 3.2: As $N \to \infty$ $z_{(n)} = E(Z_{(n)}|\underline{s}_n; N\xi_N)$ possesses a unique solution interior to $(0, \infty)$ with probability one.

Proof: In terms of ξ_N,

$$E(Z_{(n)} \mid \underline{s}_n; N\xi_N) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\xi_N + \frac{b_j}{N}} = 0(1). \tag{3.12}$$

Since

$$\sum_{j=1}^{n} \frac{1}{b_j} > \epsilon \sum_{j=1}^{n} \frac{1}{j} = \epsilon\left[\gamma + \log n + 0\left(n^{-1}\right)\right] \tag{3.13}$$

(γ = Euler's constant), $\sum_{j=1}^{n} b_j^{-1}$ diverges logarithmically as $N \to \infty$. As $Z_{(n)} \mid \underline{s}_n$ converges in probability to an atom of order one for each realizable sequence $\underline{s}_n$, $n = 1, 2, \ldots$, the event $\sum_{j=1}^{n} b_j^{-1} > Z_{(n)}$ conditional on $\underline{S}_n = \underline{s}_n$ obtains with probability one as $N \to \infty$. []

Proposition 3.3: The estimator $\widehat{\alpha}(Z_{(n)}, \underline{S}_n)$ defined as zero if $\sum_{j=1}^{n} (Y_j + \ldots + Y_n)^{-1} \leq Z_{(n)}$ and a solution to (3.5) if $\sum_{j=1}^{n} (Y_j + \ldots + Y_n)^{-1} > Z_{(n)}$ is positively biased.

Proof: Temporarily let $\hat{\alpha}(z, \underline{s}_n) = \hat{\alpha}(z)$ for fixed $\underline{s}_n$. As $\widehat{\alpha}(z) = 0$ for $z > \sum_{j=1}^{n} b_j^{-1} \equiv \beta_n$ and (3.7) obtains for $0 \leq z \leq \beta_n$, defining $H_n(z) = P\{Z_{(n)} \leq z | \underline{s}_n\}$ for $0 \leq z < \infty$,

$$E_{Z_{(n)}|\underline{s}_n} \sum_{j=1}^{n} \frac{1}{\widehat{\alpha}(Z_{(n)}) + b_j} = \beta_n P\{Z_{(n)} > \beta_n | \underline{s}_n\} + \int_0^{\beta_n} z dH_n(z). \tag{3.14}$$

Since by definition,

$$E(Z_{(n)}|\underline{s}_n; \alpha) = \sum_{j=1}^{n} \frac{1}{\alpha + b_j} = \int_0^{\infty} z dH_n(z), \tag{3.15}$$

we have

$$E_{Z_{(n)}|\underline{s}_n} \sum_{j=1}^{n} \frac{1}{\widehat{\alpha}(Z_{(n)}) + b_j} - \sum_{j=1}^{n} \frac{1}{\alpha + b_j} = \int_{\beta_n}^{\infty} [\beta_n - z] \, dH_n(z) < 0. \tag{3.16}$$

In addition, as $[\alpha + b_j]^{-1}$ is strictly convex in α on $[0, \infty)$, Jensen's inequality gives

$$\sum_{j=1}^{n} \Big[E_{Z_{(n)}|\underline{s}_n} (\widehat{\alpha}(Z_{(n)})) + b_j \Big]^{-1} < E_{Z_{(n)}|\underline{s}_n} \sum_{j=1}^{n} \frac{1}{\widehat{\alpha}(Z_{(n)}) + b_j} \tag{3.17}$$

and (3.16) and (3.17) together imply that

$$\sum_{j=1}^{n} \Big[E_{Z_{(n)}|\underline{s}_n} (\widehat{\alpha}(Z_{(n)}) + b_j \Big]^{-1} < \sum_{j=1}^{n} [\alpha + b_j]^{-1}. \tag{3.18}$$

Thus $E_{Z_{(n)}|\underline{s}_n}(\widehat{\alpha}(Z_{(n)}) > \alpha$. As $E_{Z_{(n)}|\underline{s}_n}(\widehat{\alpha}(Z_{(n)}, \underline{s}_n)) > \alpha$ for each possible $\underline{S}_n = \underline{s}_n$, $E_{Z_{(n)}, \underline{S}_n}(\widehat{\alpha}(Z_{(n)}, \underline{S}_n)) > \alpha$ as was to be proved. []

For small samples the bias of $\widehat{\alpha}(Z_{(n)}; s_n)$ can be severe; in the example of Section 5 $\underline{\mathcal{A}}_N = (1, 2, \ldots, 10)$, $n = 4$, and the bias in estimation of the sum of population magnitudes $r_N = 55$ computed by adding $\sum_{j \epsilon s_n} Y_j$ to $\widehat{\alpha}(Z_{(n)}, S_n)$ is about 26 %. We are led to consider unbiased estimators and do so in Section 4.

3.2 Unordered Samples

Consider an observational process that reveals magnitudes of the first n elements of U generated by successive sampling as defined by (3.1) without regard to order **and** the waiting time to observation of these n magnitude. That is, $Z_{(n)} = z_{(n)}$ and $S_n = s_n$ (or equivalently $\{a_j \mid j \epsilon s_n\}$) are observed.

Then

$$\begin{aligned} &Prob\{Z_{(n)} \epsilon\, dz \text{ and } S_n = s_n \mid \underline{\mathcal{A}}_N\} \\ &= D_n(z \mid s_n; \alpha(s_n))dz \times P_n(s_n \mid \alpha(s_n)) \propto e^{-\alpha(s_n)z}dz. \end{aligned} \tag{3.19}$$

As in the case of an ordered sample of magnitudes, the likelihood function for α suggested by (3.19) is not regular. However, the likelihood function for α corresponding to

$$Prob\{Z_{(n)} \epsilon\, dz \mid s_n; \underline{\mathcal{A}}_N\} = D_n(z \mid s_n; \alpha(s_n)) \propto \frac{e^{-\alpha(s_n)z}dz}{P_n(s_n \mid \alpha(s_n))}$$

is regular (Kaufman (1989)) and the efficient score function for

$$\log \mathcal{L}(\alpha \mid z_{(n)}; s_n) \equiv -\alpha z_{(n)} - \log P_n(s_n \mid \alpha) \tag{3.20}$$

using (2.9) is

$$\frac{d}{d\alpha} \log \mathcal{L}(\alpha \mid z_{(n)}; s_n) = -z_{(n)} + E(Z_{(n)} \mid s_n; \alpha). \tag{3.21}$$

In turn,

$$\frac{d^2}{d\alpha^2} \log \mathcal{L}(\alpha \mid z_{(n)}, s_n) = -\text{ Var } (Z_{(n)} \mid s_n; \alpha) < 0$$

for $0 \leq \alpha < \infty$, so if

$$E(Z_{(n)} \mid s_n; \alpha) - z_{(n)} = 0 \tag{3.22}$$

possesses a solution it must be unique. The efficient score function (3.9) is of the same form as that for the ordered sample when all of $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$ (or equivalently $\underline{T}_n$) as well as magnitudes $Y_1, Y_2, \ldots, Y_n$ in order of occurrence are observed. The only difference is that the expectation of $Z_{(n)}$ conditional on $S_n = s_n$, the **unordered** sample, replaces the expectation of $Z_{(n)}$ conditional on the **ordered** sample $\underline{S}_n = \underline{s}_n$.

We may interpret $Var(Z_{(n)}|s_n; \alpha)$ as expected Fisher information with respect to a measure (the density of $Z_{(n)}$) conditional on $S_n = s_n$. Since

$$Var(Z_{(n)} \mid s_n; \alpha) = E_{\underline{S}_n} \mid s_n Var(Z_{(n)} \mid \underline{S}_n; \alpha) + Var E_{\underline{S}_n} \mid s_n (Z_{(n)} \mid \underline{S}_n; \alpha), \tag{3.23}$$

we have that

$$E_{\underline{S}_n|s_n} Var(Z_{(n)} \mid \underline{S}_n; \alpha) < Var(Z_{(n)} \mid s_n; \alpha) \tag{3.24}$$

and inference based on observation of $(Z_{(n)}, S_n)$ is more efficient in the sense of (3.17) than inference based on observation of $(\underline{T}_n, \underline{Y}_n)$ or equivalently of $(\underline{T}_n, \underline{S}_n)$. One possible explanation is that unordering of the sample as suggested by (3.14) is a form of Rao–Blackwellization.

Paralleling our treatment of the case when $Z_{(n)} = z_{(n)}$ and $\underline{S}_n = \underline{s}_n$ are observed, for large N we define an estimator $\widehat{\xi}(Z_{(n)}, S_n) = \widehat{\alpha}(Z_{(n)}, S_n)/N$.

Proposition 3.5: As $N \to \infty$, $z_{(n)} = E(Z_{(n)}|s_n; N\xi_N)$ possesses a unique solution $\widehat{\xi}(z_{(n)}, s_n)$ interior to $(0, \infty)$ iff

$$\int_0^\infty \left\{1 - \prod_{j \epsilon s_n} \left(1 - e^{-a_j x}\right)\right\} dx > z_{(n)} \tag{3.25}$$

and as $N \to \infty$, (3.25) obtains with probability one.

Proof: The **LHS** of (3.25) is $E(Z_{(n)}|s_n; 0)$. As $\alpha = N\xi_N \to \infty$, with s_n fixed, $E(Z_{(n)} \mid s_n; \alpha) \longrightarrow 0$. Since

$$\int_0^\infty \left\{1 - \prod_{j \epsilon s_n} \left(1 - e^{-a_j x}\right)\right\} dx = 0(\log n), \tag{3.26}$$

the **LHS** diverges logarithmically as $N \to \infty$. Each realizable $S_n = s_n$ is a probability mixture of $n!$ events of the form $\underline{S}_n = \underline{s}_n$, so as $N \to \infty$ with $n/N \to f \epsilon (0, 1)$ and f bounded away from 0 or 1, $\lim_{N \to \infty} E(Z_{(n)} \mid \underline{s}_n; \alpha(s_n)) = 0(1)$ and this implies that $\lim_{N \to \infty} E(Z_{(n)} \mid s_n; \alpha(s_n)) = 0(1)$. []

4. Unbiased Estimation

Unbiased estimators of functions of magnitudes of unobserved elements U/s_n of U are intimately linked to **CMLE**'s of such functions. The linkage appears most clearly when the form of a **CMLE** for $\alpha(s_n)$ is studied in the asymptotic regime $n/N \to f\epsilon(0,1)$, f fixed and bounded away from zero and one, as $N \to \infty$. In this regime the **CMLE** is structurally similar to Murthy's (1956) and Horvitz and Thompson's (1955) unbiased estimators of finite population parameters. Bickel, Nair and Wang (1989) develop a precise theory of large sample **MLE** for successive sampling that establishes this connection when successive sampling of a fixed, finite number of magnitude classes is performed and magnitudes alone, not waiting times, are observed.

When both waiting times $\underline{Z}_n = \underline{z}_n$ and magnitudes $\underline{Y}_n = \underline{y}_n$ are observed in a sample $s_n \subset U_N$ of size n, it would seem natural to use $\hat{\pi}_k(z_{(n)}) \equiv 1 - exp\{-z_{(n)}a_k\}$, $k\epsilon s_n$, as an approximation to the unconditional probability $Prob\{k\epsilon S_n\} \equiv \pi_k(n)$ that k is in the sample and then form a Horvitz-Thompson type estimator of some property of $\mathcal{A}_N$ with the $\hat{\pi}_k(z_{(n)})s$. The resulting estimator is biased! However, if a value $z_{(n+1)}$, of the waiting time to the $(n+1)st$ observation is employed to approximate $\pi_k(n)$ as $1-exp\{-z_{(n+1)}a_k\}$, the resulting estimator is unbiased. Since $Z_{(n+1)}$ is not observed in a sample of size n, this approximation to $\pi_k(n)$ requires that the last observation in $\underline{Y}_n = \underline{y}_n$ be ignored and the sample treated as of size $n-1$. Alternatively, a correction for the (positive) bias created by use of $z_{(n)}$ in concert with $\{y_j \mid j\epsilon s_n\}$ can be computed.

Theorem 4.1: For $k \epsilon S_n$, $E_{S_n}E\{[1-\exp\{-a_kZ_{(n+1)}\}]^{-1} \mid S_n\} = \frac{1}{\pi_k(n)}$.

Proof: With $\alpha = \alpha(s_n)$, $\Pi_n(y) = \prod_{j\epsilon s_n}(1-e^{-a_j y})$ and $P(s_n|k^{\#};\alpha) \equiv$ Prob$\{S_n = s_n|k$ first in $s_n\}$,

$$E_{Z_{(n+1)}|\underline{s}_n}\left(\frac{1}{1-e^{-a_kZ_{(n+1)}}}\right) = \int_0^\infty \frac{\alpha e^{-\alpha y}}{1-e^{-a_k y}}\Pi_n(y)dy/P(s_n \mid \alpha) = \frac{P(s_n \mid k^{\#};\alpha)}{P(s_n \mid \alpha)}. \tag{4.1}$$

The right–hand side of (4.1) is a version of Murthy's estimator which has been shown by Andreatta and Kaufman (1986) to have expectation $1/\pi_k(n)$. []

Corollary 4.1: Let h be a single valued function with domain $\mathcal{A}_N$ and range $(-\infty, \infty)$ or some subset of $(-\infty, \infty)$. Define $H = \sum_{j \epsilon \mathcal{A}_N} h(a_j)$. Then when $S_n = s_n$ and $Z_{(n+1)} = z_{(n+1)}$ are observed

$$\sum_{k \epsilon s_n} h(a_k)/(1 - \exp\{-a_k z_{(n+1)}\}) \equiv \widehat{H}(s_n; z_{(n+1)}) \tag{4.2}$$

is an unbiased estimator of H.

Proof: As (4.2) has expectation equal to Murthy's estimator conditional on $S_n = s_n$ and Murthy's estimator is an unbiased estimator, (4.2) is unbiased.[]

The estimator $\widehat{H}$ defined in (4.2) is a function of **both** s_n and $z_{(n+1)}$ and is an **unordered** function of s_n. When all a_k's are identical and equal to a and h is chosen to be identically one for each $k \epsilon s_n$, $\widehat{H} = n/(1 - \exp\{-a z_{(n+1)}\})$, a familiar **MLE** for N conditional on knowledge of the parameter $p \equiv 1 - exp\{-\alpha z_{(n+1)}\}$. This is a special case of Chapman's (1951) estimator for the number of trials of a binomial probability function.

When $(s_n, z_{(n)})$ is observed but $z_{(n+1)}$ is not, the bias of an unbiased estimator based on $1 - exp\{-z_{(n)} a_k\}$ as an approximation to $\pi_k(n)$ can be unbiasedly estimated. The ensuing correction for bias takes the following form:

Corollary 4.2: An unbiased estimator of H is

$$\widehat{H}(s_n; z_{(n)}) = \sum_{k \epsilon s_n} \left[\frac{1 - p_k(s_n, z_{(n)})}{1 - e^{-z(n) a_k}} \right] h(a_k) \tag{4.3}$$

with

$$p_k(s_n, z_{(n)}) = \frac{a_k e^{-a_k z_{(n)}}/(1 - e^{-a_k z_{(n)}})}{\sum_{j \epsilon s_n} a_j e^{-a_j z_{(n)}}/(1 - e^{-a_j z_{(n)}})}. \tag{4.4}$$

Proof: With $\alpha = \alpha(s_n)$,

$$E_{Z_{(n)}|\underline{s}_n}(1-e^{-a_k Z_{(n)}})^{-1} = \int_0^\infty \frac{e^{-\alpha x}}{1-e^{-a_k x}} d\Pi_n(x)/P(s_n \mid x)$$

$$= \frac{1}{P(s_n \mid \alpha)}\left\{\int_0^\infty e^{-\alpha x}\left[\sum_{\substack{j\epsilon s_n \\ j\neq k}} \frac{a_j e^{-a_j x}}{(1-e^{-a_j x})}\right]\Pi_{\substack{j\epsilon s_n \\ j\neq k}}(1-e^{-a_j x})dx\right.$$

$$\left. + \int_0^\infty e^{-\alpha x}\left[\frac{a_k e^{-a_k x}}{(1-e^{-a_k x})}\right]\Pi_{\substack{j\epsilon s_n \\ j\neq k}}(1-e^{-a_j x})dx\right\}$$

$$= \frac{P(s_n \mid k^{\#};\alpha)}{P(s_n \mid \alpha)} + \frac{1}{P(s_n \mid \alpha)}\int_0^\infty e^{-\alpha x}\left[\frac{a_k e^{a_k x}}{(1-e^{-a_k x})^2}\right]\Pi_{j\epsilon s_n}(1-e^{-a_j x})dx \qquad (4.5)$$

Dividing and multiplying the integrand of the integral in (4.5) by $\sum_{j\epsilon s_n} a_j e^{-a_j x}/(1-e^{-a_j x})$ and summing (4.5) over $k\epsilon s_n$,

$$\sum_{k\epsilon s_n} E_{Z_{(n)}|\underline{s}_n}(1-e^{-a_k Z_{(n)}})^{-1} \qquad (4.6)$$

$$= \sum_{k\epsilon s_n}\frac{P(s_n \mid \alpha; k^{\#})}{P(s_n \mid \alpha)} + \int_0^\infty\left[\sum_{k\epsilon s_n}\frac{p_k(s_n,z)}{1-e^{-a_k z}}\right]D_n(z \mid s_n;\alpha)dz.$$

The first term on the RHS of (4.6) is Murthy's unbiased estimator of N; the second term is

$$E_{Z_{(n)}|\underline{s}_n}\sum_{k\epsilon s_n}\frac{p_k(s_n;Z_{(n)})}{1-e^{-a_k Z_{(n)}}}, \qquad (4.7)$$

so with $h(a_k) \equiv 1$, (4.3) is an unbiased estimator of N. The modifications necessary for more general choice of h are obvious.[]

4.1 Asymptotic Equivalence

To set the stage for establishing the correspondence between unbiased estimation and CMLE as $N \to \infty$, examine the score function for the log likelihood when $Z_{(n+1)} = z_{(n+1)}$ and $S_n = s_n$ is observed in place of $Z_{(n)} = z_{(n)}$ and $S_n = s_n$. From (2.7) and (2.9)

$$\frac{d}{d\alpha}\log\mathcal{L}(\alpha \mid z_{(n+1)};\underline{s}_n) = -z_{(n+1)} + E(Z_{(n+1)} \mid \underline{s}_n;\alpha) \qquad (4.8)$$

$$= -z_{(n+1)} + \frac{1}{\alpha} + E(Z_{(n)} \mid \underline{s}_n;\alpha).$$

As stated above Bickel, Nair and Wang (1990) show that when a fixed finite number of magnitude classes are successively sampled, as $N \to \infty$, unbiased estimators similar in form to (4.2) constitutes an asymptotically efficient approximation to **MLE**. In the present setting we have

Theorem 4.2: As $N \to \infty$, **CMLE** of $\alpha(s_n)$ based on observation of $(z_{(n+1)}, s_n)$ is asymptotically equivalent to unbiased estimation of $\alpha(s_n)$ using the estimator defined in Corollary 4.1.

Proof: As $N \to \infty$, $E(Z_{(n+1)} \mid s_n; \alpha) \sim w_n(\alpha)$ where $w_n(\alpha)$ satisfies

$$\sum_{j \epsilon s_n} \frac{a_j e^{-a_j w_n(\alpha)}}{1 - e^{-a_j w_n(\alpha)}} = \alpha, \tag{4.9}$$

(Andreatta and Kaufman (1986)). Thus when $Z_{(n+1)} = z_{(n+1)}$, and $S_n = s_n$ are observed, $w_n(\alpha) = z_{(n+1)}$ is a large N solution to the efficient score function for a **CMLE** of $\alpha(S_n)$.

Using (4.1) and (4.2) with $\alpha = \alpha(s_n)$,

$$E_{Z_{(n+1)}|s_n}\left\{\sum_{k \epsilon s_n} \frac{a_k}{1 - e^{-a_k Z_{(n+1)}}}\right\} = \sum_{k \epsilon s_n} \frac{a_k P(s_n|k^{\#}; \alpha)}{P(s_n|\alpha)}, \tag{4.10}$$

Murthy's unbiased estimator of $r_N \equiv \sum_{k=1}^{N} a_k$. Since

$$E_{S_n}\left[\sum_{k \epsilon S_n} \frac{a_k P(S_n|k^{\#}; \alpha)}{P(S_n|\alpha)}\right] = r_N \tag{4.11}$$

and

$$\frac{P(s_n|k^{\#}; \alpha)}{P(s_n|\alpha)} \sim \frac{1}{1 - e^{-w_n(\alpha) a_k}} \tag{4.12}$$

as $N \to \infty$ (op. cit. Theorem 5.1), **CMLE** of $\alpha(s_n)$ as $N \to \infty$ using $w_n(\alpha) = z_{(n+1)}$ yields

$$\sum_{j \epsilon s_n} \frac{a_j e^{-z_{(n+1)} a_j}}{1 - e^{-z_{(n+1)} a_j}} = \alpha(z_{(n+1)}, s_n) \tag{4.13}$$

or equivalently

$$\sum_{j \epsilon s_n} \frac{a_j}{1 - e^{-z_{(n+1)} a_j}} = \widehat{\alpha}(z_{(n+1)}, s_n) + \sum_{j \epsilon s_n} a_j \equiv \widehat{r}_N(z_{(n+1)}, s_n). \tag{4.14}$$

The estimator $\widehat{r}_N(z_{(n+1)}, s_n)$ of r_N is of precisely the same functional form as the unbiased estimator presented in Corollary 4.1 with $h(a_k) = a_k$, $k \epsilon s_n$. []

5. Numerical Study

To illustrate small sample behavior of the three types of estimators presented here, a monte carlo study of a successive sampling example from Hájek (1981) was done. In this example, a successive sample of size n = 4 is taken from $\mathcal{A}_{10} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The range 1 to 10 of A_k values is small by comparison with applications of successive sampling to oil and gas discovery in which the largest element of $\mathcal{A}_N$ can be as large as 10^3 times the smallest. [The North Sea oil and gas data taken from O'Carroll and Smith (1980) by Bickel, Nair and Wang (1989) varies from 12 to 3017 million barrels of recoverable oil equivalent]. The small range of values of $\mathcal{A}_{10}$ in Hájek's example provides a stringent test of the performance of successive sampling estimators as the expectation of Y_1 is only about 3.2 times larger than that of Y_{10}.

A computational scheme for computing inclusion probabilities based on an integral representation of $\pi_k(n)$ was used to compute Table 5.1. [Andreatta and Kaufman (1990)] It shows inclusion probabilities $\pi_k(n)$ for sample sizes $n = 1, 2, \ldots, 6$ from $\mathcal{A}_{10}$.

Table 5.1 here

Figure 5.1 is a graphical display of estimators examined in the study.

Figure 5.1 here

Results of a monte carlo study of properties of three of the four types of estimators studied here appear in Table 5.2. Means, and standard deviations of unbiased, corrected unbiased and maximum likelihood estimators for each of $r_N, \alpha(s_n)$ and N are based on 4,000 successive samples drawn from $\mathcal{A}_{10}$.

Table 5.2 here

Monte carloed properties of these estimators for $n = 4$ in Hájek's example are:

(1) Unbiased estimation based on observation of s_n and $z_{(n+1)}$ leads to estimators of r_N, $\alpha(s_n)$ and N with smaller standard deviation than that for corrected unbiased estimation based on observation of s_n and $z_{(n)}$. For example, while both unbiased and corrected unbiased estimators of r_N have monte carloed means within .43% of the true value 1.000 of r_N; the standard deviation of the former is .448 and of the latter is .514.

(2) An ordered estimator of r_N that is a solution of (3.7) based on (3.6) and observation of $\underline{s}_n$ and $z_{(n)}$ is positively biased – as proven in Proposition 3.3. The bias is about 26%. An estimator of r_N based on replacing $z_{(n)}$ with $z_{(n+1)}$ in (4.2) is negatively biased by about 5.5-7.5%. For this example, ordered estimation of $\alpha(s_n)$ led to $\hat{\alpha}(s_n) = 0$ in 32 out of 4000 cases using s_n and $z_{(n)}$ and 89 out of

4000 cases using s_n and $z_{(n+1)}$.

The behavior of these estimators for moderate sized samples is explored next in a Monte Carlo study of samples of size $n = 10$ from a population of size $N = 30$ with $a_k =$ magnitude of the $k/(N+1)st$ fractile of an exponential distribution having mean one, $k = 1, 2, \ldots, N$.

Table 5.3 here

Monte Carloed properties of estimators for this second example are:

(1) Unbiased estimation of r_N outperforms all other estimators studied here, producing smaller standard deviation and less sampling bias (about .4%) . Unbiased estimation of the successive sampling remainder yielded Monte Carlo averages $\hat{\alpha} + b_1 = .4767 + .5273 = 1.004$.

(2) An ordered estimator of α was produced for each of 1000 trials, suggesting that the probability of no solution to (3.7) $[\sum_{j=1}^{n} b_j^{-1} \leq z_{(n)}]$ is less than .001 for this example. Use of $z_{(n)}$ produced an ordered estimator with about 2% negative bias.

(3) While unbiased estimation of N over 1000 trials exhibits small bias (about .7%) the standard deviation of an estimator of N of the form (4.3) is quite large – about one-half of the true value of $N = 30$.

Prior to observing $S_n = s_n$, $\alpha(s_n) = r_n - \sum_{j \epsilon s_n} a_j$ is a rv so it is worthwhile documenting the behavior of $\hat{\alpha}(S_n) - \alpha(S_n)$, the difference between the successive sampling remainder $\alpha(S_n)$ and its estimator. To this end a decomposition of mean squared error $E_{S_n}(\hat{\alpha}(S_n) - \alpha(S_n))^2$ into its variance components $Var(\hat{\alpha}(S_n))$, $Var(\alpha(S_n))$, $Cov(\hat{\alpha}(S_n), \alpha(S_n))$ is presented in Table 5.4. The variance components just cited are dictated by formula (5.1). If X is a rv and $\hat{X}$ is a predictor of X then the mean squared error of $\hat{X}$ is

$$MSE \equiv Var(\hat{X} - X) = Var(\hat{X}) - 2Cov(\hat{X}, X) + Var(X). \tag{5.1}$$

Table 5.4 here

In Table 5.4 we see that while the bias of $\hat{\alpha}(S_n)$ is small in both examples, $Var(\hat{\alpha}(S_n) - \alpha(S_n))$ is a very large fraction of $Var(\hat{\alpha}(S_n))$.

6. Prediction of Test Effort as a Function of Number of Failures

As stated at the outset, given a history of the test process up to the time of the n^{th} failure, an estimate of the incremental time on test to the occurrence of the next $n-m$ failures is of considerable practical value to the software manager who wishes to predict testing effort as a function of number of failures.

Our objective is to provide an estimator for $Z_{(n)}$ given $Z_{(m)} = z_{(m)}$ and $S_m = s_m$ that is both easy to compute and behaves reasonably for moderate samples. Recall that with $w_n(\alpha) = z_{(n+1)}$ (4.8) yields an estimate of $\alpha(s_n)$ based on observation of $Z_{(n+1)}$ and $S_{(n)}$ that is both asymptotically CMLE, unbiased, and in addition is coincident with unbiased estimation of the sum $H = \sum_{j \epsilon A_N} h(j)$ of any single valued function $h(\cdot)$ of the $a_j s$ (See Proposition 4.1).

These facts together with (2.11) suggest a point estimator for $Z_{(n)}$, $n > m+1$, given observation of $Z_{(m+1)} = z_{(m+1)}$ and $S_m = s_m$ that is a solution $\hat{z}_{(n)}$ to

$$\sum_{j \epsilon s_m} \frac{1 - e^{-z y_j}}{1 - e^{-z_{(m+1)} y_j}} = n \tag{6.1}$$

for $z \epsilon (0, \infty)$. Specific motivation for the form of this estimator is the identity $\sum_{k=1}^{N} \pi_k(n) = n$. If we are given only inclusion probabilities $\pi_k(n)$ and $\pi_k(m), k \epsilon s_m$, then

$$\sum_{j \epsilon s_m} \frac{\pi_k(n)}{\pi_k(m)} \tag{6.2}$$

has expectation $\sum_{k=1}^{N} \pi_k(n) = n$ with respect to S_m. Replacing $1/\pi_k(m)$ with its unbiased point estimator $(1 - e^{-z_{(m+1)} y_i})^{-1}$ and $\pi_k(n)$ with Rosén's approximation $(1 - e^{-z(f) y_j})$, $z(f)$ satisfying (2.11), yields (6.1).

A unique solution to (6.1) in $(z_{(m+1)}, \infty)$ exists provided that n is chosen to be less than $\hat{N} = \sum_{j \epsilon s_m} [1 - exp\{-z_{(m+1)} y_j\}]^{-1}$, an unbiased estimate of the total number N of faults in the system at the outset of testing. This point estimator is a function of both $z_{(m+1)}$ and s_m, so prior to observing $Z_{(m+1)}$ and S_m, it is rv $\hat{z}_{(n)}(Z_{(m+1)}, S_m)$ which we abbreviate as $\hat{Z}_{(n)}$. Under mild conditions on the large N behavior of the set $\mathcal{A}_N$ of fault magnitudes, if $m/N = f_m \to g \epsilon (0,1)$ and $n/N = f_n \to f \epsilon (0,1)$ as in the asymptotic regime (A_2), $|\hat{Z}_{(n)} - z_{(n)}| = o_p(1)$; i.e. as $N \to \infty$, the absolute value of the difference

between $\hat{Z}_{(n)}$ and $z_{(n)}$ converges to zero in probability. In this sense, $\hat{Z}_{(n)}$ is a "consistent" estimator of the rv $Z_{(n)}$. [A proof will be provided elsewhere.]

The results of a small Monte Carlo study of $\hat{Z}_{(n)}$ given a sample of size $m < n$ and $Z_{(m)} = z_{(m)}$ - NOT $Z_{(m+1)} = z_{(m+1)}$ - suggest that the prediction scheme represented by (6.1) produces point estimates of $Z_{(n)}$ with small to moderate bias and increasing variability as $z_{(m)}$ increases.

Statistics describing $\hat{Z}_{(n)}$ for the Hájek and exponential examples of Section 5 are presented in Table 6.1 and 6.2 respectively. Table 6.1 is based on use of 4,000 replications of a successive sample of size $m = 4$ from $\mathcal{A}_{10} = \{1, 2, \ldots, 10\}$ to generate point estimates of $Z_{(7)}$ using (6.1). Table 6.2 is based on use of 1,000 replications of a successive sample of size $m = 10$ from $\mathcal{A}_{30}$ composed of 30 fractiles of a mean 1.0 exponential distribution to generate point estimates of $Z_{(20)}$ using (6.1).

Table 6.1 and Table 6.2 here

Even though both sample and population size for the exponential example are substantially larger for the exponential example than for the Hájek example, $n = 100, N = 30$ and $n = 4, N = 10$ respectively, $\hat{Z}_{(n)}$ behaves similarly in both cases:

(1) $\hat{Z}_{(n)}$ is positively skewed with extreme right tail outliers [Figures 6.1b and 6.2b]. By comparison $Z_{(n)}$ is substantially less positively skewed and, for the case $n = 10, N = 30$ the histogram of simulated values of $Z_{(n)}$ is reasonably approximated by a Normal distribution [Figures 6.1a and 6.2a]. (As $N \longrightarrow \infty$ in the regime (A_2) an appropriately scaled version of $\hat{Z}_{(n)}$ is $N(0, 1)$; see Andreatta and Kaufman (1990)).

(2) Variability of $\hat{Z}_{(n)}$ rapidly increases as $Z_{(n)}$ increases. [Scatterplots in Figures 6.3a and 6.3b].

(3) The expectation of the predictor $\hat{Z}_{(n)}$ given $Z_{(m)} = z_{(m)}$ increases faster than linearly as a function of $z_{(m)}$ and slower than exponentially. [Figures 6.4a and 6.4b].

(4) The mean of $\hat{Z}_{(7)}$ in the Hájek example is 1.6% greater than the mean of $Z_{(7)}$ [Table 6.1]. In the exponential example, right tail outliers boost the positive bias of $Z_{(20)}$ to 10.4% [Table 6.2a]. If the Monte Carlo sample is trimmed to exclude values of $\hat{Z}_{(20)}$ greater than four standard deviations above the maximum value of $\hat{Z}_{(20)}$ achieved in the sample, the bias of the trimmed sample is negligible. When the 16 of 868 such values are deleted the bias is -.03%.

7. CONCLUDING REMARKS

Successive sampling schemes and EOS models of software failures are intimately related to one another. This relation allows application of methods of inference and prediction developed for successive sampling to be applied to software failure sampling as demonstrated in this paper.

Further linkages between successive sampling schemes and Bayes/empirical Bayes treatment of software reliability models can be established by invoking a super-population process that describes how fault magnitudes are generated.

TABLE 5.1
INCLUSION PROBABILITIES $\pi_k(n)$

A_k	n = 1	n = 2	n = 3	n = 4	n = 5	n = 6
10	0.1818	0.3503	0.5039	0.6408	0.7588	0.8554
9	0.1636	0.3196	0.4663	0.6017	0.7231	0.8271
8	0.1455	0.2878	0.4258	0.5578	0.6811	0.7921
7	0.1273	0.2549	0.3824	0.5086	0.6317	0.7485
6	0.1091	0.2210	0.3360	0.4537	0.5737	0.6941
5	0.0909	0.1862	0.2867	0.3930	0.5059	0.6259
4	0.0727	0.1506	0.2346	0.3262	0.4274	0.4369
3	0.0545	0.1141	0.1798	0.2534	0.3377	0.5410
2	0.0364	0.0768	0.1223	0.1747	0.2365	0.3123
1	0.0182	0.0387	0.0624	0.0902	0.1239	0.1667

TABLE 5.2
[HAJEK EXAMPLE]

	Unbiased[1]	Corrected[2] Unbiased	$(\hat{\alpha} \geq 0)$[3] Ordered	$(\hat{\alpha} > 0)$[3] Ordered	$(\hat{\alpha} \geq 0)$[4] Ordered	$(\hat{\alpha} > 0)$[4] Ordered
$\hat{R}$	1.002 (.448)	.997 (.514)	1.255 (.586)	1.265 (.589)	.924 (.455)	.945 (.444)
$\hat{\alpha}$	.508 (.452)	.502 (.516)	.764 (.588)	.770 (.587)	.442 (.448)	.452 (.447)
$\hat{N}$	9.945 (1.716)	10.163 (1.843)	--	--	--	--

Based on observation of:

(1) s_n and $z_{(n+1)}$
(2) s_n and $z_{(n)}$
(3) s_n and $z_{(n)}$
(4) s_n and $z_{(n+1)}$

TABLE 5.3
[EXPONENTIAL EXAMPLE]

	Unbiased[1]	Corrected[2] Unbiased	Ordered[3]	Ordered[4]
$\hat{R}$	1.004 (.225)	.994 (.257)	.980 (.227)	1.087 (.284)
$\hat{\alpha}(5)$	.477 (.236)	.467 (.506)	.453 (.238)	.559 (.292)
$\hat{N}$	29.79 (15.48)	30.445 (16.186)		
$b_1 =$	.527 (.004)		--	--

(1) s_n and $z_{(n+1)}$

(2) s_n and $z_{(n)}$

(3) s_n and $z_{(n+1)}$ [All 1,000 trials produced an ordered estimate.]

(4) s_n and $z_{(n)}$ [All 1,000 trials produced an ordered estimate.]

(5) $\alpha(s_n)$ has mean .473 and standard deviation .004.

TABLE 5.4

MEAN SQUARE PREDICTION ERROR OF UNBIASED ESTIMATOR $\hat{\alpha}(S_n)$

(HAJEK'S EXAMPLE, N=10, n=4, 4000 TRIALS)

FRACTION $\hat{\alpha}(S_n) > 0$	= 1.0	MSE	= .1627
MEAN OF $\hat{\alpha}(S_n)$	= .508	VARIANCE OF $\hat{\alpha}(S_n)$	= .1676
MEAN OF $\alpha(S_n)$	= .505	VARIANCE OF $\alpha(S_n)$	= .0053
BIAS	= .003	COVARIANCE ($\hat{\alpha}(S_n)$, $\alpha(S_n)$	= .0051

(EXPONENTIAL EXAMPLE N=30, n=10, 1000 TRIALS)

FRACTION $\hat{\alpha}(S_n) > 0$	= 1.0	MSE	= .0509
MEAN OF $\hat{\alpha}(S_n)$	= .477	VARIANCE OF $\hat{\alpha}(S_n)$	= .0557
MEAN OF $\alpha(S_n)$	= .473	VARIANCE OF $\alpha(S_n)$	= .0040
BIAS	= .004	COVARIANCE ($\hat{\alpha}(S_n)$, $\alpha(S_n)$)	= .0044

TABLE 6.1

($\hat{Z}_{(7)}$ vs. $Z_{(7)}$ FOR HAJEK'S EXAMPLE, 4000 TRIALS)

FRACTION	= .85	MSE	= 104.31
MEAN OF $\hat{Z}_{(7)}$	= 12.65	VAR ($\hat{Z}$)	= 113.04
MEAN OF $Z_{(7)}$	= 12.45	VAR (Z)	= 28.47
BIAS	= .20	COV ($\hat{Z}$, Z)	= 18.59

TABLE 6.2a

[$\hat{Z}_{(7)}$ vs. $Z_{(7)}$ FOR EXPONENTIAL EXAMPLE]

1000 TRIALS

FRACTION	= .87	MSE	= 2073.19
MEAN OF $\hat{Z}_{(7)}$	= 57.41	VAR ($\hat{Z}$)	= 2273.95
MEAN OF $Z_{(7)}$	= 51.99	VAR (Z)	= 225.87
BIAS	= 5.42	COV ($\hat{Z}$, Z)	= 213.28

TABLE 6.2b

OUTLIERS >200 OMITTED

FRACTION	= .84	MSE	= 917.73
MEAN OF $\hat{Z}_{(7)}$	= 51.45	VAR ($\hat{Z}$)	= 992.50
MEAN OF $Z_{(7)}$	= 51.59	VAR (Z)	= 220.85
BIAS	= -.14	COV ($\hat{Z}$, Z)	= 137.81

FIGURE 5.1

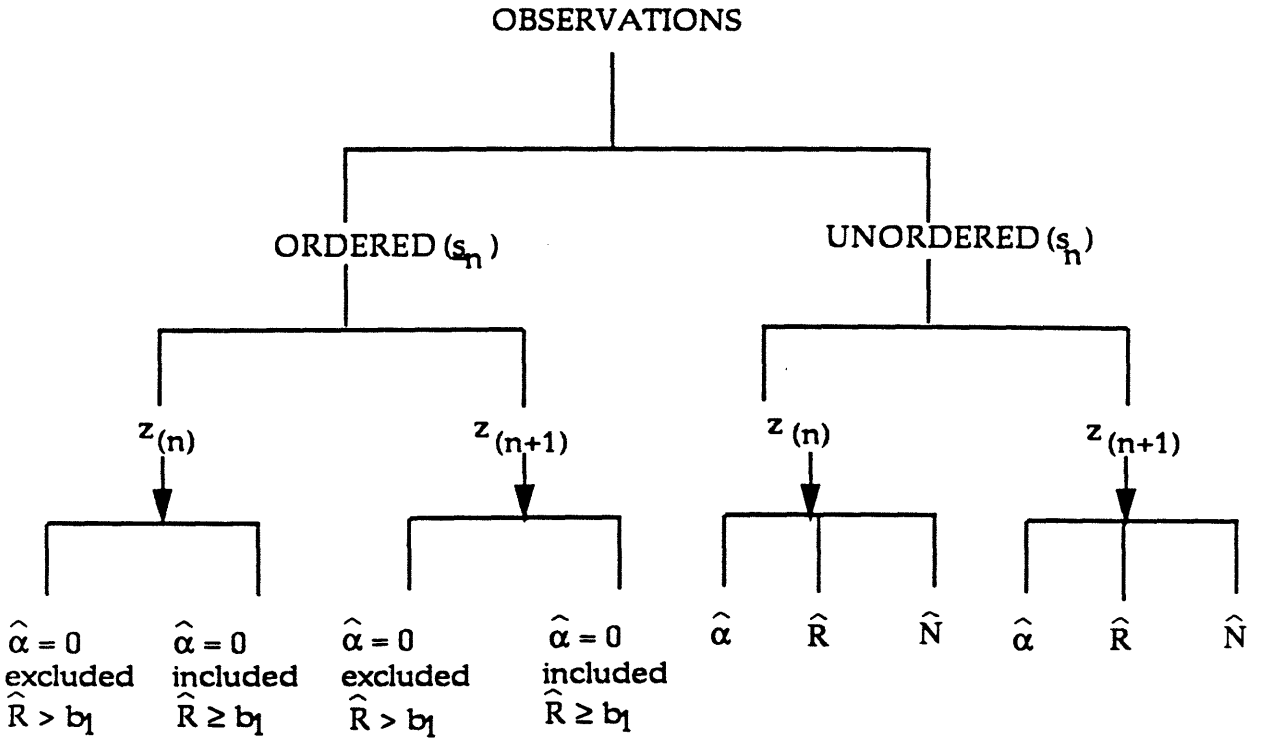

OBSERVATIONS
ORDERED (s_n)
UNORDERED (s_n)
$z_{(n)}$
$z_{(n+1)}$
$z_{(n)}$
$z_{(n+1)}$
$\hat{\alpha} = 0$ excluded $\hat{R} > b_1$
$\hat{\alpha} = 0$ included $\hat{R} \geq b_1$
$\hat{\alpha} = 0$ excluded $\hat{R} > b_1$
$\hat{\alpha} = 0$ included $\hat{R} \geq b_1$
$\hat{\alpha}$
$\hat{R}$
$\hat{N}$
$\hat{\alpha}$
$\hat{R}$
$\hat{N}$

FIGURE 6.1a
ACTUAL Z-HAJEK EXAMPLE

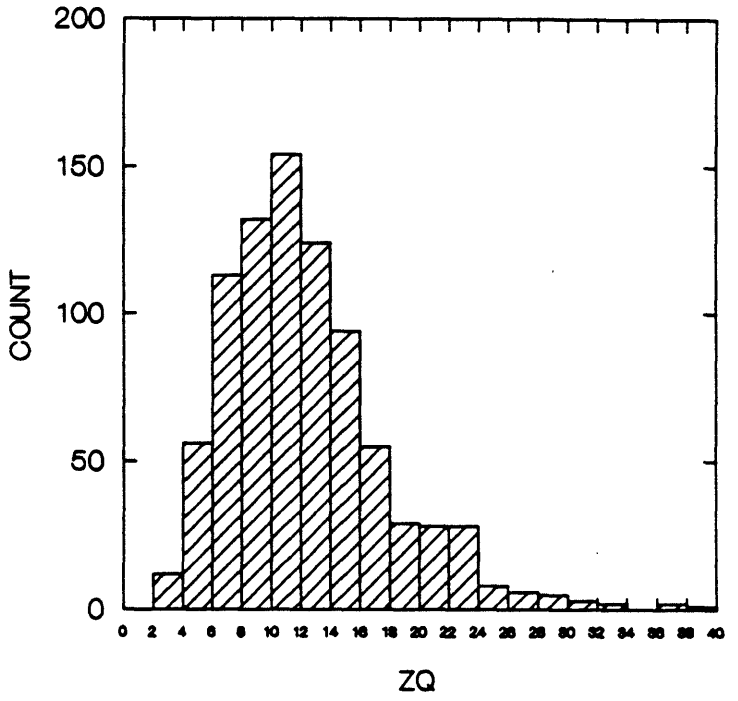


FIGURE 6.1b
ESTIMATED Z-HAJEK EXAMPLE

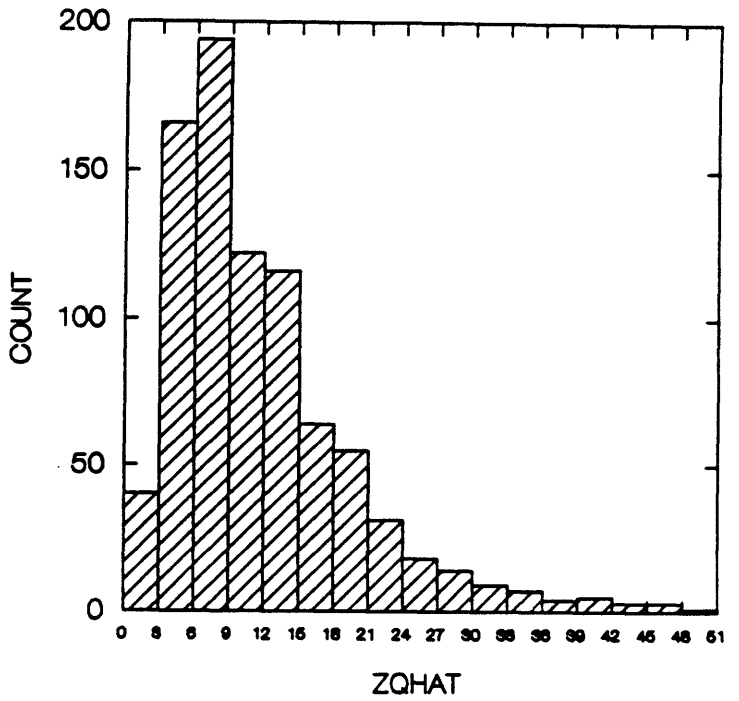


FIGURE 6.2a
ACTUAL Z-EXPONENTIAL CASE

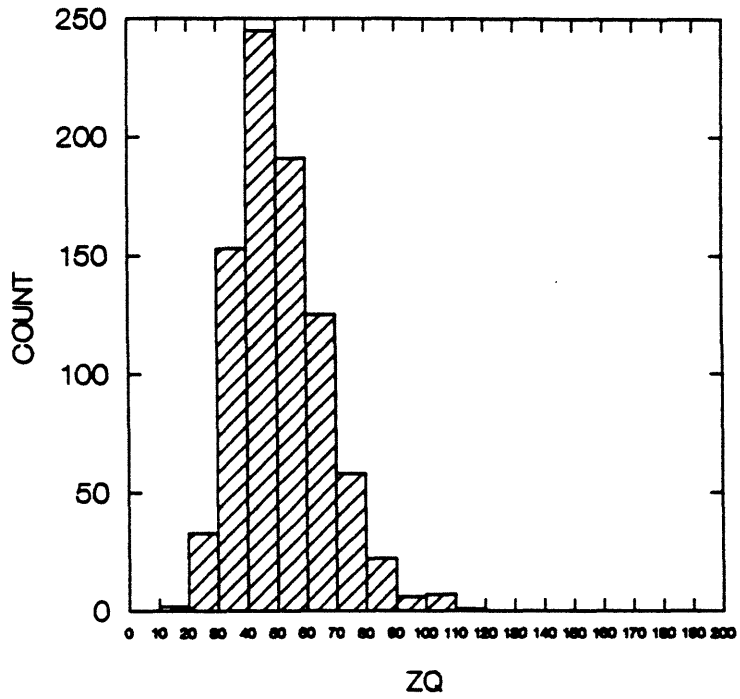


FIGURE 6.2b
ESTIMATED Z-EXPONENTIAL CASE

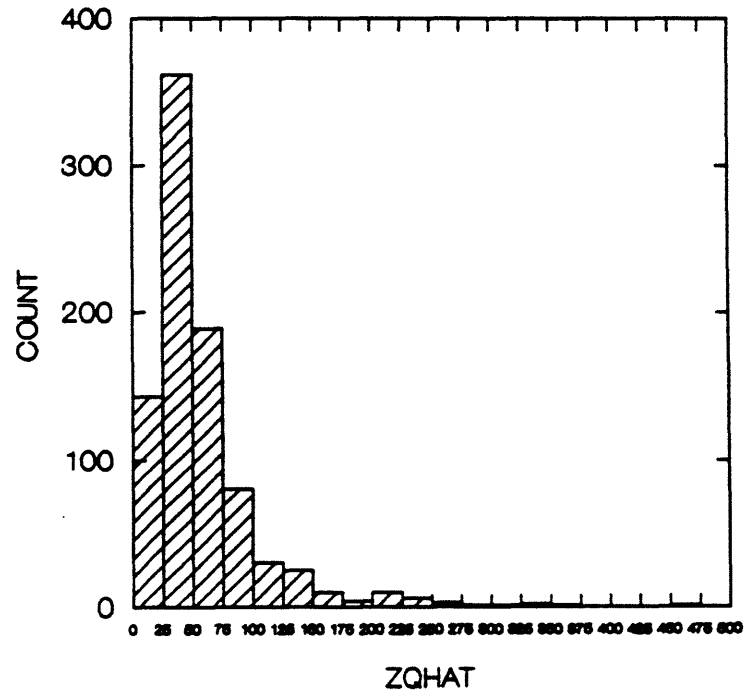


FIGURE 6.3a
ESTIMATED VS ACTUAL Z-HAJEK CASE

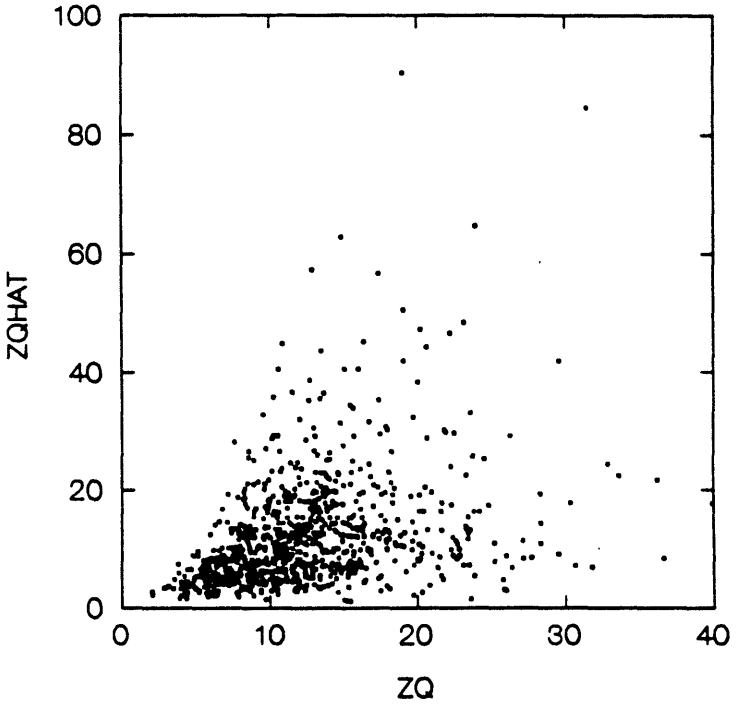


FIGURE 6.3b
ESTIMATED VS ACTUAL Z-EXPONENTIAL CASE

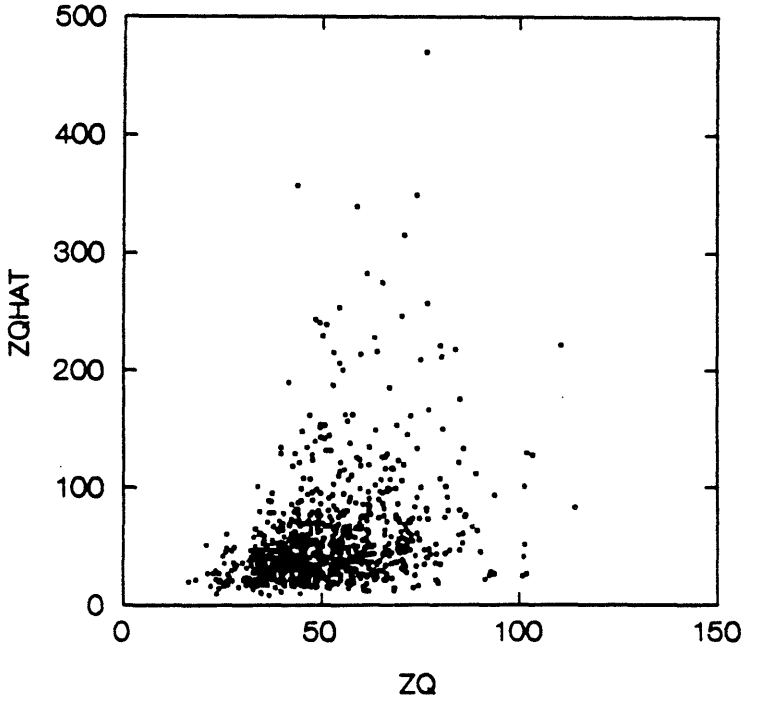


FIGURE 6.4a
$\hat{Z}_{(7)}$ vs $Z_{(4)}$ - HAJEK EXAMPLE

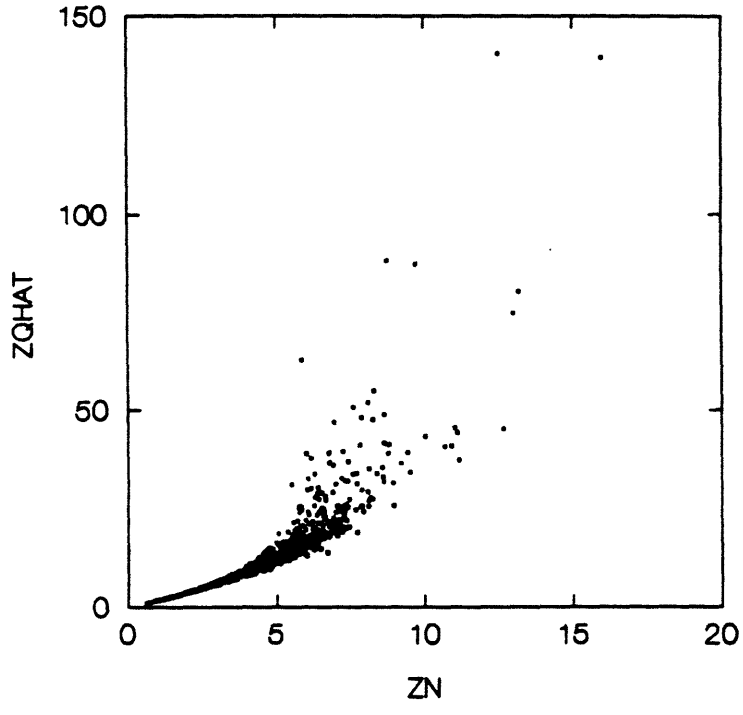
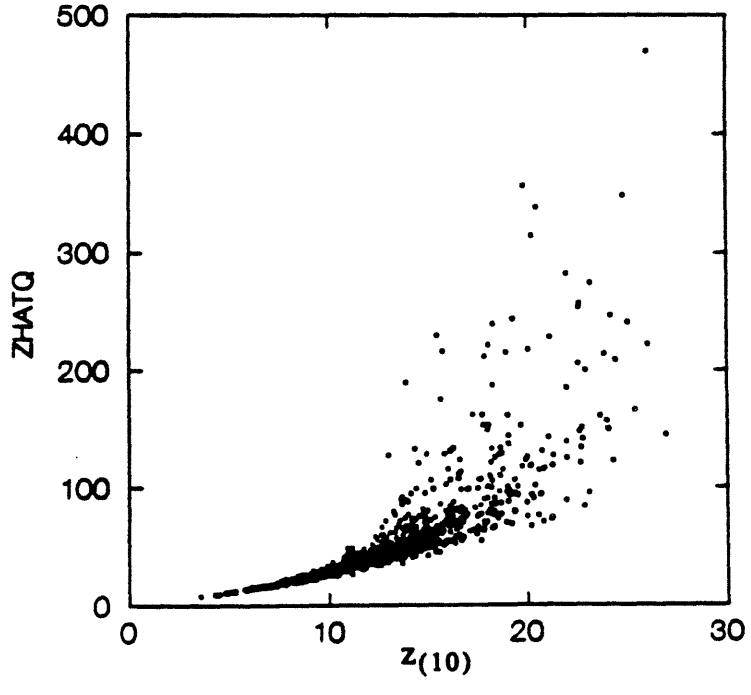


FIGURE 6.4b
$\hat{Z}_{(20)}$ vs $Z_{(10)}$ - EXPONENTIAL EXAMPLE



Bibliography

Andreatta, G., and Kaufman, G.M. (1986), "Estimation of Finite Population Properties When Sampling is Without Replacement and Proportional to Magnitude," *Journal of the American Statistical Association*, Vol. 81, pp. 657-666.

Basili, V.R., and Perricone (1982), "Software Errors and Complexity: an Empirical Investigation," *Computer Science*, University of Maryland, UOM-1195.

Basili, V.R., and Patniak, D., "A Study on Fault Prediction and Reliability Assessment in the SEL Environment," *Computer Science*, University of Maryland Technical Report, TR--1699.

Bickel, P.J.,V.N. Nair and Wang, P.C. (1989), "Nonparametric Inference Under Biased Sampling from a Finite Population," *Technical Report*, University of California at Berkeley, August 1989.

Chapman, D.G. (1951), "Some Properties of the Hypergeometric Distribution With Application to Zoological Sample Censuses," *Univeristy of California Publications in Statistics*, 1, 1313-1316.

Goel, A., "Software Reliability Models: Assumptions, Limitations, and Applicability," *IEEE Trans. Software Eng.* Vol. SE-11, No. 12 (April 1985), pp. 375-386.

Gordon, L. (1983) "Successive Sampling in Large Finite Populations," *Annals of Statistics*, Vol. 11, No. 2, pp. 702-706.

Gordon, L. (1989), "Estimation for Large Successive Samples of Unknown Inclusion Probabilities," *Advances in Applied Mathematics* (to appear).

Hájek, J. (1981), *Sampling from a Finite Population*, (New York: Marcel Decker, Inc.), 247 pp.

Horvitz, D..G., and Thompson, D.J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistics Association*, 47, 663-685.

Kaufman, G. M. (1988), "Conditional Maximum Likelihood Estimation for Successively Sampled Finite Populations," M.I.T. Sloan School of Management Working Paper.

Langberg, N. and Singpurwalla, N. (1985), " A Unification of Some Software Reliability Models,: *SIAM Journal on Scientific and Statistical Computing*, 6(3), pp. 781-790.

Littlewood, B., "Stochastic Reliability Growth: A Model for Fault-removal in Computer Programs and Hardware Designs," *IEEE* Trans. Rel. Vol. R-30 1981, pp. 313-320.

Miller, D.R., "Exponential Order Statistic Models of Software Reliability Growth," *IEEE*, Trans. Software Eng. Vol. SE-12 No. 1, Jan. 1986, pp. 12-24.

Murthy, M.N. (1957), "Ordered and Unordered Estimators in Sampling Without Replacement,"*Sankhya* 18, 378-390.

Rosén, B. (1972), "Asymptotic Theory for Successive Sampling with Varying Probabilities without Replacement," I and II, *Annals of Statistics*, Vol. 43, pp. 373-397, 748-776.

Ross, S.M. (1985), "Statistical Estimation of Software Reliability," *IEEE Transactions on Software Engineering*, Vol. SE-11, pp. 479-483.

Scholz, F.W. (1986), "Software Reliability Modeling and Analysis," *IEEE Transactions on Software Engineering*, Vol. SE-12, No. 1, January 1986.