MedCrave Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

Second order optimality of sequential designs with application in software reliability estimation

Abstract

We propose three efficient sequential designs in the software reliability estimation. The fully sequential design the multistage sequential design and the accelerated sequential design. These designs make allocation decisions dynamically throughout the testing process. We then refine these estimated reliabilities in an iterative manner as we sample. Monte Carlo simulation seems to indicate that these sequential designs are second order optimal.

Keywords: software reliability, partition testing, fully sequential design, multistage sequential design, accelerated sequential design.

Volume 2 Issue 4 - 2015

Kamel Rekab, Xing Song

Department of Mathematics and Statistics, University of Missouri-Kansas City, USA

Correspondence: Kamel Rekab, Department of Mathematics and Statistics, University of Missouri-Kansas City, PO Box 32464 Kansas City, MO 64171, USA, Tel: 816-269-4432; Email rekabk@umkc.edu

Received: April 11, 2015 | **Published:** April 29, 2015

Introduction

Reliability of a system is an important aspect of any system design since any user of the system would expect some type of guarantee that the system will function to some level of confidence. Failing to meet such guarantee will result in disastrous consequences. On the other hand, overly exceeding such guarantee level may incur additional and unnecessary expense to the developers. Moreover, for any non-trivial software system, an exhaustive testing among the entire input domain can be very expensive. By adopting the partition testing strategy, we attempt to break up the testable input domain of possible test cases into partitions, which must be non-overlapping, such that if test case i belongs to partition j, then no partition other than j will contain i. Sayre and Poore[11] have given several possible mechanics to partition the domain into finitely many subdomains,

$$X_{ij} = \begin{cases} 1, \text{if test } j \text{ taken from partition } i \text{ is processed correctly} \\ 0, \text{otherwise} \end{cases}$$, such that :

$$D = \bigcup_{i=1}^{k} D_i ; D_i \cap D_j = \varnothing, i \neq j$$

which allows us to define the system reliability by a weighted sum of reliabilities of these subdomains, i.e.

$$R = \sum_{i=1}^{k} p_i R_i$$

Where R denotes the system reliability and R_i is the reliability of each subdomain D_i; and p_i, parameters of the operational profile is the likelihood of this test case belongs to partition D_i, which are assumed to be known.[12] As mentioned above, a complete testing of any software system of non-trivial size is practically impossible, R_i are usually unknown parameters to us. So as to gain knowledge about R_i, we must distribute the k test cases among these k partitions, and generate reasonable estimates for each . Specifically, we denote $n_1, n_2, \ldots, n_k$ as sizes of the samples which are taken from sub domain $D_1, D_2, \ldots, D_k$, respectively, where $\sum_{i=1}^{k} n_i = N$.

We model the outcome of the j^{th} taken from the i^{th} partition as a Bernoulli random variable X_{ij} such that:

$$X_{ij} = \begin{cases} 1, \text{if test } j \text{ taken from partition } i \text{ is processed correctly} \\ 0, \text{otherwise} \end{cases}$$

and each X_{ij} follows a Bernoulli distribution with parameter R_i. Then, the estimate of the overall system reliability R, denoted by $\hat{R}$ can thus be defined as:

$$\hat{R} = \sum_{i=1}^{k} p_i \hat{R}_i$$

where $\hat{R}_i$ is the estimate of R_i after n_i test cases have been allocated to partition such that:

$$\hat{R}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

and

$$Var\left(\hat{R}\right) = \sum_{i=1}^{k} \frac{p_i^2 R_i \left(1 - R_i\right)}{n_i} \tag{1.1}$$

Optimal sampling scheme

Ideally, we would like to execute all possible test paths through the software and determine the true overall reliability of the system. In practice though, resources are often limited, sample test cases must be chosen and allocated strategically to attain the best reliability estimate possible given all kinds of constraints. One of the criteria of distributing test cases among the partitions, which proceeds from rewriting (1.1) as follows:

$$Var\left(\hat{R}\right) = \frac{\left[\sum_{i=1}^{k} p_i \sqrt{R_i\left(1-R_i\right)}\right]^2}{N} + \frac{1}{N} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{\left[n_i p_j \sqrt{R_j\left(1-R_j\right)} - n_j p_i \sqrt{R_i\left(1-R_i\right)}\right]^2}{n_i n_j}$$

which is bounded below by the first term:

$$Var\left(\hat{R}\right) \geq \frac{\left[\sum_{i=1}^{k} p_i \sqrt{R_i\left(1-R_i\right)}\right]^2}{N} \quad (2.2)$$

with equality of (2.1) to hold is and only if:

$$\frac{n_i}{n_j} = \frac{p_i \sqrt{R_i\left(1-R_i\right)}}{p_j \sqrt{R_j\left(1-R_i\right)}} \quad (2.3)$$

for all . Hence, the optimal allocation is determined by:

$$\frac{n_i}{N} = \frac{p_i \sqrt{R_i\left(1-R_i\right)}}{\sum_{j=1}^{k} p_j \sqrt{R_j\left(1-R_i\right)}} \quad (2.4)$$

for $i = 1, 2, \ldots, k-1$, and

$$n_k = N - \sum_{i=1}^{k-1} n_i$$

and the variance incurred by this optimal allocation is:

$$Var\left(O\right) = \frac{\left[\sum_{j=1}^{k} p_j \sqrt{R_j\left(1-R_i\right)}\right]^2}{N} \quad (2.5)$$

Note that the optimal allocation depends on the actual reliabilities $R_1, R_2, \ldots, R_k$, which are unknown. Therefore the optimal sampling scheme is not practical. It is this shortcoming that motivates us to adopt such dynamic allocation approaches that will be discussed in the following three sections.

Fully sequential sampling scheme

By adopting a fully Bayesian approach with Beta priors, Rekab, Thompson and Wei[5] proposed a fully sequential design shown to be first order optimal. The fully sequential design is defined as follows;

We first test one case from each of the partitions and estimate the reliability for each of these partitions.

Stage 1:

After cases have been allocated, $l \geq k$, the next test will be taken from partition i if for all j,

$$\frac{n_{l,i}}{n_{l,j}} < \frac{p_i \sqrt{\hat{R}_{l,i}\left(1-\hat{R}_{l,i}\right)}}{p_j \sqrt{\hat{R}_{l,j}\left(1-\hat{R}_{l,j}\right)}}$$

where $n_{l,i}$ is the cumulative test cases allocated to partition i after l tests have been allocated and the current estimated reliability for partition i is determined by:

$$\hat{R}_{l,i} = \frac{\sum_{m=1}^{n_{l,i}} X_{im}}{n_{l,i}}$$

Stage 2:

Repeat step 2 sequentially until all the test cases are allocated, and the final estimate of reliability for partition is:

$$\hat{R}_i = \frac{\sum_{m=1}^{n_{N,i}} X_{im}}{n_{N,i}}$$

And thus, the estimate of the overall reliability of the system is:

$$\hat{R} = \sum_{i=1}^{k} p_i \hat{R}_i$$

Multistage sequential sampling

By adopting a fully Bayesian approach with Beta priors, Rekab, Thompson and Wei[6] proposed a multistage sequential design shown to be first order optimal. Instead of making an allocation decision for each test at a time, the multistage sequential sampling allocates test cases among the partitions in stages in batches, where and are both pre-specified. The multistage sequential design is defined as follows:

Stage 1:

We start with an initial sample of test cases, which are allocated among the partitions with a balanced allocation scheme, such that:

$$S_{1,i} = \frac{S_1}{k}$$

and estimate the reliability for partition by:

$$\hat{R}_{1,i} = \frac{\sum_{m=1}^{s_{1,i}} X_{im}}{S_{1,i}}$$

Therefore,

$$\hat{C}_i\left(S_1\right) = \frac{p_i \sqrt{\hat{R}_{1,i}\left(1-\hat{R}_{1,i}\right)}}{\sum_{j=1}^{k} p_j \sqrt{\hat{R}_{1,j}\left(1-\hat{R}_{1,j}\right)}}$$

Stage 2 through L:

At stage j, $2 \leq j \leq L$, for partition $i = 1, 2, \ldots, k-1$, the test cases are distributed by the following way:

$$S_{j,i} = \left(\sum_{l=1}^{j} S_l\right) \hat{C}_i\left(\bar{S}_{j-1}\right);$$

and

$$S_{j,k} = \sum_{l=1}^{j} S_l - \sum_{i=1}^{k-1} S_{j,i}$$

where

$$\bar{S}_{j-1} = \sum_{y=1}^{j-1} S_y$$

At the final stage , the cumulative number of tests allocated to partition is:

$$N_i = \min\left\{N - \sum_{j=1, j\neq i}^{k-1} S_{L-1,j}, \max\left(N\hat{C}_i\left(\bar{S}_{L-1}\right), S_{L-1,i}\right)\right\}$$

and

$$N_k = N - \sum_{i=1}^{k-1} N_i$$

Therefore, the estimate for the whole system is finally obtained as:

$$\sum_{i=1}^{k} p_i \hat{R}_{L,i}$$

Among several ways of determining the number of cases at each sampling stage, the simplest one is to select:

$$S_1 = S_2 = \ldots = N / L$$

However, choosing stage sizes, especially the initial stage size, can be done by following some general criteria, a good initial stage size can be $\sqrt{N}$ for when a two stage sampling scheme is considered, and more generally, Rekab[9] shows that for a two stage procedure, must be chosen such that:

$$\lim_{N\to\infty} S_1 = \infty, \text{ and } \lim_{N\to\infty} \frac{S_1}{N} = 0 \qquad (4.1)$$

Accelerated Sampling Scheme

By adopting a fully Bayesian approach with Beta priors, Rekab, Thompson and Wei[6] proposed an accelerated sequential design shown to be first order optimal**.** The accelerated sampling scheme combines the fully sequential sampling scheme and the multistage sampling scheme. It is defined as follows:

Stage 1:

The procedure starts with an initial sample S_1, which satisfies the conditions specified in (4.1). Then, allocate equally among partitions

$$S_{1,i} = S_1 / k.$$

Stage 2 through $L-1$:

At stage j, $2 \le j \le L-1$, for partition $i = 1, 2, \ldots, k-1$, the test cases are distributed by the following way:

$$S_{j,i} = \left(\sum_{l=1}^{j} S_l\right) \hat{C}_i\left(\bar{S}_{j-1}\right);$$

and

$$S_{j,k} = \sum_{l=1}^{j} S_l - \sum_{l=1}^{k-1} S_{j,i}$$

where

$$\bar{S}_{j-1} = \sum_{y=1}^{j-1} S_y$$

At each stage, S_j must satisfy the two conditions as S_1.

Stage L:

In the final stage, we adopt a fully sequential approach by allocating one test from partition , if for all ,

$$\frac{n_{l,i}}{n_{l,j}} < \frac{p_i\sqrt{\hat{R}_{l,i}\left(1-\hat{R}_{l,i}\right)}}{p_j\sqrt{\hat{R}_{l,j}\left(1-\hat{R}_{l,j}\right)}}$$

until all the test cases have been allocated. Note that $n_{l,i}, n_{l,i}$ are the cumulative test cases allocated to partition and after the allocations of a total of test cases, where

$$\sum_{j=1}^{L-1} S_{j,i} \le l \le \sum_{j=1}^{L-1} S_{j,i} + N - \sum_{j=1}^{L-1} S_j$$

Therefore, the estimate for the whole system is finally obtained as:

$$\sum_{i=1}^{k} p_i \hat{R}_{L,i}$$

Optimality of sequential designs

First order optimality of these three sequential designs was established by Rekab, Thompson and Wu,[5,6,7] although the focus here is on minimizing the variance incurred by the sequential designs rather than minimizing the Bayes risk incurred by these designs. For estimating the mean difference of two independent normal populations, Woodroofe and Hardwick[12] adopted a quasi-Bayesian approach. They determined an asymptotic lower bound for the integrated risk and proposed a three-stage design that is second-order optimal. For estimating the mean difference of two general one -parameter exponential family, Rekab and Tahir[10] adopted a fully Bayesian approach with conjugate priors. They determined an asymptotic second order lower bound for the Bayes risk.

Monte carlo simulations

We consider the case where the test domain is partitioned into two subdomains D_1 and D_2 with reliability R_1 and R_2 respectively with equal usage probabilities p_1, p_2. Second order optimality of the three sequential designs is investigated through Monte Carlo simulations.

Table I, II, III seem to indicate that the speed $N^2 * \left(Var(\Delta) - Var(O)\right)$ is bounded.

Table I $N^2 * \left(Var(\ddot{A}) - Var(O)\right)$ by Fully Sequential Scheme

(R_1, R_2)	$N = 300$	$N = 500$	$N = 800$	$N = 2000$	$N = 5000$	$N = 8000$
0.1,0.9	28.1562	3.2104	0.7390	2.4225	9.3040	0.7133
0.5,0.2	29.0380	6.1020	0.6300	7.9375	4.6700	4.0500
0.5,0.5	1.5972	0.1460	0.2120	0.9125	0.5120	0.4933
0.5,0.9	80.7747	71.9032	19.9510	7.9600	5.7960	1.2666
0.9,0.3	64.0246	53.8008	5.5497	3.2565	15.2546	2.6517

Table 2 $N^2 * \left(Var\left(\ddot{A}\right) - Var\left(O\right)\right)$ by Multistage Scheme

$\left(R_1, R_2\right)$	$N = 300$	$N = 500$	$N = 800$	$N = 2000$	$N = 5000$	$N = 8000$
0.1,0.9	10.9835	0.0288	4.0718	55.7962	2.3320	1.8422
0.5,0.2	11.8285	19.1720	2.0555	5.5100	0.8432	7.1258
0.5,0.5	0.02749	0.0459	0.1750	0.4568	0.47324	0.1064
0.5,0.9	37.6665	31.6257	17.6417	3.6377	6.8483	9.2952
0.9,0.3	20.2034	2.86352	7.6161	3.1589	11.7869	5.2018

Table 3 $N^2 * \left(Var\left(\ddot{A}\right) - Var\left(O\right)\right)$ by Accelerated Scheme

$\left(R_1, R_2\right)$	$N = 300$	$N = 500$	$N = 800$	$N = 2000$	$N = 5000$	$N = 8000$
0.1,0.9	30.5713	6.5870	6.4568	17.4023	2.1138	1.0098
0.5,0.2	8.1424	6.1346	7.1683	0.8968	0.1462	2.1009
0.5,0.5	0.03801	0.23284	0.0319	0.7025	0.0640	0.0392
0.5,0.9	37.6665	31.6257	17.6417	3.6377	6.8483	9.2952
0.9,0.3	48.4148	13.8170	7.8473	8.4656	5.8734	1.6708

Conclusion

Second optimal designs are more efficient than the first order optimal designs especially when the total number of cases is very large. This is the main argument that led us to investigate the second order optimality of these three dynamic designs. Simulation studies seem to indicate that these designs are second order optimal. We conjecture that second order optimally may be obtained theoretically as well.

It is very common in parametric estimation to use the squared error loss. However, in reliability estimation one should distinguish between the cost of overestimating and underestimating the system reliability. Examples of practical loss functions were presented by Stüger:[2]

$$l\left(\hat{R},R\right) = c_o\left(\hat{R}-R\right)^2 I_{\left\{\hat{R}>R\right\}} + c_u\left(R-\hat{R}\right)^2 I_{\left\{\hat{R}<R\right\}}$$

and by Granger:[1]

$$l\left(\hat{R},R\right) = c_o\left(\hat{R}-R\right) I_{\left\{\hat{R}>R\right\}} + c_u\left(R-\hat{R}\right) I_{\left\{\hat{R}<R\right\}}$$

where represents the overestimation and underestimation costs, respectively.

Acknowledgement

None.

Conflict of Interest

None.

References

1. Sankaran M. The discrete Poisson-Lindley distribution. *Biometrics*. 1970;26(1):145–149.
2. Ghitany ME, Al-Mutairi DK. Estimation Methods for the discrete Poisson-Lindley distribution. *Journal of Statistical Computation and Simulation*. 2009;79(1):1–9.
3. Shanker R, Mishra A. A two-parameter Poisson-Lindley distribution. *International Journal of Statistics and Systems*. 2014;9(1):79–85.
4. Shanker R, Mishra A. A two-parameter Lindley distribution. *Statistics in Transition new Series*. 2013;14(1):45–56.
5. Shanker R, Mishra A. A quasi Poisson-Lindley distribution (submitted). 2015.
6. Shanker R, Mishra A. A quasi Lindley distribution. *African journal of Mathematics and Computer Science Research*. 2013;6(4):64–71.
7. Shanker R, Sharma S, Shanker R. A Discrete two-Parameter Poisson Lindley Distribution. *Journal of Ethiopian Statistical Association*. 2012;21:15–22.
8. Shanker R, Sharma S, Shanker R. A two-parameter Lindley distribution for modeling waiting and survival times data. *Applied Mathematics*. 2013;4:363–368.
9. Shanker R, Tekie AL. A new quasi Poisson-Lindley distribution. *International Journal of Statistics and Systems*. 2014;9(1):87–94.
10. Shanker R, Amanuel AG. A new quasi Lindley distribution. *International Journal of Statistics and Systems*. 2013;8(2):143–156.
11. Johnson NL, Kotz S, Kemp AW. Univariate Discrete Distributions, 2nd ed. John Wiley & sons Inc; 1992.
12. Fisher RA, Corpet AS, Williams CB. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*. 1943;12(1):42–58.
13. Kempton RA. A generalized form of Fisher's logarithmic series. *Biometrika*. 1975;62(1):29–38.
14. Tripathi RC, Gupta RC. A generalization of the log-series distribution. *Comm. in Stat. (Theory and Methods)*. 1985;14(8):1779–1799.

15. Mishra A, Shanker R. Generalized logarithmic series distribution-Its nature and applications. *Proceedings of the Vth International Symposium on Optimization and Statistics*. 2002:155–168.

16. Loeschke V, Kohler W. Deterministic and Stochastic models of the negative binomial distribution and the analysis of chromosomal aberrations in human leukocytes. *Biometrische Zeitschrift*. 1976;18:427–451.

17. Janardan KG, Schaeffer DJ. Models for the analysis of chromosomal aberrations in human leukocytes. *Biometrical Journal*. 1977;(8):599–612.

18. Mc Guire JU, Brindley TA, Bancroft TA. The distribution of European corn-borer larvae pyrausta in field corn. *Biometrics*. 1957;13:65–78.

19. Catcheside DG, Lea DE, Thoday JM. ypes of chromosome structural change induced by the irradiation on Tradescantia microspores. *Journal of Genetics*. 1946;47:113–136.

20. Catcheside DG, Lea DE, Thoday JM. The production of chromosome structural changes in Tradescantia microspores in relation to dosage, intensity and temperature. *Journal of Genetics*. 1946;47:137–149.

21. Lindley DV. Fiducial distributions and Bayes theorem. *Journal of Royal Statistical Society Ser. B*. 1958;20:102–107.

Citation: Rekab K, Song X. Second order optimality of sequential designs with application in software reliability estimation. *Biom Biostat Int J*. 2015;2(4):109–113. DOI: 10.15406/bbij.2015.02.00037