IMPROVING THE QUALITY OF A SOFTWARE SYSTEM WITH SEQUENTIAL REVIEWS

Young H. Chun, E. J. Ourso College of Business, Louisiana State University
Baton Rouge, LA 70803, 225-578-2120, prof@drchun.net

ABSTRACT

A complex product such as a software document is often inspected more than once in a sequential manner to ensure the product's quality. For each defect, the probability that it will be detected during each inspection cycle is usually assumed to be a known "constant". However, in many practical situations, some defects are easily detected, while others are much more difficult to identify. In this paper, we propose a "beta-geometric" inspection model in which the heterogeneity in detection probability is described by a beta distribution. In a numerical study, we show that our more realistic inspection model clearly outperforms traditional estimation methods that are based on the assumption of a constant detection probability.

INTRODUCTION

Although inspection is one of the important and effective tools that serve the task of assuring product quality, inspection error is inevitable in any inspection process. That is why some complex products are inspected multiple times in a sequential manner in order to improve the outgoing quality.

As an example of the repetitive inspection, consider a software system that contains an unknown number N of faults. The software system will be reviewed more than once in a sequential manner. For each fault, the "detection probability" is p, which is the probability that the fault will be detected during the current review cycle. After each review, the number of faults x_i detected during the review cycle i is recorded, and those faults are removed or corrected prior to the next review. After a series of k independent reviews, we have a record $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$ of the number of faults detected and corrected during each review cycle. Based on the inspection results $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$, we need to estimate the total number of faults N or, equivalently, the number of faults R_k ($=N- x_1- x_2- \ldots -x_k$) still remaining in the software system. Such multiple inspection plans have attracted considerable attention recently with various names: repetitive testing [3] [6] [8] [7], sequential defect removal sampling [1], repeat inspection [4], repeated screening [5], or sequential review or inspection [9] [11].

The detection probability p of a fault is often assumed to be (i) a known constant that is given *a priori* or (ii) an unknown constant that ought to be estimated. In many practical situations, however, each fault has a different probability of being detected; some faults can be found easily, while others are much more difficult to be detected.

The purpose of this article is to propose an improved inspection model that considers the "heterogeneity" in detection probability p. Specifically, we assume that the detection probability p is distributed as a beta distribution with parameters a and b. By changing its beta parameters, we can describe a wide variety of distributions with different shapes and scales. In a numerical analysis, we show that our "beta-geometric" model clearly outperforms traditional estimation methods such as (i) the maximum likelihood method and (ii) the conditional maximum likelihood method.

METHOD OF MAXIMUM LIKELIHOOD

At the beginning of the ith review, we still have N-s_{i-1} faults remaining in the software document. Then,

the number of faults x_i that will be discovered during the ith review follows a binomial distribution:

$$P[X_i = x_i \mid N, q] = \binom{N - s_{i-1}}{x_i} q^{N-s_i} (1-q)^{x_i}, \quad \text{for } i = 1, 2, \ldots, k. \tag{1}$$

When the inspection results $\mathbf{x}=\{x_1, x_2,\ldots, x_k\}$ are available after k reviews, the likelihood function of N and q is expressed as follows:

$$L(N,q) = \prod_{i=1}^{k} P[X_i = x_i \mid N, q] = \frac{N!}{(N-s_k)!\prod_{i=1}^{k} x_i!} q^{\sum_{i=1}^{k}(N-s_i)} (1-q)^{s_k} \tag{2}$$

The maximum likelihood estimates (MLE) of N and q are the ones that maximize the likelihood function in (2). However, the optimal values of N and q which maximize the likelihood function $L(N, q)$ in (2) also maximize its log-likelihood function $\ln L(N, q)$. Therefore, rather than maximizing the likelihood function itself, it is more convenient to maximize its natural logarithm:

$$\ln L(N,q) = \ln \frac{N!}{(N-s_k)!} + \ln q \sum_{i=1}^{k} (N - s_i) + s_k \ln(1-q). \tag{3}$$

Setting its first-order derivative equal to zero, we can derive the maximum likelihood estimator of q as follows:

$$\hat{q} = \frac{\sum_{i=1}^{k} (N - s_i)}{\sum_{i=1}^{k} (N - s_{i-1})}. \tag{4}$$

By plugging $\hat{q}$ in (4) into the log-likelihood function in (1), we can formulate the problem of finding the maximum likelihood estimate of N as a single-parameter maximization problem.

CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATOR

The so-called "conditional maximum likelihood estimator" was originally proposed by [10], who used it to estimate the number of unknown trials in a multinomial probability distribution. We modify the estimation method for the repetitive inspection model and compare its performance later with those of other estimation methods.

After k review cycles, the inspection results are $\mathbf{x}=\{x_1, x_2,\ldots, x_k\}$ and $(N\text{-}s_k)$ faults are still remaining in the software document. Thus, the likelihood function of N and q is

$$L(N,q) = \frac{N!}{(N-s_k)!\prod_{i=1}^{k} x_i!} (q^k)^{N-s_k} \prod_{i=1}^{k} [q^{i-1}(1-q)]^{x_i}. \tag{5}$$

We found that the likelihood function in (5) can be divided into two separate parts as follows:

$$L(N,q) = \left[\frac{N!}{(N-s_k)!\,s_k!} (1-q^k)^{s_k} (q^k)^{N-s_k} \right] \times \left[\frac{s_k!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \left[\frac{q^{i-1}(1-q)}{(1-q^k)} \right]^{x_i} \right]. \tag{6}$$

The first part is the likelihood based on the probability of s_k and the second part is the likelihood based on the conditional probability of x_1, x_2, …, and x_k given s_k. Since the first likelihood function in (6) is a binomial distribution, the maximum likelihood estimate of N conditional upon q is simply shown to be

$$\hat{N}(q) = \frac{s_k}{1-q^k}. \tag{7}$$

The second likelihood function in (6) is a multinomial distribution, which is independent of N. Thus, the maximum likelihood estimator of q is the one that maximizes the likelihood function in (5) or, equivalently, its log-likelihood function:

$$\ln L_2(q) = \sum_{i=1}^{k} x_i[(i-1)\ln q + \ln(1-q) - \ln(1-q^k)]. \tag{8}$$

From the first-order derivative of the log-likelihood function in (8), it can be easily shown that the maximum likelihood estimator $\hat{q}$ is the solution to the following equation:

$$\frac{\sum_{i=1}^{k} s_i}{s_k} = \frac{\sum_{i=1}^{k}(1-q^i)}{1-q^k}. \tag{9}$$

After finding $\hat{q}$ that satisfied the equation in (9), we can easily obtain $\hat{N}$ from (7).

BETA-GEOMETRIC MODEL

In the traditional estimation methods, the detection probability p (=1-q) is assumed to be an unknown *constant*. To represent the heterogeneity in detection probabilities, we now assume that the probability p of being detected during each review cycle is distributed as a beta distribution with parameters a and b:

$$f(p \mid a,b) = \frac{1}{B(a,b)}\, p^{a-1}(1-p)^{b-1}, \qquad \text{for } 0 < p < 1, \tag{10}$$

By changing the parameter values a and b in (10), we can represent a wide variety of variations in the detection probability p. If $a = b = 1$, for example, the beta distribution represents the standard uniform (or rectangular) distribution with equal probabilities over the range (0, 1). If $a = 1$ and $b = 2$ (or $a = 2$ and $b = 1$), the beta distribution becomes a triangular distribution.

For a certain fault in the product, the probability that the fault will be discovered and removed during the ith inspection cycle follows a *geometric* distribution with parameter p:

$$P[i \mid p] = (1-p)^{i-1}\, p. \tag{11}$$

Likewise, the probability that it will not be found during the first k inspection cycles and will be still remaining in the product is

$$P[i > k \mid p] = (1-p)^k. \tag{12}$$

From (11) and (12), the probability that the fault will be successfully discovered during the ith review cycle is a beta-geometric distribution as follows:

$$P[i \mid a,b] = \int_{p=0}^{1} P[i \mid p]\, f(p \mid a,b)dp = a\frac{\Gamma(b+i-1)}{\Gamma(b)}\frac{\Gamma(a+b)}{\Gamma(a+b+i)}. \tag{13}$$

Note that a gamma function in (13) has the following property: $\Gamma(c+1) = c\,\Gamma(c)$ for any constant c. Thus, it can be further simplified as

$$P[i \mid a,b] = a\frac{\prod_{j=0}^{i-2}(b+j)}{\prod_{j=0}^{i-1}(a+b+j)} = \frac{a}{b+i-1}\prod_{j=0}^{i-1}\frac{b+j}{a+b+j}. \tag{14}$$

From (12) and (14), the probability that a certain fault will still remain undetected after k inspection cycles is

$$P[i > k \mid a,b] = \int_{p=0}^{1} P[i > k \mid p]\, f(p \mid a,b)dp = \prod_{j=0}^{k-1} \frac{b+j}{a+b+j}. \quad (15)$$

The numbers of faults discovered during the k review cycles are $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$, and the number of undetected faults still remaining in the software document is N-s_k. Thus, the likelihood function of N, a, and b is

$$L(N,a,b \mid \mathbf{x}) = \frac{N!}{(N-s_k)!\prod_{i=1}^{k} x_i!} \left(\prod_{j=0}^{k-1} \frac{b+j}{a+b+j} \right)^{N-s_k} \prod_{i=1}^{k} \left(\frac{a}{b+i-1} \prod_{j=0}^{i-1} \frac{b+j}{a+b+j} \right)^{x_i}. \quad (16)$$

PERFORMANCE EVALUATION

Detection Probabilities

In the simulation study in which we compare the performance of the beta-geometric model with those of traditional estimation methods, we assume that there are N=100 faults in a software document. To represent the actual situation in which the probability p of being discovered is different from fault to fault, we consider four different cases:

(a) Rectangular distribution with $E[p]$=1/2

$$f(p) = 1 \qquad \text{for } 0 < p < 1$$

(b) Triangular distribution with $E[p]$=1/2

$$f(p) = \begin{cases} 4p & \text{for } 0 < p < 0.5 \\ 4(1-p) & \text{for } 0.5 < p < 1 \end{cases}$$

(c) Triangular distribution with $E[p]$=1/3

$$f(p) = 2(1-p) \qquad \text{for } 0 < p < 1$$

(d) Triangular distribution with $E[p]$=2/3

$$f(p) = 2p \qquad \text{for } 0 < p < 1$$

Parameter Estimation

In the simulation study, we assume that each of the 100 faults in the software document is subject to review cycle up to k=10 times. The exact detection time i of the fault with p follows a geometric distribution, and its cumulative distribution function is $1-(1-p)^i$, where i = 1, 2, …, 10. Thus, with another standard uniform random number v from Microsoft Excel, we simulated the detection time i of the fault with p as the smallest integer larger than or equal to

$$i = \left\lceil \frac{\ln(1-v)}{\ln(1-p)} \right\rceil. \quad (17)$$

Any faults with i larger than 10 have not been discovered during the k=10 review cycles.

For N=100 faults in the software document, we then counted the number of faults that have been

detected during the ith review cycle. The inspection results are simply given by $\mathbf{x} = \{x_1, x_2, \ldots, x_{10}\}$ for one simulation run. The beta-geometric model, along with other traditional estimation methods, is used to estimate the true parameter value N=100 for each simulation run. When $\mathbf{x}$ = {44, 17, 8, 6, 4, 8, 0, 1, 0, 1}, for example, the estimates of N obtained by the maximum likelihood method, conditional maximum likelihood method, and beta-geometric model are shown to be 89.00, 89.45, and 96.01, respectively.

After 100 simulation runs, we then calculated the average estimate of $\hat{N}$ and the mean absolute percentage error (MAPE) for each estimation method. The performance measures are summarized in Table 1.

Table 1. Estimates of N by three methods in four different cases
(The true parameter value is N = 100.)

Distribution of p	Performance Measures	Estimation Methods		
		MLE	Conditional MLE	Beta-Geometric
(A) Rectangular with $E[p]$=1/2	Average $\hat{N}$	89.679	89.959	99.702
	MAPE	10.321%	10.041%	4.187%
(B) Triangular with $E[p]$=1/2	Average $\hat{N}$	97.180	97.386	99.848
	MAPE	2.820%	2.635%	2.048%
(c) Triangular with $E[p]$=1/3	Average $\hat{N}$	83.999	84.723	99.756
	MAPE	16.001%	15.277%	6.867%
(d) Triangular with $E[p]$=2/3	Average $\hat{N}$	98.300	98.315	100.095
	MAPE	1.700%	1.691%	1.600%

Simulation Results

As shown in Table 1, the conditional MLE is slightly better than the MLE in terms of the average estimate and the mean absolute percentage error. However, the beta-geometric model clearly outperforms the traditional estimation methods in all four cases. When the detection probabilities are uniformly distributed between 0 and 1, for example, both the MLE and the conditional MLE severely underestimate the true number of faults N as 89.679 and 89.959, respectively. On the other hand, the average of the 100 estimates obtained by the beta-geometric model is 99.702, which is very close to the true parameter value N=100. Its mean absolute percentage error is 4.187%, which is much better than those of the MLE and the conditional MLE.

The traditional methods consistently underestimate the true number of faults when the detection probability is not a constant, but a random variable. Only the beta-geometric model handles the heterogeneity in detection probability very well, giving almost unbiased estimates in all four cases.

CONCLUDING REMARKS

In many practical situations, the probability of being detected during each inspection cycle is not the same among different types of defect. That is why we propose in the paper the beta-geometric inspection model in which the heterogeneity in detection probability is simply described as a beta

distribution. In a Monte Carlo simulation, we show that our inspection model clearly outperforms the maximum likelihood method and the conditional maximum likelihood method, predicting the total number of defects *N* with less biases and smaller variances.

Those simulation results are not unexpected, given that the maximum likelihood method is a special case of our beta-geometric model. The only drawback of the beta-geometric model is that we need to estimate three parameter values (*a*, *b*, and *N*), rather than two (*p* and *N*) as in the traditional methods. As shown in the simulation study, however, the computational complexity is not a big problem even with Microsoft Excel. Thus, in estimating the product quality after multiple inspections, there is no reason not to prefer our beta-geometric model over the traditional estimation methods.

In the paper, we focused on the problem of estimating the number of defects (i.e., non-conformities) in a complex product such as a software document. With slight modifications, our beta-geometric model can be applied directly to the problem of estimating the number of defective items (i.e., non-conforming items) in a batch of items such as IC chips [7]. In such a case, we need to estimate the average defective rate as well as the detection probability.

Another possible extension is the Bayesian estimation of the number of defects in a repetitive inspection procedure [2]. In many practical situations, the prior knowledge we have on the number of defects N can be described as a negative binomial distribution. Likewise, a beta distribution can be used as a prior distribution for the detection probability p. With those prior distributions, we can derive a posterior distribution of the number of defects still remaining after k rounds of inspection.

REFERENCES

[1] Bonett, D. G. & Woodward, J. A. Sequential defect removal sampling. *Management Science*, 1994, 40, 898-902.

[2] Chun, Y. H. Bayesian inspection model for the production process subject to a random failure. *IIE Transactions*, 2010, 42, 304-316.

[3] Ding, J. & Gong, L. The effect of testing equipment shift on optimal decisions in a repetitive testing process. *European Journal of the Operational Research*, 2008, 186, 330-350.

[4] Elshafei, M., Khan, M. & Duffuaa, S. O. Repeat inspection planning using dynamic programming. *International Journal of Production Research*, 2006, 44, 257-270.

[5] Gasparini, M., Nusser, H. & Eisele, J. Repeated screening with inspection error and no false positive results with application to pharmaceutical pill production. *Applied Statistics*, 2004, 53, 51-62.

[6] Gong, L. The effect of testing errors on a repetitive testing process. *European Journal of the Operational Research*, 2012, 220, 115-124.

[7] Greenberg, B. S. & Stokes, S. L. Repetitive testing in the presence of inspection errors. *Technometrics*, 1995, 37, 102-111.

[8] Quinino, R. C. & Ho, L. L. Repetitive tests as an economic alternative procedure to control attributes with diagnosis errors. *European Journal of the Operational Research*, 2004, 155, 209-225.

[9] Rallis, N. E. & Lansdowne, Z. F. Reliability estimation for a software system with sequential independent reviews. *IEEE Transactions on Software Engineering*, 2001, 27, 1057-1061.

[10] Sanathanan, L. Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 1972, 43, 142-152.

[11] Yao, D. D. & Zheng, S. Sequential Inspection under Capacity Constraints. *Operations Research*, 1999, 47, 410-421.