Analysis of Hypergeometric Distribution Software Reliability Model

T. Dohi[†] (01307065), N. Wakana[†], S. Osaki[†] (01002265) and Kishor S. Trivedi[‡]

[†]Hiroshima University, Higashi-Hiroshima, Japan, [‡]Duke University, NC, USA

1. INTRODUCTION

This article gives the detailed mathematical results on the hypergeometric software reliability model (HGDSRM) proposed by Thoma *et al.* [1, 2]. In the earlier paper, Thoma *et al.* [2] derived a recursive formula on the expected cumulative number of software faults detected up to ith test instance in testing phase. Since their model was based on only the mean value of the cumulative number of faults, it was impossible to estimate the software reliability as well as the other probabilistic dependability measures. By introducing the concept of cumulative trial processes, we derive the probability mass function of the number of software faults detected at ith test instance explicitly.

2. HGDSRM

Suppose that the test of a software is a set of *test instances* (such as test runs), which consists of input test data and observed test result. Define the software test by $D = \{t(i)|i = 1, 2, \cdots, \}$, where $t(i)$ is the ith test instance. Let $B = \{b(j)|j = 1, 2, \cdots, m\}$ denote a set of faults remaining in the software at the initial time ($i = 0$), where $b(j)$ means the fault labelled by j $(= 1, 2, \cdots, m)$ and m (> 0) is the initial number of faults. If a software error caused by $b(j)$ is observed at the test instance $t(i)$, the fault $b(j)$ is said to be *sensed* by the test instance $t(i)$. Suppose that a test instance $t(i)$ senses $w(i)$ software faults, where $w(i)$ is called *the sensitivity factor* and is the function of the number of test instance (or time). Make the following assumptions:

(A-1) The software faults that manifest themselves upon the application of a test instance $t(i)$ may be removed (fixed) before the next test instance $t(i+1)$ is applied.

(A-2) No new faults are introduced during the software testing. This means that the software reliability is nondecreasing as the testing progresses.

(A-3) A random set of $w(i)$ software faults are sensed by test instance $t(i)$ out of the total m initial faults.

¿From these assumptions, it is evident that the number of faults detected by the first test instance $t(1)$ is $w(1)$. However, the number of *newly* detected faults by $t(2)$ is not necessarily $w(2)$, since some of $w(2)$ faults may have been already detected and removed by $t(1)$. In general, the number of newly detected faults by the ith test instance $t(i)$, X_i, can be regarded as a positive random variable. Then, the cumulative number of newly detected faults by the test instances $t(1), \cdots, t(i)$ is $C_i = \sum_{j=1}^{i} X_j$.

With this notation, the probability that x faults can be newly detected by the test instance $t(i)$ is given by

$$\begin{aligned} P\Big\{X_i = x \mid \sum_{j=1}^{i-1} X_j = c_{i-1}\Big\} &= P\Big\{x|m, c_{i-1}, w(i)\Big\} \\ &= \frac{\binom{m-c_{i-1}}{x}\binom{c_{i-1}}{w(i)-x}}{\binom{m}{w(i)}}, \end{aligned} \tag{1}$$

where $0 \leq x \leq \min\{w(i), m-c_{i-1}\}$ and c_i is the realization of the random variable C_i. Since the above expression is the hyper-geometric pmf, the sequential model based on Eq.(1) is called the HGDSRM. From Eq.(1), the mean number of newly detected faults at the ith test instance and its variance are

$$\mathrm{E}\Big[X_i \mid \sum_{j=1}^{i-1} X_j = c_{i-1}\Big] = \Big(\frac{m - c_{i-1}}{m}\Big)w(i) \tag{2}$$

and

$$\begin{aligned} \mathrm{Var}\Big[X_i \mid \sum_{j=1}^{i-1} X_j = c_{i-1}\Big] &= \frac{(m - c_{i-1})c_{i-1}w(i)}{m^2} \\ &\quad \times\Big(\frac{m - w(i)}{m - 1}\Big), \end{aligned} \tag{3}$$

respectively. In [2], substituting $c_{i-1} = \sum_{k=1}^{i-1} x_k \approx \sum_{k=1}^{i-1} \mathrm{E}[X_k] = \mathrm{E}[C_{i-1}]$ into Eq.(2), the following recursive formula is obtained:

$$\mathrm{E}[C_i] = \mathrm{E}[C_{i-1}]\Big(1 - \frac{w(i)}{m}\Big) + w(i). \tag{4}$$

This can be solved by the induction as follows [2].

$$\begin{aligned} \mathrm{E}[C_i] &= m\Big\{1 - \Pi_{j=1}^{i}\Big(1 - \frac{w(j)}{m}\Big)\Big\} \\ &= m\Big\{1 - \exp\Big[\sum_{j=1}^{i} \log\Big(1 - \frac{w(j)}{m}\Big)\Big]\Big\}. \end{aligned} \tag{5}$$

The above equation is derived from the heuristic argument, but is correct from the independence of the Bernoulli trials.

3. FURTHER RESULTS

Suppose that the initial number of detected faults at $i = 0$ is 0. Of our interest is the derivation of the probability of the number of newly detected faults by the test instance $t(i)$. Let X_i be the number of newly detected faults at ith test instance. We make the following additional assumptions:

(B-1) The initial number of faults m remaining in the software is sufficiently larger than $w(i)$, i.e. $m \gg w(i)$ for all $i = 1, 2, 3, \cdots$.

(B-2) In the software test, it is impossible to detect all faults with probability one, i.e. $m > \lim_{i \to \infty} \sum_{j=1}^{i} x_i$.

¿From these assumptions, it can be seen that $\min\{w(i), m - c_{i-1}\} = w(i)$ has to be always satisfied. In fact, these assumptions are plausible intuitively and are often used in earlier modeling approach.

Suppose that the probability $P_1\{x_1 \mid m, w(1)\}$ that x_1 faults is detected by the test instance $t(1)$ is the hypergeometric pmf. Let $P_2\{x_2|m, w(2)\}$ denote the probability that x_2 faults are detected newly at the second test instance, i.e. $i = 2$. Then, it is straightforward to obtain

$$P_2\Big\{x_2|m, w(2)\Big\} = \sum_{x_1=0}^{w(1)} P_1\Big\{x_1|m, w(1)\Big\} \times P\Big\{x_2|m, c_1, w(2)\Big\}. \tag{6}$$

From the similar manipulation, we have

$$\begin{aligned} P_i\Big\{x_i|m, w(i)\Big\} = & \sum_{x_{i-1}=0}^{w(i-1)} \\ & \times P_{i-1}\Big\{x_{i-1}|m, w(i-1)\Big\} P\Big\{x_i|m, c_{i-1}, w(i)\Big\} \\ & \times \sum_{x_{i-1}=0}^{w(i-1)} P_{i-1}\Big\{x_{i-1}|m, w(i-1)\Big\} \\ & \times P\Big\{x_i|m, c_{i-1}, w(i)\Big\}, \end{aligned} \tag{7}$$

where the right-hand side of Eq.(7) is due to (B-1) and (B-2). Then, the problem is to solve the above recursive equation with the initial conditions:

$$P_1\Big\{x_1|m, w(1)\Big\} = \frac{\binom{m}{x_1}\binom{0}{w(1)-x_1}}{\binom{m}{w(1)}} = 1 \tag{8}$$

and

$$P_2\Big\{x_2|m, w(2)\Big\} = \frac{\binom{m-x_1}{x_2}\binom{x_1}{w(2)-x_2}}{\binom{m}{w(2)}}. \tag{9}$$

The following is the main result of this article.

Theorem 1: For the initial number of remaining software faults m, suppose that the sensitivity factor in ith test instance $t(i)$ is defined by $w(i)$. Then, the probability that x_i faults are detected newly at ith test instance, $P_i\{x_i \mid m, w(i)\}$, is

$$\begin{aligned} & \sum_{x_2=0}^{w(2)} \sum_{x_3=0}^{w(3)} \cdots \sum_{x_{i-1}=0}^{w(i-1)} \left[\prod_{n=2}^{i-1} \frac{\binom{m-\sum_{k=1}^{n-1} x_k}{x_n}\binom{\sum_{k=1}^{n-1} x_k}{w(n)-x_n}}{\binom{m}{w(n)}} \right] \\ & \times \frac{\binom{m-\sum_{k=1}^{i-1} x_k}{x_i}\binom{\sum_{k=1}^{i-1} x_k}{w(i)-x_i}}{\binom{m}{w(i)}}. \end{aligned} \tag{10}$$

Theorem 2: The mean number of newly detected faults at ith test instance is

$$\begin{aligned} \mathrm{E}[X_i] = w(i) - \frac{w(i)}{m} & \sum_{x_i=1}^{w(i)-1} \sum_{x_{i-1}=0}^{w(i-1)^*} \\ & \times \left[\sum_{k=1}^{i-1} x_k \prod_{n=2}^{i-1} \frac{\binom{m-\sum_{k=1}^{n-1} x_k}{x_n}\binom{\sum_{k=1}^{n-1} x_k.}{w(n)-x_n}}{\binom{m}{w(n)}} \right] \\ & \times \frac{\binom{m-\sum_{k=1}^{i-1} x_k - 1}{x_i - 1}\binom{\sum_{k=1}^{i-1} x_k}{w(i)-x_i}}{\binom{m-1}{w(i)-1}}, \end{aligned} \tag{11}$$

where $\sum_{x_{i-1}=0}^{w(i-1)^*} = \sum_{x_2=0}^{w(2)} \sum_{x_3=0}^{w(3)} \cdots \sum_{x_{i-1}=0}^{w(i-1)}$.

Similar to Theorem 1 and Theorem 2, we can derive analytically $\mathrm{Var}[C_i]$ and $\mathrm{Var}[X_i]$ as well as the software reliability $P\{\sum_{n=j+1}^{i} X_n = 0 \mid m, c_{i-1}, w(i)\}$. Also, it can be shown that $\mathrm{E}[C_i] = \sum_{j=1}^{i} \mathrm{E}[X_j]$ in Eq.(11). These results are never trivial and are all proved by the induction.

REFERENCES

[1] Y. Tohma, K. Tokunaga, S. Nagase and Y. Murata, Structural approach to the estimation of the number of residual software faults based on the hypergeometric distribution, *IEEE Trans. Software Eng.* **15** (3), 345–355 (1989).

[2] Y. Tohma, H. Yamano, M. Ohba and R. Jacoby, The estimation of parameters of the hypergeometric distribution and its application to the software reliability growth model, *IEEE Trans. Software Eng.* **17** (5), 483–489 (1991).