Discovery and Repair of Software Anomalies

R. C. Tausworthe

DSN Data Systems Section

This article presents a simple model which explains anomaly discovery and repair phenomena when applied to variations in work load and multiple-stage testing. The theory shows that estimates of anomaly levels and team capability can be predicted after a significant fraction of the anomalies has been found, and indicates procedures for applying these figures to schedule estimation and work-load assignment. Of particular interest is the demonstration that end-to-end testing of programs in other than the operational environment is not generally cost effective.

I. Introduction

Almost every software project enters a phase where it is "90 percent complete," in which it seemingly remains for a very disproportionate length of time before true completion. Much of this time, as it turns out, is spent discovering and repairing anomalies in the program, operations manuals, or program requirements. The number and kinds of anomalies in a software package are matters of *fact* and not matters of probability; however, since only a relatively small portion of a large program's documentation and response can ever be verified in a practical sense, the process of discovering anomalies appears to be a random process.

Repairing an anomaly requires study, software alteration, and then reverification. Because the differing kinds of anomalies exhibit a range of difficulty, and because human interaction is always required, the repair rate also appears to be a random quantity.

This article evaluates the average time required to discover and repair anomalies which appear randomly in testing. "Testing" here refers to the end-to-end tests performed after the programming is complete. Anomalies are assumed few enough that their discovery is not so rapid as to swamp out the efforts of the test team. The work bases its evaluations on a Markov model (Ref. 1) with time-independent parameters. The assumptions made about the discovery and repair process are: (1) the probability that a test will find a new anomaly is proportional to the number of undiscovered anomalies yet in the system, and (2) the probability that an anomaly remains unfixed decreases exponentially in the time since its discovery. Both of these assumptions are intuitively based, but appear to be supported well by empirical data (Refs. 2, 3, and 4).

The goal of this evaluation is not so much to model or characterize precisely the way anomalies are detected and removed, but to provide a schedule prediction tool and status

monitor for software project managers. Although statistical, the method does allow extrapolations to be made early in testing so that resource reallocations, if necessary, can be made to align completion dates with committed capabilities.

The method accommodates variations in work loads and multistage testing in different sub-operational environments.

II. Discovery of Anomalies

Let us suppose that a system possesses A (number unknown *a priori*) anomalies, and that a constant effort is being applied toward the discovery of these. Having found n of these (define this as state E_n), let us further suppose that the probability of finding the next (moving to state E_{n+1}) is proportional to the number yet remaining; that is, we shall assume that a transition from state E_n to E_{n+1} in the time interval $(t, t + \Delta t)$ will take place with probability $\lambda_n \Delta t = (A - n) \alpha \Delta t$ for some appropriate constant α. Feller (Ref. 1) gives the general form for the state probabilities $P_n(t)$ as a function of time:

$$P_0'(t) = -\lambda_0 P_0(t)$$

$$P_n'(t) = -\lambda_n P_n(t) + \lambda_{n-1} P_{n-1}(t) \tag{1}$$

The latter equation serves for $n = 1, \cdots, A$. The general solution may be found by straightforward application of the generating-function method, in which

$$\Phi(x, t) \overset{\Delta}{=} \sum_{n=0}^{A} P_n(t) x^n \tag{2}$$

yields a partial differential equation

$$\frac{\partial \Phi}{\partial t} = \alpha(x - 1)\left(A \Phi - x \frac{\partial \Phi}{\partial x}\right) \tag{3}$$

whose solution (Ref. 4) is

$$P_n(t) = \binom{A}{n} e^{-\alpha A t}(e^{\alpha t} - 1)^n \tag{4}$$

Having this expression, we may proceed to compute the mean time to discovery of the nth anomaly: The probability (density) that the nth anomaly is discovered at time T_n is the probability that $n - 1$ anomalies had been found by $T_n - \Delta t$, times the probability, $\lambda_{n-1} \Delta t$, that the transition $E_{n-1} \rightarrow E_n$

occurs during the last Δt interval. As $\Delta t \rightarrow 0$, we reach the limiting equation

$$p_n(T) = \lambda_{n-1} P_{n-1}(T) \tag{5}$$

The generalized moments of this density are then

$$\operatorname{avg}(T_n^m) = \int_0^\infty T^m p_n(T)\, dt$$

$$= \left(\frac{m!}{\alpha^m}\right)(A + 1 - n)\binom{A}{n-1}$$

$$\times \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{(-1)^k}{(A_0 + 1 - n + 1)^{m+1}} \tag{6}$$

The first two of these moments can be evaluated directly to produce mean and variance values

$$\overline{T}_n = \frac{1}{\alpha} \sum_{k=0}^{n-1} \frac{1}{A - k} = \frac{1}{\alpha}[\psi(A + 1) - \psi(A + 1 - n)]$$

$$\operatorname{var}(T_n) = \frac{1}{\alpha^2} \sum_{k=0}^{n-1} \frac{1}{(A - k)^2} \tag{7}$$

The ψ-function is the "digamma" function, Γ'/Γ, in Ref. 5. The mean-to-deviation ratio is thus

$$\frac{\operatorname{var}(T_n)}{\overline{T}_n^2} = \frac{\displaystyle\sum_{k=0}^{n-1} \frac{1}{(A - k)^2}}{\left(\displaystyle\sum_{k=0}^{n-1} \frac{1}{A - k}\right)^2} \tag{8}$$

The behavior of this ratio is shown in Fig. 1, and the normalized mean time to reach the nth anomaly is displayed in Fig. 2. The normalization in Fig. 2 is set to make the initial rates of anomaly discovery equal.

The first figure shows that the relative time variance to reach the nth anomaly decreases approximately as $1/n$, except

at values where n becomes appreciable to A, whereupon it rises, ultimately reaching the value

$$\frac{\text{var}\,(T_A)}{T_A^2} \approx \frac{1.645}{(0.57721 + \ln A)^2} \tag{9}$$

The minimum relative variance appears to occur at an n of about 80 to 85 percent of the total anomalies, and its value is about $1.5/A$.

This behavior infers that the discovery of anomalies up to about 85 percent of the total proceeds quite normally; the relative variance decreases inversely proportionally to the number found, as would be expected of a random process. The remaining 15 percent of the anomalies, however, not only require longer and longer times each to discover, but the variations in discovery times begin to grow disproportionately, as well.

The ratio $\overline{T}_n/\overline{T}_A$ represents the average relative completion status during anomaly discovery. Note in the graph of this ratio (Fig. 3), for example, that if there are 100 anomalies, then when 90 percent of them have been found, only about 44 percent of the total required time has elapsed. Therefore, if one could estimate the number of total anomalies by some figure A, he could then also estimate the time which would probably be needed to complete testing.

An approximate value for $\overline{T}_n$ is given by

$$\alpha \overline{T}_n \approx \ln x + \frac{x-1}{2\,(A+1)} + \frac{x^2-1}{12\,(A+1)^2}$$

$$- \frac{x^4-1}{120\,(A+1)^4} + \frac{x^7-1}{252\,(A+1)^6}$$

$$x \overset{\Delta}{=} \frac{A+1}{A+1-n} \tag{10}$$

A. Estimating the Unknown Parameters α and A

Upon differencing the average times to discover successive anomalies, one finds

$$\overline{t}_{n+1} \overset{\Delta}{=} \overline{T}_{n+1} - \overline{T}_n = \frac{1}{\alpha\,(A-n)} = \frac{1}{\lambda_n} \tag{11}$$

The expected time to uncover the final anomaly is $1/\alpha$, or A times as long as the discovery of the first.

By assuming a Markov process, we have thereby also assumed that the time increments t_n between discoveries are independent random variables. Given that $n-1$ anomalies have so far been uncovered, the probability function for the time required to find the nth follows (5), but may be solved in terms of t_n rather than T_n by moving the time origin to $T = T_n$ and redefining $A' = A - n$, $\lambda'_0 = \lambda_{n-1}$, in order that (1) may be used to find the new P_0 (t). Ultimately, we then find

$$p_n\,(t_n) = \lambda_{n-1}\,\exp\,[-\lambda_{n-1}\,t_n] \tag{12}$$

Because of independence, the probability density for observing the values $t_1, t_2, \cdots, t_n$ is merely the product of the corresponding densities (12), or

$$p\,(t_1, \cdots, t_n) = (\lambda_1 \cdots \lambda_{n-1})\,\exp\,[-(\lambda_0\,t_1 + \cdots + \lambda_{n-1}\,t_n)] \tag{13}$$

Equating the derivatives of (12) with respect to α and A to zero produces the maximum likelihood estimators $\hat{\alpha}$ and $\hat{A}$. These satisfy

$$\hat{A} - n\left[\sum_{k=0}^{n-1}\frac{1}{\hat{A}-k}\right]^{-1} = \frac{\displaystyle\sum_{k=1}^{n}(k-1)\,t_k}{\displaystyle\sum_{k=1}^{n}t_k}$$

$$\hat{\alpha} = \frac{\displaystyle\sum_{k=0}^{n-1}\frac{1}{\hat{A}-K}}{\displaystyle\sum_{k=1}^{n}t_k} \tag{14}$$

The solutions for $\hat{A}$ at several given values of n are displayed in Fig. 4 as a function of computations based on observed values t_k. Use of this $\hat{A}$ and the observed data yields $\hat{\alpha}$. These estimators were previously evaluated by Jelinski and Moranda (Ref. 6).

Note from the curves that when $\hat{A} \gg n$ there is very little precision in estimating $\hat{A}$; small variations in the computed ratio yield very large uncertainty regions in $\hat{A}$. Thus, the predictors $\hat{\alpha}$ and $\hat{A}$ are not very accurate unless A is fairly large and n is appreciable to A.

Wolverton and Schick (Ref. 3) report the amazing fidelity with which these estimators predicted failures in an actual application. After $n = 26$ anomalies had been discovered, they computed via (14) the estimates $\hat{A} = 31$, $\hat{\alpha} = 0.007/\text{day}$. They thus estimated that there were 5 errors remaining and that the time to discover the next would be about $1/5\hat{\alpha} = 24$ days. They record the fact that 5 errors were then found in later testing and that, after these, the estimates became $\hat{A} = 31.6$ and $\hat{\alpha} = 0.006/\text{day}$. They concluded with the projection that another error may yet remain in the system, but, if so, it would probably require $1/0.006 = 167$ days at the same level of effort to detect. They do not record, however, whether this error was ever actually found or not.

III. Repair-Only Model

Before going into a concurrent find-and-fix model, let us first address the repair-process statistics for a simplified situation in which the number of anomalies discovered is fixed. That is, we presume that there are n initially known anomalies. The rate at which anomalies are repaired is presumed to be constant and independent of the number yet open, as long as there are anomalies left to work on; more precisely, we assume that a transition from state E_{m-1} (i.e., $m - 1$ have been repaired) at time $t - \Delta t$ to state E_m (i.e., m repaired) at time t will take place with probability $\mu \Delta t$ for some appropriately chosen constant μ, so long as $m \leq n$. Feller (Ref. 1) gives the governing equation for the state probabilities as a function of time:

$$P'_m (t) = -\mu P_m (t) + \mu P_{m-1} (t) \quad \text{for } m = 0, \cdots, n - 1$$

$$P'_n (t) = \mu P_{n-1} (t) \tag{15}$$

The solution of these is

$$P_m (t) = (\mu t)^m e^{-\mu t} / m! \quad \text{for } m = 0, \cdots, n - 1$$

$$P_n (t) = e^{-\mu t} \sum_{k=n}^{\infty} (\mu t)^k / k! \tag{16}$$

The equation for the probability (density) that the nth anomaly will be repaired at time t is similar to (5):

$$p_m (T) = \mu P_{m-1} (t) \tag{17}$$

The mean time $\overline{T}_m$ to repair m anomalies and the variance about this value are straightforwardly found by integration similar to (6), yielding

$$\overline{T}_m = m/\mu$$

$$\text{var} (T_m) = m/\mu^2$$

$$\frac{\text{var} (T_m)}{(\overline{T}_m)^2} = 1/m \tag{18}$$

This model thus predicts that all n of the known anomalies will be repaired uniformly in time with growing absolute, but decreasing relative, uncertainty. Each anomaly requires an average time $1/\mu$ to repair, with variance $1/\mu^2$.

We may argue that this model has validity for a certain span of time. Whenever the initial find-rate (λ_0 in Section II) exceeds μ, the constant fix-rate, this span of time extends from time zero until such time as there is a significant likelihood that all anomalies so far discovered will have been cleared. That is, as long as there are anomalies to work on, the rate remains unaffected by their number.

A. Estimation of the μ Parameter

The μ parameter is, of course, generally unknown *a priori*, and therefore must be estimated to be of practical utility. We expect, subject to the assumptions above, that the times t_k between repairs will take the cumulative form $\mu \overline{T}_m = m$. Having measured

$$T_m = \sum_{k=1}^{m} t_k \quad \text{for } m = 0, \cdots, M$$

we may evaluate parameters $\hat{\mu}$ and b for the line $\hat{T}_m = m/\hat{\mu} + b$ with least mean-square error from the observed values:

$$\hat{\mu}^{-1} = \frac{6}{m (m + 1) (m + 2)} \sum_{j=1}^{n} (m + 1 - j) t_j \tag{19}$$

$$b = \frac{1}{(m + 1) (m + 2)} \sum_{j=1}^{m} (m + 1 - j) (m + 2 - 3j) t_j$$

These estimators are unbiased (i.e., $E(b) = E(1/\hat{\mu}) = 0$) and the variation in $1/\hat{\mu}$ is reflected by

$$\text{var}(\mu/\hat{\mu}) = (6/5)\frac{m^3 + 4m^2 + 6m + 4}{m(m+1)(m+2)^2} \qquad (20)$$

$$\sim \frac{6}{5m} \text{ for large } m$$

B. Zero-Defects Estimation

The condition in which all known anomalies have been fixed is known as "zero defects." Its more precise characterization when discovery and repair are concurrent will be addressed a little later; for the moment, however, we may estimate when the first zero-defects condition will occur, on the average, using the models of this and the preceding sections. It will occur for $n < A$ when the repair time for n anomalies equals the discovery time for these anomalies plus the extra time needed to fix the last-discovered anomaly, so long as this extra time is not long enough to discover the next anomaly. This situation is described by

$$\frac{A}{(n-1)} \sum_{k=0}^{n-1} \frac{1}{A-k} = \frac{A\alpha}{\mu} < \frac{A}{A-n} \qquad (21)$$

Estimated values for $\hat{\alpha}$, $\hat{A}$, and $\hat{\mu}$ thus provide a means for predicting the first zero-defects time. The prediction is perhaps most easily determined graphically, as indicated in Fig. 5. On reaching zero defects, there remain $A - n$ anomalies yet to be found. The ratio $r \triangleq (A - n)/A = 1 - n/A$, i.e., the fraction of the anomalies undiscovered at the time of zero defects, is a function of the initial find-fix-rate ratio $\lambda_0/\mu = A \, \alpha/\mu$, governed by (21). The behavior of r vs λ_0/μ is shown in Fig. 6.

IV. Concurrent Discovery and Repair

This section formulates the equations which govern discovery and concurrent repair of anomalies; however, closed-form solutions for such things as the average number of open anomalies at any time, the average time to a zero-defect condition, etc., are not yet known. The equations which describe the probability $P_{n,m}(t)$ that n of A anomalies have been discovered by time t, and m have been fixed, are a generalization of (1) and (15):

$$P'_{0,0} = -\lambda_0 P_{0,0}$$

$$P'_{n,0} = -(\lambda_0 + \mu - \alpha n) P_{n,0} + [\lambda_0 - \alpha(n-1)] P_{n-1,0}$$

$$\text{for } n > 0$$

$$P'_{n,m} = -(\lambda_0 + \mu - \alpha n) P_{n,m} + [\lambda_0 - \alpha(n-1)] P_{n-1,m}$$

$$+ \mu P_{n,m-1} \text{ for } 0 < m < n < A$$

$$P'_{n,n} = -(\lambda_0 - \alpha n) P_{n,n} + \mu P_{n,n-1} \text{ for } n > 0 \qquad (22)$$

Of course, these equations can be solved recursively as, for example,

$$P_{0,0}(t) = e^{-\lambda_0 t}$$

$$P_{1,0}(t) = \lambda_0 e^{-\lambda_0 t} [1 - e^{-(\mu-\alpha)t}]/(\mu - \alpha) \qquad (23)$$

and so on.

By extending the generating functions described by (2) to accommodate the present case,

$$\Phi(x,y,t) \triangleq \sum_{n=0}^{A} \sum_{m=0}^{n} P_{n,m}(t) x^n y^m$$

$$\Delta(x,y,t) \triangleq \sum_{n=0}^{A} P_{n,n}(t) x^n y^n \qquad (24)$$

Then the system displayed in (22) can be written in the partial differential equation form

$$\frac{\partial \Phi}{\partial t} = \alpha(x-1)\left[A\Phi - x\frac{\partial \Phi}{\partial x}\right] + \mu(y-1)(\Phi - \Delta) \qquad (25)$$

As may be noted, the case with $y = 1$ is identical to Eq. (3), since

$$P_n = \sum_{m=0}^{n} P_{n,m}$$

However, a direct solution to (25) has not been forthcoming since Δ is unknown until Φ is found and vice versa. Thus,

53

further work is needed to solve or approximate the statistics of this more general case.

V. Effects of Variation in Effort

The models so far described have assumed that constant levels of effort were being applied to finding and fixing of anomalies. This assumption is tantamount to equating time and expended effort in the equations previously derived. In actuality, the effort profile may be variable for a number of reasons, among which are manpower phasing, availability of computer resources, and availability of software resources.

It is typical that effort during the early testing is at a rather lower level than later on, because "things are getting up to speed." Effort is being put into planning, coordination, and resource acquisition rather than actual testing. Toward the end, effort may also drop again, to the level supported by sustaining personnel. This phenomenon is illustrated in Fig. 7. In this figure, one person is applied for 7 days, 7 for 25, 4 for 22, and 2 for 30 days.

These efforts tend to distort the anomaly vs time curves shown previously. However, we may compensate for these by replacing the time variable t in previous calculations by the integral of the work-level profile, $\int w(t)\, dt$. An illustration of this principle appears in Fig. 8.

Conversely, one may compensate in reverse; that is, the anomaly vs effort behavior may be plotted and analyzed using the previous estimators, then translated via the projected work-level profile to produce estimates of anomaly status at future dates.

VI. Cascaded Testing

Even when work profile effects are factored in, it is frequently the case that the discovery of anomalies takes place in varying environments for supposed economic reasons. Certain anomalies may not be discoverable in one environment, but perhaps in another, due to the differing software or hardware configurations.

For example, if a set of real-time programs are first tested outside the real-time environment, those anomalies which are due to the real-time interaction among programs are perhaps undiscoverable until the programs are integrated into their true operational environment.

In such situations, only a portion A_1 of the total anomalies will be found during the first testing, no matter how long testing goes on. If the second stage takes place in the opera-tional environment, then $A - n$ become discoverable during this stage ($n \leqslant A_1$ being the number discovered in stage 1). Typical projects generally find only about half of the total anomalies in this first state.

This phenomenon is illustrated in Fig. 9, assuming the same constant level of effort α, 100 total anomalies, and $A_1 = 50$ findable during the first stage. Switching to the operational environment takes place at 45 errors, or 90 percent of those that can be found. As may be seen, multistage testing may take a significantly longer time (31 percent in the illustrated case) to find all anomalies.

Moreover, the test time requirements on usage of the operational facility are about the same (52 vs 46 days) in either case; that is, there is probably no significant savings in the use of the more expensive facility! Therefore, unless there are other overriding constraints which mandate multistage testing, this form of "bottom-up" anomaly discovery plan cannot be cost effective.

VII. Conclusions

Although a general solution to the discovery and concurrent repair model may not be known at this time, solutions to the simpler underlying component models are given, valid when discovery and repair processes are non-interacting. These solutions permit the estimation of model parameters and the subsequent forecast of project completion dates. The accuracy of forecasting is also obtainable, to evaluate the need for contingency planning.

Inasmuch as the models studied here probably represent a fairly optimistic view of anomaly data, the results of this article tend to point out those things which *cannot* be gleaned from such data, perhaps more than revealing what things can accurately be deduced.

For example, the curves in Fig. 4 allow predictions of the total number of anomalies to be made. The fundamental truth contained in that figure is that an accurate estimation of the total number of anomalies requires testing a system which has many anomalies to begin with, of which a significant fraction must already have been found. Figure 2 echoes this fact.

Having estimated that a certain significant fraction, say, 90 percent, of the anomalies have already been found, Fig. 3 permits one to estimate how much time is yet required to find the remaining 10 percent. If there are 1000 estimated in all, the remaining 100 will require more than twice as much time to uncover as has already been spent! But the lesson here is that, although disproportionate, that time requirement is not unnatural or unreasonable.

Figure 1 shows that once 80 to 85 percent of the anomalies have been found, the expected discovery rate is subject to wider and wider variations; if schedules are being set or forecasts being made requiring contingencies or reapportionment of resources, such variances need to be taken into account in order to avert disasters. The lesson from this figure is that it is not realistic to believe the last 15 percent of anomalies will even progress as smoothly as the first 85 percent, in addition to requiring a disproportionate time.

This article has shown how the effects of a variable work profile can be used to influence the rate of discovery or repair.

Also, it has shown that cascaded-stage testing is probably more expensive than single-stage end-to-end tests in the final operational environment.

Figure 10 is an actual anomaly history; unfortunately, the work profile is not available, so a detailed comparison with the theory presented here is not possible. However, the reader will note that all of the predicted elements are present: the effects of low-level effort during the start of testing, the decreasing rate of anomaly discovery in the sub-operational testing prior to transfer, the increased rate thereafter, and the ultimate leveling off as the testing continued.

References

1. Feller, W., *An Introduction to Probability and Statistics,* John Wiley & Sons, Inc., New York, 1957.

2. Craig, G. R., *et al.,* "Software Reliability Study," TRW Systems Group Report AD-787-784 for Rome Air Development Ctr, Griffiss AFB, N.Y., Oct. 1974.

3. Wolverton, G., and Schick, R., "Assessment of Software Reliability," Proc. 11th Meeting, German Operations Research Society, Hamburg, Sept. 1972.

4. Musa, J. D., "A Theory of Software Reliability and Its Application," *IEEE Trans. Software Eng.*, Vol. SE-1, No. 3, pp. 312-327, Sept. 1975.

5. Abramowitz, M. (ed.), *Handbook of Mathematical Functions,* National Bureau of Standards, U.S. Dept. of Commerce, Applied Math Series 55, June 1964.

6. Jelinski, Z., and Moranda, P. B., "Software Reliability Research," Conf. on Statistical Methods for the Evaluation of Computer Systems, Brown University, Providence, R.I., Nov. 1971.
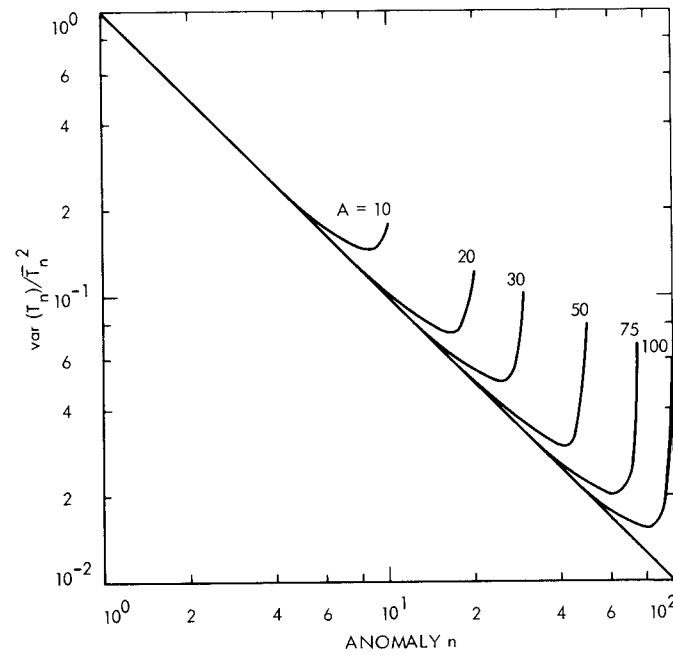
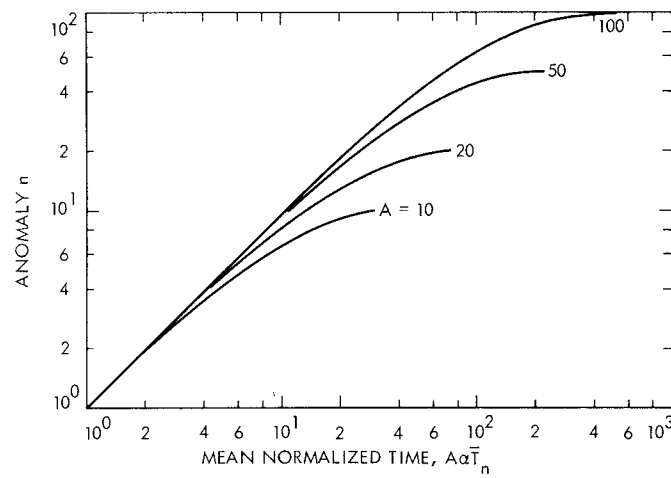**Fig. 1. Normalized variance in time to detect nth anomaly**

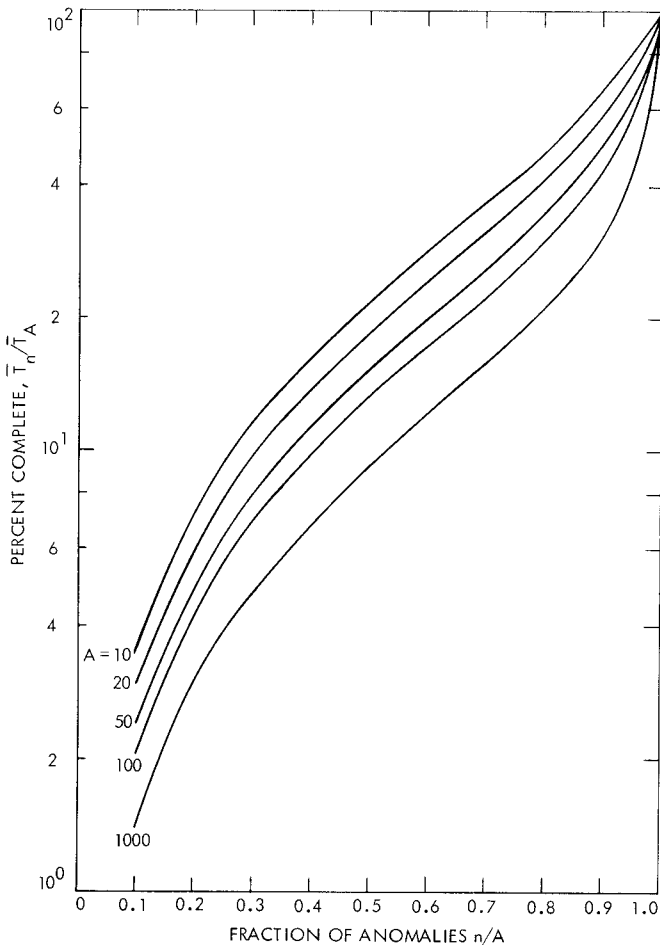**Fig. 2. Normalized mean time to reach nth anomaly**

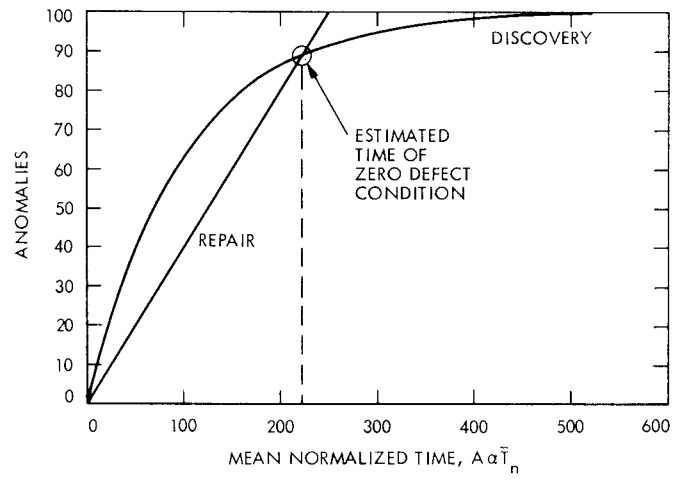Fig. 3. Normalized expected completion ratio vs fraction of anomalies discovered

Fig. 4. Measurement ratio vs maximum likelihood estimator for number of anomalies, given the number n of anomalies discovered so far

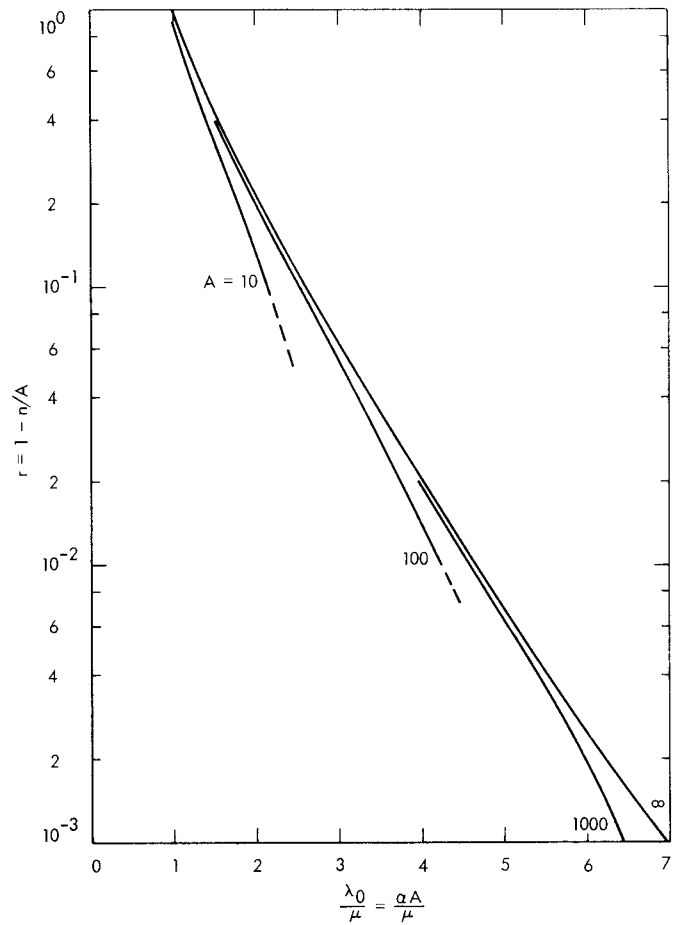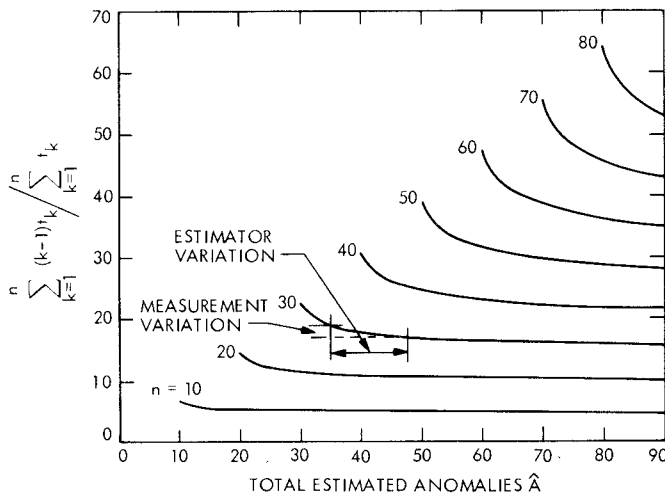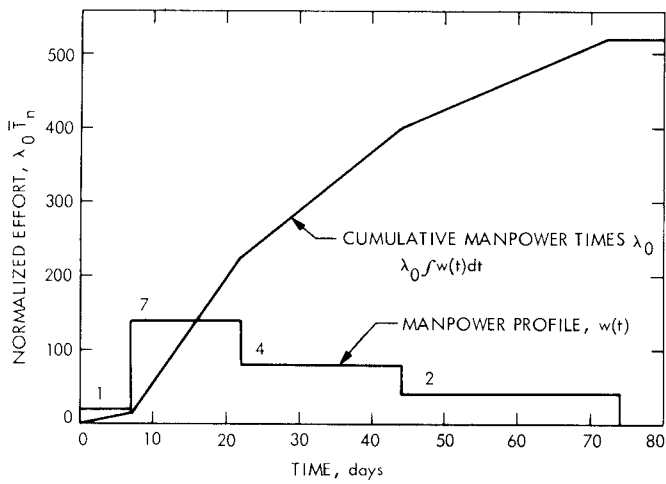Fig. 5. Plot of Eq. (7) for $A = 100$ and Eq. (18) to estimate time of zero defects for $\lambda_0/\mu = 2.5$

Fig. 6. Mean time intersection approximation to fraction of anomalies remaining at time of first zero-defect condition

Fig. 7. Work-level profile and cumulative effort normalized by initial discovery rate of $\lambda_0 = 2$ anomalies per man-day
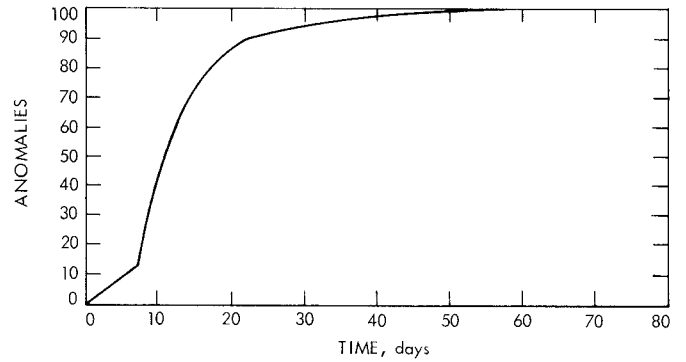

Fig. 8. Effect of work-level profile on discovery of anomalies for in Fig. 7 and $\lambda_0 = 2$ anomalies per man-day
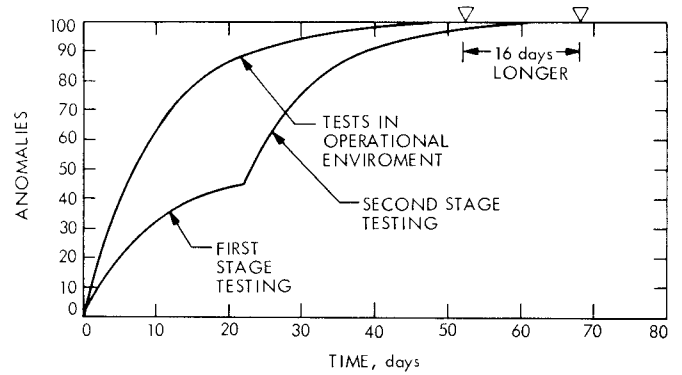

Fig. 9. Cascaded discovery of 100 anomalies in 2 stages where 50 anomalies only were visible during first stage for same $\alpha = 0.1$
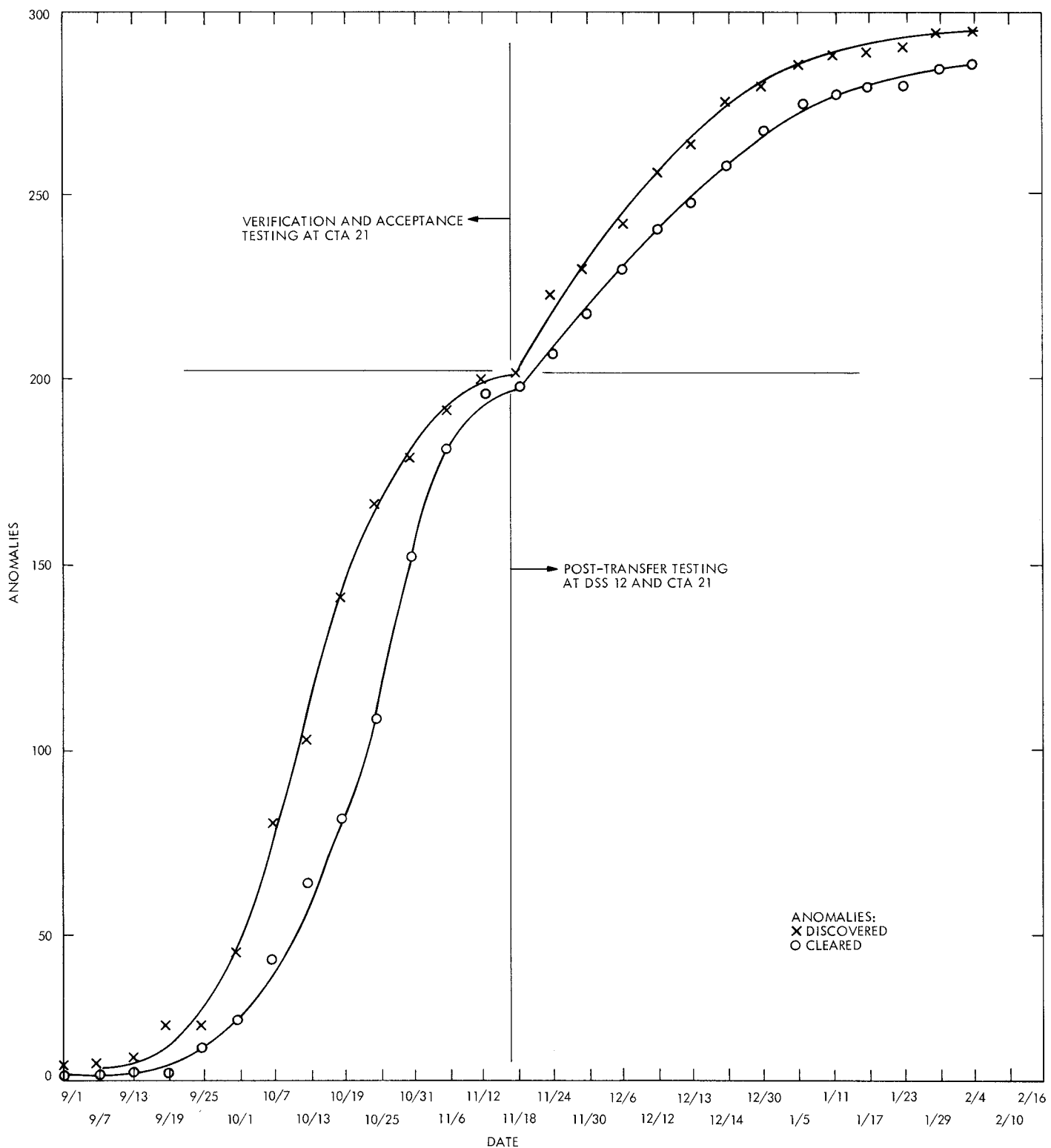
Fig. 10. Mark III Data System software anomaly history