Stochastic Models for Software Project Management

R. C. Tausworthe
DSN Data Systems Section

This article presents a method for determining the number and characteristics of milestones to be achieved during a development project in order that effective monitors of progress can be provided. Projections of progress data lead to estimates of the completion with determinable accuracy, but accuracy imposes a requirement that the number of milestones be inversely proportional to the estimate-error variance, and that the milestones themselves be defined in such a way that each represents approximately the same level of effort to complete.

I. Introduction

The progress in development of a piece of computer software (a program and its documentation) is, in many ways, like the classical random-walk problems associated with birth and death processes (Ref. 1). If a project has identified a number of milestones M to be achieved during the course of development, then the number of milestones achieved by a certain date can be modeled by a birth process in which the population never exceeds M. During acceptance testing of a "completed" program, anomalies are discovered in a similar birth process, whose population never exceeds some number A, the total number of anomalies in the program. As anomalies are repaired, the joint process describing anomalies found versus anomalies repaired is a birth-death process whose limiting condition is (hopefully) zero unrepaired anomalies.

During the course of development, project management requires effective monitors on which the health of the project can be assessed and corrective action initiated, should that assessment so indicate. Such monitors as cumulative milestones and anomaly status have been used with success in the past, not only in software developments, but probably in almost all endeavors involving development activity.

This article explores the behavior of milestone-completion processes, and a later article will discuss anomaly discovery/repair processes. Both will be simplified for the sake of analysis, in that they will assume that uniform Markovian statistics apply. That is, the factors which influence the processes are not time-origin-dependent, and future statistics depend only on the current completion status of the process; i.e., at a given status, the remaining behavior of the process does not depend on any of the past history up to that status point. Should statistics change (by management decisions, for example), the remaining process to completion can be analyzed using only the new statistical parameters.

Uniformity of statistics depends on inertia over a project lifetime; it discounts such things as improvement of progress by learning and degradation of progress by attrition, as factors that average out. The assumption of uniformity is one that permits statements to be made with predefined precision in the form, "If the team keeps progressing as it has so far, then . . .".

These monitors allow projections of completion dates to be made rather handily and accurately with only a minimum number of assumptions necessary on the underlying causal relationships within the development process. This is due, in a large development, to the multitude of factors which combine to make the progress appear stochastic in the first place. By the central limit theorem (Ref. 1, pp. 228–233), each such process appears normally distributed.

These are not necessarily optimistic assumptions to be making about processes involving expenditures of large amounts of money and other resources. Reviewing the progress via such monitors on a regular basis will reveal departures from theory rather dramatically, and thereby permit the management function to act appropriately. In fact, it is this feedback and corrective action on the part of management that tends to align the model with reality.

For example, should the series appear to exhibit some early adverse nonuniform statistics, management can take corrective action to bring the progress back into uniformity, so as to fall within the negotiated limits for the projected completion. Should the series appear to be favorably nonuniform, then management can again restore the uniformity by removal of resources, if appropriate to do so.

The point is that monitors based on theoretical models are quantitative tools that can be applied effectively in addition to the qualitative judgements normally necessary for management.

Models permit management in planning to make certain assumptions concerning the productivity of a team, thereby arriving at a preliminary schedule. During early development, the assumptions can be calibrated by actual measurements, and more realistic schedules drawn up. Management can thus preplan activities with known precision and can either utilize leverage as needed to maintain the plan, or renegotiate plans and capabilities.

As a final point in this introductory material, the theoretical model aids in defining the types of milestones to be monitored, and their number. Since events to be monitored are presumed by the theory to have certain statistical properties, then the accuracy of the results will be influenced by the accuracy with which the actual events conform to these assumed statistics. The assumptions concerning the statistics are simple: normally distributed events, with uniform, history-independent time behavior. The law of large numbers helps keep the normality assumption approximately true, and project inertia and feedback tend to keep the process uniform. The proper definition of events for history-independence then remains as the principal challenge to the event definition process.

II. Schedule Prediction Model

For schedule prediction, let us assume that it is known *a priori* that the project will be completed after M milestones have been achieved. These milestones correspond to all the various tasks which have to be accomplished, and once accomplished, are finished forever (that is, some later activity does not reopen an already completed task; if such is the case, however, it can be accommodated by making M larger, to include all such milestones as separate events). The number M, of course, may not be known precisely *a priori*, but may be estimated via a preliminary design phase. Any uncertainty in the value of M will translate to an uncertainty in the estimated completion date, and we will treat this possibility a little later.

Let us now further suppose that at regular ΔT intervals (e.g., weekly, biweekly, or monthly) the numbers of milestones k reported as being achieved follows a time-independent statistical distribution function of the binomial form (Ref. 1, pp. 136–142);

$$P(k) = b(k; n, p) \overset{\Delta}{=} \binom{n}{k} p^k (1 - p)^{n-k} \tag{1}$$

The reported number k of milestones achieved each ΔT period is then a random variable whose mean value m and variance σ^2 are given by well-known formulas (Ref. 1, pp. 209, 214):

$$\begin{aligned} m &= pn \\ \sigma^2 &= np(1 - p) = mq \end{aligned} \tag{2}$$

where we use $q \overset{\Delta}{=} 1 - p$ hereafter.

We use the binomial distribution function (1) above for two reasons: First, for very nominal values of m and σ^2, the binomial distribution well approximates the normal distribution function (Ref. 1, pp. 168–173):

$$P(k) \approx \frac{1}{\sigma (2\pi)^{1/2}} \exp\left\{\frac{-(k-m)^2}{2\sigma^2}\right\} \tag{3}$$

thus fulfilling the intuitive requirement previously mentioned. The second reason is that the distribution (1) describes the probability of achieving exactly k out of n equally likely goals in which the figure p is associated with the success of each event.

Thus, if a set of M milestones can be defined for a project, a maximum of n to be achieved each ΔT reporting period, if each milestone represents the accomplishment of a task with approximately the same degree of difficulty, and if milestones scheduled in one ΔT period, but missed, can be rescheduled for future ΔT periods without altering the statistics, then the $P(k)$ form supposed above is a faithful description of the progress achievement process in the project.

The binomial distribution lends itself easily to solution and to interpretations. It is exactly the same formula which governs the statistics of obtaining k heads out of n coin tosses, using a coin that turns up heads with probability p each toss (the average number of heads being thus $m = np$). All the theoretical results known for coin tossings thus apply to our scheduling model, suitably interpreted. The "fine structure" of a productivity model for achieving milestones is thus simulated by corresponding a "toss" with a "trial for achieving a milestone," each with probability p which can in turn be related back to the m and σ^2 of the normal distribution. Each trial or *step* will require an average time

$$\Delta t = \frac{\Delta T}{n} = \frac{\Delta T\,(m - \sigma^2)}{m^2} \tag{4}$$

III. Progress Averages

The correspondence to coin tosses permits us to state a number of known results immediately: First, the time T_k to reach the kth milestone has average value

$$\bar{T}_k = k/p \text{ steps}$$

$$= k\Delta t/p = k\Delta T/pn \text{ units of time} \tag{5}$$

and a variance about this value of

$$\operatorname{var}(T_k) = \bar{T}_k^2\, q/k \tag{6}$$

We may also compute the average cumulative progress $\bar{\pi}_s$ in milestones achieved after any particular number of steps s as the expression

$$\bar{\pi}_s = \sum_{k=0}^{M-1} kb\,(k; s, p) + Mp \sum_{t=0}^{s-M} b\,(M-1; M-1+t, p) \tag{7}$$

The first terms above represent the progress value k weighted by the probability that progress is at the kth milestone after s steps; the final terms represent the progress value M (completion) weighted by the probability that the M milestones were accomplished on or before step s (being the sum of the probabilities that M milestones were first reached on the $(M+t)$th step, for $t = 0, \ldots, s - M$ and $s \geqslant M$).

A closed-form formula for $\bar{\pi}_s$ is not known in the case $s > M$, but the sum may be readily approximated using the normal approximation (3) and integrating, rather than summing, to yield

$$\bar{\pi}_s = sp \text{ for } s \leqslant M$$

$$\approx \frac{sp}{2}\operatorname{erfc}\left[\frac{sp-(M-1)}{(2pqs)^{1/2}}\right] - \left(\frac{pqs}{2\pi}\right)^{1/2} \exp\left[-(M-sp)^2/2pqs\right]$$

$$+ M\left\{1 - (1/2)\operatorname{erfc}\left[\frac{sp-(M-1)}{(2q\,(M-1))^{1/2}}\right]\right\} \text{ for } s > M > 10 \tag{8}$$

At $s = M/p$ (the average time to project completion) the average progress is approximately

$$\bar{\pi}_{M/p} \approx M - \left(\frac{Mq}{2\pi}\right)^{1/2} = M\left[1 - \left(\frac{q}{2\pi M}\right)^{1/2}\right] \tag{9}$$

From (9) we may note that for $M \geqslant 16$, the average progress (taken over many projects) will show only at least 90% completion, even though the average project will have completed by this time!

The variance on π_s is

$$\operatorname{var}(\pi_s) = q\bar{\pi}_s \text{ for } s \leqslant M \tag{10}$$

The expression for this variance when $s > M$ is too complicated to be enlightening.

IV. Scheduling With Accuracy When *p* Is Known

The ability to project schedules with accuracy based on the foregoing milestone-achievement model requires only two constants, m and p, and the definition of M milestones, which may be achieved in increments having the same likelihood distributions proposed in (1) or (3). No assumptions have been made relative to the precedence of milestones, or how individual milestones are placed on the schedule. It is only the cumulative number which enters the picture so far. Figure 1 shows what the typical achievement chart should look like.

If we suppose that p and m are known exactly, *a priori*, then from Eqs. (5) and (6), the relative accuracy with which this model predicts project completion time (as a one-sigma event) is $\epsilon = (q/M)^{1/2}$. Thus, to predict a completion time T_M within a factor ϵ requires

$$M > q/\epsilon^2 = \sigma^2/m\epsilon^2 \qquad (11)$$

Recall that it was earlier mentioned that the assumed milestone achievement distribution is the same as if a maximum of n milestones were scheduled to be completed during each ΔT period, but only some lesser number k with average value $m = np$ will actually be achieved as scheduled. Milestones scheduled but unachieved in one ΔT period are then "slipped," the schedule reorganized with n milestones again scheduled for accomplishment in the next ΔT period, and so on, until completion. This type of schedule will be referred to as "maximum performance" schedule. The parameter p is the probability that a given milestone will be achieved in ΔT, and q is the probability that it will slip, to be rescheduled in some future period.

The original (unslipped) maximum-performance schedule shows the completion date after only M steps of length $\Delta t = \Delta T/n$. However, slippages lengthen this to $\bar{T}_M$ on the average and to $\bar{T}_M\ [1 + (q/M)^{1/2}]$ as a "1 standard deviation" event. In order that an original schedule based on maximum performance be correct within a relative precision factor ϵ, it is necessary that

$$M \geqslant \frac{1 - p}{[(1+\epsilon)p - 1]^2} = \frac{q}{[\epsilon - (1+\epsilon)]q^2}$$

$$p > \frac{1}{1+\epsilon}$$

$$q < \frac{\epsilon}{1+\epsilon} < \epsilon \qquad (12)$$

That is, the slip probability q must be no larger than about ϵ, and there must be enough milestones so as to make predictions fall within the desired precision. These relationships are shown in Fig. 2.

V. Estimation of the *pn* Parameter

It is seldom the case that m and σ^2 (or n and p) are known before a project begins, although the process of generating an initial schedule makes an implicit estimation of these parameters as a matter of course. More accurate values may be estimated once the project has begun by tabulating the progress in cumulative milestones achieved, as depicted in Fig. 3. If we let k_i for $i = 1, \ldots, r$ be the individual accomplishments for each of the ΔT reporting periods up to the rth, then the best-fit line (least-square-error) to the cumulative progress up to that time is

$$\hat{\pi}_r \overset{\Delta}{=} \hat{p}nr + b \qquad (13)$$

in which the parameters $\hat{p}$ and r are to be computed from observed data by

$$\hat{p} = \frac{6}{nr(r+1)(r+2)} \sum_{j=1}^{r} (r+1-j)jk_j$$

$$b = \frac{1}{(r+1)(r+2)} \sum_{j=1}^{r} (r+1-j)(r+2-3j)\,k_j \qquad (14)$$

These parameters have mean values given by

$$E(\hat{p}) = p$$

$$E(b) = 0 \qquad (15)$$

so that $\hat{\pi}_r$ is an unbiased estimator of the mean time to achieve a given milestone progress. The mean and estimated-mean time to completion, $\bar{T}_M$ and $\hat{T}_M$ then satisfy

$$M = p\bar{T}_M = \hat{p}\hat{T}_M + b \qquad (16)$$

which provides the approximate estimation-error value

$$\hat{\epsilon} \overset{\Delta}{=} \frac{\bar{T}_M - \hat{T}_M}{\bar{T}_M} = \left(\frac{\hat{p} - p}{p}\right)\frac{\hat{T}_M}{\bar{T}_M} + \frac{b}{M} \approx \left(\frac{\hat{p} - p}{p}\right) + b/M \qquad (17)$$

within first-order effects. The average estimation error is zero (within first-order effects). The variance computation for $\hat{\epsilon}$ is straightforward, though somewhat lengthy, leading to bounds that are independent of M:

$$\frac{q}{pnr} F_{\min}(r) < \text{var}(\hat{\epsilon}) < \frac{q}{pnr} F_{\max}(r) \qquad (18)$$

Both $F_{\min}$ and $F_{\max}$ are only slightly more than unity; these bounds and the ratio $F_{\min}/F_{\max}$ are shown in Fig. 4.

The estimated completion date cannot be estimated with very high accuracy early in the project (when r is small) unless $pn = m$, the average number of milestones achieved per ΔT period, is large, or unless q is very small. The denominator value, pnr, is the expected number of accomplished milestones up to and including the rth reporting period. The accuracy in estimating the time to accomplish the kth milestone, given by (6), is thus about the same as the accuracy for drawing the best-fit line through the observed data (within the factor F).

VI. Scheduling With Accuracy Using Estimates of *p*

The uncertainty with which the completion date can be predicted at the rth report springs from two sources: variance in the value of $\hat{p}$ to be used to estimate p and variance in the completion date due to p being other than unity:

$$\text{var}(T_M) = \bar{T}_m^2 (q/M)(R/r)\,[F(r) + r/R]$$

$$< 2.2\,(\bar{T}_M^2\, q/M)(R/r) \qquad (19)$$

where $R \triangleq M/m$, the average number of reporting periods to completion. When $r << R$, the prediction accuracy is, of course, dominated by the first term of the two; at any report r, the completion date variance is bounded according to the relation given above.

If it is therefore required, as before, to estimate the completion date to within an error factor ϵ, by a given report period r_0, then the total number of milestones M must satisfy

$$M > (q/\epsilon^2)(R/r_0)\,[F(r_0) + r_0/R] \qquad (20)$$

For good measure, we should probably have used the upper bound in (19), or $M > 2.2(q/\epsilon^2)\,(R/r_0)$ to set the number of milestones. The value of q to be used in (19) and (20) is a guess, and therefore should be more pessimistic (larger) than actual. In case no knowledge of q is to be had, we can always require $M \geqslant (2.2/\epsilon^2)\,(R/r_0)$ total milestones.

The refinement to the considerations given previously concerning estimating the deviation from an original maximum-performance schedule to within a factor of ϵ takes the same form as (12), except that M must be increased by a factor $[F(r_0) + r_0/R]\;(R/r_0)$:

$$M > \frac{q\,(R/r_0)\;[F(r_0) + r_0/R]}{[\epsilon - (1+\epsilon)\,q]^2}$$

$$q < \frac{1}{1+\epsilon} \qquad (21)$$

Figure 3 thus illustrates the constraints on M and ϵ when M is properly scaled.

Note that q is really needed only to judge how many milestones will be needed at the outset. A best-fit line to cumulative slip-statistics, using a proper reinterpretation of (14) will yield a $\widehat{qn}$, if desired; together with $\widehat{pn}$, all three parameters, $\hat{p}$, $\hat{q}$, and n (actually $\hat{n}$) can be found.

VII. Effect of Uncertainties in *M*

Up to this point, we have assumed M was known and fixed; in actuality, only an estimate of the total number of milestones may be known, perhaps by way of a preliminary, or architectural, design phase. The translation of a ΔM to a ΔT_M along the $p\bar{T}_k$ mean-time-to-completion-of-k-milestones line, coupled with estimation uncertainty in $\hat{p}$ and random fluctuations in the progress, leads to a first-order-effect approximate value for the total relative error in the time to completion, estimated at the rth report:

$$\text{var}(T_M/\bar{T}_M) \approx (\Delta M/\bar{M})^2 + (q/\bar{M})\,(R/r)\;[F(r) + r/R] \qquad (22)$$

In (22), ΔM is the standard deviation of M, and $\bar{M}$ is the mean value of M. Because $(\Delta M/\bar{M})$ appears squared in this expression its effect may not be felt so directly as the other terms contributing to schedule variance.

VIII. Conclusion

The progress of a development team is characterized by milestones achieved; whenever milestones can be defined in such a way that the expected number and variance of accomplishments is the same each for each status report, then the model explored in this article applies.

The main conclusion of this article is that schedule prediction accuracy is attainable only when a sufficient number of milestones to be achieved have been defined. The number of milestones needed is at least inversely proportional to the desired estimation error variance, and even more drastic than this if conformance to a maximum-performance schedule is attempted. It is therefore both necessary and important to refine tasks and to generate a detailed work-breakdown structure (WBS) rather carefully, if monitoring accuracy is the aim.

The generation of a schedule from the WBS should then proceed to allocate a constant number m of milestones for each reporting interval, m being the believed mean achievability during such intervals.

Reference

1. Feller, W., *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, 1950.

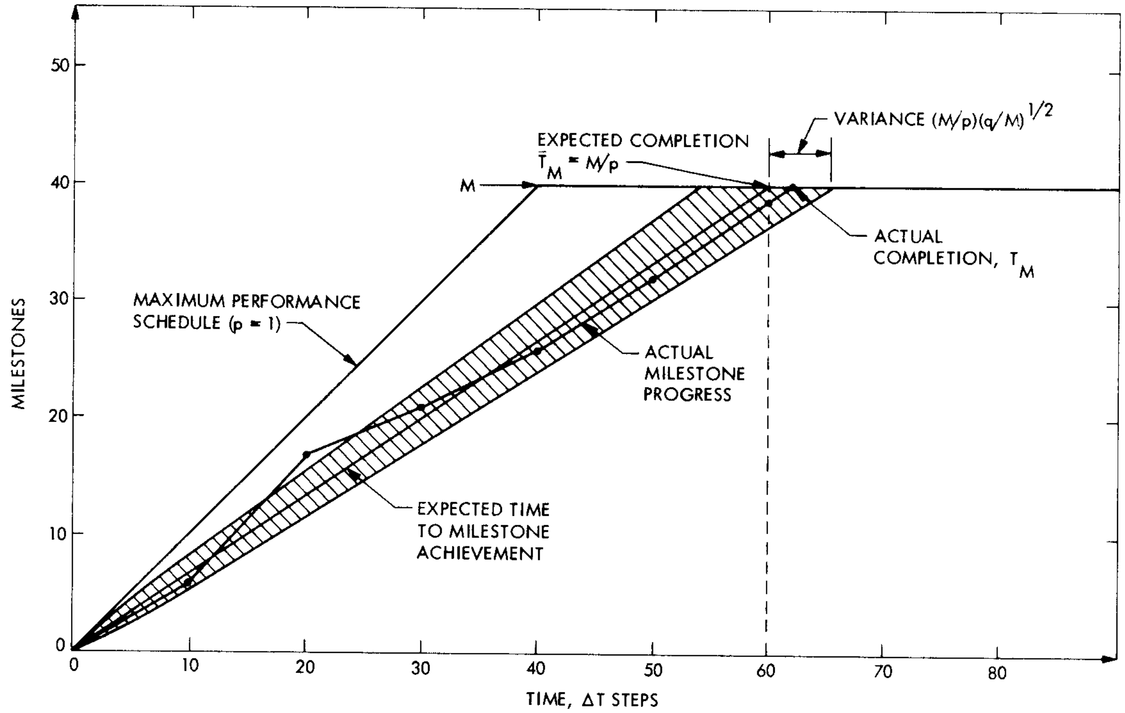

Fig. 1. Cumulative milestones achieved as a function of time. Case illustrated has parameters $n = 10$, $p = 2/3$, $M = 40$

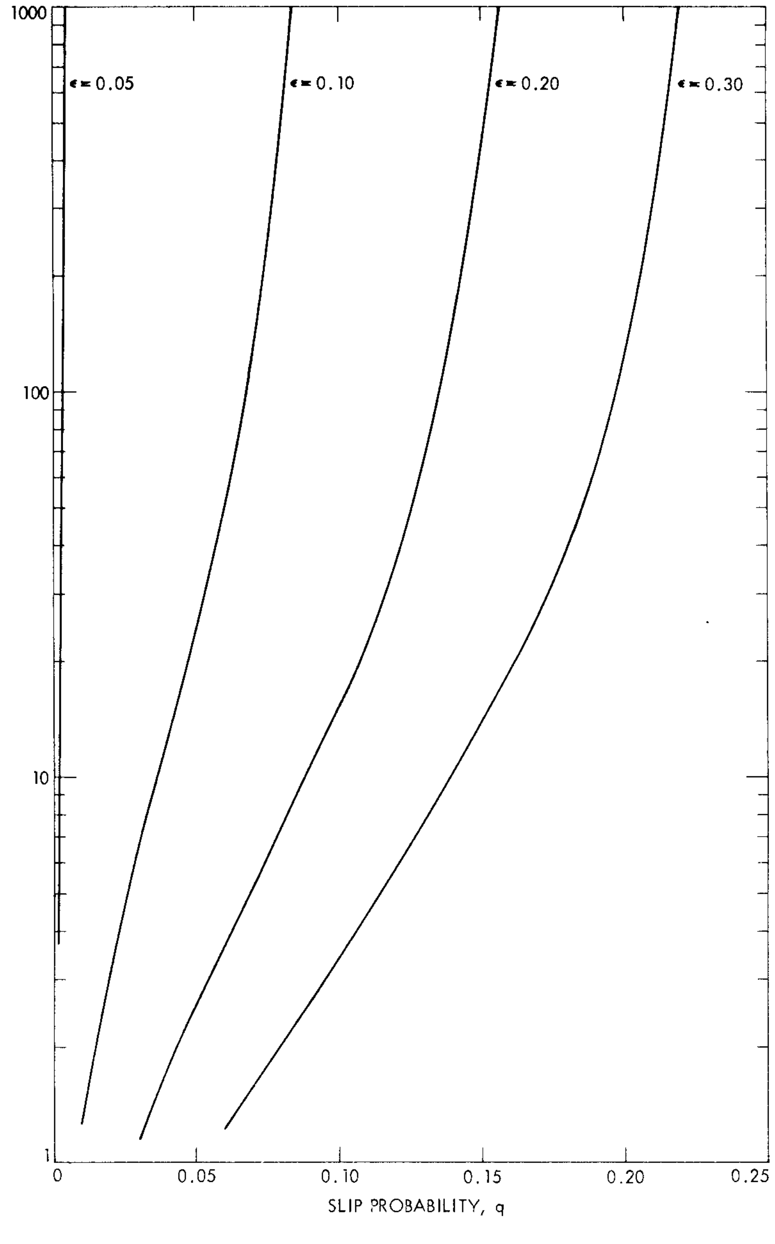

Fig. 2. Minimum number of schedule milestones required to achieve 1 = ε estimation accuracy a given slip probability *q* from original maximum performance schedule

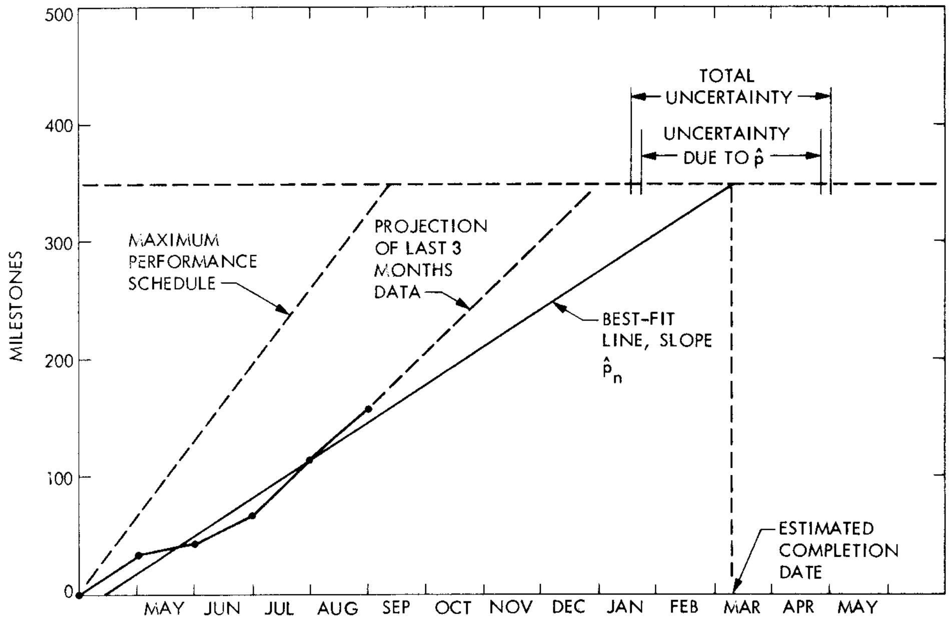

Fig. 3. Schedule prediction and uncertainty intervals based on 5 months' data

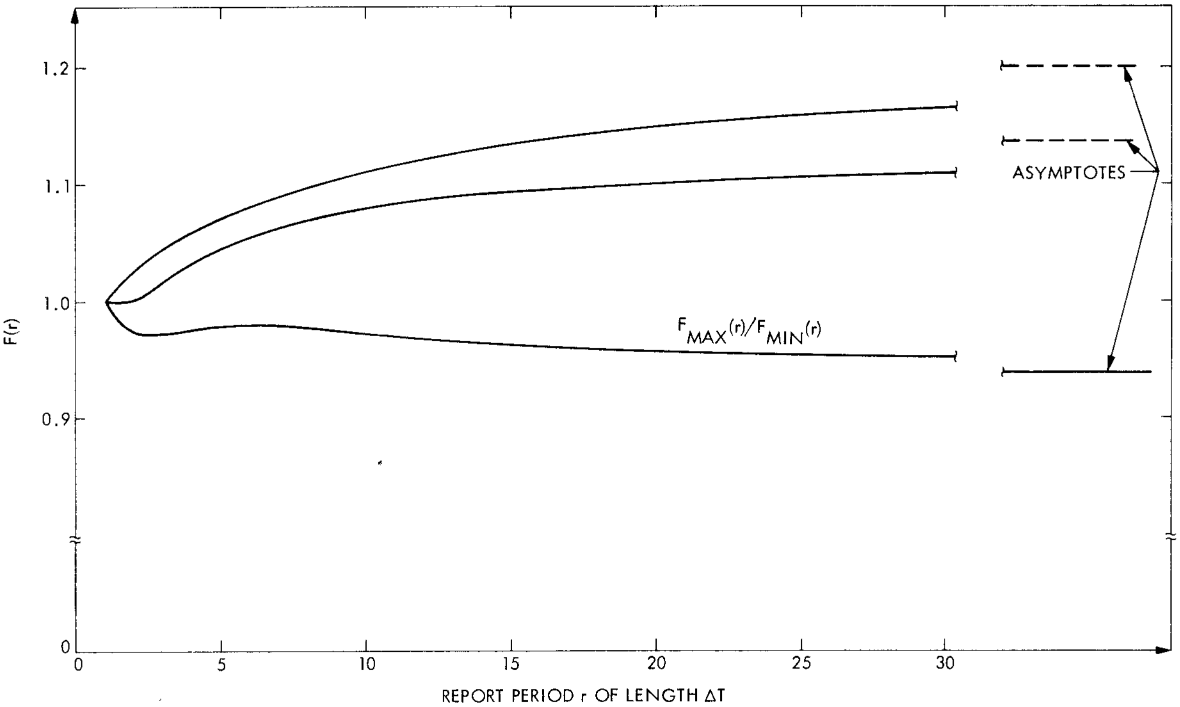

Fig. 4. Variance coefficient of completion date estimation from reported milestone data