Properties of Rare Fail-States and Extreme Values of *TTF* in a Markov Chain Model for Software Reliability

Michael G. Thomason* and James A. Whittaker†

April 19, 1999

Abstract

Numerous probability distributions are used in the estimation of software reliability. Typically, these distributions arise from assumptions together with actual success/failure data collected during testing or field-usage. *Statistics of the extremes* has been applied by Kaufman *et al* for software reliability analysis when failure is an infrequent, unlikely occurrence—a so-called *rare event*. When applicable, these statistics give powerful results concerning limiting distributions without the assumption of initial distributions often imposed in existing reliability models. This paper combines (*i*) results in rare events and extreme values with (*ii*) a finite-state, discrete-parameter, recurrent Markov chain which incorporates both the failures as rare events (as transitions to a rare fail-state) and usage of the software between failures (as transitions among ordinary usage-states not involving the fail-state). Four distributions arise naturally as approximations for the chain, namely: the "Poisson law of small numbers" gives an explicit error-bound on a *Poisson Approximation* for count of occurrences of a rare fail-state in long intervals of software usage; the interoccurrence time of the rare fail-state (the time-to-failure or *TTF*) has an approximately *exponential* distribution; and the *Weibull* and *Gumble* distributions, respectively, are the limiting distributions of the minimum and maximum values (the *extreme values*) in independent samples of the *TTF*. Results are illustrated by examples, including χ^2 goodness-of-fit tests for samples and approximate distributions.

Keywords: exponential, extreme values, Gumbel, Markov chain, Poisson, rare event, software reliability, Weibull

1 Introduction

The field of software reliability deals with the probabilistic estimation and prediction of software quality by combining stochastic models with statistical analysis of testing and field-

*M. G. Thomason is with the Department of Computer Science, University of Tennessee, Knoxville, TN 37996 USA. thomason@cs.utk.edu

†J. A. Whittaker is with the Center for Software Engineering Research, Computer Science Program, Florida Institute of Technology, Melbourne, FL 32901 USA. jw@cs.fit.edu

usage data. Such data typically includes *successful executions* in which the software's functionality was exercised without deviation from the specification, and *unsuccessful* or *faulty runs* in which the software performed one or more of its tasks incorrectly. Performing a task incorrectly is called a *failure event*. A software reliability model encapsulates observed success/failure data or other estimates of component reliability within the framework of probability models in order to predict patterns of future performance. The probability distributions for the number of failures within a specified time span, the failure interoccurrence time, and other relevant random variables are often adapted from hardware reliability theory [9] or justified empirically [1, 17].

Rare events and extreme values are topics in probability and statistics with application in several fields of science and engineering (cf. [3, 5, 6, 13, 20]). When the necessary conditions are met, these topics lead to well-defined distributions as approximate probability laws that are relevant in restricted, but important, areas of software reliability analysis. *Statistics of the extremes* has been applied to software reliability by Kaufman *et al* [14, 15] for an analysis of rare event data (for infrequent, unlikely failures) without requiring *a priori* knowledge of its distribution. An application for safety critical software systems is described in [14].

By calling failure a *rare event* [3, 13], we mean here that its probability of occurrence is greater than zero but smaller by at least several orders of magnitude than non-failure events in software execution intervals. In this situation, the mean time-to-failure ($MTTF$) is a large number. *Extreme values* [5, 6, 13] are the minimum and maximum values among independent[1], identically distributed (*iid*) random variables; thus, analysis of the time-to-failure (TTF) may look not only at the $MTTF$ but also at extreme variation above and below this mean value in multiple, *iid* executions of software. This paper concentrates on describing the TTF when failure is a rare event.

Intuitive definitions and convenient computations are attributes of Markov chains for work with rare events and extreme values. In this paper, a discrete-parameter, finite-state Markov chain [24] is used to represent both software failures (as transitions to a rare fail-state) and usage of the software between failures (as transitions among the ordinary usage states not involving the fail-state). Other work with Markov chains in software reliability includes computation of the sensitivity of system reliability to a module's reliability via a discrete-parameter Markov chain [8], generation of test cases for software systems based on a birth-death Markov chain [2], and versions of discrete- and continuous-parameter Markov chains as structural models in software reliability prediction [12]. Markov chains can be developed hierarchically for different levels of representation of software [23], can be used to model the interconnection of components [12], facilitate bookkeeping with frequency counts in empirical data [24], and offer an intuitive definition of rare events in terms of a recurrent chain's stationary probability distribution. Generic results in rare events and extreme values are not restricted to Markov chains and may, with suitable formulation, apply to other stochastic models [3, 13]. Other models are not explored here.

The requirement that failure be a rare event is not met in all software development activity; however, it is often the case that software evolves into a phase in which it rarely fails but its past history of failure with nonzero relative frequency, or an estimate of less-than-perfect reliability of some component(s), should not be completely discounted. For

[1] *Independent* means *stochastically independent* thoroughout this paper.

example, failures may be rare events in situations such as the following:

- Post-release "beta" versions of software, from which failures have been removed previously, can lead to failure as a rare event. If users of beta versions build up significant histories of successful execution intervals and few faulty runs, then failures may be rare events when compared with many non-failure execution intervals.

- Software components with defects may be embedded in end-user applications. Software developers work around known problems, making failure a rare event. We interpret this to mean that the usage distribution is adjusted to avoid a known defect. If this is achieved, the relative frequency of failure in the total usage history may approach 0.

- Growth in usage of widely-distributed, commercial software has been modeled as a power function of calendar time [17]. If the defects that cause customer-reported failures are corrected immediately for all users, the software may reach a phase in which failure is a rare event in execution intervals for a large, active group of independent users.

Thus, software which receives heavy usage according to established probabilities is a candidate for the treatment of its failures as rare events. If the likelihood of failure is indeed very small, then results in rare events and extreme values [3, 13] lead to four probability distributions *as approximations* under quite general conditions:

- The probability law for counts of failures in long execution intervals may be approximated as a *Poisson distribution*. Results of this kind are sometimes said to be a consequence of the "Poisson law of small numbers" [3]. The Poisson distribution has been introduced in other ways in the software reliability literature and already plays an important role in several software reliability models [9].

- The *exponential distribution* may approximate the probability law for the interoccurrence time of rare failures in long intervals. This distribution is sometimes assumed for the *TTF* in software reliability computations because it is implied by certain empirical studies (cf. [1]).

- Given a set of random samples of the *TTF*, the minimum value in the set is a random variable bounded away from 0 and its limiting distribution (for number of samples $N \to \infty$) is *Weibull*. The Weibull distribution itself has been used or suggested for other phenomena (cf. [17]).

- If the distribution of the *TTF* falls off fast enough for large values (see subsection 4.1), a counterpart to Weibull for the minimum *TTF* is *Gumbel* as the limiting distribution for the maximum *TTF*.

When conditions for rare events are met, reliability analysis in greater detail with fewer assumptions may be possible, there may be additional justification for using popular Poisson and exponential distributions, and Weibull and Gumbel distributions may also be applicable. This paper discusses aspects of rare events and extreme values in the Markov chain model

M described in the next section. Section 3 describes a Poisson Approximation for counts of a rare fail-state F in realizations of **M** and indicates approximation of interoccurrence times of F by an exponential distribution. Weibull and Gumbel distributions are discussed in section 4 for the extreme values of interoccurrence times of F, and section 5 illustrates computation based on the Weibull as the approximate distribution for the minimum TTF.

2 Markov Chain Model

A Markov chain can provide not only a convenient definition of software failure as a rare event (as a visit to an abnormal fail-state) but also a direct measure of rarity (using the steady-state probabilities of a recurrent chain). We will use the following model and assumptions.

Both the software usage distribution and the likelihood of failure are represented by a finite-state, discrete-parameter, time-homogeneous Markov chain **M** [10, 16, 24]. **M** has at least three states: starting-state S, terminate or end-state H, and fail-state F. State sequences are *realizations* of **M**. A realization from S to first occurrence of H represents a single successful execution cycle through the software. A transition from any state to F represents a failure of the software at that location in a realization; thus, occurrences of state F are identified with the failure events.

The probabilities on arcs in **M** may be actual relative frequencies from test or usage data [24] or may be values estimated for the usage distribution and component reliabilities expected in practice. Note that if no failure occurs in observed data with real software, then arcs to F with frequency count 1 are conservative assumptions in this sense: if state i, $i \neq F$, has been visited n_i times and exited without failure, and if the data-based relative frequency on each arc is close to the value defined by the expected usage, then this data implies that the probability of failure at i is not greater than $1/(n_i + 1)$. A conditional case study could set the count on an arc from i to F to 1 to give probability $p_{iF} = 1/(n_i + 1)$, renormalize other p_{ij} accordingly, and compute values for this hypothetical situation. If the relative frequencies p_{iF} make F a rare state, then the results in this paper are applicable.

We require that Markov chain **M** have (i) all states, including F, reachable from S by paths with nonzero probability, and (ii) arcs from both F and H to S with $p_{FS} = p_{HS} = 1$ so that both a successful termination in H and an unsuccessful run to F cause an immediate reset/restart in initial state S. This makes **M** a *recurrent chain* for studying long-term properties [10, 16]: the probability of eventually reaching every state from every state is 1. A long realization of **M** will ultimately include all states and, as its length increases, will become *statistically typical* of the entire chain in the frequencies of occurrence of all states and arcs. Let $\mathbf{P} = [p_{ij}]$ denote **M**'s transition probability matrix. **M**'s *steady-state probability distribution* $\Pi = [\pi_S, ..., \pi_F]$ is the unique solution of $\Pi = \Pi\mathbf{P}$ where $\sum_i \pi_i = 1$ and $\pi_i > 0$ is the limiting relative frequency of occurrence of state i as a count of transitions[2]. The *mean recurrence time* of state i is $m_{ii} = 1/\pi_i$ and the mean number of occurrences of state j between visits to state i is π_j/π_i [16].

[2]Count of transitions is a natural reference for computation with chain **M**. Standard terminology uses the word *time* instead of the phrase *count of transitions*, e.g., recurrence *time*, interoccurrence *time*, *time* of first occurrence.

A visit to fail-state F must be a rare event, that is, π_F must be smaller by at least several orders of magnitude than π_i for any other state $i \neq F$. Put another way, mean recurrence time $m_{FF} = 1/\pi_F$ must be larger by orders of magnitude than m_{ii} for all $i \neq F$. m_{FF} is large but finite because π_F is small but nonzero.

To focus on execution intervals between failures, we designate a realization of $\mathbf{M}$ from F to first recurrence of F a *run*. The run-length random variable is the interoccurrence time of F; this is the *TTF* and its mean value m_{FF} is the *MTTF*. Since $p_{FS} = 1$, runs are equivalent to realizations from starting-state S to first occurrence of F except for the initial transition F-to-S. The ensemble of all runs is infinite but has a well-defined, discrete probability distribution assigned by chain $\mathbf{M}$.

3 The Poisson Approximation: Counts of Rare Fail-State F

Two basic and enduring number laws in probability are the well-known Gaussian Approximation which is a "Central Limit Theorem" [10] for sample averages, and the less accessible Poisson Approximation which is a "Law of Small Numbers" [3] for rare, uncommon, unlikely events. Extensive discussion of the Poisson Approximation in numerous applications is given in [3]. Po(λ) denotes the Poisson distribution with parameter λ, that is, the discrete probability mass function

$$p(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$$

for $k = 0, 1, 2, \ldots$. Po(λ) has mean and variance equal to λ. When k is the largest integer not greater than λ, $p(k;\lambda)$ has as large a value as it does for any k.

Visits to rare state F in recurrent Markov chain $\mathbf{M}$ are crucial events. Let random variable n_F be the number of visits to F in a randomly generated realization of n transitions starting in state S. Let $\mathcal{L}(n_F)$ be the true probability law of n_F. Write $\lambda = E(n_F)$ for the mean value. Developments in small number laws stress that Po(λ) is an approximation of $\mathcal{L}(n_F)$ and compute an upper bound on a measure of distance between them. The *total variational distance* $d_{TV}[\mathcal{L}(n_F)$,Po(λ)] is defined as

$$d_{TV}[\mathcal{L}(n_F), \text{Po}(\lambda)] = \sup_A |\mathcal{L}(n_F)(A) - \text{Po}(\lambda)(A)|$$

for events (subsets) A in the sample space [3]. The value is $0 \leq d_{TV} \leq 1$. With reference to λ, π_F, and j-step transition probabilities $p_{FF}^{(j)}$, an explicit bound derived in [3] is

$$d_{TV}[\mathcal{L}(n_F), \text{Po}(\lambda)] \leq (1 - e^{-\lambda}) \left(\pi_F + 2 \sum_{j \geq 1} |p_{FF}^{(j)} - \pi_F| \right).$$

Since π_F equals the limiting relative frequency of state F, for large n

$$\pi_F \approx \frac{E(n_F)}{n} = \frac{\lambda}{n} \quad \text{and} \quad \lambda \approx n\pi_F.$$

The approximation

$$\mathcal{L}(n_F)(k) \approx \frac{(n\pi_F)^k}{k!} e^{-n\pi_F}$$

is for large n and small π_F. $\lambda \approx n\pi_F$ is the approximate parameter for a full sequence of n transitions, not per transition, and Poisson $p(k; n\pi_F)$ is the approximate probability of k visits to F in a full sequence.

The interoccurrence times of rare state F in $\mathbf{M}$ are *iid* (due to the Markov property) according to a discrete distribution with mean value m_{FF}. The exponential probability law Exp(β) has a continuous cumulative distribution function (Cdf) of the form

$$F_E(t) = \text{Prob}(T \leq t) = 1 - \exp\left\{-\frac{t}{\beta}\right\}$$

with mean value β for $T \geq 0$. If the number of transitions n is very large relative to m_{FF}, the discrete distribution for the TTF can be approximated by an exponential model as follows. Take any long, random realization of n transitions and mark the visits to non-rare state S. This partitions the realization into subsequences with mean length m_{SS}. Since state F has either 0 or 1 occurrences in each subsequence, there is an independent Bernoulli trial [10] for each subsequence with probabilities

$$p = \text{Prob}(1 \text{ occurrence of } F), \quad q = \text{Prob}(0 \text{ occurrences of } F) = 1 - p$$

where $p << q < 1$ because F is rare but S is not. But the mean number of occurrences of F between occurrences of S equals p, that is, for the 0-1 Bernoulli random variable

$$\frac{\pi_F}{\pi_S} = 1 \cdot p + 0 \cdot q = p.$$

Given that F has just occurred, let ν be the number of subsequences until its next recurrence. ν has a geometric distribution [10]. Cumulative geometric probabilities P_g and their exponential approximations are:

$$\begin{aligned} P_g(\nu \leq 1) &= p \approx 1 - e^{-p} = F_E(1), \\ P_g(\nu \leq 2) &= p + qp \approx 1 - e^{-2p} = F_E(2), \\ P_g(\nu \leq 3) &= p + qp + q^2 p \approx 1 - e^{-3p} = F_E(3), \\ &\vdots \end{aligned}$$

where $F_E(m)$ is the exponential Cdf, $1 - e^{-mp}$. These approximations are based on the exponential series [18]

$$e^{-mp} = 1 - mp + \frac{(mp)^2}{2!} + \cdots$$

and inequalities $p >> p^2 >> p^3 >> \cdots$. The expected value $E(\nu)$ for the geometric distribution and for its approximation is $1/p = \pi_S/\pi_F = m_{FF}/m_{SS}$, the mean number of occurrences of S between occurrences of F. Since the mean subsequence length is m_{SS} transitions, the expected count of transitions between visits to F is

$$m_{SS} E(\nu) = m_{SS} \frac{m_{FF}}{m_{SS}} = m_{FF}$$

as required. This exponential approximation for the TTF does not depend on the detailed structure of $\mathbf{M}$ beyond the requirement for a 0-1 count of F's between visits to S.

Example 1: A high level, five-state Markov chain for operation of three windows in a graphical user-interface is given in [23]. Here, the fail-state F shown in figure 1 has been added as a sixth state to represent failure events associated with state C, e.g., as failures actually observed during testing or as an estimate of less-than-perfect reliability of a software component. The transition probabilities p_{ij} for the ordinary usage-states $i \neq F, j \neq F$ represent the usage distribution, except that $p_{CA} = 1$ is reduced by small probability p_{CF}. For the range $0 < p_{CF} \leq 10^{-3}$, we compute $0 < \pi_F \leq 1.43 \times 10^{-4}$ and find that all other π's have relatively stable values. Specifically for $p_{CF} = 10^{-4}$, the stationary distribution vector is

$$\begin{aligned} \Pi &= [\pi_S, \pi_A, \pi_B, \pi_C, \pi_H, \pi_F] \\ &= [0.14278, 0.42843, 0.14283, 0.14283, 0.14276, 1.42689 \times 10^{-5}], \end{aligned}$$

the vector of mean recurrence times is

$$\begin{aligned} [m_{SS}, ..., m_{FF}] &= [\frac{1}{\pi_S}, ..., \frac{1}{\pi_F}] \\ &= [7.00400, 2.33412, 7.00119, 7.00120, 7.00470, 7.00821 \times 10^4], \end{aligned}$$

the vector of the expected number of occurrences of states between visits to non-rare state S is

$$\frac{\Pi}{\pi_S} = [1, 3.0007, 1.0004, 1.0004, 0.9999, 0.0001],$$

and the vector of the expected number of occurrences between visits to rare state F is

$$\frac{\Pi}{\pi_F} = 10^4 \times [1.0006, 3.0025, 1.001, 1.001, 1.0005, 0.0001].$$

For $p_{CF} = 10^{-5}$, these vectors are

$$\Pi = [0.14277, 0.42844, 0.14284, 0.14284, 0.14276, 1.42693 \times 10^{-6}],$$

$$[m_{SS}, ..., m_{FF}] = [7.00445, 2.33406, 7.00102, 7.00102, 7.00452, 7.00803 \times 10^5],$$

$$\frac{\Pi}{\pi_S} = [1, 3.001, 1.0005, 1.0005, 0.99999, 9.9949 \times 10^{-6}],$$

and

$$\frac{\Pi}{\pi_F} = 10^5 \times [1.0005, 3.0025, 1.001, 1.001, 1.0005, 0.00001]$$

respectively.

The mean recurrence time of state S is $m_{SS} \approx 7$. $n = 7 \times 10^4$ transitions correspond to about 10^4 average sequences from S to S. We fix n at 7×10^4 and evaluate the Poisson Approximation of $\mathcal{L}(n_F)$ for four values of p_{CF}: 10^{-4}, 10^{-5}, 10^{-6}, and 10^{-7}.

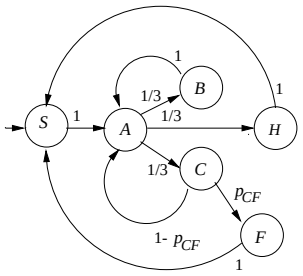

Figure 1: 6-state Markov chain **M**.

p_{46}	λ	$d_{TV} \leq$	$p(0;\lambda)$	$p(1;\lambda)$	$p(2;\lambda)$	$p(3;\lambda)$	$p(4;\lambda)$
10^{-4}	1	9.03×10^{-6}	0.3679	0.3679	0.1839	0.0613	0.0153
10^{-5}	0.1	1.36×10^{-7}	0.9048	0.0905	0.0045	1.51×10^{-4}	3.77×10^{-6}
10^{-6}	0.01	1.42×10^{-9}	0.99	0.0099	4.95×10^{-5}	1.65×10^{-7}	4.13×10^{-10}
10^{-7}	0.001	1.43×10^{-11}	0.999	9.99×10^{-4}	5×10^{-7}	1.67×10^{-10}	4.16×10^{-14}

Table 1: Poisson Approximation for **Example 1**

Table 1 lists parameter $\lambda \approx 7\pi_F \times 10^4$, the upper bound on d_{TV}, and Poisson $p(k;\lambda)$ for $k = 0, 1, 2, 3, 4$. For $p_{CF} = 10^{-4}$, rare state F has steady-state probability $\pi_F \approx 1.43 \times 10^{-5}$ and the upper bound 0.903×10^{-5} is about the same order of magnitude. For $p_{CF} = 10^{-5}$, visits to F are more rare and the upper bound 1.36×10^{-7} is an order of magnitude smaller than $\pi_F \approx 1.43 \times 10^{-6}$. The $MTTF$ ranges from $m_{FF} \approx 7 \times 10^4$ for $p_{CF} = 10^{-4}$ to $m_{FF} \approx 7 \times 10^7$ for $p_{CF} = 10^{-7}$. ■

The upper bound on d_{TV} is one kind of measure of the suitability of the Poisson Approximation. χ^2 goodness-of-fit tests can be applied to sample data and approximate distributions as a different kind of measure. See the Appendix for a brief summary of this classic statistical test. The approximate distributions we are studying are hypothesized in the tests here. The binary outcome of each test is solely *to reject* or *not to reject* the hypothesis based on the statistic computed and the level of significance specified.

Example 2: Sample data was obtained by using random-number generator RAND in MATLAB (R) to simulate **M** for 7×10^4 transitions starting in state S. Visits to state F were then counted. The experiment was repeated 1000 times for $p_{CF} = 10^{-4}$ and for $p_{CF} = 10^{-5}$.

For $p_{CF} = 10^{-4}$, the 1000 samples were grouped into five classes: C_1 for count 0, C_2 for count 1, C_3 for count 2, C_4 for count 3, and C_5 for count ≥ 4. Table 3 in the Appendix lists observed frequencies f_i, expected frequencies $1000p_i$, and χ^2-residuals for Po(1). The χ^2 critical value for level of significance 0.1 and 4 (=5-1) degrees of freedom is 7.78. The statistic 1.2071 is less than the critical value; therefore, in terms of testing for goodness of fit at significance level 0.1, there is no reason to reject the hypothesis that the Poisson Po(1) models the phenomenon that gave rise to the 1000 samples.

For $p_{CF} = 10^{-5}$, we combine classes at higher counts where expected frequencies are low.

A rule-of-thumb for the χ^2 test is to define the classes so that most expected counts exceed 5, but one or two can be less than 5 [6]. Classes are C_1 for count 0, C_2 for count 1, and C_3 for count ≥ 2. Table 4 presents the numbers. The χ^2 critical value is 4.61 for 2 (=3-1) degrees of freedom and level of significance 0.1. The statistic 1.4229 is less than 4.61, so the hypothesis is not rejected at significance level 0.1.

The *TTF* in this software model is the interoccurrence time of rare fail-state F. The χ^2 goodness-of-fit test can also be applied to the hypothesis that this interoccurrence time has an exponential distribution. A test can be computed in the following way without an explicit value for the parameter β of Exp(β) [11]. Given that N visits to F occur in time interval $(0, T)$, let random variables T_1, T_2, $\ldots, T_N$ be the interoccurrence times and let random variables

$$A_1 = T_1, \quad A_2 = T_1 + T_2, \quad \ldots, \quad A_N = T_1 + \cdots + T_N$$

be the waiting times. If the interoccurrence times T_i are *iid* according to an exponential distribution, the waiting times A_i, taken as N unordered samples, are independently and uniformly distributed in $(0, T)$ [4]. Thus, the interval $(0, T)$ can be partitioned into k equal subintervals, each subinterval having length T/k and (due to uniformity) each having N/k as the expected number of visits. Let f_i be the number of visits actually observed in subinterval i; then the statistic

$$\sum_{i=1}^{k} \frac{\left(f_i - \frac{N}{k}\right)^2}{\frac{N}{k}}$$

is approximately χ^2 with $k-1$ degrees of freedom, provided that $N \geq 20$ and each f_i is sufficiently large [11]. If the statistic is less than the appropriate critical value, the hypothesis of exponential interoccurrence times is not rejected; otherwise, it is rejected.

1000 samples of interoccurrence times of rare fail-state F in $\mathbf{M}$ in figure 1 were generated for $p_{CF} = 10^{-4}$ and for $p_{CF} = 10^{-5}$. Waiting times $A_1, \ldots, A_{1000}$ were then computed. Partitioning into $k = 20$ equal subintervals gives an expected frequency of $1000p_i = 50$ per subinterval. See table 5 in the Appendix for the data. The χ^2 critical value for level of significance 0.1 and 19 (=20-1) degrees of freedom is 27.20. As shown in the table, the statistic is 23.56 for $p_{CF} = 10^{-4}$ and is 20.04 for $p_{CF} = 10^{-5}$. At significance level 0.1, the hypothesis of exponential interoccurrence times is not rejected for either value of p_{CF}. ■

4 Distributions of Extreme Values

Estimation of the *MTTF* is included in many models of software reliability [19]. Results in extreme values concern the *minimum* and *maximum* values of the *TTF* as random variables in their own right. Scenarios in which stochastic properties of one or both extreme values are of interest include the following.

- Evaluation of software for safety critical applications is an important issue. A software reliability model based on statistics of the extremes for safety critical systems is developed in [14].

- Suppose widely-distributed software has a large base of independent users, the users have the same usage probabilities, and failure is a rare event. Long-term, statistical characteristics of both shortest and longest TTF provide information about expected performance encountered by the ensemble of users as a group.

- Suppose a long computation executes in parallel on a network of computers as follows. There is a time requirement: the parallel computation must have uninterrupted time interval τ simultaneously at every processor in order to complete. Processors use the same software and have the same usage distribution which includes interaction with the network. Tasks at different processors are loosely-coupled and are, to a first approximation, independent. A software failure at a processor is a rare event but may be a consequence of local data or the network. Any software failure before τ forces the entire computation to restart; but due to randomness in network characteristics, a restarted computation is not likely to be identical to previous attempts, even though the same local data is processed again. A long-run characteristic of interest is the number of times on average a computation must restart due to a failure in time less than τ.

4.1 Weibull and Gumbel Distributions in the Limit

Consider Markov chain **M** again as a source of *iid* runs from F to the first recurrence of F. In extreme value analysis, the distribution of these runs is called the *initial distribution*. Given a set of runs, let $\mathcal{F}_{min}$ be the minimum run-length and $\mathcal{F}_{max}$ the maximum run-length. $\mathcal{F}_{min}$ and $\mathcal{F}_{max}$ are called *extreme value random variables*.

$\mathcal{F}_{min}$ is nonnegative and has an initial distribution bounded away from 0 in the direction of the extreme by the minimum-length run from F to F. $\mathcal{F}_{max}$ is nonnegative and has an initial distribution that is unbounded in the direction of the extreme but, by virtue of the exponential approximation in section 3, has an exponential upper bound (that is, the initial distribution is of the exponential type [6, 13]). Generic results in limiting distributions for $\mathcal{F}_{min}$ and $\mathcal{F}_{max}$ in our model are as follows [5, 6, 13].

- The probability law of $\mathcal{F}_{min}$ approaches a *Weibull distribution* Weib(α, β) with Cdf

$$F_W(x) = 1 - \exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\}, \alpha > 0, \beta > 0.$$

α is the shape parameter: larger (smaller) α implies more (less) peaked density function. Weib(1,β) with $\alpha = 1$ is in fact the exponential Exp(β) with mean value β. β is the scale parameter: the distribution depends on β and x through the ratio x/β.

- The probability law of $\mathcal{F}_{max}$ approaches a *Gumbel distribution* Gumb(μ, ρ) with Cdf

$$F_G(x) = \exp\left\{-\exp\left\{-\frac{x-\mu}{\rho}\right\}\right\}, \infty > \mu > -\infty, \rho > 0.$$

μ is the location parameter: association of x with μ is of the form $x - \mu$. ρ is the scale parameter.

- In the special case of N samples of extreme values for which the initial distribution for

recurrences of rare state F in Markov chain $\mathbf{M}$ is *exactly* the exponential $\text{Exp}(m_{FF})$, then $\text{Gumb}(\mu, \rho)$ for $\mathcal{F}_{max}$ has parameters [6]

$$\mu = m_{FF} \ln(N) \text{ and } \rho = m_{FF}$$

and $\text{Weib}(\alpha, \beta)$ for $\mathcal{F}_{min}$ reduces to an exponential with parameters [6]

$$\alpha = 1 \text{ and } \beta = \frac{m_{FF}}{N}.$$

4.2 Maximum-Likelihood Estimates from Samples

Estimation of parameters of Weibull and Gumbel Cdf's from samples is of interest for comparisons with the theoretical limits described above. In general, finite sets of samples do not fit the asymptotic formulations perfectly. This section describes maximum-likelihood (ML) estimation and illustrates with data from Markov chain $\mathbf{M}$ in figure 1.

Given samples $\{x_1, x_2, \ldots, x_N\}$ of $\mathcal{F}_{min}$, ML-estimates of α and β in $\text{Weib}(\alpha, \beta)$ satisfy the equations [6]

$$\frac{1}{N}\sum_{i=1}^{N} \ln x_i = \left[\sum_{i=1}^{N} x_i^{\hat{\alpha}} \ln x_i\right] \left[\sum_{i=1}^{N} x_i^{\hat{\alpha}}\right]^{-1} - \frac{1}{\hat{\alpha}}$$

and

$$\hat{\beta} = \left[\frac{1}{N}\sum_{i=1}^{N} x_i^{\hat{\alpha}}\right]^{-\hat{\alpha}}.$$

$\hat{\alpha}$ may be obtained from the first equation by iteration, then $\hat{\beta}$ computed by the second equation. Small N biases the estimate $\hat{\alpha}$ upwards [6], that is, an unbiased estimate would be smaller. An unbiasing factor is available [21] and ranges from 0.669 to 0.99 as N ranges from 5 to 120. Recall that $\alpha = 1$ in a Weibull Cdf coincides with an exponential Cdf; but even when $\alpha = 1$ should be the case, an ML-estimate $\hat{\alpha}$ precisely equal to 1 is an unlikely prospect for finite samples of real data.

Given samples $\{y_1, \ldots, y_N\}$ of $\mathcal{F}_{max}$, the ML-estimates of μ and ρ in $\text{Gumb}(\mu, \rho)$ satisfy [6]

$$\hat{\rho} = \bar{y} - \left[\sum_{i=1}^{N} y_i \exp\left\{-\frac{y_i}{\hat{\rho}}\right\}\right] \left[\sum_{i=1}^{N} \exp\left\{-\frac{y_i}{\hat{\rho}}\right\}\right]^{-1}$$

and

$$\hat{\mu} = -\hat{\rho} \ln\left[\frac{1}{N}\sum_{i=1}^{N} \exp\left\{-\frac{y_i}{\hat{\rho}}\right\}\right]$$

where $\bar{y} = \sum_{i=1}^{N} y_i/N$ is the sample mean. Small N biases the estimate $\hat{\rho}$ downwards [6], that is, an unbiased estimate would be larger.

Example 3: Assume that a group of 100 users have runs *iid* according to Markov chain $\mathbf{M}$ in figure 1. $\mathbf{M}$ was simulated using RAND in MATLAB (R) to obtain a set of 100 runs, and $\mathcal{F}_{min}$ and $\mathcal{F}_{max}$ for the set were recorded. The experiment was repeated 64 times to obtain $N = 64$ samples of both extreme values for $p_{CF} = 10^{-4}$ and for $p_{CF} = 10^{-5}$. Details of the χ^2 goodness-of-fit tests are described in the Appendix. In summary:

- For $p_{CF} = 10^{-4}$ and samples of $\mathcal{F}_{min}$, Weibull W_1 with ML-estimates of parameters is not rejected at level of significance 0.1, but Weibull W_2 based on Exp(m_{FF}) as the exact initial distribution is rejected at significance levels 0.1 and 0.05.

- For $p_{CF} = 10^{-5}$ and $\mathcal{F}_{min}$, neither Weibull W_3 with ML-estimates nor Weibull W_4 based on Exp(m_{FF}) is rejected at level of significance 0.1.

- For $p_{CF} = 10^{-4}$ and $\mathcal{F}_{max}$, Gumbel G_1 with ML-estimates of parameters is not rejected at level of significance 0.1, but Gumbel G_2 based on Exp(m_{FF}) is rejected at significance levels 0.1 and 0.05.

- For $p_{CF} = 10^{-5}$ and $\mathcal{F}_{max}$, Gumbel G_3 with ML-estimates is not rejected at level of significance 0.1. Gumbel G_4 based on Exp(m_{FF}) is rejected at level of significance 0.1 but is not rejected at reduced significance level 0.05.

None of the distributions W_1, W_3, G_1, G_3 with ML-estimates of parameters from samples is rejected at level of significance 0.1. Distributions W_2, W_4, G_2, G_4 are based on an assumption that an exponential Exp(m_{FF}) is *exactly* the initial distribution. W_2 and G_2 are obtained for state F being less rare; W_4 and G_4 are obtained for state F being more rare. At significance level 0.1, W_2 is rejected but W_4 is not; at significance level 0.05, G_2 is rejected but G_4 is not. We conjecture that one factor in the outcomes of these tests is less error in the exponential approximation of the TTF's distribution for more rare F than for less rare F. ■

5 Extreme Value $\mathcal{F}_{min}$ of Software TTF

Analysis of extreme values is rich in details. We will use the Weibull F_W as the approximate Cdf for $\mathcal{F}_{min}$ to compute *how often on average the minimum TTF experienced within a group of runs that are iid according to the same probablity model is not less than* τ where τ is a fixed reference.

An occurrence of $\mathcal{F}_{min} > \tau$ is called an *exceedance of* τ [3, 6] and its complement $\mathcal{F}_{min} \leq \tau$ is called a *nonexceedance*. As shown in figure 2, the binary events *exceedance* and *nonexceedance* define a Bernoulli trial [10] with probabilities

$$\text{Prob}(\mathcal{F}_{min} > \tau) = 1 - F_W(\tau),\ \text{Prob}(\mathcal{F}_{min} \leq \tau) = F_W(\tau).$$

Let random variable n_τ be the number of independent trials from one nonexceedance of τ until the next. A standard result [10] is that n_τ is geometrically distributed with mean value $E_W(n_\tau) = 1/F_W(\tau)$. The *mean number of times* τ *is exceeded in between nonexceedances* is

$$\begin{aligned} E_W(n_\tau) - 1 &= \frac{1}{F_W(\tau)} - 1 \\ &= \frac{1 - F_W(\tau)}{F_W(\tau)} \\ &= \frac{\text{Prob}(\mathcal{F}_{min} > \tau)}{\text{Prob}(\mathcal{F}_{min} \leq \tau)}. \end{aligned}$$

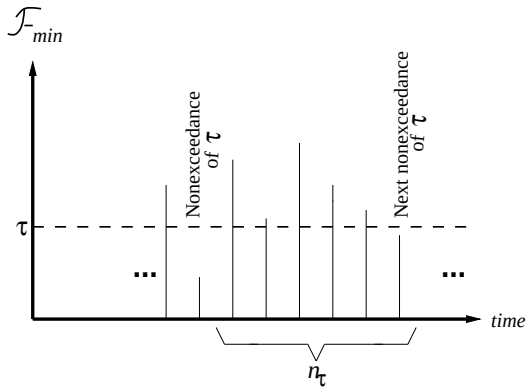

Figure 2: Illustrating a time-series of *iid* samples of $\mathcal{F}_{min}$

Example 4: Weibulls W_1 and W_3 above are for a group of 100 runs, say, as experienced by 100 independent users with the same usage distribution. W_1 is obtained for $MTTF = 70082$ and W_3 for $MTTF$ about 10 times longer. For $\tau = 500$,

$$F_{W_1}(500) = 0.6076, \quad E_{W_1}(n_{500}) = \frac{1}{0.6076} = 1.6457$$

$$F_{W_3}(500) = 0.0325, \quad E_{W_3}(n_{500}) = 30.7630.$$

Based on W_1 for $\mathcal{F}_{min}$ in the group of 100 users, on average the group experiences 0.6457 runs with minimum $TTF > 500$ in between occurrences of minimum $TTF \leq 500$. Based on W_3 for $\mathcal{F}_{min}$, the comparable figure is 29.763 runs. Doubling τ to 1000 and again to 2000 gives

$$E_{W_1}(n_{1000}) = 1.1959, \quad E_{W_3}(n_{1000}) = 14.2553$$

$$E_{W_1}(n_{2000}) = 1.0312, \quad E_{W_3}(n_{2000}) = 6.76056.$$

Rather than a group of different users, W_1 or W_3 could describe $\mathcal{F}_{min}$ for blocks of 100 runs by a single user. Suppose W_3 applies for a single user $\mathcal{U}$ who starts up a block of 100 tasks every day, each of which executes *iid* according to **M** in figure 1 and each of which must have execution-time exceeding τ. Then user $\mathcal{U}$ will encounter a day with $\mathcal{F}_{min} \leq \tau$ on average about once each month ($\sim$30.76 blocks) for $\tau = 500$, about once each fortnight ($\sim$14.26 blocks) for $\tau = 1000$, or about once each week ($\sim$6.76 blocks) for $\tau = 2000$. If user $\mathcal{U}$ voluntarily stops the failure-free executions after τ is reached in one day and restarts all tasks the next day, the next sample of $\mathcal{F}_{min}$ is generated independently according to the same Cdf F_{W_3} (there is no parameter-estimation with data that is censored by the non-failure stops of software execution [7]). Given that the event $\mathcal{F}_{min} \leq \tau$ does occur, the conditional mean values of $\mathcal{F}_{min}$ for the three values of τ are

$$E_{W_3}(\mathcal{F}_{min}|\mathcal{F}_{min} \leq 500) = 264.9, \quad E_{W_3}(\mathcal{F}_{min}|\mathcal{F}_{min} \leq 1000) = 526.5,$$

$$E_{W_3}(\mathcal{F}_{min}|\mathcal{F}_{min} \leq 2000) = 1038.7.$$

■

If nonexceedance is a rare event for a given τ, then $F_W(\tau) = 1/E_W(n_\tau)$ is small and the exponential series [18] gives as an approximation

$$\begin{aligned} F_W(\tau) &= 1 - \exp\left\{-\left(\frac{\tau}{\beta}\right)^\alpha\right\} \\ &\approx \left(\frac{\tau}{\beta}\right)^\alpha. \end{aligned}$$

Example 5: For W_3 in the example above:

$$F_{W_3}(1000) = 0.0701 \text{ and } \left(\frac{1000}{\beta}\right)^\alpha = 0.0727,$$

$$F_{W_3}(500) = 0.0325 \text{ and } \left(\frac{500}{\beta}\right)^\alpha = 0.0330,$$

$$F_{W_3}(100) = 0.0053 \text{ and } \left(\frac{100}{\beta}\right)^\alpha = 0.0053.$$

Suppose user $\mathcal{U}$ in that example wants τ as large as possible to average a full year of successful days (that is, an average of 365 blocks having $\mathcal{F}_{min} > \tau$ in between blocks having $\mathcal{F}_{min} \leq \tau$). An estimate of τ for $E_{W_3}(n_\tau) = 366$ is

$$\tau \approx \beta\left(\frac{1}{366}\right)^{\frac{1}{\alpha}} = 55.945$$

where, in fact, $1/F_{W_3}(55.945) = 366.5$. ■

Example 6: As a final example, suppose that (*i*) a continuously-executing application consists of many tasks distributed on a network of processors, (*ii*) task executions are *iid* according to the same Markov chain model with rare fail-state, and (*iii*) all current executions in the block of parallel tasks continue until one fails, at which time the system resets and all tasks immediately restart in synchronism. Then W=Weib(α, β), which is the limiting probability law of $\mathcal{F}_{min}$, also approximately describes the *system-restart interoccurrence times*. As approximations, the probability that a system-restart does not occur before τ is $1 - F_W(\tau) = \text{Prob}(\mathcal{F}_{min} > \tau)$ and the expected number of exceedances of τ in between nonexceedances is $E_W(n_\tau) - 1$. ■

6 Summary and Conclusions

Markov chains have diverse uses in software reliability computations (cf, [2, 8, 12, 24]). Rare events and statistics of the extremes are applicable to restricted, but important, aspects

of software reliability analysis (cf, [14, 15]). This paper combines results in rare events and extreme values with software reliability computations based on a finite-state, discrete-parameter, recurrent Markov chain **M**. The chain provides a convenient definition of failure as a rare event, namely, as a fail-state F for which the steady-state probability π_F is orders of magnitude smaller than π_i for ordinary usage states $i \neq F$.

When applicable, results in rare events and extreme values may provide information about software reliability, especially in details of likely performance in the long-run. Four distributions arise naturally *as approximations* when software failures correspond to visits to a rare fail-state F in **M**:

- Poisson for count of visits to F in a long realization.
- Exponential Cdf for interoccurrence time of F (the *TTF*).
- Weibull as limiting Cdf for minimum recurrence time of F.
- Gumbel as limiting Cdf for maximum recurrence time of F.

An upper bound on d_{TV} was computed for a Poisson Approximation for Markov chain **M** in figure 1. χ^2 goodness-of-fit tests were computed for approximate distributions and random samples from chain **M**. For extreme values, the parameters for Weib(α, β) and Gumb(μ, ρ) were computed by ML-estimation from samples and also by assuming an initial exponential Exp(m_{FF}) for the *TTF*. Outcomes of these tests are compatible with the general proposition that the more rare the event, the better the approximation. Although our example used a small Markov chain to demonstrate various computations, the methods apply to chains of arbitrarily large size.

One significance of the Markov chain model is that it structurally represents the software system under test or in use in the field, and indirectly represents the software complexity. If the transition probabilities are obtained from actual frequency counts during testing, an up-to-date representation of the amount and completeness of testing is provided.

Important contributions of rare events and extreme values are the four distributions listed above as approximations in the analysis of *TTF*, given the basic assumptions set forth. We believe that the combination of the Markov model and these results is a powerful tool for aspects of software reliability analysis.

7 Appendix

7.1 The χ^2 Goodness-of-Fit Test

The test is described in numerous references (cf. [6, 11]). H_0 is the hypothesis that independent samples are from a candidate distribution. A statistic is computed to measure how much the samples disagree with the hypothesis. If disagreement is "significant," H_0 is rejected; otherwise, it is not.

The binary outcome of a test is solely *to reject* or *not to reject* H_0. The χ^2 distribution provides the *critical value* for this binary decision. This distribution's parameter is the

number of degrees of freedom. If there are N samples and s parameters for the distribution under test are estimated from the data, there are $N-s-1$ degrees of freedom; if no parameters are estimated, there are $N-1$ degrees of freedom. The critical value is computed for a given *level of significance*, which is the probability of rejecting H_0 when H_0 is true. A higher significance level defines a more critical test than a lower one. Significance levels 0.1, 0.05, and 0.01 are commonly used in practice.

The set of possible values of samples is partitioned into k classes $C_1, \ldots, C_k$ where p_i is the probability of an observation in C_i according to the hypothesized distribution, Np_i is the expected number of observations in C_i in N samples, and f_i is the frequency of occurrence of actual samples in C_i. Two rules-of-thumb [6] are (i) the sample size N should be greater than 20, and (ii) the classes should be defined so that most expected frequencies Np_i exceed 5, but one or two can be less than 5. The test statistic

$$\sum_{i=1}^{k} \frac{(f_i - Np_i)^2}{Np_i}$$

is approximately χ^2 with the appropriate degrees of freedom. If the statistic is less than the appropriate χ^2 critical value, H_0 is not rejected; otherwise, H_0 is rejected because a statistic larger than the critical value implies, at the given significance level, that the observed data are not random samples from the hypothesized model.

For reference in examples, Table 2 lists χ^2 critical values for two levels of significance and several degrees of freedom.

level					degrees	of	freedom				
of sig.	1	2	3	4	5	6	7	8	9	10	19
0.10	2.71	4.61	6.25	7.78	9.24	10.65	12.02	13.36	14.68	15.99	27.20
0.05	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31	30.14

Table 2: χ^2 critical values (rounded values from MATLAB (R) function CHI2INV)

7.2 Tables for Examples 1 and 2 (Section 2)

The following tables are referenced in the χ^2 goodness-of-fit tests in section 2.

Class C_i	Observed frequency f_i	Expected frequency $1000p_i$	χ^2-residual $\frac{(f_i-1000p_i)^2}{1000p_i}$
C_1: count 0	366	367.9	0.0098
C_2: count 1	379	367.9	0.3349
C_3: count 2	183	183.9	0.0044
C_4: count 3	55	61.3	0.6475
C_5: count ≥ 4	17	19.0	0.2105
sum	1000	1000.0	1.2071

Table 3: Calculation of χ^2 statistic for Poisson for $p_{CF} = 10^{-4}$

Class C_i	Observed frequency f_i	Expected frequency $1000p_i$	χ^2-residual $\frac{(f_i-1000p_i)^2}{1000p_i}$
C_1: count 0	895	904.8	0.1061
C_2: count 1	101	90.5	1.2182
C_3: count ≥ 2	4	4.7	0.0985
sum	1000	1000.0	1.4229

Table 4: Calculation of χ^2 statistic for Poisson for $p_{CF} = 10^{-5}$

Subinterval	Observed frequency f_i	χ^2-residual $\frac{(f_i-50)^2}{50}$
1	59/53	1.62/0.18
2	42/39	1.28/2.42
3	58/45	1.28/0.50
4	38/49	2.88/0.02
5	57/42	0.98/1.28
6	49/60	0.02/2.00
7	54/61	0.32/2.42
8	53/47	0.18/0.18
9	44/38	0.72/2.88
10	38/49	2.88/0.02
11	48/54	0.08/0.32
12	54/52	0.32/0.08
13	61/48	2.42/0.08
14	56/38	0.72/2.88
15	43/56	0.98/0.72
16	59/62	1.62/2.88
17	59/53	1.62/0.18
18	44/53	0.72/0.18
19	39/46	2.42/0.32
20	45/55	0.50/0.50
sum	1000/1000	23.56/20.04

Table 5: Calculation of χ^2 statistics for Uniform Waiting Times for $p_{CF} = 10^{-4}/p_{CF} = 10^{-5}$

7.3 $\mathcal{F}_{min}$ and Weibull Distributions in Example 3 (Section 4)

Class C_i	Observed frequency f_i	Expected frequency $64p_i$	χ^2-residual $\frac{(f_i-64p_i)^2}{64p_i}$
C_1: ≤ 100	11	11.7253/5.5856	0.0449/5.2484
C_2: 100^+ to 300	14	16.2790/9.7514	0.3191/1.8511
C_3: 300^+ to 500	12	10.8846/8.1235	0.1143/1.8498
C_4: 500^+ to 700	12	7.4646/6.7674	2.7509/4.0459
C_5: 700^+ to 1100	6	8.8135/10.3343	0.8981/1.8178
C_6: > 1100	9	8.8300/23.4377	0.0033/8.8937
sum	64	64.0000/64.0000	4.1353/23.7068

Table 6: Calculations of χ^2 statistics for Weibulls W_1/W_2 for $p_{CF} = 10^{-4}$

Class C_i	Observed frequency f_i	Expected frequency $64p_i$	χ^2-residual $\frac{(f_i-64p_i)^2}{64p_i}$
C_1: $\leq 2 \times 10^3$	7	9.4663/10.6841	0.6426/1.2704
C_2: $(2^+$ to $4)\times 10^3$	13	9.5367/8.9005	1.2577/1.8882
C_3: $(4^+$ to $6)\times 10^3$	7	8.3981/7.4146	0.2328/0.0232
C_4: $(6^+$ to $8)\times 10^3$	11	7.1243/6.1769	2.1084/3.7660
C_5: $(8^+$ to $10)\times 10^3$	6	5.9191/5.1457	0.0011/0.1418
C_6: $(1^+$ to $1.4)\times 10^4$	4	8.7797/7.8577	2.6021/1.8939
C_7: $> 1.4 \times 10^4$	16	14.7759/17.8205	0.1014/0.1860
sum	64	64.0000/64.0000	6.9461/9.1695

Table 7: Calculations of χ^2 statistics for Weibulls W_3/W_4 for $p_{CF} = 10^{-5}$

For $p_{CF} = 10^{-4}$, the 64 samples of $\mathcal{F}_{min}$ have range [5,2912], mean 548.5, and standard deviation 545.05. Iteration with these samples gives $\hat{\alpha} = 0.9708$ as ML-estimate of the Weibull parameter. The unbiasing factor for 64 samples is 0.98 [21], so the unbiased ML-estimate is $\hat{\alpha} = 0.9513$. We obtain $\hat{\beta} = 536.26$ and the distribution W_1=Weib$(\hat{\alpha}, \hat{\beta})$ with mean 548.45 and standard deviation 576.74.

Using Exp(m_{FF}) as the initial distribution for interoccurrence time of rare fail-state F gives $\alpha = 1$, $\beta = 1095$, and W_2= Weib(1,1095) = Exp(1095) with mean value about twice the sample mean.

χ^2 goodness-of-fit tests were applied to the samples of $\mathcal{F}_{min}$ and the separate Weibulls W_1/W_2. (This is not a test of one distribution vs. another as alternate hypotheses.) Samples were grouped into six classes by value. See Table 6. Since two parameters are estimated from the data for W_1, there are 3 (=6-2-1) degrees of freedom. Consider level of significance 0.1. The χ^2 critical value 6.25 exceeds the statistic 4.1353 for W_1; therefore, at significance level 0.1, there is no reason to reject the hypothesis that W_1 models the phenomenon that gave rise to the samples. Since no parameters are estimated from the data for W_2, there are 5 (=6-1) degrees of freedom; the critical value 9.24 is less than the statistic 23.7068 for

W_2, so W_2 is rejected. In fact, W_2 is also rejected at the reduced level of significance 0.05 because the χ^2 critical value 11.07 is less than its statistic 23.7068.

For $p_{CF} = 10^{-5}$, the 64 samples of $\mathcal{F}_{min}$ have range [23,35712], mean 9559.9, and standard deviation 8301.7. ML-estimate is $\hat{\alpha} = 1.1613$, adjusted to 1.1381 by the unbiasing factor and giving $\hat{\beta} = 10004$. W_3=Weib(1.1381,10004) has mean 9550.3 and standard deviation 8410.4. Assuming initial distribution Exp(m_{FF}) gives $\alpha = 1$, $\beta = 10950$, and W_4=Exp(10950) with mean value fairly close to the sample mean.

The 64 samples were grouped into seven classes. See Table 7. Consider level of significance 0.1. The critical values are 7.78 for 4 (=7-2-1) degrees of freedom and 10.65 for 6 (=7-1) degrees of freedom. Critical value 7.78 exceeds W_3's statistic 6.9461 and critical value 10.65 exceeds W_4's statistic 9.1695. Neither W_3 nor W_4 is rejected. ■

7.4 $\mathcal{F}_{max}$ and Gumbel Distributions in Example 3 (Section 4)

Class C_i	Observed frequency f_i	Expected frequency $64p_i$	χ^2-residual $\frac{(f_i-64p_i)^2}{64p_i}$
C_1: $\leq 2.7 \times 10^5$	5	6.2966/16.4534	0.2670/7.9728
C_2: $(2.7^+$ to $3)\times 10^5$	7	6.3376/9.9517	0.0692/0.8755
C_3: $(3^+$ to $3.3)\times 10^5$	8	7.9324/9.5351	0.0006/0.2471
C_4: $(3.3^+$ to $3.6)\times 10^5$	11	8.3552/7.9985	0.8372/1.1263
C_5: $(3.6^+$ to $3.9)\times 10^5$	8	7.7914/6.1483	0.0056/0.5577
C_6: $(3.9^+$ to $4.2)\times 10^5$	7	6.6685/4.4631	0.0165/1.4420
C_7: $(4.2^+$ to $4.8)\times 10^5$	5	9.5246/5.2503	2.1474/0.0119
C_8: $> 4.8 \times 10^5$	13	11.0937/4.1993	0.3275/18.4441
sum	64	64.0000/64.0000	3.6730/30.6775

Table 8: Calculations of χ^2 statistics for Gumbels G_1/G_2 for $p_{CF} = 10^{-4}$

Class C_i	Observed frequency f_i	Expected frequency $64p_i$	χ^2-residual $\frac{(f_i-64p_i)^2}{64p_i}$
C_1: $\leq 2.5 \times 10^6$	4	3.1986/10.5076	0.2008/4.0303
C_2: $(2.5^+$ to $3)\times 10^6$	13	12.6651/15.9003	0.0089/0.5290
C_3: $(3^+$ to $3.5)\times 10^6$	15	17.5685/15.0706	0.3755/0.0003
C_4: $(3.5^+$ to $4)\times 10^6$	18	13.8709/10.2697	1.2292/5.8188
C_5: $(4^+$ to $4.5)\times 10^6$	6	8.2948/5.9238	0.6349/0.0010
C_6: $> 4.5 \times 10^6$	8	8.4021/6.3378	0.0192/0.4419
sum	64	64.0000/64.0000	2.4684/10.8214

Table 9: Calculations of χ^2 statistics for Gumbels G_3/G_4 for $p_{CF} = 10^{-5}$

For $p_{CF} = 10^{-4}$, the 64 samples of $\mathcal{F}_{max}$ have range [203311,667105], mean 3.896×10^5, and standard deviation 1.063×10^5. ML-estimates for parameters of the Gumbel distribution are $\hat{\rho} = 84002.79$ and $\hat{\mu} = 3.407 \times 10^5$. G_1=Gumb($\hat{\mu}, \hat{\rho}$) has mean 3.891×10^5 and standard

deviation 1.077×10^5. Computing parameters as if Exp(m_{FF}) were the initial distribution for interoccurrence time for rare fail-state F gives $\rho = 70082$ and $\mu = 2.9146 \times 10^5$, in which case G_2=Gumb(μ, ρ) has mean 3.3192×10^5 and standard deviation 8.9884×10^4.

The samples were grouped into eight classes by value. See Table 8. Consider significance level 0.1. G_1 has ML-estimation of two parameters from the sample data. The critical value for 5 (=8-2-1) degrees of freedom is 9.24. χ^2 statistic 3.6730 for G_1 is less than 9.24, so G_1 is not rejected. The second Gumbel G_2 is based on an assumption that the initial distribution is exactly the exponential Exp(m_{FF}). No parameters are estimated from the data. The critical value 12.02, for 7 (=8-1) degrees of freedom, is less than the G_2 statistic 30.6775; therefore, statistical evidence rejects G_2 at significance level 0.1. G_2 is also rejected at the reduced level of significance 0.05.

For $p_{CF} = 10^{-5}$, the 64 samples have range [2244658,5950128], mean 3.585×10^6, and standard deviation 7.917×10^5. ML-estimates yield G_3 = Gumb(3.218×10^6,653981.33) which has mean 3.595×10^6 and standard deviation 8.388×10^5. An initial exponential Exp(m_{FF}) yields G_4=Gumb(2.9146×10^6,700800) which has mean 3.3191×10^6 and standard deviation 8.9881×10^5.

The samples were grouped into six classes. See Table 9. Consider significance level 0.1. The critical values are 6.25 for 3 (=6-2-1), and 9.24 for 5 (=6-1), degrees of freedom. G_3 is not rejected but G_4 is rejected; however, G_4 is not rejected at the reduced level of significance 0.05 because its statistic 10.8214 is less than the critical value 11.07. ■

References

[1] E. N. Adams. Optimizing Preventive Service of Software Products, *IBM Jour. Res. Dev.*, vol. 28, no. 1, 1984, pp. 2-14.

[2] A. Avritzer and E. J. Weyuker. The Automatic Generation of Load Test Suites and the Assessment of the Resulting Software, *IEEE Trans. Software Engrg.*, vol. 21, no. 9, 1995, pp. 705-716.

[3] A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation,* Clarendon Press (Oxford University), Oxford, UK, 1992.

[4] R. E. Barlow and F. Proschan. *Mathematical Theory of Reliability*, John Wiley & Sons, Inc., NY, 1965. Republished by SIAM, Philadelphia, PA, 1996.

[5] R. E. Barlow and F. Proschan. *Statistical Theory of Reliability and Life Testing*, Holt, Reinhart and Winston, Inc., NY, 1975. Republished by TO BEGIN WITH, Silver Spring, MD, 1981.

[6] K. V. Bury. *Statistical Models in Applied Science*, John Wiley & Sons, Inc., NY, 1975. Republished by Robert E. Krieger Pub. Co., Inc., Malabar, FL, 1986.

[7] K.-Y. Cai. Censored Software-Reliability Models, *IEEE Trans. Reliab.*, vol. 46, no. 1, 1997, pp. 69-75.

[8] R. C. Cheung. A User-Oriented Software Reliability Model, *IEEE Trans. Software Engrg.*, vol. 6, no. 2, 1980, pp. 118-125.

[9] W. Farr. Software Reliability Modeling Survey, in *Handbook of Software Reliability Engineering* (M. R. Lyu, Ed.), IEEE Computer Society Press, NJ, 1996, pp. 71-115.

[10] W. Feller. *An Introduction to Probability Theory and Its Applications*, Vol. 1, John Wiley & Sons, Inc., NY, 1968.

[11] B. V. Gnedenko, Yu. K. Belyayev, and A. D. Solovyev. *Mathematical Methods of Reliability Theory*, Academic Press, NY and London, 1969.

[12] S. S. Gokhale and K. S. Trivedi. Structure-Based Software Reliability Prediction, *Proc. of Advanced Computing (ADCOMP) '97*, Channai, India, 1997, pp. 1-6. Available as TR 98/19, Center for Advanced Computing and Communication, NCSU/Duke, Raleigh/Durham, NC. **http://www.ece.ncsu.edu/cacc**

[13] E. J. Gumbel. *Statistics of Extremes*, Columbia University Press, NY, 1958.

[14] L. M. Kaufman, J. B. Dugan, and B. W. Johnson. Using Statistics of the Extremes for Software Reliability Analysis of Safety Critical Systems, *Proc. 1998 Int'l. Symp. Software Reliab. (ISSRE)*, Paderborn, Germany, Nov, 1998, pp. 355-363.

[15] L. M. Kaufman, D. T. Smith, J. B. Dugan, and B. W. Johnson. Software Reliability Analysis Using Statistics of the Extremes, *Proc. 1997 Annual Reliab. and Maint. Symp.* Philadelphia, PA, Jan., 1997, pp. 175-180. (Extension to appear in *IEEE Trans. Reliab.*)

[16] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*, Van Nostrand, Princeton, NJ, 1960. Republished by Springer-Verlag, NY, 1976.

[17] G. Q. Kenney. Estimating Defects in Commercial Software During Operational Use, *IEEE Trans. Reliab.*, vol. 42, no. 1, 1993, pp. 107-115.

[18] K. Knopp. *Theory and Application of Infinite Series*, Blackie and Son Ltd., London and Glasgow, 1951. Republished by Dover Publications, NY, 1990.

[19] M. R. Lyu (Ed.) *Handbook of Software Reliability Engineering*, IEEE Computer Society Press, NJ, 1996.

[20] R. Matthews. Far Out Forecasting, *New Scientist*, pp. 37-40, 12 Oct. 1996.

[21] D. R. Thoman, L. J. Bain, and C. E. Antle. Inferences on the Parameters of the Weibull Distribution, *Technometrics*, vol. 11, no. 3, 1969, pp. 445-460.

[22] W. Weibull. A Statistical Distribution of Wide Applicability. *J. Appl. Mech.*, vol. 18, 1951, pp. 292-297.

[23] J. A. Whittaker. Stochastic Software Testing, *Ann. of Software Engrg.*, vol. 4, 1997, pp. 115-131.

[24] J. A. Whittaker and M. G. Thomason. A Markov Chain Model for Statistical Software Testing, *IEEE Trans. Software Engrg.*, vol. 20, no. 10, 1994, pp. 812-824.