Cramer–von Mises Tests for the Compatibility of Two Software Operating Environments

D. R. JESKE

Department of Statistics
University of California
Riverside, CA 92521
(*daniel.jeske@ucr.edu*)

R. A. LOCKHART

Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC Canada
(*lockhart@sfu.ca*)

M. A. STEPHENS

Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC Canada
(*stephens@stat.sfu.ca*)

Q. ZHANG

Department of Statistics
University of California
Riverside, CA 92521
(*qzhan007@student.ucr.edu*)

Higher-precision inferences about impending software failures can be achieved when the same software reliability model that fits failure data from the test interval also fits data from the field interval. If the test and field environments differ significantly in terms of how the software is used, then a single model for the pooled data may not be adequate. In this article we formulate the hypothesis of compatible test and field environments in terms of a statistical hypothesis and develop a Cramer–von Mises (CvM) test procedure within the context of a well-known nonhomogeneous Poisson process software reliability model. The CvM test has a compelling advantage over a previously proposed likelihood ratio test (LRT_0), because it does not require specification of the class of alternatives, which are frequently unknown for real-life problems. Moreover, although there are existence issues with LRT_0, the CvM test always exists. An asymptotic approximation for the p value of the CvM test is derived, and an algorithm for a small-sample bootstrap approximation is presented. A simulation study shows that the CvM test works well for the class of alternatives for which LRT_0 also would work well and continues to work well for other alternatives for which LRT_0 has no statistical motivation or otherwise has existence problems. Data from a real software project are used to illustrate the hypothesis testing procedures.

KEY WORDS: Cramer–von Mises statistic; Likelihood ratio; Software reliability.

1. INTRODUCTION

The reliability of software can be a key product differentiator and also can be a defining characteristic of a company's image. A classic approach for modeling software reliability is to suppose that the cumulative number of failures observed through cumulative exposure time t, say $N(t)$, follows a nonhomogeneous Poisson process (NHPP) with mean value function $M(t;\theta)$, where θ is a vector of unknown parameters. Discussions of the assumptions and practicality of a vast number of NHPP software reliability models have been provided by Xie (1991), Lyu (1996), and Pham (1999). Most NHPP models assume a parametric form for the mean value function. Goel and Okumoto (1979; GO hereinafter) proposed a widely used NHPP model that assumes that the mean value function is $M(t;\theta) = a(1 - e^{-bt})$. Here $\theta = (a, b)$, with a representing the expected number of initial faults in the software and b representing the failure rate of an individual fault. The failure rate function is the derivative of the mean value function, which is $\lambda(t;\theta) = abe^{-bt}$ for the GO model. The underlying parameters of NHPP models are typically estimated using maximum likelihood procedures based on grouped failure counts.

Figure 1 provides a conceptual view of the failure data corresponding to a software system, where the points $(u_i, Y_i)_{i=1}^{l}$ correspond to the cumulative exposure time and cumulative number of failures through the ith measurement epoch of the test interval. Similarly, the points $(v_i, Z_i)_{i=1}^{m}$ correspond to the cumulative exposure time and cumulative number of failures through the ith measurement epoch of the field interval. During the test interval, the software is executed in a way that ideally mimics how actual users will use it in the field. As failures are observed, the software is modified to remove the underlying faults. The debugging process continues throughout the test interval, after which the software is released into the field. With the software in the field, the "find-and-fix" process continues, as depicted in the right half of Figure 1.

Each point in Figure 1 corresponds to a time epoch in which the cumulative number of failures and the cumulative exposure time are observed. The time epochs might be weekly measurements but in general are arbitrary. Note that even if the time epochs are equally spaced, then the corresponding cumulative exposure times may not be so, due to varying levels of software use among the time epochs. A typical use for the data illustrated by Figure 1 is to fit a curve to the points and then use the derivative of the fitted curve as an estimate of the failure rate function of the software. Due to the increasing but bounded nature of the fitted curve, the failure rate $\lambda(t;\theta)$ will eventually (if not immediately) be decreasing toward 0, and extrapolations can be made for the failure rate at future time epochs.

If software is tested in a manner that mirrors how users will use it in a field environment, then inferences about future fail-

TECHNOMETRICS, FEBRUARY 2008, VOL. 50, NO. 1
DOI 10.1198/004017007000000335

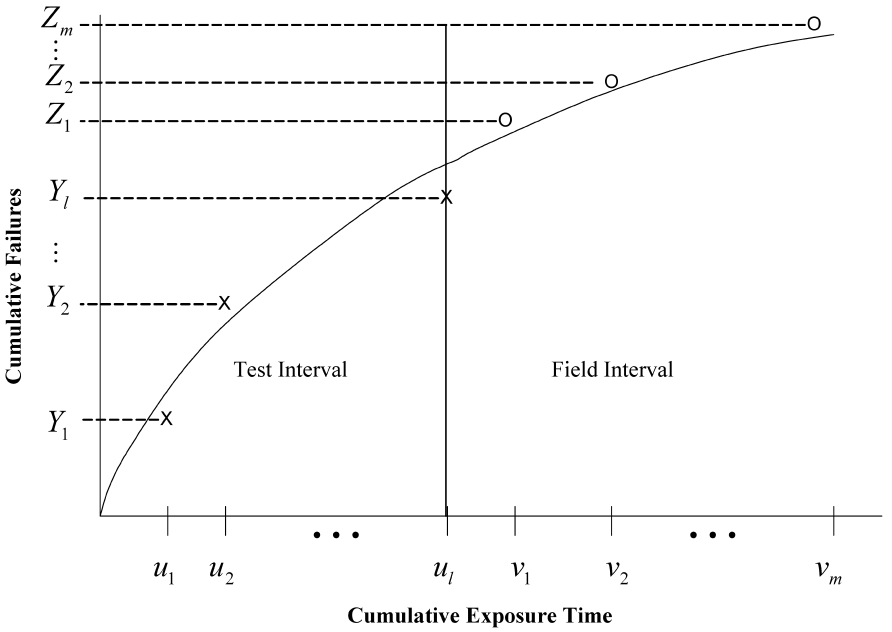

Figure 1. Schematic of fundamental software failure data.

ures obtained based on the test data will be valid from a user's point of view. When the test and field environments differ, the failure data from the field is an important source of information to use in correcting the otherwise invalid inferences obtained from the test data. One way in which the environments could differ is if faults were discovered in the field more slowly than the rate at which they were discovered during the test period. This can happen, because users are merely using the software for its intended function, whereas testers are often purposely trying to break the code. On the other hand, if the software is tested only lightly, then the fault discovery rate in the field can be higher than in the test period. For similar reasons, the user-perceived number of faults can differ between the test and field environments. For example, if the testing profile is more uniform than the usage profile, then testers may discover faults in the code that users will not experience. The software operational profile (see, e.g., Musa 1975) defines how users interact with the software. If testing does not follow the operational profile, then the software could appear reliable from the tester's perspective but unreliable from the user's perspective. The opposite situation—the software appearing unreliable from the tester's perspective but reliable from the user's perspective—is also possible. Unfortunately, operational profile testing is rarely done, because it is expensive and difficult.

The use of calibration factors has been proposed as a possible way to reconcile differences between test and field environments (see, e.g., Huang, Kuo, Lyu, and Lo 2000; Zhang, Jeske, and Pham 2002; Jeske and Zhang 2004). Jeske, Zhang, and Pham (2005) proposed applying calibration factors to the GO model by specifying a mean value function of the form

$$M(t; a, b, K_1, K_2) = \begin{cases} a(1 - e^{-bt}), & t \le u_l \\ a(1 - e^{-bu_l}) & \\ \quad + K_1 a e^{-bu_l}\left(1 - e^{-K_2 b(t-u_l)}\right), & t > u_l. \end{cases} \quad (1)$$

In this model a denotes the expected number of initial faults in the software, and b represents the failure rate of an individual fault. The calibration factors are shown explicitly as (K_1, K_2) and can be interpreted as follows. The expected number of faults remaining in the code at the end of the test interval is $a - a(1 - e^{-bu_l}) = ae^{-bu_l}$, indicating that K_1 is being used to scale the residual faults appropriately so that $K_1 ae^{-bu_l}$ can be used as the number of initial faults once the field interval begins. Similarly, K_2 is being used to scale the average failure rate of a fault in the test interval to the appropriate value, $K_2 b$, in the field environment. It is clear from (1) that under $H_0 : (K_1, K_2) = (1, 1)$, we have $M(t; a, b) = a(1 - e^{-bt})$ for all t. Thus when the calibration factors are unity, the two-part mean value function in (1) reduces to a one-part mean value function. The one-part model effectively combines the test and field failure data and advantageously improves the precision of inferences concerning future failures during the field interval.

To illustrate the advantage of being able to use the one-part model associated with (1), consider Figure 2, which shows the cumulative number of failures versus the cumulative exposure time for a software system developed for a Brazilian telephonic switching system (see Martini, Kanoun, and de Souza 1990). The units of exposure time are 10-day periods of calendar time. The vertical line in the figure denotes the transition of the software into the field environment, separating periods 1–42, which correspond to the test interval, and periods 43–81, which correspond to the field interval. It will be shown formally in a subsequent section that the one-part model is not rejected for these data.

Define $S(v_j)$ to be the time to next failure, given that the process has been observed through field time v_j. Then $P[S(v_j) > s] = \exp\{-[M(v_j + s; a, b, K_1, K_2) - M(v_j; a, b, K_1, K_2)]\}$, and because (1) is bounded, it follows that the probability of the event $S(v_j) = \infty$ is nonzero. Consequently, the distribution of $S(v_j)$ does not have a mean, and a more useful summary of the distribution is a quantile, for example, the median $Q_{.5}(v_j) =$

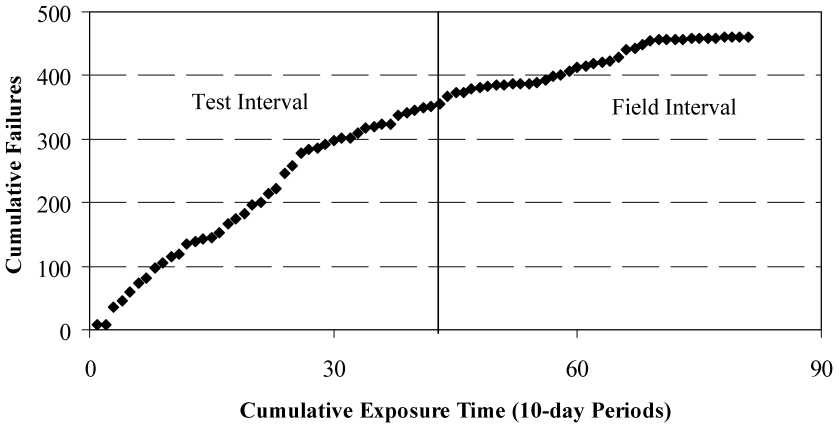

Figure 2. Software failure data for the Brazilian switching system.

$-\log[1 - \log 2/\{K_1 a e^{-bu_l} e^{-K_2 b(v_j - u_l)}\}]/(K_2 b)$, which exists provided that $K_1 a e^{-bu_l} e^{-K_2 b(v_j - u_l)} > \log 2$.

Under the one-part model, $Q_{.5}(v_j)$ simplifies to $-\log[1 - \log 2/(ae^{-bv_j})]/b$, and fitting the model sequentially using the cumulative failure data (test and field) available through $v_1, \ldots, v_{39}$ leads to the estimated median times to next failure shown in Figure 3. Also shown are pointwise 90% confidence intervals (CIs) for the median failure time computed using the formula that is detailed in Appendix B. For example, a 90% CI for $Q_{.5}(v_{39})$ is (.31 , .56) or, equivalently, 3.1 to 5.6 days.

In contrast, if the two-part model is (inefficiently) used, then difficulties arise when using it to estimate the median times to next failure. Figure 4 is analogous to Figure 3, except that the two-part model is used at each field period. First, it can be seen that the median estimates in the field interval are not available for periods 1–3 or periods 24–29. The difficulty in periods 1–3 is that there are insufficient field observations to estimate the calibration parameters (K_1, K_2), and the difficulty in periods 24–29 is that necessary and sufficient conditions on the field observations for their maximum likelihood estimates to exist are not fulfilled. Comparing Figures 3 and 4 clearly shows that the CIs for the medians, when they do exist in the two-part model analysis, are significantly wider. This is especially true for the early field periods, but even during the last field period, the one-part model shrinks the width of the confidence interval by about 50%.

Our example illustrates the advantage of being able to use a one-part model. At the same time, inappropriate use of a one-part model clearly will give biased estimates of important quantities concerning future failures. Applications need a statistical method to test whether or not a one-part model is appropriate to either take advantage of situations when it is appropriate or avoid adverse consequences when it is not. Although plots such as Figure 2 provide some intuition as to the validity of a one-part model, they are not adequate substitutes for an objective statistical procedure.

The LRT of H_0 under (1), say LRT_0, is a natural test to consider for testing the compatibility of the test and field environments. The appropriate asymptotic theory for LRT_0 follows from results of van Pul (1992) by letting the expected number of faults (a in our model) in the software increase. As the expected number of faults in the software increases, the number of observed failures also increases, and the more conventional interpretation of asymptotic theory follows. Note that with modern software architecture and development practices, it would not be unusual for the expected number of faults to be large; however, using LRT_0 requires a sufficiently large number of observed failures from *both* environments before its approximate null distribution will be valid, and furthermore, the maximum likelihood estimators (MLEs) of the two-part model needed to implement the test will not exist at the very beginning of the field interval, and even later individual field observations can

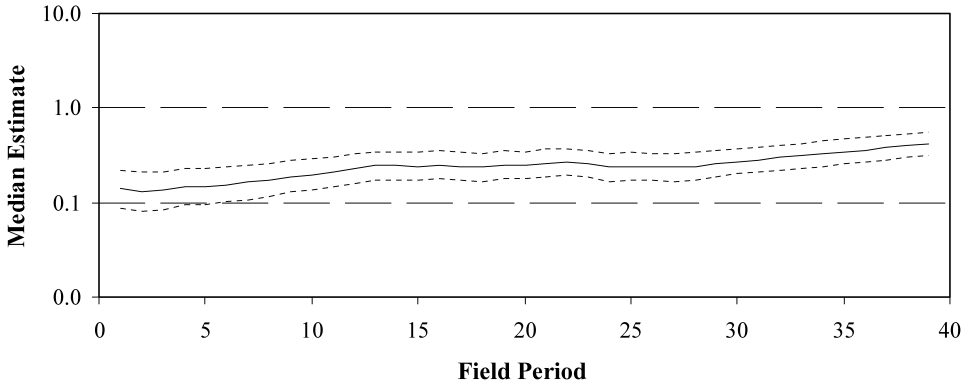

Figure 3. Median times to next failure and 90% CI limits for the Brazilian switching system based on a one-part model.

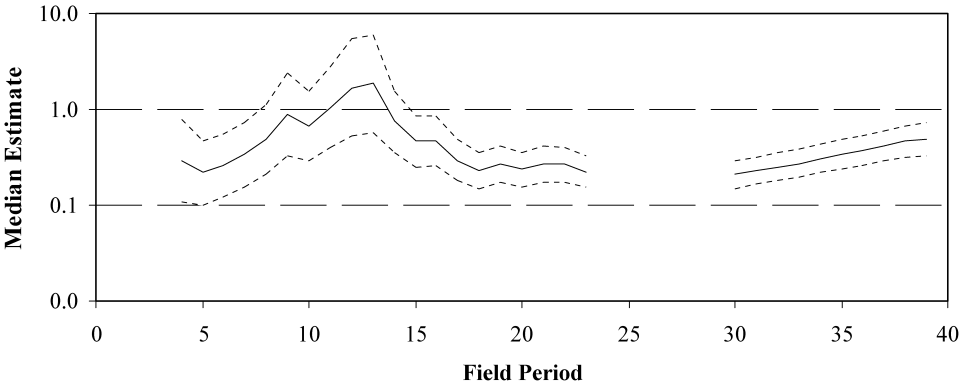

Figure 4. Median times to next failure and 90% CI limits for the Brazilian switching system based on a two-part model.

have an appreciable influence on whether the conditions that ensure their existence are satisfied.

In this article we derive a Cramer–von Mises (CvM) test as an alternative test to LRT_0. The CvM test has two advantages over LRT_0. First, in cases where (1) is a suitable representation of the alternatives to the null model, the CvM test is always computable, thus eliminating the problems associated with the existence of LRT_0 in the early stages of the field interval. Second, in cases where the alternatives are not contained within the family of models described by (1), the CvM test is a well-motivated and applicable test, whereas LRT_0, which explicitly relies on (1) capturing the possible alternatives, has no statistical justification.

The rest of the article is organized as follows. In Section 2 we review the development of LRT_0, and in Section 3 we develop the CvM test. An asymptotic procedure and a bootstrap resampling approach for approximating the p value of the CvM test are proposed, and the CvM test and the two approximations for its p value are illustrated in the context of a real software failure data set. In Section 4 we report results from a simulation experiment used to evaluate the size and power of the LRT_0 and CvM tests. Alternatives described by (1), as well as a class of alternatives not captured by (1), are used to study the power of the tests. We conclude with a summary in Section 5.

2. LIKELIHOOD RATIO STATISTIC

2.1 Computation

Define $(u_0, y_0) = (0, 0)$ and $(v_0, z_0) = (u_l, y_l)$. The likelihood function for the NHPP model with the two-part mean value function given in (1) is then

$$\begin{aligned} &L(a, b, K_1, K_2) \\ &\quad \propto e^{-a(1-e^{-bu_l})} \prod_{i=1}^{l} \frac{[a(e^{-bu_{i-1}} - e^{-bu_i})]^{y_i - y_{i-1}}}{(y_i - y_{i-1})!} \\ &\qquad \times e^{-K_1 a e^{-bu_l}(1-e^{-K_2 b(v_m - u_l)})} \\ &\qquad \times \prod_{i=1}^{m} \frac{[K_1 a e^{-bu_l}(e^{-K_2 b(v_{i-1}-u_l)} - e^{-K_2 b(v_i - u_l)})]^{z_i - z_{i-1}}}{(z_i - z_{i-1})!}. \end{aligned}$$

The reparameterization (a, b, c, d), where $c = K_1 a e^{-bu_l}$ and $d = K_2 b$, simplifies the optimization problem for obtaining the MLE $(\hat{a}, \hat{b}, \hat{K}_1, \hat{K}_2)$. The reparameterized likelihood is

$$\begin{aligned} &L(a, b, c, d) \\ &\quad \propto e^{-a(1-e^{-bu_l})} \prod_{i=1}^{l} \frac{[a(e^{-bu_{i-1}} - e^{-bu_i})]^{y_i - y_{i-1}}}{(y_i - y_{i-1})!} \\ &\qquad \times e^{-c(1-e^{-d(v_m - u_l)})} \\ &\qquad \times \prod_{i=1}^{m} \frac{[c(e^{-d(v_{i-1}-u_l)} - e^{-d(v_i - u_l)})]^{z_i - z_{i-1}}}{(z_i - z_{i-1})!}, \end{aligned}$$

from which it is evident that the MLE $(\hat{a}, \hat{b}, \hat{c}, \hat{d})$ can be obtained by first finding $(\hat{a}, \hat{b})$ that maximizes the first term of the right side (which represents the contribution of the test data to the full likelihood), and then finding $(\hat{c}, \hat{d})$ that maximizes the second term of the right side. The MLE of (K_1, K_2) is then obtained from $\hat{K}_1 = \hat{c} e^{\hat{b} u_l} / \hat{a}$ and $\hat{K}_2 = \hat{d}/\hat{b}$. Moreover, it can be shown that $(\hat{a}, \hat{b})$ can be found by first finding $\hat{b}$ as the solution to the equation

$$\frac{y_l u_l e^{-\hat{b}u_l}}{1 - e^{-\hat{b}u_l}} = \sum_{i=1}^{l} (y_i - y_{i-1}) \frac{u_i e^{-\hat{b}u_i} - u_{i-1} e^{-\hat{b}u_{i-1}}}{e^{-\hat{b}u_{i-1}} - e^{-\hat{b}u_i}}, \qquad (2)$$

and then computing $\hat{a} = y_l/(1 - e^{-\hat{b}u_l})$. Similarly, $(\hat{c}, \hat{d})$ can be found by first finding $\hat{d}$ as the solution to the equation

$$\begin{aligned} &\frac{(z_m - y_l)(v_m - u_l) e^{-\hat{d}(v_m - u_l)}}{1 - e^{-\hat{d}(v_m - u_l)}} \\ &\quad = \sum_{i=1}^{m} (z_i - z_{i-1}) \\ &\qquad \times \frac{(v_i - u_l) e^{-\hat{d}(v_i - u_l)} - (v_{i-1} - u_l) e^{-\hat{d}(v_{i-1} - u_l)}}{e^{-\hat{d}(v_{i-1} - u_l)} - e^{-\hat{d}(v_i - u_l)}}, \end{aligned} \qquad (3)$$

and then computing $\hat{c} = (z_m - y_l)/(1 - e^{-\hat{d}(v_m - u_l)})$. Simple bisection routines can be used to find the solutions of (2) and (3).

The likelihood function for the NHPP model corresponding to the one-part mean value function [i.e., the likelihood function

reduced by $H_0 : (K_1, K_2) = (1, 1)$] is

$$L_{H_0}(a, b) \propto e^{-a(1-e^{-bv_m})} \prod_{i=1}^{l} \frac{[a(e^{-bu_{i-1}} - e^{-bu_i})]^{y_i - y_{i-1}}}{(y_i - y_{i-1})!} \times \prod_{i=1}^{m} \frac{[a(e^{-bv_{i-1}} - e^{-bv_i})]^{z_i - z_{i-1}}}{(z_i - z_{i-1})!}.$$

The constrained MLE $(\tilde{a}, \tilde{b})$ is obtained by first finding $\tilde{b}$ as the solution to

$$\frac{z_m v_m e^{-\tilde{b}v_m}}{1 - e^{-\tilde{b}v_m}} = \sum_{i=1}^{l} (y_i - y_{i-1}) \frac{u_i e^{-\tilde{b}u_i} - u_{i-1} e^{-\tilde{b}u_{i-1}}}{e^{-\tilde{b}u_{i-1}} - e^{-\tilde{b}u_i}} + \sum_{i=1}^{m} (z_i - z_{i-1}) \frac{v_i e^{-\tilde{b}v_i} - v_{i-1} e^{-\tilde{b}v_{i-1}}}{e^{-\tilde{b}v_{i-1}} - e^{-\tilde{b}v_i}},$$

and then computing $\tilde{a} = z_m/(1 - e^{-\tilde{b}v_m})$. The LRT statistic of H_0 is then $\mathrm{LRT}_0 = -2 \log \Lambda$, where $\Lambda = L_{H_0}(\tilde{a}, \tilde{b})/L(\hat{a}, \hat{b}, \hat{K}_1, \hat{K}_2)$, and the test rejects if $\mathrm{LRT}_0 > \chi^2_{2,\alpha}$.

The existence of the solutions $\hat{b}, \hat{d}$, and $\tilde{b}$ is not automatic. However, when they do exist, they trivially imply the existence of $\hat{a}, \hat{c}$, and $\tilde{a}$. Jean (1998) showed that a necessary and sufficient condition for $\hat{b}$ to exist is

$$\sum_{i=1}^{l} (y_i - y_{i-1})(u_i + u_{i-1}) < y_l u_l. \tag{4}$$

Similarly, it can be shown that a necessary and sufficient condition for $\hat{d}$ to exist is

$$\sum_{i=1}^{m} (z_i - z_{i-1})(v_i + v_{i-1}) < (z_m - y_l)(v_m - u_l), \tag{5}$$

and that $\tilde{b}$ exists if and only if

$$\sum_{i=1}^{l} (y_i - y_{i-1})(u_i + u_{i-1}) + \sum_{i=1}^{m} (z_i - z_{i-1})(v_i + v_{i-1}) < z_m v_m. \tag{6}$$

Thus it follows that LRT_0 of H_0 will exist if and only if all three conditions (4)–(6) are satisfied. One of the problems that will because evident later is that if the alternatives in the field do not follow the two-part mean value function specified by (1), then LRT_0 often will not exist due to failure of either (5) or (6). Moreover, even when the two-part model specified by (1) is correct, there can be times during the data collection from the field during which (5) and/or (6) is violated, and LRT_0 will not exist.

2.2 Computation of *p* Values

In what follows, the p value is computed by comparing LRT_0 to its asymptotic chi-squared distribution with 2 degrees of freedom. It also would be possible to use the parametric bootstrap in the way discussed in Section 3.2.2 for the CvM test to compute a p value.

2.3 Example

Returning to the data in Figure 2 and following the aforementioned sequence of steps results in $\hat{a} = 586.25$, $\hat{b} = 2.184 \times 10^{-2}$, $\hat{K}_1 = .660$, and $\hat{K}_2 = 1.434$. In addition, $(\tilde{a}, \tilde{b})$ is $(524.98, 2.60 \times 10^{-2})$, which leads to $\mathrm{LRT}_0 = 1.34$. The p value of the test statistic is .512 and the LRT_0-based test does not reject H_0. Thus, as Figure 2 suggests, it would be acceptable to fit a one-part model to both the test and field data for the purpose of making inference about future failures, such as estimating their median failure times. In particular, the median of the next failure time based on all of the data through time v_{39} is $Q_{.5}(v_{39}) = -\log[1 - \log 2/(ae^{-bv_{39}})]/b$, and using $(\tilde{a}, \tilde{b})$, its estimate is .42. The results in Appendix B can be used to compute an approximate 90% confidence interval for $Q_{.5}(v_{39})$, namely (.31, .56), as was reported in Section 1.

3. CRAMER–VON MISES STATISTIC

As discussed previously, the use of LRT_0 depends on the correct specification of the class of field alternatives. If the field alternatives are correctly specified by (1), then LRT_0 is an optimal test; however, if the field alternatives are not correctly specified by (1), then the use of LRT_0 lacks statistical justification. When little is known about what type of alternative to a one-part model can be realized in the field, it is preferable to use a test for the adequacy of the one-part model that does not depend on specifying a class of alternatives.

An early detection test of H_0 was derived by Jeske and Zhang (2006) that is independent of the field part of the mean value function. But, their test was intended primarily as a gap solution for use between the time at which the field interval starts and the time at which a sufficient number of observations from the field are available, to ensure that LRT_0 exists and can be used. Using the early detection test with a large number of field observations would be computationally tedious.

In this section we derive another test, the CvM test, which is a competitor to LRT_0 in the sense that it provides a test for the adequacy of the one-part mean value function and is not necessarily intended to be an early detection test. On the other hand, the CvM test has two apparent advantages over LRT_0: It is easier to compute and is always available, and it has motivation independent of the field part of the mean value function and thus is an omnibus test that may be expected to have satisfactory power against an arbitrary alternative.

3.1 Derivation

Numerous common tests of goodness of fit for a sample of data, are based on a comparison of the empirical cumulative distribution function $F_n(x)$, defined by

$$F_n(x) = \frac{\#\{\text{observations} \le x\}}{n},$$

and a theoretical cumulative distribution function $F(x)$. In the situation considered here, there is a strong analogy between the cumulative process $N(t)$ and its null expected value $M(t; \theta)$ on the one hand and $F_n(x)$ and $F(x)$ on the other hand. Various statistics have been suggested to summarize the discrepancy between $F_n(x)$ and $F(x)$ (see, e.g., Stephens 1986 for a survey).

Extensive Monte Carlo experience suggests that quadratic statistics, such as CvM and others, have superior power properties in the goodness-of-fit context, and we are led to suggest an analog of the CvM test for our situation. The classic CvM statistic is defined as

$$W^2 = n\int_{-\infty}^{\infty}\{F_n(x) - F(x)\}^2\, dF(x).$$

The analog for an NHPP would be a multiple of

$$\int_0^T \{N(t) - M(t;\hat\theta)\}^2\, dM(t;\hat\theta)$$

where T is the end of the observation period and $\hat\theta$ is an appropriate estimator of θ.

When adapting this analogy to our context, several points arise:

a. The hypothesized form of $M(t;\theta)$ is $M(t;a,b) = a(1 - e^{-bt})$, and we are concerned only about its shape for $t > u_l$, because we assume that the model is correctly specified during the test interval.

b. We have data only at the time points $\{u_i\}_{i=1}^{l}$ and $\{v_i\}_{i=1}^{m}$.

c. An appropriate estimator (a,b) is the MLE based exclusively on the test data, because in that interval the model is assumed to be correctly specified, and there is typically a sufficient amount of test data to yield an accurate estimate.

d. As mentioned previously, the total amount of data will be large only if the parameter a is large, and thus a must play the role of n in the asymptotic theory of the test.

These points lead us to approximate the integral as a sum, start the integration at u_l, and compare $N(t) - N(u_l)$ to the hypothesized form of $M(t;\theta) - M(u_l;\theta) = a(e^{-bu_l} - e^{-bt})$. Furthermore, although $(\hat a, \hat b)$ has been previously defined as the MLE of (a,b) under the two-part model, examining (2) demonstrates that it also can be viewed as the MLE of (a,b) based only on the test data. Consequently, our proposed statistic in standardized form (necessary for subsequent asymptotic considerations) becomes

$$\text{CvM} = \sum_{i=1}^{m}(e^{-\hat b v_{i-1}} - e^{-\hat b v_i})[Z_i - Y_l - \hat a(e^{-\hat b u_l} - e^{-\hat b v_i})]^2/\hat a. \tag{7}$$

The form of CvM is strongly analogous to suggestions made by Choulakian, Lockhart, and Stephens (1994) and Spinelli and Stephens (1997) for goodness-of-fit testing for discrete distributions in general and the Poisson distribution in particular.

Although we propose (7) as the CvM statistic, other possibilities could be considered. For example, we could compare $N(t) - N(u_l)$ to the hypothesized form of $M(t;\theta) - M(u_l;\theta)$ over both the test and field intervals. We choose not to do this, because our assumption is that the model is correct in the test interval, and thus it seems that including the test data in the computation of the CvM test statistic would only mute its sensitivity to alternatives. Another possibility would be to estimate (a,b) by fitting the null model to the pooled test and field data, because the null distribution of the CvM statistic is ultimately desired. However, this estimation approach would have practical problems in applications in which the null model is not correct. Very often in these cases, for example, the MLE of (a,b) based on the pooled data would not exist, and the practitioner would be left without a usable test procedure.

Test statistics could be based on other comparisons of $N(t)$ to $M(t;\theta)$. Our CvM statistics and the variants suggested in the previous paragraph are quadratic in nature. We also might consider linear statistics of the form $\int_0^T w(t;\hat\theta)\{N(dt) - M(dt;\hat\theta)\}$ or, equivalently, $\int_0^T w(t;\hat\theta)\{N(dt) - \lambda(t;\hat\theta)\,dt\}$, where $w(t;\hat\theta)$ is a suitably chosen weight function. Although such statistics certainly deserve consideration, we do not pursue them here.

3.2 Approximate p Values for CvM

In this section we derive an asymptotic approximation for the p value of the CvM statistic, and also describe a bootstrap approximation that can be used when the conditions for the asymptotic approach are suspect.

3.2.1 Asymptotic Approximation. It is possible to derive the asymptotic null distribution of the statistic CvM as a tends to infinity. Define $\delta = (\delta_T^t, \delta_F^t)^t$ to be the column vector with entries

$$e^{-bu_0} - e^{-bu_1}, \ldots, e^{-bu_{l-1}} - e^{-bu_l},$$
$$e^{-bv_0} - e^{-bv_1}, \ldots, e^{-bv_{m-1}} - e^{-bv_m}.$$

To state the asymptotic theory, we need several definitions. Define

$$I^* = \begin{bmatrix} \sum_{i=1}^{l}\delta_i & \sum_{i=1}^{l}\partial\delta_i/\partial b \\ \sum_{i=1}^{l}\partial\delta_i/\partial b & \sum_{i=1}^{l}(\partial\delta_i/\partial b)^2/\delta_i \end{bmatrix}$$

and

$$A = (I^*)^{-1}\begin{bmatrix} 1_{1\times l} & 0_{1\times m} \\ \nabla\log\delta_T^t & 0_{1\times m}\end{bmatrix},$$

where $\nabla\log\delta_T$ is the $l\times 1$ vector whose ith element is $(\partial\delta_{T,i}/\partial b)/\delta_{T,i}$. Next, let

$$M = I_{(l+m)\times(l+m)} - \begin{bmatrix}\delta & \dfrac{\partial\delta}{\partial b}\end{bmatrix}A \qquad \text{and} \qquad Q = S^t\Delta^2 S,$$

where

$$S = \begin{bmatrix} 0_{l\times l} & 0_{l\times m} \\ 0_{m\times l} & S^*\end{bmatrix},$$

with S^* the $m\times m$ matrix with elements 0 above the diagonal and elements unity on and below the diagonal, and where Δ is a diagonal matrix with elements $\{\sqrt{\delta_i}\}_{i=1}^{l+m}$. The following proposition is proved in Appendix A.

Proposition. The limiting null distribution, as $a\to\infty$, of the CvM statistic defined in (7) is the distribution of $\sum_{i=1}^{l+m}\lambda_i\chi_i^2$, where the χ_i^2 are independent, 1-degree of freedom, chi-squared random variables and the λ_i are the eigenvalues of the matrix $\Delta M^t Q M\Delta$.

The proposition facilitates the computation of approximate p values corresponding to an observed value of the CvM test, say CvM_{obs}, as follows:

1. Obtain $\hat a$ and $\hat b$ using only the test data.
2. Use $\hat a$ and $\hat b$ to compute the entries in the matrices M, Q, and Δ.
3. Find the eigenvalues $\lambda_1, \ldots, \lambda_{l+m}$ of $\Delta M^t Q M\Delta$.

For $k = 1$ to B:
Simulate $(u_i, Y_i^*)_{i=1}^{l}$ from $Y_i^*|Y_{i-1}^* \sim Y_{i-1}^*$ + Poisson $[M_0(u_i; \hat{a}, \hat{b}, 1, 1) - M_0(u_{i-1}; \hat{a}, \hat{b}, 1, 1)]$.
Simulate $(v_i, Z_i^*)_{i=1}^{m}$ from $Z_i^*|Z_{i-1}^* \sim Z_{i-1}^*$ + Poisson $[M_0(v_i; \hat{a}, \hat{b}, 1, 1) - M_0(v_{i-1}; \hat{a}, \hat{b}, 1, 1)]$.
Compute the MLE $(\hat{a}^*, \hat{b}^*)$ of (a, b) using the booststrap sample $(u_i, Y_i^*)_{i=1}^{l}$.
Compute the $\text{CvM}^*(k) = \sum_{i=1}^{m}(e^{-\hat{b}^* v_i} - e^{-\hat{b}^* v_{i-1}})/\hat{a}^*[Z_i^* - Y_l^* - \hat{a}^*(e^{-\hat{b}^* u_l} - e^{-\hat{b}^* v_i})]^2$.
Continue.
Estimate the p value as the [number of $\text{CvM}^*(k)$ values $> \text{CvM}_{\text{obs}}$]/B.

Figure 5. Bootstrap algorithm for approximating the p value of the CvM statistic.

4. Compute an asymptotic p value as $p \approx P(\sum_{i=1}^{l+m} \lambda_i \chi_i^2 > \text{CvM}_{\text{obs}})$, using either the numerical inversion algorithm of Imhof (1961) or an approximation to the distribution of $\sum_{i=1}^{l+m} \lambda_i \chi_i^2$ of the form $c_1 + c_2 \chi_{c_3}^2$, where the constants $\{c_i\}_{i=1}^{3}$ are found by matching the first three moments. Genest, Lockhart, and Stephens (2002) found that the three-moment approximation is very accurate in the upper tail, where p values are usually required.

3.2.2 Bootstrap Approximation. An alternative approach to approximating the p value is to use a parametric bootstrap approach. Figure 5 shows an algorithm for implementing this approach.

3.3 Example

When computing LRT_0 for the example in Section 2, we found $\hat{a} = 586.25$ and $\hat{b} = 2.184 \times 10^{-2}$. It is an easy calculation to show that the CvM statistic in (7) evaluates to .0728. Straightforward calculations using the R programming language show that the first 10 eigenvalues of $\Delta M^t Q M \Delta$ are 9.28×10^{-2}, 3.08×10^{-3}, 9.28×10^{-4}, 4.65×10^{-4}, 2.78×10^{-4}, 1.87×10^{-4}, 1.34×10^{-4}, 1.01×10^{-4}, 7.96×10^{-5}, and 6.44×10^{-5}. Using Imhof's (1961) inversion algorithm gives an asymptotic p value of .394. On the other hand, using the bootstrap approximation with $B = 10{,}000$ gives an approximate p value of .388. The agreement between the two p values is quite good. Consistent with the conclusions drawn from LRT_0, the adequacy of the one-part model is not rejected by the CvM test.

4. SIZE AND POWER OF TEST STATISTICS

To examine the size and power of LRT_0 and CvM, we consider a mean value function of the form

$$M_c(t; a, b, K_1, K_2) = \begin{cases} a(1 - e^{-bt}), & t \le u_l \\ a(1 - e^{-bu_l}) + \dfrac{K_1 a e^{-bu_l}[1 - e^{-K_2 b(t-u_l)}]}{1 + ce^{-K_2 b(t-u_l)}}, & t > u_l. \end{cases} \quad (8)$$

Here (a, b, K_1, K_2) have the same interpretation as given previously and $c \ge 0$ is an additional parameter to be discussed in what follows. Note that $M_0(t; a, b, K_1, K_2)$ is the class of mean value functions corresponding to the two-part mean value function defined by (1), and $M_0(t; a, b, 1, 1)$ corresponds to the reduced one-part mean value function.

LRT_0 is well motivated for testing the adequacy of $M_0(t; a, b, 1, 1)$ within the class $M_0(t; a, b, K_1, K_2)$, and we would expect that finding a test of the same size that has better power to not be an easy task. However, for $c > 0$, LRT_0 lacks motivation and is no longer a valid test for the adequacy of $M_0(t; a, b, 1, 1)$ within the class $M_c(t; a, b, K_1, K_2)$, for $c > 0$. On the other hand, the CvM test defined in Section 3 does not explicitly require specification of the form of the alternative in the field. The CvM test is valid for testing $M_0(t; a, b, 1, 1)$ versus *any* unspecified alternative. The only assumption made by the CvM test is that the mean value function in the test interval is $M_0(t; a, b, 1, 1)$. In particular, it is a valid test for the adequacy of $M_0(t; a, b, 1, 1)$ versus $M_c(t; a, b, K_1, K_2)$, for $c > 0$.

The parameter c is referred to as a learning parameter. When c is nonzero, the term $1 + ce^{-K_2 b(t-u_l)}$ slows down the initial rate at which faults are discovered in the field and gives the mean value function a convex shape in the early portion of the field environment. As time in the field environment increases, $1 + ce^{-K_2 b(t-u_l)}$ approaches unity, and the mean value function transitions to a concave shape. For $c > 0$, the field portion of $M_c(t; a, b, K_1, K_2)$ takes on a stretched "S" shape. S-shaped mean value functions are viable alternatives to concave shapes (see, e.g., Yamada, Ohba, and Osaki 1983), and as such, the alternatives $M_c(t; a, b, K_1, K_2)$, $c > 0$, are plausible and provide an interesting context for evaluating the power of the CvM test of Section 3 and to demonstrate the difficulties associated with using LRT_0.

4.1 Case 1: $c = 0$

If a practitioner prefers to use the CvM test on the basis of not having any specific knowledge of the type of alternatives that could arise in the field, then he or she might be interested in knowing how well it does compared with LRT_0 when LRT_0 is known to be optimal. In this section we examine the size and power of the CvM and LRT_0 tests of the adequacy of $M_0(t; a, b, 1, 1)$ within the class $M_0(t; a, b, K_1, K_2)$. We investigate this question through simulations using the R programming language.

Taking u and v to be the vectors associated with the BSS application, we simulated sample paths $\{(u_i, Y_i)_{i=1}^{42}, (v_i, Z_i)_{i=1}^{39}\}$ from an NHPP with mean value function given by $M_0(t; a, b, 1, 1)$ for various choices of (a, b). Values for a, the expected number of initial faults, came from the set $\{100, 200, 400\}$. The choices for a reflect the number of faults found in the actual BSS data set, as well as consideration of the size of

Table 1. Simulation estimates of power for nominal $\gamma = .1$ tests using (u, v) from the Brazilian switching system application

		$(K_1, K_2) = (.5, .5)$		$(K_1, K_2) = (.5, 1.0)$		$(K_1, K_2) = (.5, 1.5)$	
a	*FRE*	LRT_0	CvM	LRT_0	CvM	LRT_0	CvM
100	2/3	.50	.54	.23	.31	.18	.20
	3/4	.48	.46	.27	.28	.25	.19
	4/5	.48	.38	.25	.21	.25	.16
200	2/3	.83	.86	.47	.52	.33	.32
	3/4	.80	.82	.48	.47	.37	.28
	4/5	.78	.77	.46	.42	.38	.27
400	2/3	.99	.99	.76	.79	.51	.46
	3/4	.98	.99	.79	.79	.63	.50
	4/5	.99	.99	.74	.78	.64	.50

new software development efforts. Values for b, the failure rate of an individual fault in the test environment, were driven by consideration of the fault removal efficiency (FRE), defined as the fraction of faults removed during the test interval. For the models $M_c(t; a, b, K_1, K_2)$, we have $FRE = 1 - e^{-bu_l}$. For the BSS application, $u_l = 42$, and by choosing b from the set {.02615744, .03300701, .03831995} values of 2/3, 3/4, and 4/5 are achieved for the FRE.

For each choice of (a, b), we generated 1,600 sample paths from the model $M_0(t; a, b, K_1, K_2)$ and calculated the LRT_0 and CvM tests of $H_0 : K_1 = K_2 = 1$ for each of them. We examined the power of the CvM test against alternatives that are equally spaced around the null value $(K_1, K_2) = (1, 1)$. The first three alternatives, (.5, .5), (.5, 1.0), and (1.0, .5), all correspond to alternatives where the fault discovery process "slows down" in the field. The next three alternatives, (1.0, 1.5), (1.5, 1.0), and (1.5, 1.5), correspond to "speed up" alternatives. The remaining two alternatives, (1.5, .5) and (.5, 1.5), are "conflicting" in the sense that one parameter of the failure rate decreases while the other increases. The cutoff value for the CvM test was determined using the Monte Carlo method described in Section 3.2, and, that for the LRT_0 test was determined using the nominal asymptotic chi-squared cutoff point $\chi^2_{2,\gamma}$.

Tables 1–3 show the size and power of nominal 10% LRT_0 and CvM tests. The sizes of the tests are shown in the middle two columns of Table 3, which corresponds to $(K_1, K_2) = (1, 1)$. It appears that LRT_0 is a slightly conservative test (actual size smaller than nominal) for small a and that CvM is a slightly liberal test (actual size is larger than nominal) for small FRE. As a result, the power comparisons must be interpreted with care, because a liberal test will naturally reject more often. Nevertheless, the power results in Tables 1–3 indicate that the CvM test competes surprisingly well with the LRT_0 test even in this setting, in which the alternatives are those for which the LRT was designed. The simulated powers for the two tests are quite comparable, with neither test dominating the other. For the speed-up alternatives, however, the CvM test often has power equal to or greater than LRT_0 for the cases considered. The low power for the case $(K_1, K_2) = (1.5, .5)$ is a consequence of the fact that for the values of a and *FRE* considered here, this alternative is much closer to the null than the other alternatives considered; this may be expected due to the fact that $K_1 > 1$ and $K_2 < 1$ have conflicting effects on the mean function.

4.2 Case 2: $c > 0$

For the class of models $M_c(t; a, b, K_1, K_2)$, where $c > 0$, LRT_0 loses its motivation and optimal property. Unbeknown to the practitioner, the correct LRT should be derived from the likelihood based on $M_c(t; a, b, K_1, K_2)$, and the adequacy of a one-part model would have to be tested according to $H_0 : K_1 = K_2 = 1, c = 0$. If the data follow $M_c(t; a, b, K_1, K_2)$, with $c > 0$, then we could expect that LRT_0 may not exist, because the assumed model is not correct. Figure 6 illustrates this problem by plotting simulation estimates of the probability that LRT_0 exists versus $c \in \{1, 2, 4\}$ for the models $M_c(t; a_1, b_1, 1, 1)$ and

Table 2. Simulation estimates of power for nominal $\gamma = .1$ tests using (u, v) from the Brazilian switching system application

		$(K_1, K_2) = (1.0, .5)$		$(K_1, K_2) = (1.0, 1.0)$		$(K_1, K_2) = (1.0, 1.5)$	
a	*FRE*	LRT_0	CvM	LRT_0	CvM	LRT_0	CvM
100	2/3	.15	.24	.070	.13	.23	.26
	3/4	.13	.20	.075	.14	.22	.24
	4/5	.11	.14	.074	.11	.18	.20
200	2/3	.30	.40	.078	.14	.32	.29
	3/4	.32	.32	.084	.12	.35	.32
	4/5	.28	.27	.091	.10	.35	.26
400	2/3	.65	.60	.098	.13	.55	.43
	3/4	.66	.58	.11	.11	.51	.34
	4/5	.64	.52	.10	.11	.53	.34

Table 3. Simulation estimates of power for nominal $\gamma = .1$ tests using (u, v) from the Brazilian switching system application

		$(K_1, K_2) = (1.5, .5)$		$(K_1, K_2) = (1.5, 1.0)$		$(K_1, K_2) = (1.5, 1.5)$	
a	*FRE*	LRT_0	CvM	LRT_0	CvM	LRT_0	CvM
100	2/3	.051	.12	.22	.40	.58	.67
	3/4	.064	.13	.20	.36	.55	.61
	4/5	.072	.11	.19	.33	.52	.57
200	2/3	.076	.15	.33	.50	.84	.83
	3/4	.13	.11	.31	.47	.79	.79
	4/5	.13	.09	.29	.47	.77	.76
400	2/3	.27	.15	.48	.67	.99	.97
	3/4	.33	.12	.53	.65	.97	.96
	4/5	.37	.095	.50	.65	.96	.93

$M_c(t; a_3, b_3, 1, 1)$. The estimates were obtained by simulating 800 data sets according to each model and checking what fraction of them satisfy all three conditions given by (4)–(6). Figure 6 shows results for $m \in \{5, 39\}$. It is evident that the probability that LRT_0 exists is low in all cases and decreases as c increases.

The only assumption that the CvM test makes is that $M_0(t; a, b, 1, 1)$ is the correct model in the test environment. Irrespective of the mean value function in the field environment, the CvM test is a valid test of the adequacy of $M_0(t; a, b, 1, 1)$ as a one-part model. Table 4 gives simulation estimates of the power of the CvM test against alternatives of the form $M_c(t; a, b, 1, 1)$ with $c > 0$. The results confirm intuition that the power increases as c grows larger. Note, however, that the power for $m = 39$ is not significantly larger than the power for $m = 5$. While this result is surprising at first, the reason for it is that sufficient learning has taken place by the time $m = 5$, so that the field part of the mean value function from that point onward looks very much like the corresponding part of $M_0(t; a, b, 1, 1)$. Thus the additional observations provide no further evidence against the adequacy of the one part model.

To further explore the effect of increasing m, Tables 5 and 6 give simulation estimates of the power of the CvM test against alternatives of the form $M_c(t; a, b, .5, 1)$ and $M_c(t; a, b, 1, .5)$, with $c > 0$. The results in these tables show that power is significantly greater for the case where $m = 39$, due to the fact that even after the learning period has expired, the mean value functions of the alternative models $M_c(t; a, b, .5, 1)$ and $M_c(t; a, b, 1, .5)$ look different than that of $M_0(t; a, b, 1, 1)$.

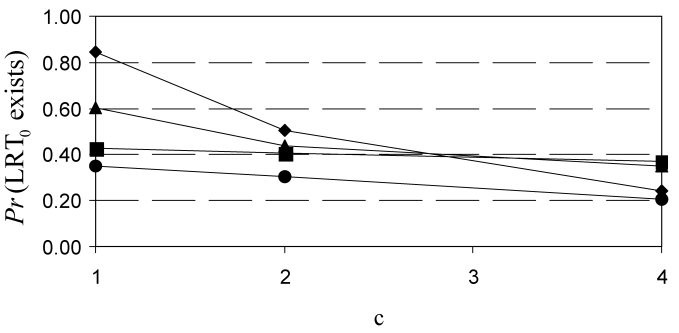

Figure 6. Probability that LRT_0 exists for various $M_c(t; a, b, 1, 1)$ models [—◆— (a_3, b_3), $m = 39$; —▲— (a_1, b_1), $m = 39$; —■— (a_3, b_3), $m = 5$; —●— (a_1, b_1), $m = 5$].

5. SUMMARY

Significant gains in inference associated with future failures can be realized when the one-part model is valid. But the one-part model is valid only when the test and field environments are identical in terms of how the software is used during operation. To take advantage of the increased precision in inference procedures that are possible when the two environments are identical, while at the same time avoiding the adverse consequences of fitting a one-part model that is not valid, a statistical test of whether the two environments are compatible is needed. We have proposed a CvM test for this purpose and have shown that it works well (i.e., has the right size and has satisfactory power) in situations where LRT_0 also would work well, but that it continues to work well in situations where LRT_0 has no statistical justification or where LRT_0 is justified but fails to exist due to the nonexistence of the two-part MLEs. The most attractive feature of the CvM test compared with LRT_0 is the fact that it is a valid statistical test no matter what the alternative to compatible environments might be. The CvM test, unlike LRT_0, does not require that the class of alternatives be specified. As such, the CvM test is very practical and useful test for software reliability engineers. Finally, we note that our proposed data analysis strategy is a two-stage methodology, with the hypothesis test for compatible environments constituting the first stage and fitting the appropriate one-part or two-part model constituting the

Table 4. Simulation estimates of CvM test power for nominal $\gamma = .1$ tests using (u, v) from the Brazilian switching system application and $(K_1, K_2) = (1, 1)$

		$c = 1$		$c = 2$		$c = 3$	
a	*FRE*	$m = 5$	$m = 39$	$m = 5$	$m = 39$	$m = 5$	$m = 39$
100	2/3	.14	.19	.20	.33	.32	.50
	3/4	.14	.16	.19	.26	.27	.39
	4/5	.11	.11	.16	.18	.23	.31
200	2/3	.27	.30	.50	.53	.71	.80
	3/4	.25	.21	.43	.42	.67	.73
	4/5	.21	.19	.38	.34	.62	.65
400	2/3	.51	.45	.80	.83	.95	.99
	3/4	.48	.38	.77	.75	.96	.98
	4/5	.41	.30	.71	.66	.92	.94

Table 5. Simulation estimates of CvM test power for nominal $\gamma = .1$ tests using (u, v) from the Brazilian switching system application and $(K_1, K_2) = (.5, 1)$

		$c = 1$		$c = 2$		$c = 3$	
a	*FRE*	$m = 5$	$m = 39$	$m = 5$	$m = 39$	$m = 5$	$m = 39$
100	2/3	.27	.50	.35	.62	.42	.72
	3/4	.25	.41	.32	.54	.39	.65
	4/5	.18	.31	.22	.43	.31	.55
200	2/3	.63	.80	.79	.90	.88	.97
	3/4	.59	.74	.74	.87	.85	.96
	4/5	.52	.67	.67	.84	.79	.94
400	2/3	.92	.98	.98	1.0	1.0	1.0
	3/4	.90	.97	.97	1.0	1.0	1.0
	4/5	.87	.96	.96	.99	.99	1.0

second stage. When making inferences (e.g., prediction intervals) from the model fit in the second stage, we have not provided guidance on how to adjust the methods used for the added variability introduced by the first-stage hypothesis test. An integrated data analysis strategy that explicitly does this will be the subject of future research.

APPENDIX A: PROOF OF THE PROPOSITION

In this appendix we sketch a proof of the proposition given in Section 3.2.1; refer back to that section for the notation definitions. We use the symbol $\approx$ to denote "has the same asymptotic distribution as." We introduce some partitioned matrices and vectors to present the asymptotic results. Let Y be the column vector $(Y_1, \ldots, Y_l)^t$, Z be the column vector $(Z_1, \ldots, Z_m)^t$, and $W = (Y^t, Z^t)^t$. Let D be the column vector with entries $D_i = W_i - W_{i-1}$, where $W_0 = 0$, and partition D as $D = (D_T^t, D_F^t)^t$. Our first step, which can be verified by direct multiplication, is to express the CvM statistic in (5) as

$$\text{CvM} = \left(\frac{D - \hat{a}\hat{\delta}}{\sqrt{\hat{a}}}\right)^t \hat{Q}\left(\frac{D - \hat{a}\hat{\delta}}{\sqrt{\hat{a}}}\right), \tag{A.1}$$

where $\hat{Q}$ is the matrix Q with b replaced by $\hat{b}$. The key to proving the proposition is our second step, in which we develop an asymptotically equivalent representation for $(D - \hat{a}\hat{\delta})/\sqrt{\hat{a}}$.

Table 6. Simulation estimates of CvM test power for nominal $\gamma = .1$ tests using (u, v) from the Brazilian switching system application and $(K_1, K_2) = (1, .5)$

		$c = 1$		$c = 2$		$c = 3$	
a	*FRE*	$m = 5$	$m = 39$	$m = 5$	$m = 39$	$m = 5$	$m = 39$
100	2/3	.27	.49	.36	.63	.44	.74
	3/4	.26	.39	.33	.55	.39	.67
	4/5	.19	.29	.24	.44	.31	.58
200	2/3	.64	.79	.78	.91	.90	.98
	3/4	.55	.72	.74	.88	.85	.97
	4/5	.52	.64	.67	.86	.8	.95
400	2/3	.92	.98	.99	1.0	1.0	1.0
	3/4	.90	.97	.98	1.0	1.0	1.0
	4/5	.86	.95	.96	1.0	1.0	1.0

Suppose that the hypothesized model holds for all $t \le v_m$, that is, $M(t; \theta) = a(1 - e^{-bt})$ for all $t \le v_m$. Then it is well known from the normal approximation to the Poisson distribution that $a^{-1/2}\{D - a\delta\}$ has a limiting multivariate normal distribution with mean 0 and variance–covariance matrix Δ^2. The limiting distribution of CvM can be deduced from this fundamental result by applications of Slutsky's theorem and Taylor expansion, as we now describe.

First, note that the MLE of (a, b) obtained from the two-part model maximizes the likelihood constructed from only the test data,

$$L_{Test}(a, b) \propto e^{-a(1-e^{-bu_i})} \prod_{i=1}^{l} \frac{[a(e^{-bu_{i-1}} - e^{-bu_i})]^{y_i - y_{i-1}}}{(y_i - y_{i-1})!}.$$

The score vector corresponding to $L_{Test}(a, b)$, namely $U(a, b) = [\frac{\partial \log L_{Test}}{\partial a}, \frac{\partial \log L_{Test}}{\partial b}]^t$, has the form

$$U(a, b) = \begin{bmatrix} a^{-1}1_{1\times l} & 0_{1\times m} \\ \nabla \log \delta_T^t & 0_{1\times m} \end{bmatrix} (D - a\delta).$$

We know that $U(a, b)$ has a large-sample distribution that is bivariate normal with mean 0 and variance–covariance matrix equal to the Fisher information matrix derived from $L_{Test}(a, b)$, say $I(a, b)$. Rather than computing the elements of $I(a, b)$ as expected values of the mixed second-order partial derivatives, we instead use the equivalent form $I(a, b) = E\{U(a, b)U(a, b)^t\}$ and find that

$$\begin{aligned} I(a, b) &= \begin{bmatrix} a^{-1}1_{1\times l} & 0_{1\times m} \\ \nabla \log \delta_T^t & 0_{1\times m} \end{bmatrix} \\ &\quad \times a\Delta^2 \begin{bmatrix} a^{-1}1_{l\times 1} & \nabla \log \delta_T \\ 0_{m\times 1} & 0_{m\times 1} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{l} \delta_i/a & \sum_{i=1}^{l} \partial\delta_i/\partial b \\ \sum_{i=1}^{l} \partial\delta_i/\partial b & a\sum_{i=1}^{l} (\partial\delta_i/\partial b)^2/\delta_i \end{bmatrix}. \end{aligned} \tag{A.2}$$

We also know that $(\hat{a}, \hat{b})^t$ is asymptotically bivariate normal with mean $(a, b)^t$ and variance–covariance matrix $I^{-1}(a, b)$. Therefore, we have $(\hat{a} - a, \hat{b} - b)^t \approx I^{-1}(a, b)U(a, b)$ and

$$\begin{aligned} \begin{bmatrix} (\hat{a} - a)/a \\ \hat{b} - b \end{bmatrix} &\approx \begin{bmatrix} 1/a & 0 \\ 0 & 1 \end{bmatrix} I^{-1}(a, b)U(a, b) \\ &= \left\{ I(a, b) \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \right\}^{-1} U(a, b) \\ &= \left\{ \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} I^* \right\}^{-1} U(a, b) \\ &= (I^*)^{-1} \begin{bmatrix} 1 & 0 \\ 0 & a^{-1} \end{bmatrix} \\ &\quad \times \begin{bmatrix} a^{-1}1_{1\times l} & 0_{1\times m} \\ \nabla \log \delta_T^t & 0_{1\times m} \end{bmatrix} (D - a\delta) \\ &= (I^*)^{-1} \begin{bmatrix} 1_{1\times l} & 0_{1\times m} \\ \nabla \log \delta_T^t & 0_{1\times m} \end{bmatrix} \\ &\quad \times \begin{bmatrix} a^{-1} & 0 \\ 0 & 1 \end{bmatrix} (D - a\delta) \\ &= (I^*)^{-1} \begin{bmatrix} 1_{1\times l} & 0_{1\times m} \\ \nabla \log \delta_T^t & 0_{1\times m} \end{bmatrix} \end{aligned}$$

$$\times \begin{bmatrix} a^{-1} & 0 \\ 0 & a^{-1} \end{bmatrix} (D - a\delta)$$

$$= A\frac{D - a\delta}{a}. \quad \text{(A.3)}$$

Next, using a first-order Taylor expansion about (a, b) of the elements in $\hat{a}\hat{\delta}$, we have

$$\frac{\hat{a}\hat{\delta} - a\delta}{\sqrt{a}} = \sqrt{a}\begin{bmatrix} \delta & \frac{\partial \delta}{\partial b} \end{bmatrix}\begin{bmatrix} (\hat{a} - a)/a \\ \hat{b} - b \end{bmatrix}. \quad \text{(A.4)}$$

Applying Slutzky's theorem and combining (A.3) and (A.4), we find that

$$\frac{D - \hat{a}\hat{\delta}}{\sqrt{\hat{a}}} \approx \frac{D - \hat{a}\hat{\delta}}{\sqrt{a}}$$

$$= M\frac{D - a\delta}{\sqrt{a}}. \quad \text{(A.5)}$$

Combining (A.1) with (A.5) and again using Slutzky's theorem gives

$$\text{CvM} \approx \left(\frac{D - a\delta}{\sqrt{a}}\right)^t M^t \hat{Q} M \left(\frac{D - a\delta}{\sqrt{a}}\right)$$

$$\approx Z^t \Delta M^t Q M \Delta Z, \quad \text{(A.6)}$$

where Z has a standard $(l+m)$-dimensional multivariate normal distribution. The proposition follows trivially from (A.6).

APPENDIX B: APPROXIMATE CONFIDENCE INTERVALS

In this appendix we derive an approximate confidence interval for the median of the time to next failure under the one-part model associated with (1). The confidence limits shown in Figure 3 were computed from this formula. A derivation very similar to what is shown here produces a formula for approximate confidence limits under the two-part model associated with (1), and we used that formula to compute the confidence limits shown in Figure 4.

Under the one-part model associated with (1), the median time to next failure given that the process has been observed through field time v_j is $Q_{.5}(v_j) = -\log[1 - \log 2/(ae^{-bv_j})]/b$. It can be shown that

$$\frac{\partial \log Q_{.5}(v_j)}{\partial a} = -\frac{1}{Q_{.5}(v_j)} \frac{\log 2}{a^2 b e^{-bv_j}[1 - \log 2/(ae^{bv_j})]}$$

and

$$\frac{\partial \log Q_{.5}(v_j)}{\partial b} = \frac{1}{Q_{.5}(v_j)}\left\{\frac{1 - \log 2/(ae^{bv_j})}{b^2} + \frac{v_j \log 2}{abe^{-bv_j}[1 - \log 2/(ae^{bv_j})]}\right\}.$$

Denote the MLE of a and b based on the cumulative failure data observed in $[0, v_j]$ by $\tilde{a}(j)$ and $\tilde{b}(j)$. The MLEs of $Q_{.5}(v_j)$ and $\log Q_{.5}(v_j)$ are then $\tilde{Q}_{.5}(v_j) = -\log[1 - \log 2/\{\tilde{a}(j)e^{-\tilde{b}(j)t}\}]/\tilde{b}(j)$ and $\log \tilde{Q}_{.5}(v_j)$, and using the delta method, we have the following approximation:

$$\text{Var}\{\log \tilde{Q}_{.5}(v_j)\} \approx \begin{bmatrix} \frac{\partial \log Q_{.5}(v_j)}{\partial a} & \frac{\partial \log Q_{.5}(v_j)}{\partial b} \end{bmatrix} \times I_j^{-1}(a, b)\begin{bmatrix} \frac{\partial \log Q_{.5}(v_j)}{\partial a} & \frac{\partial \log Q_{.5}(v_j)}{\partial b} \end{bmatrix}',$$

$$\equiv \sigma^2(j)$$

where $I_j(a, b)$ denotes the Fisher information matrix in (A.2) but with l replaced by $l + j$. With $\tilde{\sigma}^2(j)$ denoting $\sigma^2(j)$ but with $[\tilde{a}(j), \tilde{b}(j)]$ replacing (a, b), an approximate $100(1-\alpha)\%$ confidence interval for $\log Q_{.5}(v_j)$ is thus $\log \tilde{Q}_{.5}(v_j) \pm z_{\alpha/2}\tilde{\sigma}(j)$, and therefore an approximate $100(1-\alpha)\%$ confidence interval for $Q_{.5}(v_j)$ is $[\tilde{Q}_{.5}(v_j)e^{-z_{\alpha/2}\tilde{\sigma}(j)}, \tilde{Q}_{.5}(v_j)e^{z_{\alpha/2}\tilde{\sigma}(j)}]$.

[Received July 2006. Revised May 2007.]

REFERENCES

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994), "Cramer–von Mises Statistics for Discrete Distributions," *Canadian Journal of Statistics*, 22, 125–137.

Genest, C., Lockhart, R. A., and Stephens, M. A. (2002), "χ^2 and the Lottery," *The Statistician*, 51, 243–257.

Goel, A. L., and Okumoto, K. (1979), "Time-Dependent Fault Detection Rate Model for Software and Other Performance Measures," *IEEE Transactions on Reliability*, 28, 206–211.

Huang, C., Kuo, S., Lyu, M. R., and Lo, J. (2000), "Quantitative Software Reliability Modeling From Testing to Operation," in *Proceedings of the* 11th *International Symposium on Software Reliability Engineering*, San Jose, CA, pp. 72–82.

Imhof, J. P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika*, 48, 419–426.

Jean, J. (1998), "Inference in Nonhomogenous Poisson Process Models, With Applications to Software Reliability," unpublished doctoral dissertation, University of Waterloo, Ontario, Canada.

Jeske, D. R., and Zhang, X. (2004), "Some Successful Approaches to Software Reliability Modeling in Industry, *Journal of Systems and Software*, 74, 85–99.

Jeske, D. R., and Zhang, Q. (2006), "Assessing the Validity of One-Part Software Reliability Models Using Likelihood Ratio and Early Detection Tests," *Journal of Systems and Software*, 80, 205–216.

Jeske, D. R., Zhang, X., and Pham, L. (2005), "Adjusting Software Failure Rates That Are Estimated From Test Data," *IEEE Transactions on Reliability*, 54, 107–114.

Lyu, M. (ed.) (1996), *Handbook on Software Reliability Engineering*, New York: McGraw-Hill.

Martini, M. R., Kanoun, K., and de Souza, J. M. (1990), "Software-Reliability Evaluation of the TROPICO–R Switching System," *IEEE Transactions on Reliability*, 39, 369–379.

Musa, J. D. (1975), "A Theory of Software Reliability and Its Applications," *IEEE Transactions on Software Engineering*, 18, 423–433.

Pham, H. (1999), *Software Reliability*, Singapore: Springer.

Spinelli, J. J., and Stephens, M. A., (1997), "Cramer–von Mises Tests of Fit for the Poisson Distribution," *Canadian Journal of Statistics*, 25, 257–268.

Stephens, M. A. (1986), "Tests Based on EDF Statistics," in *Goodness-of-Fit Techniques*, eds. R B D'Agostino and M. A. Stephens, New York: Marcel-Dekker, pp. 93–197.

van Pul, M. C. (1992), "Asymptotic Properties of a Class of Statistical Models in Software Reliability," *Scandinavian Journal of Statistics*, 19, 235–253.

Xie, M. (1991), *Software Reliability Engineering*, Singapore: World Scientific.

Yamada, S., Ohba, M., and Osaki, S. (1983), "S–Shaped Reliability Growth Modeling for Software Error Detection," *IEEE Transactions on Reliability*, R-32, 475–478.

Zhang, X., Jeske, D. R., and Pham, H. (2002), "Calibrating Software Reliability Models When Test Data Does Not Reflect the User Operational Profile," *Journal of Applied Stochastic Models in Business and Industry*, 18, 87–99.