The Power of Alternative Kolmogorov-Smirnov Tests Based on Transformations of the Data

SONG-HEE KIM, Yale University
WARD WHITT, Columbia University

The Kolmogorov-Smirnov (KS) statistical test is commonly used to determine if data can be regarded as a sample from a sequence of independent and identically distributed (i.i.d.) random variables with specified continuous cumulative distribution function (cdf) F, but with small samples it can have insufficient power, that is, its probability of rejecting natural alternatives can be too low. However, in 1961, Durbin showed that the power of the KS test often can be increased, for a given significance level, by a well-chosen transformation of the data. Simulation experiments reported here show that the power can often be more consistently and substantially increased by a different transformation. We first transform the given sequence to a sequence of mean-1 exponential random variables, which is equivalent to a rate-1 Poisson process. We then apply the classical conditional-uniform transformation to convert the arrival times into i.i.d. random variables uniformly distributed on [0, 1]. And then, after those two preliminary steps, we apply the original Durbin transformation. Since these KS tests assume a fully specified cdf, we also investigate the consequence of having to estimate parameters of the cdf.

Categories and Subject Descriptors: I.6.5 [**Simulation and Modeling**]: Model Development

General Terms: Theory

Additional Key Words and Phrases: Hypothesis tests, Kolmogorov-Smirnov statistical test, power, data transformations

ACM Reference Format:
Song-Hee Kim and Ward Whitt. 2015. The power of alternative Kolmogorov-Smirnov tests based on transformations of the data. ACM Trans. Model. Comput. Simul. 25, 4, Article 24 (May 2015), 22 pages.
DOI: http://dx.doi.org/10.1145/2699716

1. INTRODUCTION

We are pleased to contribute to this special issue honoring Donald L. Iglehart, our academic grandfather and father, respectively. Don deserves recognition in this journal because of the research he and his students have done on simulation methodology, for example, Crane and Iglehart [1974a, 1974b, 1975], Glynn and Iglehart [1989], and Heidelberger and Iglehart [1979].

We consider the basic statistical problem of testing whether observations can be regarded as a sample from a sequence of independent and identically distributed (i.i.d.) random variables with a specified cumulative distribution function (cdf). Such testing commonly should be done in simulation input modeling, for example, to judge whether

This work is supported by the U.S. National Science Foundation grants CMMI 1066372 and 1265070 and by the Samsung Foundation.
Authors' addresses: S.-H. Kim, Yale School of Management, Yale University, New Haven, CT 06520; email: hailey.songhee.kim@gmail.com; W. Whitt, Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027; email: ww2040@columbia.edu.

DOI: http://dx.doi.org/10.1145/2699716

customer service-time data from service systems are consistent with a particular distribution. The testing often is also appropriate for simulation output analysis.

A common way to determine if data can be regarded as a sample from a sequence of i.i.d. random variables $\{X_k : k \geq 1\}$, each distributed as a random variable X with a specified continuous cdf $F(x) \equiv P(X \leq x)$, $x \in \mathbb{R}$, is to apply the Kolmogorov-Smirnov (KS) statistical test. The KS test is based on the maximum difference D_n between the empirical cdf (ecdf)

$$F_n(x) \equiv \frac{1}{n} \sum_{k=1}^{n} 1_{\{X_k \leq x\}}, \quad x \in \mathbb{R}, \tag{1}$$

and the underlying cdf F, where n is the sample size, 1_A is an indicator function, equal to 1 if the event A occurs, and equal to 0 otherwise, that is,

$$D_n \equiv \sup_x \{|F_n(x) - F(x)|\}, \tag{2}$$

which has a distribution that is independent of the cdf F, provided that the cdf is continuous. The null hypothesis in the KS test is that the data indeed comes from a sequence of i.i.d. random variables $\{X_k : k \geq 1\}$, each distributed as F.

For any observed value y of the maximum difference D_n from a sample of size n, we compute the p-value $P(D_n > y|H_0)$ under the null hypothesis H_0 that the sequence is i.i.d. with cdf F, for example, by using the Matlab program *ksstat*, and compare it to the significance level α, that is, for specified probability of rejecting the null hypothesis when it is in fact correct (type I error), which we take to be $\alpha = 0.05$. For $n \geq 35$, $P(D_n > 1.36/\sqrt{n}|H_0) \approx 0.05$. Sometimes it is preferable to use corresponding one-sided KS tests, but we will concentrate on the two-sided test. See Simard and L'Ecuyer [2011] and Shorack and Wellner [2009] for additional background and references on the KS test.

Alternative KS tests can be obtained by considering various transformations of the data, based on transformations of the hypothesized sequence of i.i.d. random variables $\{X_k : k \geq 1\}$ with continuous cdf F into a new sequence of i.i.d. random variables $\{Y_k : k \geq 1\}$ with continuous cdf G, while keeping the significance level α unchanged. Since the KS test applies in both settings, we should prefer the new test based on the transformed data if it has substantially greater statistical power for contemplated alternatives, that is, if it has a higher probability of rejecting the null hypothesis when the null hypothesis is false. Specifically, for specified significance criterion α, the power of a specified alternative is the probability $1 - \beta$, where $\beta \equiv \beta(\alpha)$ is the probability of incorrectly accepting the null hypothesis (type II error) when it is false (which of course depends on the alternative as well as α).

Durbin [1961] suggested transforming the data to increase the power of the KS test (without altering the distribution under the null hypothesis) and proposed a specific transformation for that purpose. In this article we study the issue further. We conclude that a good data transformation can indeed significantly increase the power of the KS test, but that a modification of the Durbin [1961] transformation consistently has even more power. Given the null hypothesis of an i.i.d. sequence $\{X_k\}$ with cdf F, our proposed test starts by transforming the given random variables X_k into i.i.d. mean-1 exponential random variables through the transformation $Y_k \equiv -\log_e\{1 - F(X_k)\}$, which can be regarded as the interarrival times of a rate-1 Poisson Process (PP). Then we apply a statistical test of a PP proposed by Lewis [1965]. The first step is to apply the classical conditional-uniform transformation to the associated arrival times $T_k \equiv Y_1 + \cdots + Y_k$, $1 \leq k \leq n$ of the PP; that is, under the null hypothesis, we obtain a new sequence of i.i.d. random variables T_k/T_n that are i.i.d. random variables uniformly distributed on the interval [0, 1]; for example, see Section 2.3 of Ross [1996]. After those two steps have been completed, we apply the original Durbin [1961]

transformation. While the component transformations that we use are not new, to the best of our knowledge, this combination of transformations has not been considered before. The idea of considering this alternative KS test came to us while working on ways to test if service-system arrival process data can be modeled as a nonhomogeneous PP, which is reported in Kim and Whitt [2014a, 2014b] and Kim et al. [2015]; we elaborate after we define the alternative tests that we examine.

We close this introduction by indicating how the rest of the article is organized. We start in Section 2 by carefully defining the six different KS tests we consider. Next, in Section 3, we elaborate on our motivation and explain why the new method should be promising. In Section 4 we describe our first simulation experiment, which is a fixed-sample-size discrete-time stationary-sequence analog of the fixed-interval-length continuous-time stationary point process experiment, aimed at studying tests of a PP, conducted in Kim and Whitt [2014b]. In addition to the natural null hypothesis of i.i.d. exponential random variables, we also consider i.i.d. nonexponential sequences with Erlang, hyperexponential, and lognormal marginal cdf's. We report the results in Section 5, which surprisingly show that the original Durbin [1961] method performs poorly, but we consider different models than those in Durbin [1961]. In contrast, our new method, which we call the Lewis test because it is based on an idea from Lewis [1965], performs well, providing increased power. However, Durbin [1961] considered different examples. Motivated by the good results found for a standard normal null hypothesis by Durbin [1961], in Section 6 we consider a second experiment to test for a sequence of i.i.d. standard normal random variables. Consistent with Durbin [1961], we find that the original Durbin [1961] method performs much better for the standard normal null hypothesis, but again the new version of the Lewis [1965] test also performs well. In Section 7 we discuss the common problem that we typically must estimate parameters when we apply the KS test. We draw conclusions in Section 8. Additional information appears in the online Appendix.

2. THE ALTERNATIVE KS TESTS

We consider the following six KS tests of the null hypothesis H_0 that n observations X_k, $1 \le k \le n$, can be considered a sample from a sequence of i.i.d. random variables having a continuous cdf F. We start by forming the associated variables $U_k \equiv F(X_k)$, which are i.i.d. uniform variables on [0, 1] under the null hypothesis.

Standard Test. We use the standard KS test based on (2) to test whether $U_k \equiv F(X_k)$, $1 \le k \le n$, can be considered to be i.i.d. random variables uniformly distributed on [0, 1].

Sort-Log Test. Starting with the n random variables U_k, $1 \le k \le n$, in the standard test, let $U_{(j)}$ be the j^{th} smallest of these, so that $U_{(1)} < \cdots < U_{(n)}$. As in Section 3.1 of Brown et al. [2005], we use the fact that under the null hypothesis

$$Y_j^{(L)} \equiv -j \log_e \left(U_{(j)}/U_{(j+1)}\right), \quad 1 \le j \le n-1,$$

are $n-1$ i.i.d. mean-1 exponential random variables; a proof is given in Section 2.2 of Kim and Whitt [2014c]. We then apply the KS test with n replaced by $n-1$ and the mean-1 exponential cdf.

Durbin (≡ Sort-Durbin) Test. This is the original test proposed by Durbin [1961], which also starts with $U_k \equiv F(X_k)$ and $U_{(k)}$ with $U_{(1)} < \cdots < U_{(n)}$ as previously. In this context, look at the successive *intervals* between these ordered observations:

$$C_1 \equiv U_{(1)},\ C_j \equiv U_{(j)} - U_{(j-1)}, \quad 2 \le j \le n+1, \quad \text{and} \quad C_{n+1} \equiv 1 - U_{(n)}.$$

Then let $C_{(j)}$ be the j^{th} smallest of these intervals, $1 \le j \le n$, so that $0 < C_{(1)} < \cdots < C_{(n+1)} < 1$. Now let Z_j be scaled versions of the intervals between these new

ordered intervals, that is, let

$$Z_j = (n+2-j)(C_{(j)} - C_{(j-1)}), \quad 1 \le j \le n+1, \quad (\text{with} \quad C_{(0)} \equiv 0). \tag{3}$$

Remarkably, Durbin [1961] showed (by a simple direct argument giving explicit expressions for the joint density functions, exploiting the transformation of random vectors by a function) that, under the null hypothesis, the random vector $(Z_1, \ldots, Z_n)$ is distributed the same as the random vector $(C_1, \ldots, C_n)$. Hence, again under the null hypothesis, the vector of associated partial sums $(S_1, \ldots, S_n)$, where $S_k \equiv Z_1 + \cdots + Z_k$, $1 \le k \le n$, has the same distribution as the original random vector $(U_{(1)}, \ldots, U_{(n)})$ of ordered uniform random variables. Hence, we can apply the KS test with the ecdf

$$F_n(x) \equiv n^{-1} \sum_{k=1}^{n} 1_{\{S_k \le x\}}, \quad 0 \le x \le 1,$$

for S_k above, comparing it to the uniform cdf on [0, 1].

CU (Conditional-Uniform ≡ Exp+CU) Test. We start with $Y_k \equiv -\log_e \{1 - F(X_k)\}$, $1 \le k \le n$, which are i.i.d. mean-1 exponential random variables under the null hypothesis. Thus, the cumulative sums $T_k \equiv Y_1 + \cdots + Y_k$, $1 \le k \le n$, are the arrival times of a rate-1 PP. In this context, the conditional-uniform property states that, under the null hypothesis, T_k/T_n, $1 \le k \le n-1$, are distributed as the order statistics of $n-1$ i.i.d. random variables uniformly distributed on [0, 1]. Thus we can apply the KS statistic with the ecdf

$$F_n^{(CU)}(x) \equiv \frac{1}{n-1} \sum_{k=1}^{n-1} 1_{\{(T_k/T_n) \le x\}}, \quad 0 \le x \le 1, \tag{4}$$

and the underlying uniform cdf on [0, 1].

Log (Exp+CU+Log) Test. We start with the partial sums T_k, $1 \le k \le n$, used in the CU test, which are the arrival times of a rate-1 PP under the null hypothesis. We again use the conditional-uniform property for fixed sample size to conclude that, under the null hypothesis, T_k/T_n, $1 \le k \le n-1$, are distributed as $U_{(k)}$, the order statistics of $n-1$ random variables, with $U_{(1)} < \cdots < U_{(n-1)}$. Hence, just as in the previous Sort-Log test,

$$Y_j^{(L)} \equiv -j \log_e \left(T_j/T_{j+1}\right), \quad 1 \le j \le n-1,$$

should be $n-1$ i.i.d. rate-1 exponential random variables, to which we can apply the KS test.

Lewis (Exp+CU+Durbin) Test. We again start with the partial sums T_k, $1 \le k \le n$, used in the CU test, which are the arrivals times of a rate-1 PP under the null hypothesis. We again use the conditional-uniform property for fixed sample size to conclude that, under the null hypothesis, T_k/T_n, $1 \le k \le n-1$, are distributed as $U_{(k)}$, the order statistics of $n-1$ random variables uniformly distributed on [0, 1], with $U_{(1)} < \cdots < U_{(n-1)}$. From this point, we apply the previously mentioned Durbin [1961] test with n replaced by $n-1$, just as Lewis [1965] did in his test of a PP.

3. MOTIVATION AND EXPLANATION

In this section, we describe our motivation for considering these new KS tests and we explain why the good performance we find in our experiments might be anticipated.

3.1. Testing if Arrival Processes Can be Regarded as Nonhomogeneous Poisson Processes

Our research was motivated by the desire to fit stochastic queueing models to data from large-scale service systems, such as telephone call centers and hospital emergency rooms, as discussed in Brown et al. [2005] and Armony et al. [2011]. These queueing models typically possess at least two stochastic elements that might be tested: arrival processes and service times. We started by looking at the arrival processes.

Since the arrival rate typically varies strongly by time of day in these service systems, the natural arrival process model is a nonhomogeneous PP (NHPP). The Poisson property arises from many people acting independently, each of whom uses the service system infrequently. Mathematical support is provided by the Poisson superposition theorem (see Section 9.8 of Whitt [2002], and references therein).

However, as emphasized by Brown et al. [2005], it is important to perform statistical tests on arrival data to see if the NHPP model is appropriate. For that purpose, Brown et al. [2005] proposed a variant of the Log KS test. First, Brown et al. [2005] assumed that the arrival rate function can be approximated by a Piecewise-Constant (PC) arrival rate function, which is often reasonable, because the arrival rate evidently changes relatively slowly. (We investigate how the subintervals should be chosen in Kim and Whitt [2014a].) Under the PC NHPP null hypothesis, the NHPP is then equivalent to a PP over each subinterval where the rate is constant. Then Brown et al. [2005] applied the CU transformation over each of these subintervals. Since the CU transformation is independent of the rate of the PP, the CU transformation can be applied to each interval where the rate is constant, and then all the data can be combined into a single sequence of i.i.d. random variables uniformly distributed on [0, 1].

For a PC NHPP, we strongly exploit the fact that the CU transformation eliminates all nuisance parameters. We need not estimate the rate on each of the many subintervals. As a consequence, however, the KS test after applying the CU transformation does not support any given arrival rate, and even allows it to be random. Thus, as discussed in Kim and Whitt [2014a, 2014b], the KS test might also be regarded as being for a Cox process, that is, a PP with a rate function that is a stochastic process. However, the possible rate stochastic processes are greatly restricted by the requirement that the rate be constant over each subinterval over which the CU property is applied.

After applying the CU transformation in that way to the PC NHPP, it is possible to apply the standard KS test directly, but Brown et al. [2005] did not do that. Instead, they performed the Log test. They then justified an NHPP model for the banking call center arrival data they were studying by showing that they could not reject the PP hypothesis with their Log KS test.

We wondered why Brown et al. [2005] applied the Log test with the additional logarithmic transformation instead of applying the CU KS test. As we presumed must be the case, we found that the CU KS test of a PP has remarkably little power against common alternative hypotheses such as renewal processes with nonexponential interarrival time distributions. We present theoretical support via asymptotic analysis and empirical evidence from extensive simulation experiments in Kim and Whitt [2014b].

We also found that there is a substantial history in the statistical literature. First, Lewis [1965] made a significant contribution for testing a PP, recognizing that the Durbin [1961] transformation could be effectively applied after the CU transformation. Second, from Lewis [1965] we discovered that the direct CU KS test of a PP was evidently first proposed by Barnard [1953]; and Lewis [1965] showed that it had little power.

Upon discovering Lewis [1965], we first supposed that the Log KS test of Brown et al. [2005] would turn out to be equivalent to the Lewis [1965] transformation and that the KS test proposed by Lewis [1965], drawing upon Durbin [1961], would coincide with the KS test given in Durbin [1961], but neither is the case. Thus, this past work suggests

several different KS tests. In Kim and Whitt [2014b], we concluded that the Lewis test of a PP has the most power against stationary point processes having nonexponential interarrival distributions, providing a significant improvement over the Log KS test.

On the other hand, we also found that none of the KS tests has much power against stationary point processes with dependent exponential interarrival times, that is, which differ from a PP only through the dependence. In fact, for those alternative hypotheses, we found the CU KS test tended to be most effective.

3.2. The Explanation

The key insight is the observation that the uniform random variables in the CU KS test are very different from the uniform random variables in the Standard KS test. Under the null hypothesis of i.i.d. exponential variables, these exponential variables directly correspond to the *interarrival* times of a PP. The uniform random variables in the standard test are direct transformations of these interarrival times, one by one.

In contrast, the uniform random variables produced by the CU transformation applied to the PP correspond to the successive *arrival times* in the PP, that is, the *cumulative sums of the interarrival times*. As a consequence, the CU KS test is evidently less able to detect differences in the interarrival-time distribution. In Section 7 of Kim and Whitt [2014b] we provide mathematical support by proving that the ecdf in Equation (1) converges to the uniform cdf as the sample size n increases for *any* rate-1 stationary ergodic point process, that is, for any stationary point process satisfying a strong law of large numbers. Thus, to first order, asymptotically, the CU KS test has no power at all against any of the alternatives in this large class.

This insight also helps explain why the Lewis test does so much better. It applies the Durbin transformation after performing the CU transformation. However, the first step of the Durbin transformation is to focus on the interarrival times and put them in ascending order. Thus, the Durbin transformation strongly brings the focus back to the interarrival times.

This advantage of the Lewis test is well illustrated by the problem of data rounding, which is studied in Kim and Whitt [2014a]. In applications, the data are often rounded, for example, to the nearest second. With large datasets, this produces zero-length interarrival times. Before applying the Durbin transformation, these are spread out throughout the data, so that they tend not to be detected by the KS test. On the other hand, the Durbin transformation shifts all these zero-length interarrival times to the left end of the distribution, leading to rejection. This is easy to see in the plots of the ecdfs.

The reordering property of the Durbin transformation also helps explain why the CU KS test tends to do relatively well against dependent exponential sequences. The reordering of the interarrival times, which is helpful for identifying nonexponential distributions, tends to dissipate the dependence among dependent exponential random variables. The cumulative impact of the dependence evidently can best be seen through the cumulative sums of the interarrival times, that is, the arrival times, without reordering.

4. THE FIRST EXPONENTIAL EXPERIMENT

Our first simulation experiment is for the discrete-time analog of the experiment for testing the continuous-time PP in Kim and Whitt [2014b]. To study the alternative KS tests of a PP, in Kim and Whitt [2014b] we let the null hypothesis in the base case be a rate-1 PP observed over the time interval [0, 200], so that the expected sample size was 200, but we also considered the longer time interval [0, 2,000].

Hence, closely paralleling that experimental design, our null hypothesis here in the base case is a sample of size $n = 200$ i.i.d. mean-1 exponential random variables, but

to see the impact of the sample size, we also give results for the larger sample size of $n = 2{,}000$.

Closely linking the experiments helps make insightful comparisons. From an applications perspective, the exponential distribution is also a natural reference case, because the exponential distribution is often assumed for service times as well as interarrival times in queueing models in order that associated stochastic processes, such as the number of customers in the system, will be Markov processes. We are thus developing statistical tests of Markov model components.

4.1. The Cases Considered

We use the same alternative hypotheses to the continuous-time PP used in Kim and Whitt [2014b], except that we replace the time intervals of fixed length t by sample sizes of fixed size n. That is, we now consider stationary sequences of mean-1 random variables. There are nine cases, each with from one to five subcases, yielding 29 cases in all. Again, using the same cases as before facilitates comparison.

The first five cases involve i.i.d. mean-1 random variables; the last four cases involve dependent identically distributed mean-1 random variables. The first i.i.d. case is our null hypothesis with exponential random variables. The other i.i.d. cases have nonexponential random variables. Cases 2 and 3 contain Erlang and hyperexponential random variables, which are, respectively, stochastically less variable and stochastically more variable than the exponential distribution in convex stochastic order, as in Section 9.5 of Ross [1996]. Thus, they have *squared coefficient of variation* (scv; variance divided by the square of the mean, denoted by c^2), $c^2 < 1$ and $c^2 > 1$, respectively. These distributions show deviations from the exponential distribution in their variability. They are special phase-type distributions, which are also often assumed in order to obtain Markov process models (that are more complicated than when the distribution is exponential); for example, see Neuts [1981].

Cases 4 and 5 contain other i.i.d. sequences with nonexponential cdfs. Case 4 contains a nonexponential distribution with the same scv $c^2 = 1$ as the exponential distribution, as well as $E[X] = 1$, while Case 5 contains lognormal distributions, with four different scvs. Lognormal distributions often have been found to fit service-time data well (e.g., see Brown et al. [2005]).

Case 1, Exponential. The null hypothesis with i.i.d. mean-1 exponential random variables (Base Case).

Case 2, Erlang, $\mathbf{E_k}$. Erlang-k (E_k) random variables, a sum of k i.i.d. exponentials for $k = 2, 4, 6$ with $c_X^2 \equiv c_k^2 = 1/k$.

Case 3, Hyperexponential, $\mathbf{H_2}$. Hyperexponential-2 (H_2) random variables, a mixture of two exponential cdfs with $c_X^2 = 1.25$, 1.5, 2, 4, and 10 (five cases). The cdf is $P(X \le x) \equiv 1 - p_1 e^{-\lambda_1 x} - p_2 e^{-\lambda_2 x}$. We further assume balanced means ($p_1 \lambda_1^{-1} = p_2 \lambda_2^{-1}$) as in (3.7) of Whitt [1982] so that given the value of c_X^2, $p_i = [1 \pm \sqrt{(c_X^2 - 1)/(c_X^2 + 1)}]/2$ and $\lambda_i = 2p_i$.

Case 4, mixture with $\mathbf{c_X^2 = 1}$. A mixture of a more variable cdf and a less variable cdf so that the $c_X^2 = 1$; $P(X = Y) = p = 1 - P(X = Z)$, where Y is H_2 with $c_Y^2 = 4$, Z is E_2 with $c_Z^2 = 1/2$, and $p = 1/7$.

Case 5, lognormal, $\mathbf{LN}$. Lognormal ($LN(1, \sigma^2)$) random variables with mean 1 and variance σ^2 for $\sigma^2 = c_X^2 = 0.25$, 1.0, 4.0, 10.0 (four cases).

Cases 6 and 7 are dependent stationary sequences that deviate from the null hypothesis (Case 1) only through dependence among successive variables, each exponentially distributed with mean 1. It is not customary to test for dependence among successive service times in applications, but see Gans et al. [2010]. We think that it deserves more attention. Toward that end, we consider the two cases:

Case 6, RRI, dependent exponential interarrival times. Randomly Repeated Interarrival (RRI) times with exponential interarrival times, constructed by letting each successive interarrival time be a mixture of the previous interarrival time with probability p or a new independent interarrival time from an exponential distribution with mean 1, with probability $1-p$ (a special case of a first-order Discrete Autoregressive process, DAR(1), studied by Jacobs and Lewis [1978, 1983]). Its serial correlation is $Corr(X_j, X_{j+k}) = p^k$. We consider three values of p: 0.1, 0.5, and 0.9.

Case 7, EARMA, dependent exponential interarrival times. A stationary sequence of dependent exponential interarrival times with the correlation structure of an autoregressive-moving average process, called EARMA(1,1) in Jacobs and Lewis [1977]. Starting from three independent sequences of i.i.d. random variables $\{X_n : n \geq 0\}$, $\{U_n : n \geq 1\}$, and $\{V_n : n \geq 1\}$, where Y_0 and X_n, $n \geq 1$, are exponentially distributed with mean $m = 1$, while

$$P(U_n = 0) = 1 - P(U_n = 1) = \beta \quad \text{and} \quad P(V_n = 0) = 1 - P(V_n = 1) = \rho, \tag{5}$$

the EARMA sequence $\{S_n : n \geq 1\}$ is defined recursively by

$$\begin{aligned} S_n &= \beta X_n + U_n Y_{n-1}, \\ Y_n &= \rho Y_{n-1} + V_n X_n, \quad n \geq 1. \end{aligned} \tag{6}$$

Its serial correlation is $Corr(S_j, S_{j+k}) = \gamma\rho^{k-1}$, where $\gamma = \beta(1-\beta)(1-\rho)+(1-\beta)^2\rho$. We consider five cases of (β, ρ): (0.75, 0.50), (0.5, 0.5), (0.5, 0.75), (0.00, 0.75), and (0.25, 0.90) so that the cumulative correlations $\sum_{k=1}^{\infty} Corr(S_j, S_{j+k})$ increase: 0.25, 0.50, 1.00, 3.00, and 5.25. For more details, see Pang and Whitt [2012]. We specify these cases by these cumulative correlations.

The final two cases are stationary sequences that have *both* nonexponential marginal distributions and dependence among successive variables:

Case 8, $\mathbf{mH_2}$*, superposition of* $\mathbf{m}$ *i.i.d.* $\mathbf{H_2}$ *renewal processes.* A stationary sequence of interarrival times from a superposition of m i.i.d. equilibrium renewal processes, where the times between renewals (interarrival times) in each renewal process has a hyperexponential (H_2) distribution with $c_a^2 = 4$ ($m\,H_2$). As the number m of component renewal processes increases, the superposition process converges to a PP, and thus looks locally more like a PP, with the interarrival distribution approaching exponential and the lag-k correlations approaching 0, but small correlations extending further across time, so that the superposition process retains an asymptotic variability parameter, $c_A^2 = 4$. We consider four values of m: 2, 5, 10, and 20.

Case 9, RRI ($\mathbf{H_2}$)*, dependent* $\mathbf{H_2}$ *interarrival times with* $\mathbf{c^2 = 4}$. RRI times with H_2 interarrival times, each having mean 1, $c^2 = 4$ and balanced means (as specified in Case 3). The repetition is done just as in Case 6. We again consider three values of p: 0.1, 0.5, and 0.9.

Cases 6 and 7 have short-range dependence, whereas Case 8 for large m tends to have nearly exponential interarrival times, but longer-range dependence. For small

m, the mH_2 superposition process should behave much like the H_2 renewal process in Case 3 with the component $c^2 = 4$; for large m, the mH_2 superposition process should behave more like Cases 6 and 7 with dependence and exponential interarrival times.

Since the new KS tests apply to i.i.d. sequences with arbitrary continuous cdfs, we also consider alternative null hypotheses. In particular, here we report results for E_2, H_2 (with $c^2 = 2$), and lognormal $LN(1, 4)$ (with $c^2 = 4$) marginal cdfs having mean 1 as well as the exponential base case.

4.2. Simulation Design

For each case, we simulated 10^4 replications of 3,000 interarrival times. We generate much more data than needed in order to get rid of any initial effects. We are supposing that we observe a stationary sequence. There is, of course, no problem if the sequence is i.i.d. However, for the dependent sequences, stationarity is achieved approximately by having the system operate for some time before collecting data. The initial effect was observed to matter for the cases with dependent interarrival times and relatively small sample sizes.

We use this simulation output to generate sample sizes of a fixed size n. With fixed sample size $n = 200$, in each replication of the 10^4 simulated interarrival times we use interarrival times from the 10^3th interarrival time to the $10^3 + 200$th interarrival time. To consider large sample sizes, we increased n from 200 to 2,000. We then consider the interarrival times from the 10^3th interarrival time to the $10^3 + 2{,}000$th interarrival time to observe the effect of larger sample size. This choice leaves little doubt about the stationarity assumption.

For each sample, we checked our simulation results by estimating the mean and scv of each interarrival-time cdf both before and after transformations; tables of the results and plots of the average of the ecdfs appear in the online Appendix.

5. RESULTS OF THE FIRST EXPERIMENT

The online Appendix contains detailed results of the experiments; we present a summary here. First, we found that the sort-Log and Log tests were consistently dominated by the Durbin [1961] test or the Lewis [1965] test, so we do not present detailed results for those two Log cases here. For the CU, CU+Log, and Lewis tests, we considered variants based on the exponential variables $-\log_e\{F(X)\}$ and well as $-\log_e\{1 - F(X)\}$, but we did not find great differences, so we do not report those either. Thus, we present the results of four KS tests: (i) the standard test, using the variables $U_k \equiv F(X_k)$, (ii) the Durbin [1961] test, (iii) the CU test, and (iv) the Lewis [1965] test, as specified in Section 2. Under the null hypotheses, the cdf in all four cases is uniform on [0, 1].

5.1. The Base Case: i.i.d. Mean-1 Exponential Variables

For our base case, we let the null hypothesis H_0 be that the data are from i.i.d. mean-1 exponential variables. We report the number of KS tests passed (not rejected) out of 10,000 replications as well as the average p-value with associated 95% confidence intervals. Thus, the estimate of the power is 1−(number passed/10,000). The p-value is the significance level below which the hypothesis would be rejected. Thus low p-values indicate greater power. Just as in Table 1 of Kim and Whitt [2014b], the differences in the tests is striking for the middle H_2 alternative with $c^2 = 2.0$, as shown in Table I here. The results for the Lewis, standard, and CU tests are very similar to those for the corresponding KS tests of a PP in Table 1 of Kim and Whitt [2014b], but the results for the Durbin [1961] test are new, and surprisingly bad.

The results for all 29 cases are given in Table II. The first "exponential" case is the i.i.d. exponential null hypothesis. The results show that all tests behave properly for the i.i.d. exponential null hypothesis. The results also show that the tests perform quite

Table I. The Power of Alternative KS Tests of the Null Hypothesis that Data are i.i.d. Mean-1 Exponential Variables for the Sample Size $n = 200$ with Significance Level $\alpha = 0.05$: The Alternative Hypothesis of i.i.d. H_2 Interarrival Times having $c_X^2 = 2$

KS test	Lewis	Standard	CU	Durbin
Power	0.93	0.64	0.28	0.14
Average p-value	0.02	0.09	0.24	0.40

Table II. The Power of Alternative KS Tests of the Null Hypothesis that Data are i.i.d. Mean-1 Exponential Variables for the Sample Size $n = 200$ for Various Alternative Hypotheses: Number of KS Tests Passed (denoted by #P) at Significance Level 0.05 out of 10,000 Replications and the Average p-Values (denoted by E[p-*value*]) with Associated 95% Confidence Intervals

		Standard		Durbin		CU		Lewis	
Case	*Subcase*	#P	E[p-value]	#P	E[p-value]	#P	E[p-value]	#P	E[p-value]
Exp	–	9487	0.50 ± 0.0057	9515	0.50 ± 0.0056	9511	0.50 ± 0.0056	9493	0.50 ± 0.0057
E_k	$k = 2$	28	0.00 ± 0.0001	3320	0.08 ± 0.0029	9985	0.78 ± 0.0045	0	0.00 ± 0.0000
	$k = 4$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10,000	0.94 ± 0.0021	0	0.00 ± 0.0000
	$k = 6$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10,000	0.98 ± 0.0011	0	0.00 ± 0.0000
H_2	$c^2 = 1.25$	8843	0.42 ± 0.0058	9451	0.49 ± 0.0057	8956	0.41 ± 0.0056	7501	0.30 ± 0.0056
	$c^2 = 1.5$	7204	0.27 ± 0.0053	9331	0.48 ± 0.0058	8418	0.33 ± 0.0053	3966	0.12 ± 0.0039
	$c^2 = 2$	3603	0.09 ± 0.0032	8667	0.40 ± 0.0058	7186	0.24 ± 0.0046	695	0.02 ± 0.0013
	$c^2 = 4$	90	0.00 ± 0.0003	4569	0.13 ± 0.0039	3648	0.08 ± 0.0027	22	0.00 ± 0.0003
	$c^2 = 10$	0	0.00 ± 0.0000	878	0.02 ± 0.0012	928	0.02 ± 0.0014	67	0.00 ± 0.0006
Mixture	–	1200	0.02 ± 0.0009	7016	0.26 ± 0.0053	9438	0.57 ± 0.0061	187	0.00 ± 0.0004
LN	(1, 0.25)	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10,000	0.94 ± 0.0022	0	0.00 ± 0.0000
	(1, 1)	98	0.00 ± 0.0002	3482	0.08 ± 0.0025	9517	0.53 ± 0.0058	24	0.00 ± 0.0001
	(1, 4)	176	0.00 ± 0.0005	5542	0.18 ± 0.0047	4742	0.13 ± 0.0036	28	0.00 ± 0.0002
	(1, 10)	0	0.00 ± 0.0000	353	0.01 ± 0.0008	2024	0.04 ± 0.0019	0	0.00 ± 0.0000
RRI	$p = 0.1$	9048	0.41 ± 0.0055	1911	0.03 ± 0.0012	9044	0.42 ± 0.0056	9121	0.41 ± 0.0054
	$p = 0.5$	4659	0.11 ± 0.0030	0	0.00 ± 0.0000	5587	0.16 ± 0.0039	4624	0.11 ± 0.0030
	$p = 0.9$	16	0.00 ± 0.0001	0	0.00 ± 0.0000	701	0.01 ± 0.0011	13	0.00 ± 0.0001
EARMA	0.25	9284	0.47 ± 0.0058	9475	0.50 ± 0.0057	8564	0.36 ± 0.0055	9498	0.50 ± 0.0057
	0.5	8865	0.43 ± 0.0059	9516	0.50 ± 0.0057	7519	0.27 ± 0.0050	9393	0.49 ± 0.0058
	1	8178	0.37 ± 0.0059	9419	0.50 ± 0.0057	6009	0.19 ± 0.0043	8964	0.44 ± 0.0059
	3	5209	0.21 ± 0.0055	6356	0.23 ± 0.0050	1896	0.04 ± 0.0018	6796	0.30 ± 0.0061
	5.25	4100	0.14 ± 0.0044	8215	0.38 ± 0.0061	1598	0.03 ± 0.0018	5680	0.21 ± 0.0051
mH_2	$m = 2$	4398	0.14 ± 0.0044	8871	0.42 ± 0.0058	4355	0.11 ± 0.0032	1546	0.04 ± 0.0024
	$m = 5$	7514	0.32 ± 0.0058	9363	0.48 ± 0.0057	5400	0.17 ± 0.0043	7228	0.29 ± 0.0057
	$m = 10$	7818	0.35 ± 0.0060	9423	0.49 ± 0.0057	6562	0.24 ± 0.0051	9004	0.44 ± 0.0059
	$m = 20$	7996	0.37 ± 0.0060	9457	0.50 ± 0.0057	7804	0.33 ± 0.0057	9431	0.49 ± 0.0057
$RRI(H_2)$	$p = 0.1$	104	0.00 ± 0.0003	126	0.00 ± 0.0003	2987	0.07 ± 0.0024	37	0.00 ± 0.0003
	$p = 0.5$	253	0.00 ± 0.0005	0	0.00 ± 0.0000	1105	0.02 ± 0.0013	215	0.00 ± 0.0006
	$p = 0.9$	4	0.00 ± 0.0000	0	0.00 ± 0.0000	229	0.00 ± 0.0005	5	0.00 ± 0.0000

differently for the alternative hypotheses. Table II shows that the standard and Lewis tests all perform reasonably well for the i.i.d. cases with nonexponential interarrival-time cdfs, in marked contrast to the CU and Durbin tests. Table II also shows that the Lewis test is consistently most powerful for these cases. The ordering remains for H_2 cdfs with both lower and higher scvs.

Just as in Kim and Whitt [2014b], the story is more complicated for the dependent sequences. The Durbin KS test performs remarkably well for the RRI cases, far better than all others. Upon further reflection, this makes sense, because the RRI sequence produces strings of identical observations. When the random variables are

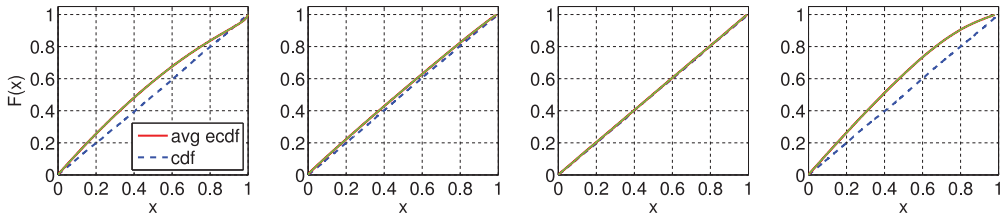

Fig. 1. Comparison of the average ecdf based on H_2 ($c^2 = 2$) data for 10^4 replications and $n = 200$ with the cdf of the exponential null hypothesis: Standard, Durbin, CU, and Lewis tests (from left to right).

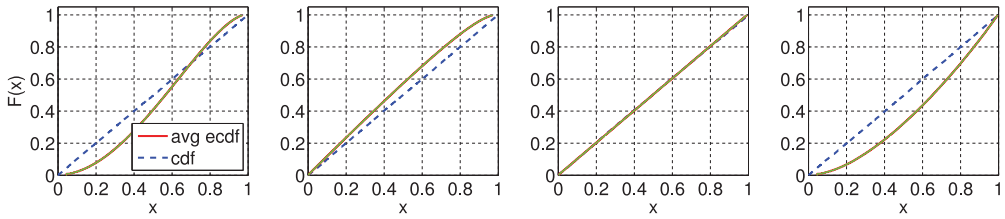

Fig. 2. Comparison of the average ecdf based on E_2 data for 10^4 replications and $n = 200$ with the cdf of the exponential null hypothesis: Standard, Durbin, CU, and Lewis tests (from left to right).

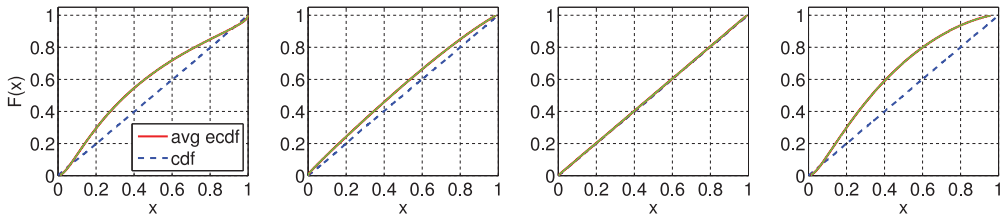

Fig. 3. Comparison of the average ecdf based on $LN(1, 4)$ data for 10^4 replications and $n = 200$ with the cdf of the exponential null hypothesis: Standard, Durbin, CU, and Lewis tests (from left to right).

ordered in ascending order, all repeated values will remain next to each other. And then, afterwards, when the Durbin transformation looks at the intervals between the ordered variables, these intervals will all be 0's. Hence, all the repetitions will be converted to 0's by the Durbin transformation. That in turn increases $\bar{F}_n(0)$ for the ecdf $\bar{F}_n$ in Equation (1), which typically increases the KS statistic D_n in Equation (2). It is evident that this property is not achieved by any of the other KS tests.

For the RRI(H_2) cases, all tests except CU perform very well. Hence, the Lewis test is consistently superior against nonexponential marginals. As in Kim and Whitt [2014b], none of the tests has much power against the EARMA alternatives, but the CU test has the most power.

5.2. Plots of the Average Empirical Distributions

As in Kim and Whitt [2014b], we find that useful insight is provided by plots comparing the average of the ecdfs over all 10,000 replications to the cdf associated with the null hypothesis, which is uniform in each case here. Figures 1–4 illustrate for the i.i.d. variables having cdfs H_2 with $c^2 = 2$, E_2, and LN(1, 4), and for the dependent RRI(0.5) variables with $n = 200$. These figures show that the transformation in the Lewis KS test provides greater separation between the average ecdf and the cdf in the i.i.d. cases.

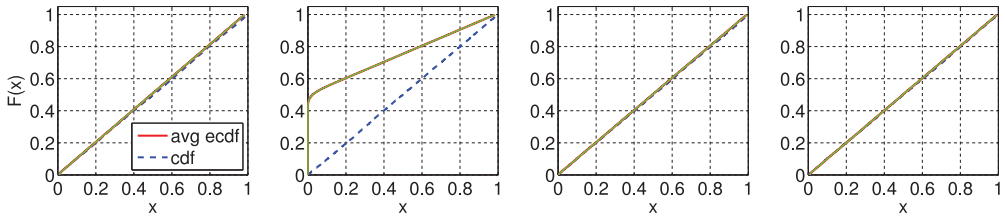

Fig. 4. Comparison of the average ecdf based on $RRI(0.5)$ data for 10^4 replications and $n = 200$ with the cdf of the exponential null hypothesis: Standard, Durbin, CU, and Lewis tests (from left to right).

In each case, the Durbin and Lewis tests tend to produce stochastic order compared to the uniform cdf, whereas the ecdf crosses over for the standard KS test, which is especially evident for E_2.

We have already observed that the Durbin test excels for RRI because it converts the repetitions into 0's. For RRI with $p = 0.5$, half of the variables are repetitions. Hence, half of the variables will be transformed into 0's. That is confirmed by the ecdf associated with the Durbin test in Figure 4.

5.3. Erlang, Hyperexponential, and Lognormal Null Hypotheses

We now consider three different i.i.d. null hypotheses: E_2, H_2 with $c^2 = 2$, and $LN(1, 4)$; lognormal hypotheses are especially interesting for service systems, for example, Brown et al. [2005]. The results are shown for the same 29 cases in the following Tables III–V for the base case of $n = 200$. As before, all tests perform properly for the null hypotheses. The ordering of the tests by power when we consider the i.i.d. exponential alternative hypothesis is the same as before. Overall, these tables show that the previous conclusions for the i.i.d. exponential null hypothesis conclusions extend to i.i.d. null hypotheses with other marginal cdfs.

As with the exponential null hypothesis, the Durbin test performs especially well for the RRI, because the repetitions are converted to 0's, but for these other null hypotheses, the standard and Lewis tests have almost equal power.

5.4. Larger Sample Sizes

Tables II–V clearly show how the power decreases as the alternative gets closer to the i.i.d. null hypothesis. For the i.i.d. exponential null hypothesis and the i.i.d. alternative hypotheses, we see this as the scv c_X^2 approaches 1; for the dependent exponential sequences, we see this as the degree of dependence decreases. However, all of these are for the sample size $n = 200$. The power also increases as we increase the sample size, as we now illustrate by considering the case $n = 2{,}000$ for the exponential null hypothesis in Table VI. Corresponding results for Erlang, hyperexponential, and lognormal null hypotheses appear in the online Appendix. When the sample size is increased to $n = 2{,}000$, all the tests except the CU test reject the alternative hypotheses in all 10^4 replications for most of the alternatives. Nevertheless, the superiority of the Lewis test for nonexponential marginals is evident from the H_2 case with $c^2 = 1.25$, the superiority of the Durbin test for the RRI cases is evident, and the superiority of the CU test for the EARMA cases is evident, consistent with the previous results for $n = 200$.

6. THE SECOND NORMAL EXPERIMENT

The poor results for the Durbin [1961] test for the i.i.d. cases in Section 5 seem inconsistent with the results in Durbin [1961] and the enthusiastic endorsement by Lewis [1965], so we decided to repeat some of the experiments actually performed by Durbin [1961]. We now consider the same four KS tests applied to the i.i.d. standard normal

Table III. The Power of Alternative KS Tests of the Null Hypothesis that Data are i.i.d. E_2 variables for the Sample Size $n = 200$ for Various Alternative Hypotheses: Number of KS Tests Passed (denoted by #*P*) at Significance Level 0.05 out of 10,000 Replications and the Average *p*-Values (denoted by E[*p-value*]) with Associated 95% Confidence Intervals

		Standard		Durbin		CU		Lewis	
Case	*Subcase*	#*P*	E[*p*-value]	#*P*	E[*p*-value]	#*P*	E[*p*-value]	#*P*	E[*p*-value]
Exp	–	129	0.00 ± 0.0003	2596	0.06 ± 0.0027	7421	0.24 ± 0.0046	0	0.00 ± 0.0000
E_k	$k = 2$	9492	0.50 ± 0.0056	9500	0.49 ± 0.0057	9497	0.50 ± 0.0057	9506	0.50 ± 0.0057
	$k = 4$	155	0.00 ± 0.0003	4100	0.11 ± 0.0034	9977	0.77 ± 0.0046	0	0.00 ± 0.0000
	$k = 6$	0	0.00 ± 0.0000	7	0.00 ± 0.0001	9999	0.88 ± 0.0033	0	0.00 ± 0.0000
H_2	$c^2 = 1.25$	17	0.00 ± 0.0001	1181	0.03 ± 0.0016	6106	0.17 ± 0.0040	0	0.00 ± 0.0000
	$c^2 = 1.5$	0	0.00 ± 0.0000	539	0.01 ± 0.0008	4905	0.12 ± 0.0033	0	0.00 ± 0.0000
	$c^2 = 2$	0	0.00 ± 0.0000	129	0.00 ± 0.0004	3336	0.07 ± 0.0024	0	0.00 ± 0.0000
	$c^2 = 4$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	752	0.01 ± 0.0009	0	0.00 ± 0.0000
	$c^2 = 10$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	67	0.00 ± 0.0004	0	0.00 ± 0.0000
Mixture	–	8069	0.32 ± 0.0054	9286	0.46 ± 0.0058	7152	0.28 ± 0.0054	4466	0.15 ± 0.0046
LN	$(1, 0.25)$	0	0.00 ± 0.0000	425	0.01 ± 0.0006	9973	0.75 ± 0.0048	0	0.00 ± 0.0000
	$(1, 1)$	3086	0.07 ± 0.0027	8424	0.37 ± 0.0058	6809	0.22 ± 0.0045	331	0.01 ± 0.0009
	$(1, 4)$	0	0.00 ± 0.0000	3	0.00 ± 0.0000	1507	0.03 ± 0.0014	0	0.00 ± 0.0000
	$(1, 10)$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	408	0.01 ± 0.0006	0	0.00 ± 0.0000
RRI	$p = 0.1$	135	0.00 ± 0.0003	24	0.00 ± 0.0001	6455	0.19 ± 0.0042	5	0.00 ± 0.0000
	$p = 0.5$	164	0.00 ± 0.0004	0	0.00 ± 0.0000	2429	0.05 ± 0.0020	45	0.00 ± 0.0002
	$p = 0.9$	3	0.00 ± 0.0000	0	0.00 ± 0.0000	142	0.00 ± 0.0004	3	0.00 ± 0.0000
EARMA	0.25	108	0.00 ± 0.0002	2552	0.06 ± 0.0027	5494	0.15 ± 0.0037	1	0.00 ± 0.0000
	0.5	114	0.00 ± 0.0003	2614	0.07 ± 0.0027	4064	0.10 ± 0.0029	0	0.00 ± 0.0000
	1	135	0.00 ± 0.0003	2597	0.07 ± 0.0028	2670	0.06 ± 0.0022	6	0.00 ± 0.0001
	3	918	0.02 ± 0.0015	3573	0.12 ± 0.0043	508	0.01 ± 0.0008	585	0.02 ± 0.0018
	5.25	432	0.01 ± 0.0007	2347	0.07 ± 0.0032	374	0.01 ± 0.0006	339	0.01 ± 0.0007
mH_2	$m = 2$	0	0.00 ± 0.0000	289	0.01 ± 0.0007	1248	0.02 ± 0.0013	0	0.00 ± 0.0000
	$m = 5$	23	0.00 ± 0.0001	1179	0.03 ± 0.0015	2356	0.05 ± 0.0022	0	0.00 ± 0.0000
	$m = 10$	63	0.00 ± 0.0002	1684	0.04 ± 0.0020	3581	0.09 ± 0.0031	0	0.00 ± 0.0000
	$m = 20$	96	0.00 ± 0.0002	2070	0.05 ± 0.0024	4884	0.14 ± 0.0038	0	0.00 ± 0.0000
$RRI(H_2)$	$p = 0.1$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	557	0.01 ± 0.0007	0	0.00 ± 0.0000
	$p = 0.5$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	151	0.00 ± 0.0003	0	0.00 ± 0.0000
	$p = 0.9$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	23	0.00 ± 0.0002	1	0.00 ± 0.0000

$(N(0, 1))$ null hypothesis. To keep the same mean equal to 0 for all alternatives, we consider all the previous 29 cases after subtracting 1 to make them all have mean 0. Indeed, the first alternative considered by Durbin [1961] was an i.i.d. sequence of random variables distributed as $Y - 1$, where Y is a mean-1 exponential variable; it has the same mean and variance as $N(0, 1)$. We summarize the results for this alternative with the sample size $n = 50$ used by Durbin [1961] in Table VII. Table VII shows that now the Durbin [1961] and Lewis [1965] have essentially the same power, which is far greater than for the standard and CU tests.

Table VIII shows all the results for our original 29 cases with $n = 50$. Since those alternatives have quite a different shape from the symmetric $N(0, 1)$ distributions, we also considered i.i.d. sequences of random variables distributed as $Z_k - 1 + \sqrt{1 - (1/k)}N(0, 1)$, where Z_k has an E_k cdf, for $k = 2, 4, 6$. These have the same first two moments and approximately the same shape. The new base case is the i.i.d. standard normal null hypothesis; it appears just below the previous alternatives in Table VIII. Just as in the previous tables, the results show that all tests behave properly for the standard normal null hypothesis. Overall, Table VIII shows that the

Table IV. The Power of Alternative KS Tests of the Null Hypothesis that Data are i.i.d. H_2 with $c^2 = 2$ Variables for the Sample Size $n = 200$ for Various Alternative Hypotheses: Number of KS Tests Passed (denoted by #*P*) at Significance Level 0.05 out of 10,000 Replications and the Average *p*-Values (denoted by *E* [*p-value*]) with Associated 95% Confidence Intervals

		Standard		Durbin		CU		Lewis	
Case	*Subcase*	#*P*	*E*[*p*-value]	#*P*	*E*[*p*-value]	#*P*	*E*[*p*-value]	#*P*	*E*[*p*-value]
Exp	–	3661	0.10 ± 0.0034	8951	0.43 ± 0.0058	9935	0.69 ± 0.0051	1613	0.03 ± 0.0014
E_k	$k = 2$	0	0.00 ± 0.0000	92	0.00 ± 0.0003	10000	0.89 ± 0.0032	0	0.00 ± 0.0000
	$k = 4$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10000	0.98 ± 0.0012	0	0.00 ± 0.0000
	$k = 6$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10000	0.99 ± 0.0005	0	0.00 ± 0.0000
H_2	$c^2 = 1.25$	6574	0.23 ± 0.0052	9433	0.49 ± 0.0057	9850	0.63 ± 0.0055	5543	0.15 ± 0.0038
	$c^2 = 1.5$	8530	0.39 ± 0.0059	9497	0.50 ± 0.0057	9750	0.58 ± 0.0056	8307	0.34 ± 0.0055
	$c^2 = 2$	9511	0.50 ± 0.0056	9482	0.50 ± 0.0057	9482	0.50 ± 0.0057	9507	0.50 ± 0.0056
	$c^2 = 4$	4983	0.14 ± 0.0040	9107	0.44 ± 0.0058	8143	0.31 ± 0.0052	3888	0.11 ± 0.0038
	$c^2 = 10$	269	0.01 ± 0.0005	6142	0.19 ± 0.0046	5098	0.15 ± 0.0039	1221	0.04 ± 0.0024
Mixture	–	0	0.00 ± 0.0000	1932	0.04 ± 0.0021	9989	0.80 ± 0.0043	0	0.00 ± 0.0000
LN	(1, 0.25)	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10000	0.98 ± 0.0011	0	0.00 ± 0.0000
	(1, 1)	0	0.00 ± 0.0000	585	0.01 ± 0.0006	9982	0.77 ± 0.0046	0	0.00 ± 0.0000
	(1, 4)	5685	0.18 ± 0.0045	9051	0.44 ± 0.0059	8493	0.36 ± 0.0055	5281	0.16 ± 0.0043
	(1, 10)	13	0.00 ± 0.0001	4888	0.15 ± 0.0043	5824	0.17 ± 0.0042	11	0.00 ± 0.0001
RRI	$p = 0.1$	3400	0.09 ± 0.0032	1352	0.02 ± 0.0010	9804	0.61 ± 0.0056	1410	0.03 ± 0.0013
	$p = 0.5$	2058	0.05 ± 0.0020	0	0.00 ± 0.0000	7608	0.28 ± 0.0050	883	0.02 ± 0.0012
	$p = 0.9$	9	0.00 ± 0.0001	0	0.00 ± 0.0000	1282	0.03 ± 0.0017	6	0.00 ± 0.0000
EARMA	0.25	3697	0.10 ± 0.0035	8922	0.43 ± 0.0058	9684	0.56 ± 0.0056	1577	0.03 ± 0.0014
	0.5	3839	0.11 ± 0.0037	8872	0.42 ± 0.0059	9216	0.45 ± 0.0057	1630	0.03 ± 0.0015
	1	3755	0.11 ± 0.0037	8629	0.40 ± 0.0059	8364	0.34 ± 0.0055	1607	0.03 ± 0.0017
	3	3607	0.13 ± 0.0044	5683	0.19 ± 0.0047	3333	0.08 ± 0.0028	2577	0.07 ± 0.0032
	5.25	2770	0.08 ± 0.0032	6642	0.27 ± 0.0056	3118	0.08 ± 0.0029	1690	0.05 ± 0.0025
mH_2	$m = 2$	8771	0.42 ± 0.0058	9466	0.49 ± 0.0057	7788	0.29 ± 0.0052	9091	0.43 ± 0.0057
	$m = 5$	6227	0.24 ± 0.0053	9290	0.47 ± 0.0058	7974	0.33 ± 0.0056	5465	0.16 ± 0.0041
	$m = 10$	5052	0.18 ± 0.0047	9032	0.44 ± 0.0058	8543	0.40 ± 0.0061	3210	0.07 ± 0.0025
	$m = 20$	4598	0.15 ± 0.0044	9013	0.43 ± 0.0058	9265	0.50 ± 0.0061	2263	0.05 ± 0.0018
$RRI(H_2)$	$p = 0.1$	4641	0.14 ± 0.0040	1227	0.02 ± 0.0010	7377	0.26 ± 0.0048	3720	0.11 ± 0.0037
	$p = 0.5$	2542	0.05 ± 0.0022	0	0.00 ± 0.0000	3586	0.09 ± 0.0029	2467	0.05 ± 0.0022
	$p = 0.9$	13	0.00 ± 0.0001	0	0.00 ± 0.0000	440	0.01 ± 0.0008	9	0.00 ± 0.0001

Durbin [1961] test performs much better now, just as originally reported. In this case *both* the Durbin [1961] and Lewis [1965] KS tests perform much better than the standard and CU alternatives. An exception is the set of three modified Erlang cases, with the same shape and first two moments as $N(0, 1)$. The Lewis test has the most power, but all four tests have low power for these cases.

As in Section 5, the power increases as the sample size increases; see the Appendix for the test results for the larger sample size $n = 200$. In that case, we observe that all tests except CU have estimated perfect power except in the last three modified Erlang cases, where the Lewis test stands out with power 0.375 for the modified E_2 case compared to 0.130 for standard and CU, and only 0.055 for Durbin. Figures 5 and 6 show that the reason can be seen in the average of the ecdfs of the transformed data.

7. ESTIMATING PARAMETERS

The KS test assumes a fully specified cdf, which is rarely the case in applications. In this section we investigate the consequence of having to estimate the parameters of the cdf in the null hypothesis. Before doing so, we observe that there is one case in

Table V. The Power of Alternative KS Tests of the Null Hypothesis that Data are i.i.d. $LN(1, 4)$ Variables for the Sample Size $n = 200$ for Various Alternative Hypotheses: Number of KS Tests Passed (denoted by #*P*) at Significance Level 0.05 out of 10,000 Replications and the Average *p*-Values (denoted by E [*p-value*]) with Associated 95% Confidence Intervals

		Standard		Durbin		CU		Lewis	
Case	*Subcase*	#*P*	E[*p*-value]	#*P*	E[*p*-value]	#*P*	E[*p*-value]	#*P*	E[*p*-value]
Exp	–	181	0.00 ± 0.0005	5509	0.18 ± 0.0046	9972	0.75 ± 0.0047	38	0.00 ± 0.0002
E_k	$k = 2$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10000	0.93 ± 0.0024	0	0.00 ± 0.0000
	$k = 4$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10000	0.99 ± 0.0007	0	0.00 ± 0.0000
	$k = 6$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10000	1.00 ± 0.0003	0	0.00 ± 0.0000
H_2	$c^2 = 1.25$	811	0.02 ± 0.0012	7382	0.29 ± 0.0056	9939	0.70 ± 0.0051	513	0.01 ± 0.0007
	$c^2 = 1.5$	2340	0.05 ± 0.0023	8354	0.37 ± 0.0058	9895	0.66 ± 0.0053	2255	0.05 ± 0.0020
	$c^2 = 2$	5665	0.17 ± 0.0043	9006	0.43 ± 0.0058	9788	0.59 ± 0.0055	6140	0.19 ± 0.0043
	$c^2 = 4$	9164	0.36 ± 0.0048	8864	0.41 ± 0.0058	9294	0.46 ± 0.0056	8783	0.31 ± 0.0046
	$c^2 = 10$	3774	0.08 ± 0.0023	6700	0.23 ± 0.0050	8538	0.35 ± 0.0054	5450	0.13 ± 0.0032
Mixture	–	0	0.00 ± 0.0000	196	0.00 ± 0.0005	10,000	0.87 ± 0.0034	0	0.00 ± 0.0000
LN	$(1, 0.25)$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10,000	0.99 ± 0.0005	0	0.00 ± 0.0000
	$(1, 1)$	0	0.00 ± 0.0000	90	0.00 ± 0.0003	9999	0.85 ± 0.0037	0	0.00 ± 0.0000
	$(1, 4)$	9508	0.50 ± 0.0056	9508	0.50 ± 0.0056	9508	0.50 ± 0.0057	9490	0.50 ± 0.0057
	$(1, 10)$	232	0.01 ± 0.0005	6261	0.22 ± 0.0051	8094	0.30 ± 0.0051	185	0.00 ± 0.0004
RRI	$p = 0.1$	193	0.00 ± 0.0005	346	0.01 ± 0.0004	9921	0.68 ± 0.0053	47	0.00 ± 0.0001
	$p = 0.5$	408	0.01 ± 0.0007	0	0.00 ± 0.0000	8255	0.34 ± 0.0054	120	0.00 ± 0.0003
	$p = 0.9$	13	0.00 ± 0.0001	0	0.00 ± 0.0000	1738	0.04 ± 0.0021	3	0.00 ± 0.0001
EARMA	0.25	206	0.00 ± 0.0006	5443	0.18 ± 0.0046	9866	0.64 ± 0.0054	34	0.00 ± 0.0001
	0.5	312	0.01 ± 0.0007	5388	0.17 ± 0.0045	9571	0.53 ± 0.0058	44	0.00 ± 0.0002
	1	436	0.01 ± 0.0009	5032	0.16 ± 0.0045	9023	0.43 ± 0.0058	72	0.00 ± 0.0003
	3	1594	0.04 ± 0.0024	4073	0.13 ± 0.0041	4018	0.10 ± 0.0033	647	0.01 ± 0.0012
	5.25	1220	0.03 ± 0.0019	3612	0.12 ± 0.0042	4027	0.11 ± 0.0036	469	0.01 ± 0.0013
mH_2	$m = 2$	4930	0.15 ± 0.0040	8640	0.39 ± 0.0058	8786	0.39 ± 0.0057	4425	0.12 ± 0.0035
	$m = 5$	1706	0.04 ± 0.0022	7193	0.27 ± 0.0055	8677	0.40 ± 0.0059	606	0.01 ± 0.0008
	$m = 10$	1083	0.03 ± 0.0017	6179	0.22 ± 0.0051	9085	0.48 ± 0.0062	178	0.00 ± 0.0004
	$m = 20$	808	0.02 ± 0.0013	5752	0.19 ± 0.0049	9572	0.57 ± 0.0060	79	0.00 ± 0.0002
$RRI(H_2)$	$p = 0.1$	8581	0.29 ± 0.0046	834	0.02 ± 0.0008	8830	0.39 ± 0.0055	8117	0.26 ± 0.0044
	$p = 0.5$	3857	0.08 ± 0.0024	0	0.00 ± 0.0000	5080	0.14 ± 0.0036	3547	0.07 ± 0.0024
	$p = 0.9$	17	0.00 ± 0.0001	0	0.00 ± 0.0000	658	0.01 ± 0.0010	5	0.00 ± 0.0001

which we do *not* need to estimate any parameters. That fortunate situation occurs with exponential cdfs. Exponential cdfs can be regarded as the interarrival times of a PP with a rate equal to the reciprocal of its mean. However, we do not need to know that mean, because the conditional-uniform transformation is independent of the rate of the PP. Thus, the new KS tests of an i.i.d. sequence with an exponential cdf that exploit the CU property have the advantage that they do not require estimating the mean.

Having to estimate the parameters can have a big influence. For example, in Kim and Whitt [2014b] (see Section 6 of its online Appendix [Kim and Whitt 2014c] for further details), we found that in a standard KS test of a mean-1 exponential cdf, if we use the KS test with the estimated mean and act as if it is the known mean, then it is necessary to increase the nominal significance level from 0.05 to 0.18 with a sample size of $n = 200$ in order for the actual significance level to be $\alpha = 0.05$. The resulting statistical test with estimated mean then coincides with the Lilliefors [1969] test.

To examine the impact of estimating the parameters, we consider testing for lognormal and normal distributions with estimated parameters, using the maximum likelihood estimators. (See Section F of the Appendix for further information on how we estimated the parameters and selected the nominal significance levels.) The nominal

Table VI. The Power of Alternative KS Tests of the Null Hypothesis that Data are i.i.d. Mean-1 Exponential Variables for the Sample Sze $n = 2{,}000$ for Various Alternative Hypotheses: Number of KS Tests Passed (denoted by #*P*) at Significance Level 0.05 out of 10,000 Replications and the Average *p*-Values (denoted by *E* [*p-value*]) with Associated 95% Confidence Intervals

		Standard		Durbin		CU		Lewis	
Case	*Subcase*	#*P*	*E*[*p*-value]	#*P*	*E*[*p*-value]	#*P*	*E*[*p*-value]	#*P*	*E*[*p*-value]
Exp	–	9515	0.50 ± 0.0056	9495	0.50 ± 0.0057	9481	0.50 ± 0.0057	9495	0.50 ± 0.0057
E_k	$k = 2$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	9985	0.79 ± 0.0044	0	0.00 ± 0.0000
	$k = 4$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10,000	0.95 ± 0.0019	0	0.00 ± 0.0000
	$k = 6$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10,000	0.98 ± 0.0009	0	0.00 ± 0.0000
H_2	$c^2 = 1.25$	3380	0.08 ± 0.0029	9360	0.48 ± 0.0057	8957	0.40 ± 0.0055	281	0.01 ± 0.0006
	$c^2 = 1.5$	68	0.00 ± 0.0002	8320	0.36 ± 0.0059	8313	0.32 ± 0.0051	0	0.00 ± 0.0000
	$c^2 = 2$	0	0.00 ± 0.0000	3425	0.08 ± 0.0030	6893	0.21 ± 0.0043	0	0.00 ± 0.0000
	$c^2 = 4$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	2788	0.05 ± 0.0019	0	0.00 ± 0.0000
	$c^2 = 10$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	34	0.00 ± 0.0002	0	0.00 ± 0.0000
Mixture	–	0	0.00 ± 0.0000	4	0.00 ± 0.0001	9450	0.52 ± 0.0058	0	0.00 ± 0.0000
LN	(1, 0.25)	0	0.00 ± 0.0000	0	0.00 ± 0.0000	10,000	0.95 ± 0.0019	0	0.00 ± 0.0000
	(1, 1)	0	0.00 ± 0.0000	0	0.00 ± 0.0000	9501	0.51 ± 0.0057	0	0.00 ± 0.0000
	(1, 4)	0	0.00 ± 0.0000	0	0.00 ± 0.0000	2610	0.06 ± 0.0023	0	0.00 ± 0.0000
	(1, 10)	0	0.00 ± 0.0000	0	0.00 ± 0.0000	242	0.00 ± 0.0005	0	0.00 ± 0.0000
RRI	$p = 0.1$	9010	0.41 ± 0.0055	0	0.00 ± 0.0000	9129	0.41 ± 0.0055	9014	0.40 ± 0.0055
	$p = 0.5$	4410	0.10 ± 0.0028	0	0.00 ± 0.0000	4666	0.11 ± 0.0030	4531	0.10 ± 0.0028
	$p = 0.9$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	25	0.00 ± 0.0001	0	0.00 ± 0.0000
EARMA	0.25	9336	0.47 ± 0.0057	9483	0.50 ± 0.0057	8326	0.33 ± 0.0052	9429	0.49 ± 0.0057
	0.5	8806	0.42 ± 0.0059	9505	0.50 ± 0.0057	7063	0.22 ± 0.0044	9408	0.49 ± 0.0057
	1	8210	0.37 ± 0.0059	9488	0.50 ± 0.0057	4722	0.12 ± 0.0031	8901	0.43 ± 0.0058
	3	5247	0.21 ± 0.0054	6406	0.22 ± 0.0049	822	0.01 ± 0.0008	6715	0.29 ± 0.0061
	5.25	4111	0.14 ± 0.0045	9290	0.47 ± 0.0058	193	0.00 ± 0.0003	5769	0.21 ± 0.0051
mH_2	$m = 2$	0	0.00 ± 0.0000	5272	0.16 ± 0.0042	3029	0.06 ± 0.0022	0	0.00 ± 0.0000
	$m = 5$	3135	0.09 ± 0.0032	9281	0.46 ± 0.0058	3434	0.07 ± 0.0024	182	0.00 ± 0.0004
	$m = 10$	6428	0.25 ± 0.0054	9471	0.49 ± 0.0057	3732	0.09 ± 0.0027	4432	0.13 ± 0.0040
	$m = 20$	7364	0.31 ± 0.0058	9470	0.50 ± 0.0057	4365	0.11 ± 0.0033	8127	0.35 ± 0.0058
$RRI(H_2)$	$p = 0.1$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	1897	0.03 ± 0.0015	0	0.00 ± 0.0000
	$p = 0.5$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	177	0.00 ± 0.0003	0	0.00 ± 0.0000
	$p = 0.9$	0	0.00 ± 0.0000	0	0.00 ± 0.0000	0	0.00 ± 0.0000	0	0.00 ± 0.0000

Table VII. The Power of Alternative KS Tests of the Null Hypothesis that Data are i.i.d. Standard Normal $N(0, 1)$ Variables for the Sample Size $n = 50$ with Significance Level $\alpha = 0.05$: The Alternative Hypothesis of i.i.d. Random Variables Distributed as $Y - 1$, where Y is a Mean-1 Exponential Random Variable

KS test	Lewis	Standard	CU	Durbin
Power	0.885	0.443	0.328	0.813
Average *p*-value	0.02	0.07	0.23	0.04

significance level had to be increased from 0.05 to 0.38 for the Standard test and to 0.16 for the Lewis test in order for the actual significance level to be $\alpha = 0.05$. Tables IX–XI provide the test results.

The four lognormal distributions with different variances can each be regarded as the null hypothesis when the KS tests of a lognormal null hypothesis are applied with estimated parameters, because they will have the appropriate estimated parameters. Table IX for $n = 200$ and Table X for $n = 2{,}000$ shows that, after the adjustments described earlier, they all have the correct significance level. Comparing Table IX to Table V, we see that the relative performance of the four KS tests is about the same: The

Table VIII. The Power of Alternative KSTests of the Null Hypothesis that Data are i.i.d. $N(0, 1)$ Variables for the Sample Size $n = 50$ for Various Alternative Hypotheses: Number of KS Tests Passed (denoted by $\#P$) at Significance Level 0.05 out of 10,000 Replications and the Average p-Values (denoted by $E[p\text{-value}]$) with Associated 95% Confidence Intervals. The First Nine Alternative Hypotheses have Mean 0 by Subtracting 1 from the Previous Mean-1 Cases

		Standard		Durbin		CU		Lewis	
Case	*Subcase*	$\#P$	$E[p\text{-value}]$	$\#P$	$E[p\text{-value}]$	$\#P$	$E[p\text{-value}]$	$\#P$	$E[p\text{-value}]$
Exp	–	5576	0.07 ± 0.0010	1871	0.04 ± 0.0016	6716	0.23 ± 0.0049	1154	0.02 ± 0.0006
E_k	$k = 2$	3813	0.04 ± 0.0006	2953	0.05 ± 0.0018	9364	0.52 ± 0.0059	376	0.01 ± 0.0004
	$k = 4$	20	0.01 ± 0.0002	336	0.01 ± 0.0004	9977	0.81 ± 0.0043	0	0.00 ± 0.0000
	$k = 6$	0	0.00 ± 0.0000	5	0.00 ± 0.0000	10,000	0.92 ± 0.0026	0	0.00 ± 0.0000
H_2	$c^2 = 1.25$	4188	0.05 ± 0.0010	1004	0.02 ± 0.0011	5051	0.16 ± 0.0043	417	0.01 ± 0.0004
	$c^2 = 1.5$	3100	0.04 ± 0.0009	629	0.01 ± 0.0009	4022	0.12 ± 0.0039	174	0.00 ± 0.0003
	$c^2 = 2$	1747	0.02 ± 0.0008	221	0.00 ± 0.0004	2639	0.07 ± 0.0031	36	0.00 ± 0.0001
	$c^2 = 4$	222	0.00 ± 0.0003	17	0.00 ± 0.0001	1237	0.04 ± 0.0027	1	0.00 ± 0.0000
	$c^2 = 10$	7	0.00 ± 0.0001	1	0.00 ± 0.0000	1870	0.09 ± 0.0046	0	0.00 ± 0.0000
$Mixture$	–	4836	0.05 ± 0.0008	2671	0.05 ± 0.0018	7273	0.37 ± 0.0065	533	0.01 ± 0.0004
LN	$(1, 0.25)$	0	0.00 ± 0.0001	41	0.00 ± 0.0001	9915	0.76 ± 0.0050	0	0.00 ± 0.0000
	$(1, 1)$	1722	0.03 ± 0.0005	700	0.01 ± 0.0007	5971	0.24 ± 0.0055	89	0.00 ± 0.0002
	$(1, 4)$	460	0.01 ± 0.0004	31	0.00 ± 0.0002	2027	0.06 ± 0.0028	5	0.00 ± 0.0000
	$(1, 10)$	24	0.00 ± 0.0001	0	0.00 ± 0.0000	1168	0.03 ± 0.0021	1	0.00 ± 0.0000
RRI	$p = 0.1$	5219	0.06 ± 0.0010	763	0.01 ± 0.0009	6239	0.21 ± 0.0049	1152	0.02 ± 0.0007
	$p = 0.5$	2791	0.03 ± 0.0008	0	0.00 ± 0.0000	4283	0.13 ± 0.0039	788	0.01 ± 0.0007
	$p = 0.9$	62	0.00 ± 0.0001	0	0.00 ± 0.0000	3696	0.17 ± 0.0057	15	0.00 ± 0.0001
$EARMA$	0.25	5395	0.07 ± 0.0010	1813	0.04 ± 0.0016	5820	0.20 ± 0.0048	1120	0.02 ± 0.0007
	0.5	5296	0.06 ± 0.0011	1872	0.04 ± 0.0016	5140	0.17 ± 0.0045	1192	0.02 ± 0.0008
	1	5028	0.06 ± 0.0011	1884	0.04 ± 0.0017	4883	0.17 ± 0.0047	1370	0.02 ± 0.0010
	3	3034	0.03 ± 0.0008	2492	0.05 ± 0.0019	2970	0.09 ± 0.0034	1474	0.03 ± 0.0014
	5.25	3446	0.04 ± 0.0010	2049	0.05 ± 0.0020	4275	0.17 ± 0.0053	2115	0.05 ± 0.0023
mH_2	$m = 2$	2363	0.03 ± 0.0009	460	0.01 ± 0.0007	2777	0.09 ± 0.0038	76	0.00 ± 0.0002
	$m = 5$	4045	0.05 ± 0.0010	1109	0.02 ± 0.0012	4591	0.16 ± 0.0046	421	0.01 ± 0.0005
	$m = 10$	4667	0.06 ± 0.0010	1477	0.03 ± 0.0015	5682	0.20 ± 0.0049	706	0.01 ± 0.0006
	$m = 20$	4932	0.06 ± 0.0010	1636	0.03 ± 0.0015	6361	0.23 ± 0.0050	891	0.02 ± 0.0007
$RRI(H_2)$	$p = 0.1$	302	0.01 ± 0.0003	3	0.00 ± 0.0001	1306	0.04 ± 0.0028	3	0.00 ± 0.0000
	$p = 0.5$	454	0.01 ± 0.0004	0	0.00 ± 0.0000	2009	0.07 ± 0.0037	12	0.00 ± 0.0001
	$p = 0.9$	17	0.00 ± 0.0001	0	0.00 ± 0.0000	4063	0.21 ± 0.0065	1	0.00 ± 0.0000
$N(0, 1)$	–	9447	0.50 ± 0.0057	9460	0.50 ± 0.0057	9501	0.50 ± 0.0056	9492	0.50 ± 0.0057
$E_k - 1$	$k = 2$	9336	0.47 ± 0.0057	9472	0.49 ± 0.0057	8782	0.40 ± 0.0057	8393	0.38 ± 0.0058
$+ \sqrt{1 - 1/k}$	$k = 4$	9526	0.51 ± 0.0056	9493	0.50 ± 0.0057	9330	0.47 ± 0.0057	9410	0.48 ± 0.0057
$\times N(0, 1)$	$k = 6$	9503	0.50 ± 0.0057	9476	0.50 ± 0.0057	9427	0.49 ± 0.0057	9445	0.49 ± 0.0057

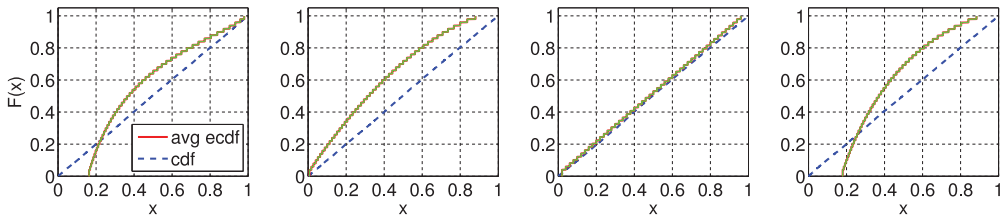

Fig. 5. Comparison of the average ecdf based on $Exp - 1$ data for 10^4 replications with $n = 200$ with the $N(0, 1)$ null hypothesis: Standard, Durbin, CU, and Lewis tests (from left to right).

Table IX. Performance of Alternative KS Tests of i.i.d. Lognormal Variables with Estimated Mean and Variance for the Sample Size $n = 200$: Number of KS Tests Passed at Significance Level 0.05 out of 10,000 Replications The nominal significance levels are increased from 0.05 to 0.38 for the Standard test and to 0.16 for the Lewis test in order for the actual significance level to be 0.05.

Case	Subcase	Standard	Durbin	CU	Lewis
Exp	–	432	7238	9950	46
E_k	$k = 2$	2712	9001	9896	1030
	$k = 4$	5782	9381	9833	3858
	$k = 6$	7027	9427	9781	5636
H_2	$c^2 = 1.25$	926	8152	9931	197
	$c^2 = 1.5$	1716	8689	9915	598
	$c^2 = 2$	2616	9121	9877	1467
	$c^2 = 4$	3358	9036	9695	2823
	$c^2 = 10$	1607	7609	9496	1503
Mixture	–	1182	8049	9888	483
LN	(1, 0.25)	9497	9496	9493	9454
	(1, 1)	9501	9525	9505	9477
	(1, 4)	9535	9519	9516	9530
	(1, 10)	9499	9479	9505	9484
RRI	$p = 0.1$	396	638	9872	61
	$p = 0.5$	223	0	7823	123
	$p = 0.9$	0	0	830	2
EARMA	0.25	435	7313	9795	41
	0.5	384	7214	9440	33
	1	432	7033	8721	61
	3	2606	5846	2862	1369
	5.25	712	5716	3318	223
mH_2	$m = 2$	1952	8888	9008	950
	$m = 5$	959	8145	8604	228
	$m = 10$	653	7728	8917	114
	$m = 20$	460	7517	9448	63
$RRI(H_2)$	$p = 0.1$	2874	1173	9437	2644
	$p = 0.5$	1041	0	6128	1408
	$p = 0.9$	1	0	463	4

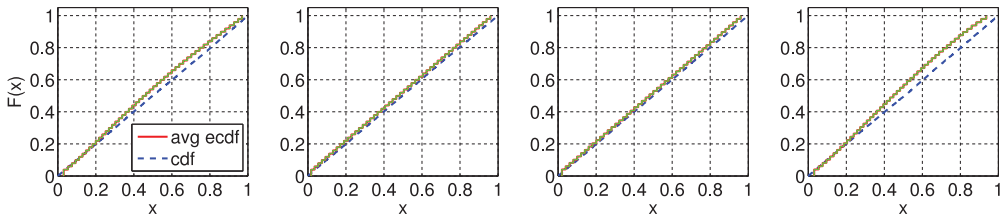

Fig. 6. Comparison of the average ecdf based on $E_2 - 1 + \sqrt{1 - 1/k}N(0, 1)$ data for 10^4 replications and $n = 200$ with the $N(0, 1)$ null hypothesis: Standard, Durbin, CU, and Lewis tests (from left to right).

Lewis KS test is best, with the standard KS test close behind, and both significantly better than the other two.

We also find that the relative performance of the KS tests of the normal null hypothesis when we estimate parameters is about the same as for specified parameters. Table XI shows the results for $n = 50$. Comparing Table XI to Table VIII, we see that in some cases the parameter estimation causes a loss in power, but the degradation is least for the Lewis KS test. As in Section 6, we consider i.i.d. sequences of random

Table X. Performance of Alternative KS Tests of i.i.d. Lognormal Variables with Estimated Mean and Variance for the Sample Size $n = 2{,}000$: Number of KS Tests Passed at Significance Level 0.05 out of 10,000 Replications The nominal significance levels are increased from 0.05 to 0.38 for the Standard test and to 0.16 for the Lewis test in order for the actual significance level to be 0.05.

Case	Subcase	Standard	Durbin	CU	Lewis
Exp	–	0	28	9968	0
E_k	$k = 2$	0	3214	9907	0
	$k = 4$	0	7743	9838	0
	$k = 6$	24	8744	9805	0
H_2	$c^2 = 1.25$	0	510	9955	0
	$c^2 = 1.5$	0	1857	9922	0
	$c^2 = 2$	0	4445	9870	0
	$c^2 = 4$	0	4774	9693	0
	$c^2 = 10$	0	267	9422	0
Mixture	–	0	675	9889	0
LN	(1, 0.25)	9522	9496	9521	9500
	(1, 1)	9506	9477	9512	9494
	(1, 4)	9455	9513	9517	9452
	(1, 10)	9482	9494	9483	9488
RRI	$p = 0.1$	0	0	9892	0
	$p = 0.5$	0	0	7717	0
	$p = 0.9$	0	0	203	0
EARMA	0.25	0	22	9797	0
	0.5	0	36	9351	0
	1	0	21	8295	0
	3	0	575	2272	0
	5.25	0	95	1322	0
mH_2	$m = 2$	0	3325	8834	0
	$m = 5$	0	625	7830	0
	$m = 10$	0	160	7522	0
	$m = 20$	0	62	7694	0
$RRI(H_2)$	$p = 0.1$	0	0	9445	0
	$p = 0.5$	0	0	5555	0
	$p = 0.9$	0	0	43	0

variables distributed as $E_k - 1 + \sqrt{1 - (1/k)}N(0, 1)$ for $k = 2, 4, 6$, which have the same first two moments and approximately the same shape as $N(0, 1)$. For this relatively challenging example, we see from Tables VIII and XI that the Lewis KS test has greater power than for known parameters, and outperforms the other KS tests. Similar results for $n = 200$ can be found in the Appendix.

8. CONCLUSIONS

We have conducted simulation experiments to study the power of alternative KS statistical tests of the null hypothesis that observations come from a sequence if i.i.d. random variables with continuous cdf F, focusing on the exponential and standard normal null hypotheses. Our analysis strongly supports the data-transformation approach proposed by Durbin [1961] for the normal null hypothesis, but we find that it performs poorly for the exponential distribution and related distributions of nonnegative random variables.

Our main conclusion is that a new KS test, involving a new combination of transformations, has more power. In particular, we find that it is better to transform the given hypothesized sequence to i.i.d. exponential random variables (under the null

Table XI. Performance of Alternative KS Tests of i.i.d. Normal Variables with Estimated Mean and Variance for the Sample Size $n = 50$: Number of KS Tests Passed at Significance Level 0.05 out of 10,000 Replications
The first nine alternative hypotheses have mean 0 by subtracting 1 from the previous mean-1 cases. (The nominal significance levels are increased from 0.05 to 0.38 for the Standard test and to 0.16 for the Lewis test in order for the actual significance level to be 0.05.)

Case	Subcase	Standard	Durbin	CU	Lewis
Exp	–	431	2511	6425	170
E_k	$k = 2$	3083	7517	7417	1465
	$k = 4$	6063	9185	8168	4001
	$k = 6$	7199	9400	8468	5392
H_2	$c^2 = 1.25$	141	1293	5490	52
	$c^2 = 1.5$	58	774	4753	22
	$c^2 = 2$	21	339	3778	7
	$c^2 = 4$	14	167	2197	7
	$c^2 = 10$	68	424	1659	24
Mixture	–	2053	5699	5795	925
LN	(1, 0.25)	3409	8109	7089	1618
	(1, 1)	268	2154	5009	80
	(1, 4)	3	75	3110	0
	(1, 10)	0	5	2299	0
RRI	$p = 0.1$	464	1161	5785	200
	$p = 0.5$	276	0	2962	260
	$p = 0.9$	2	0	742	5
EARMA	0.25	482	2663	5272	180
	0.5	604	2924	4411	251
	1	924	3456	3819	440
	3	1668	4500	1244	896
	5.25	3282	5718	1771	2534
mH_2	$m = 2$	68	667	2826	19
	$m = 5$	222	1680	4277	72
	$m = 10$	343	2181	5287	117
	$m = 20$	380	2419	6046	142
$RRI(H_2)$	$p = 0.1$	26	77	1958	10
	$p = 0.5$	54	0	1265	51
	$p = 0.9$	1	0	575	4
$N(0, 1)$	–	9489	9477	9516	9437
$E_k - 1$	$k = 2$	8670	9489	8826	7729
$+ \sqrt{1 - (1/k)}$	$k = 4$	9469	9504	9348	9314
$\times N(0, 1)$	$k = 6$	9526	9465	9439	9415

hypothesis), which can be regarded as the interarrival times of a PP. Then we apply the CU transformation to the PP and afterwards apply the Durbin transformation, as was done by Lewis [1965] in his test of a PP. This new KS test, which we call the Lewis test, is usually superior, often markedly so. Thus, we recommend the Lewis test, implemented as described in Section 2.

The Lewis test was found to have more power against alternatives with nonexponential distributions. To test against dependent exponential alternative hypotheses, we recommend also applying the CU KS test. None of the KS tests has much power against dependent exponential alternatives, but the CU KS test seems best.

The realized performance of these different KS tests can be better understood by recognizing that the CU transformation converts the arrival times into uniform random variables, whereas the standard KS test focuses on the interarrival times. The Durbin [1961] transformation is so effective after the CU transformation because it starts by

reordering the interarrival times in ascending order. That explains the performance of the CU and Lewis KS tests.

We have observed that the original Durbin KS test has exceptional power against the RRI alternative hypotheses and we have explained why. That occurs because the repeated values are converted into 0's. That repetition structure makes the model relatively easy to analyze, but it probably is not often a realistic model of dependence.

Since there is some variation in the results, we recommend applying simulation as we have done in this article, if there is the opportunity, in order to assess what KS test has the most power and what that power should be in a new setting of interest. The tables and plots based on 10^4 replications give a very clear picture. As we saw in Section 7, simulation plays an important role in determining the appropriate KS test with estimated parameters. As in the Lilliefors [1969] test of an i.i.d. sequence of exponential random variables using the estimated mean, it is often necessary to increase the nominal significance level substantially in order that the KS test with estimated parameters has the final desired significance level. For example, to achieve the target significance level of $\alpha = 0.05$ in the standard KS test of the exponential null hypothesis, we found that it is necessary to increase the nominal significance level to 0.18.

Both in Kim and Whitt [2014b] and here we have focused on the two-sided KS test, but we also conducted one-sided KS tests. We found that the one-sided test can further increase power when it is justified (see Kim and Whitt [2014c]). As usual with statistical tests, the power increases with the sample size, so that some sample sizes may be too small to have any power, whereas other sample sizes may be too large to accept even the slightest deviation from a null hypothesis. Thus, as many have discovered before, judgment is required in the use of statistical tests.

ACKNOWLEDGMENTS

This work was done at Columbia University while Song-Hee Kim was a doctoral student. Support was received from U.S. National Science Foundation grants CMMI 1066372 and 1265070 and by the Samsung Foundation.

REFERENCES

M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov. 2011. Patient Flow in Hospitals: A Data-based Queueing-Science Perspective. New York University, http://www.stern.nyu.edu/om/faculty/armony.

G. A. Barnard. 1953. Time Intervals Between Accidents–A Note on Maguire, Pearson & Wynn's Paper. *Biometrika* 40 (1953), 212–213.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Am. Stat. Assoc.* 100 (2005), 36–50.

M. A. Crane and D. L. Iglehart. 1974a. Simulating stable stochastic systems, I: General multiserver queues. *J. ACM* 21, 1 (1974), 103–113.

M. A. Crane and D. L. Iglehart. 1974b. Simulating stable stochastic systems, II: Markov chains. *J. ACM* 23, 1 (1974), 114–123.

M. A. Crane and D. L. Iglehart. 1975. Simulating stable stochastic systems: III. Regenerative processes and discrete-event simulations. *Oper. Res.* 23, 1 (1975), 33–45.

J. Durbin. 1961. Some methods for constructing exact tests. *Biometrika* 48, 1 (1961), 41–55.

N. Gans, N. Liu, A. Mandelbaum, H. Shen, and H. Ye. 2010. Service times in call centers: Agent heterogeneity and learning with some operational consequences. In *Borrowing Strength: Theory Powering Applications, A Festschrift for Lawrence D. Brown*. Institute of Mathematical Statistics, New York.

P. W. Glynn and D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Manage. Sci.* 35, 11 (1989), 1367–1392.

P. Heidelberger and D. L. Iglehart. 1979. Comparing stochastic systems using regenerative simulation with common random numbers. *Adv. Appl. Prob.* 11 (1979), 804–819.

P. A. Jacobs and P. A. W. Lewis. 1977. A Mixed autoregressive-moving average exponential sequence and point process (EARMA 1,1). *Adv. Appl. Prob.* 9, 9 (1977), 87–104.

P. A. Jacobs and P. A. W. Lewis. 1978. Discrete time series generated by mixtures. I: Correlational and runs properties. *J. Roy. Stat. Soc. B* 40, 1 (1978), 94–105.

P. A. Jacobs and P. A. W. Lewis. 1983. Stationary autoregressive discrete moving average time series generated by mixtures. *J. Roy. Stat. Soc. B* 4, 1 (1983), 19–36.

S.-H. Kim and W. Whitt. 2014a. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes. *Manufact. Service Oper. Manage.* 16, 3 (2014), 464–480.

S.-H. Kim and W. Whitt. 2014b. Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Nav. Res. Logist.* 61, 1 (2014), 66–90.

S.-H. Kim and W. Whitt. 2014c. Choosing Arrival Process Models for Service Systems: Tests of a Nonhomogeneous Poisson Process: Appendix. Columbia University. Retrieved from http://www.columbia.edu/∼ww2040/allpapers.html.

S.-H. Kim, P. Vel, W. Whitt, and W. C. Cha. 2015. Poisson and non-Poisson properties in appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters* 43 (2015), 247–253.

P. A. W. Lewis. 1965. Some results on tests for Poisson processes. *Biometrika* 52, 1 (1965), 67–77.

H. W. Lilliefors. 1969. On the Kolmogorov-Smimov test for the exponential distribution with mean unknown. *J. Am. Statist. Assoc.* 64 (1969), 387–389.

M. F. Neuts. 1981. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University, Baltimore.

G. Pang and W. Whitt. 2012. The impact of dependent service times on large-scale service systems. *Manufact. Service Oper. Manage.* 14, 2 (2012), 262–278.

S. M. Ross. 1996. *Stochastic Processes* (2nd. ed.). Wiley, New York.

G. R. Shorack and J. A. Wellner. 2009. *Empirical Processes with Applications in Statistics* (SIAM Classics in Applied Mathematics ed.). SIAM, Philadelphia.

R. Simard and P. L'Ecuyer. 2011. Computing the two-sided Kolmogorov-Smirnov distribution. *J. Statist. Softw.* 39, 11 (2011), 1–18.

W. Whitt. 1982. Approximating a point process by a renewal process: Two basic methods. *Oper. Res.* 30 (1982), 125–147.

W. Whitt. 2002. *Stochastic-Process Limits*. Springer, New York.

Received August 2013; revised July 2014; accepted November 2014