

Choosing Arrival Process Models for Service Systems: Tests of a Nonhomogeneous Poisson Process

Song-Hee Kim, Ward Whitt

*Department of Industrial Engineering and Operations Research, Columbia University,
New York, New York 10027*

Received 26 February 2013; revised 11 December 2013; accepted 12 December 2013

DOI 10.1002/nav.21568

Published online 10 January 2014 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Service systems such as call centers and hospital emergency rooms typically have strongly time-varying arrival rates. Thus, a nonhomogeneous Poisson process (NHPP) is a natural model for the arrival process in a queueing model for performance analysis. Nevertheless, it is important to perform statistical tests with service system data to confirm that an NHPP is actually appropriate, as emphasized by Brown et al. [8]. They suggested a specific statistical test based on the Kolmogorov–Smirnov (KS) statistic after exploiting the conditional-uniform (CU) property to transform the NHPP into a sequence of i.i.d. random variables uniformly distributed on $[0, 1]$ and then performing a logarithmic transformation of the data. We investigate why it is important to perform the final data transformation and consider what form it should take. We conduct extensive simulation experiments to study the power of these alternative statistical tests. We conclude that the general approach of Brown et al. [8] is excellent, but that an alternative data transformation proposed by Lewis [22], drawing upon Durbin [10], produces a test of an NHPP test with consistently greater power. We also conclude that the KS test after the CU transformation, without any additional data transformation, tends to be best to test against alternative hypotheses that primarily differ from an NHPP only through stochastic and time dependence. © 2014 Wiley Periodicals, Inc. *Naval Research Logistics* 61: 66–90, 2014

Keywords: nonhomogeneous Poisson process; power of statistical tests; data transformations for statistical tests; service systems; arrival processes; time-varying arrival rate

1. INTRODUCTION

This research was motivated by the desire to develop and apply appropriate stochastic queueing models to set capacities and analyze the performance of service systems, such as telephone call centers and hospital emergency rooms; see [1, 2] for background. Since the arrival rate in these service systems typically varies strongly by time of day, a natural arrival-process model is a nonhomogeneous Poisson process (NHPP). The Poisson property arises from the independent decisions of many people, each of whom uses the service system only rarely. There is a supporting limit theorem, often called the Poisson superposition theorem or the Poisson law of rare events; for example, see [4], section 11.2 of [9] and section 9.8 of [38].

1.1. The Conditional-Uniform Transformation

Despite the theoretical basis, it is important to apply statistical tests with service system data to determine if an NHPP

is actually appropriate in any instance. A specific statistical test for this purpose was proposed and applied by Brown et al. [8]. Their approach effectively addresses the complicated time-varying arrival rate by converting the problem into a standard statistical test to determine whether data can be regarded as a sample from a sequence of independent and identically distributed (i.i.d.) random variables with a specified distribution.

The first step in [8] is to approximate the NHPP by a piecewise-constant (PC) NHPP, which is usually reasonable because the arrival rate in a service system typically changes relatively slowly compared to the overall arrival rate. We make the same assumption here, but we separately evaluate the impact of the PC NHPP approximation on the statistical tests in [20].

The second step in [8] is to apply the classical conditional-uniform (CU) transformation to transform the PC NHPP into a sequence of i.i.d. random variables uniformly distributed on $[0, 1]$. The PC rate approximation is an important prerequisite, because a PC NHPP can be regarded as a homogeneous Poisson process (PP) over each subinterval. For a PP on an interval $[0, T]$, it is well known that, conditional on the total

Correspondence to: S.-H. Kim (sk3116@columbia.edu)

number of arrivals in that interval, the arrival times divided by T are distributed as the order statistics of i.i.d. random variables uniformly distributed on $[0, 1]$; for example, see section 2.3 of [30]. With that CU transformation, the data from all the subintervals can be combined to obtain a single sequence of i.i.d. random variables uniformly distributed on $[0, 1]$.

There are significant costs and benefits associated with the CU transformation. An important benefit of the CU transformation is that it eliminates all nuisance parameters; the final representation is independent of the rate of the PP on each subinterval. That is crucial for testing a PC NHPP, because it allows us to combine the data from separate intervals with different rates on each interval. The test is thus the same as if it were for a PP with specified rate.

However, there is also a cost of using the CU transformation. With the CU transformation, the constant rate on each subinterval could be random. A good test result does not support any candidate rate or imply that the rate on each subinterval is deterministic. Thus, those issues remain to be addressed. For dynamic time-varying estimation needed for staffing, that can present a challenging forecasting problem, as reviewed in [14] and references therein.

On the positive side, this limitation of the CU transformation can be an advantage. By applying the CU transformation to different days separately, as well as to different subintervals within each day as needed to warrant the PC rate approximation, this method accommodates the commonly occurring phenomenon of day-to-day variation, in which the rate of the PP randomly varies over different days; see, for example, [3, 14, 18]; that is, when the CU transformation is applied to different days separately, it will not reject because of such day-to-day variation. Indeed, the CU transformation applied in that way makes the statistical test actually be for a Cox process, that is, for a doubly stochastic PP, where the random rate is constant over each subinterval over which the CU transformation is applied. Having a test for a Cox process can be useful too, because where Cox processes have been used, as in [6, 14], it has been assumed that the Cox process is appropriate.

It is significant that we have a choice: If we do not want to reject the NHPP hypothesis when there is day-to-day variation, then we can apply the CU transformation to different days separately; if we do want to reject the NHPP hypothesis when there is day-to-day variation, then we simply combine the data over multiple days.

1.2. The Kolmogorov–Smirnov Statistical Test

Given the approximation by a PC NHPP and the CU transformation, the test of an NHPP is transformed into a test of a sequence of i.i.d. random variables with a specified distribution. Hence, many standard statistical tests can be applied.

Brown et al. [8] elected to apply the standard *Kolmogorov–Smirnov* (KS) test, so we will focus on the KS test here. However, we also have considered the one-sided KS tests and the Anderson-Darling (AD) test, for example, see [35], and give comparative results for them in the appendix [19]. The results for the AD test are similar to the results for the KS test; we elaborate in section 5.1.

The KS test determines if n observations can be regarded as a sample from a sequence of i.i.d. random variables $\{X_n : n \geq 1\}$, each distributed as a random variable X with a specified continuous cumulative distribution function (cdf) $F(x) \equiv P(X \leq x), x \in \mathbb{R}$. Just like many of the other related tests, the KS test is based on the difference between the empirical cdf (ecdf)

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

and the underlying cdf F , where 1_A is an indicator function, equal to 1 if the event A occurs, and equal to 0 otherwise. The KS test focuses on the maximum difference

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

which has a distribution that is independent of the cdf F (provided that the cdf is continuous), for example, see [32]. For any observed maximum y from a sample of size n , we compute the P -value $P(D_n > y)$ 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 $\alpha = 0.05$. The other standard tests also look at the difference between the ecdf in (1) and the cdf F , but using different distance measures than the uniform distance in (2).

We call the KS test after applying the CU transformation the CU KS test. However, Brown et al. [8] did not stop with the CU KS test, but instead proposed a (rather complicated, scaled) logarithmic transformation of the data (see section 3.2), which under the Poisson null hypothesis produces a single sequence of i.i.d. mean-1 exponential random variables. Then, they applied the KS test (2) with $F(x) \equiv 1 - e^{-x}$; thus, we call their test the Log test. They found that their banking call center data passed their Log test; this reference has since been highly cited. This article studies the Log KS test of a NHPP and various alternatives.

1.3. Why Is the Final Logarithmic Data Transformation Needed?

We conducted this research because we wondered why the additional logarithmic data transformation was used in [8]. Why not just use the CU KS test?

In this article, we investigate why an additional data transformation is needed and what form it should take. Since we

approximate by a PC NHPP, it suffices to study alternative tests of a PP, and we frame the problem that way, but we also investigate the consequences of combining data from subintervals because of the PC NHPP approximation of a general NHPP; see section 5.5.

Our experiments show that the CU KS test of a PP has remarkably little power against standard alternatives, such as renewal processes with nonexponential interrenewal distributions. In contrast, the Log data transformation from [8] produces a KS test of a PP with much greater power against these alternatives. Moreover, we find another data transformation that produces a KS test of a PP that has even greater power. [Recall that the power is the 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 failing to reject the null hypothesis (type II error) when it is false, which of course depends on the alternative.]

After study, we conclude that the problem with the CU KS test of a PP is that it is actually subtly different from a standard KS test of a PP, which focuses on the PP via its interarrival times. Since a PP is characterized by having its successive interarrival times be a sequence of i.i.d. exponentially distributed random variables, the standard KS test of a PP is the KS test of the exponential distribution, and this standard KS test does have reasonable power against renewal processes with nonexponential interrenewal times, as we show. However, the standard KS test is unattractive for the extension to NHPP's, because we would need to estimate the rate of the PC NHPP over each of the subintervals over which the rate is assumed to be constant.

Just like the standard KS test of a PP, the CU KS test of a PP determines if observations can be regarded as coming from an i.i.d. sequence of random variables with a specified distribution, but these random variables have a different interpretation. If there are n arrival times T_j in the interval $[0, T]$, the ecdf is just as in (1) with $X_j \equiv T_j/T$ and the CU KS statistic is just as in (2).

The subtle difference between the CU KS test and the standard KS test is that the CU property of a PP over an interval $[0, T]$ produces i.i.d. random variables uniformly distributed over the entire interval $[0, T]$. These variables T_j correspond to the arrival times (the successive partial sums of the interarrival times), rather than the interarrival times themselves, which are tested in the standard KS test. As a consequence, the CU KS test evidently is less able to detect differences in the interarrival-time distribution.

We provide mathematical support by proving that the ecdf in (1) used in the CU KS test converges to the uniform cdf as the sample size n increases for any rate-1 stationary ergodic point process, that is, for any rate-1 stationary point process satisfying a strong law of large numbers (LLN) see sections 7

and 8. As a consequence, as the sample size n increases, to first order, there is asymptotically no power at all against any of the alternatives in this enormous class.

1.4. Early Papers by Durbin [10] and Lewis [22]

Our study not only strongly supports Brown et al. [8], but we also found that there is relevant history in the statistical literature. In particular, Lewis [22] made a significant contribution for testing a PP, recognizing that a transformation proposed by Durbin [10] could be effectively applied after the CU transformation to obtain a new KS test; we call that the Lewis test. From [22], we also discovered that the direct CU test of a PP had been proposed by Barnard [5], and that the CU test was known to have little power.

On discovering [22], we first supposed that the Log test would turn out to be equivalent to the Lewis test, and that the Lewis test would coincide with the original KS test in Durbin [10], but neither is the case. Thus, this past work suggests several different KS tests, providing even more motivation for our study. The experimental results do not all agree, but overall we conclude that the Lewis test is the best; it consistently has higher power than the Log test. However, the difference is small compared to the advantage over the virtually powerless CU KS test against renewal process alternatives with nonexponential interrenewal times.

1.5. Why Does the Lewis KS Test Have More Power?

We have explained that the CU KS test of a PP has relatively low power against alternatives with nonexponential interarrival times, because the CU transformation focuses on the arrival times instead of the interarrival times, and thus produces i.i.d. random variables that are uniformly distributed over the entire interval. However, the Lewis KS test of a PP starts with that same CU transformation. Why does it have significantly greater power against alternatives with nonexponential interarrival times?

Fortunately, there seems to be a simple explanation: The Durbin [10] transformation used by Lewis [22] after the CU transformation focuses directly on the interarrival times instead of the arrival times themselves. Indeed, the first step of the Durbin transformation in (3) of section 3 is to reorder the interarrival times in ascending order. Of course, the story must be somewhat more complicated because, under the null hypothesis, the final transformed random variables are again i.i.d. uniform random variables on $[0, 1]$, just as in the CU KS test. However, if the interarrival times in the alternative hypothesis are substantially different from those of a PP, then the difference in those interarrival times is likely to be highlighted by this transformation.

1.6. But the CU KS Test Is Good for Dependence Alone

We conclude that the Lewis test of a PP tends to have relatively high power for testing the assumed interarrival-time cdf, as in a non-PP renewal process. However, we also conclude that none of the KS tests of a PP has much power for alternatives that differ only because of dependence among successive exponential interarrival times.

Surprisingly, given the previous results, we find that the CU KS test of a PP has more power against alternative hypotheses involving dependent exponential interarrival times than the other KS tests. More generally, we find that the CU test of an NHPP has more power against both time dependence and stochastic dependence. This is important for queueing systems, because these forms of dependence commonly occur, as we illustrate for departure processes from many-server queues in section 6.

We thus ask: Why does the CU KS test of a PP have more power against alternative hypotheses with dependent exponential interarrival times?

Fortunately, the same reason seems to apply: For such alternative hypotheses, there is no difference in the interarrival-time cdf, so there should be no gain from reordering the interarrival times, which is at the heart of the Durbin transformation. In fact, the dependence is likely to cause a difference from a PP through the cumulative impact of several consecutive interarrival times. This impact may well be reduced by reordering the interarrival times, because the original order has been disturbed.

1.7. Our Contributions

Our main contribution is to expose: (i) why the CU KS test of a PP has remarkably low power against alternative hypotheses with nonexponential interarrival times, (ii) why the Lewis test of a PP should have much greater power against alternative hypotheses with nonexponential interarrival times, and (iii) why the CU KS test of a PP may be more effective than the other KS tests against dependent exponential interarrival times. We also substantiate these insights with extensive simulation experiments to study how the data transformations perform.

As usual, the power increases as the distance between the hypotheses (appropriately defined) and the sample size increases. To provide insight, we considered a range of natural alternatives and a range of interval lengths for a rate-1 PP, which translates immediately into a range of expected sample sizes. Our base case has the interval $[0, 200]$ and thus expected sample size 200, which is consistent with call center examples in which the arrival rate might be regarded as approximately constant over half hours. (A call center with arrival rate 200 per half hour, handling calls of average duration 6 min, should have a little over 40 agents responding to the calls.)

We go beyond Lewis [22] in several important ways. First, instead of actual data sets, we use simulation to have

controlled experiments. Second, we explore the power as a function of the way the alternative hypothesis differs from the null hypothesis (all within the setting of stationary point processes with the same rate). We focus on two important different aspects of the PP null hypothesis: (i) the assumed interarrival-time cdf, and (ii) the assumed independence of successive interarrival times. For the interarrival-time cdf's, we primarily consider renewal processes with interarrival-time cdf's that are both more variable and less variable than the exponential cdf. We also consider non-Poisson renewal processes in which the nonexponential interarrival-time cdf's have the same mean and variance as the exponential cdf associated with the null hypothesis. In addition, we consider nonrenewal processes with nonexponential interarrival times.

To explore the impact of dependence, we use basic models of dependent sequences with exponential interarrival times, as in [15–17]. We find that none of the KS tests has much power against alternatives with dependent exponential distribution. Consistent with the discussion above, we find that the CU test can be useful to detect deviations from a PP due to dependence alone, but that tends to require very large sample sizes.

Since none of the KS tests of a PP have much power against alternatives that differ from the PP null hypothesis only via dependence among exponential interarrival times, one might well wonder whether the KS test can detect such departures from a PP with any amount of data. (Indeed, it is common to ignore the dependence, and regard the KS test as a test only of the interarrival-time cdf.) However, the KS test can indeed detect dependence, but it usually requires much more data. This can be explained by asymptotics for the KS statistic as the sample size increases, which we review in sections 7 and 8. A difference between the interarrival-time cdf's of the two hypotheses is revealed in the LLN, yielding a deviation in the ecdf's (which are scaled by dividing by n) of order $O(1)$ as the sample size n increases, whereas dependence among exponential interarrival times is only revealed in the central limit theorem (CLT), but it is revealed there, but yielding a deviation in the ecdf's of only of order $O(1/\sqrt{n})$ as the sample size n increases, which is the same order as the KS statistic under the null hypothesis. Thus, the dependence can be detected, but it will require a large sample size n .

Because we are primarily interested in testing for an NHPP with nonstationary data, we also studied the consequence of combining data from the separate subintervals of a PC NHPP, over each of which the rate is assumed to be constant. First, it is not difficult to understand the consequence of the subintervals. It causes the terminal intervals to be truncated, but of potentially greater importance, it eliminates any dependence extending beyond the subinterval boundary. Thus, we should anticipate what we see: the use of subintervals has little impact for renewal process alternative hypotheses with nonexponential distributions, but it significantly degrades the

performance of the CU KS test, with dependent exponential interarrival times, evidently because the use of subintervals eliminates dependence that extends beyond the subinterval boundaries.

Because we find that different tests can have advantages for different alternatives, we also studied composite tests, obtained by combining the individual tests with fixed overall significance level, using the Bonferroni inequality and refined methods in [13, 33]; see [27] for background. For the most part, we conclude that the Lewis test of a PP alone has highest power against the arrival processes with different marginal interarrival-time cdf's, whereas other tests have higher power for various forms of dependence among the interarrival times with the correct exponential cdf, but the differences are small for dependence. We find that the CU KS test alone is relatively good for these cases. We put the details of this investigation in the appendix [19].

1.8. Organization of the Paper

We now indicate how the rest of the article is organized. We start in section 2 by carefully defining what is meant by an admissible data transformation and what is meant for it to be effective. In section 2.3 we also indicate how to see whether or not an admissible data transformation will be effective against a candidate alternative hypothesis through appropriate plots. Next, in section 3 we carefully define the different KS tests we consider. In section 4, we describe our simulation experiment. In section 5, we report our results. In section 6, we present a separate example of the KS tests applied to the departure process from a many-server queue, which shows the possible advantages of the CU KS test of an NHPP, where the alternative has both time and stochastic dependence. In section 7, we review asymptotic results for the KS statistic. In section 8, we show that the CU test has low power because it has no asymptotic power in the LLN scaling. In section 9, we draw conclusions. Additional details, including many more tables and plots, are given in an appendix [19], which is available from the authors' web pages.

2. ADMISSIBLE DATA TRANSFORMATIONS FOR KS TESTS

In section 3, we define the two data transformations used after the CU transformation from [8] and [22]. To place them in perspective, we now formally define what it means to be an admissible data transformation for a KS test, and specify what it means for the admissible data transformation to be effective. That leads naturally to various optimization problems, which so far remain unsolved.

Finally, we show that we can see if an admissible data transformation will be effective (have significant power) against

any specified alternative hypothesis of interest by estimating the average ecdf in (1) for that alternative hypothesis, based on the sample average from many replications of a simulation experiment, and comparing that estimated ecdf to the theoretical cdf. While doing that, we can estimate the distribution of the KS test statistic in (2). However, the visual plots dramatically show the power of alternative KS tests.

As will be clear from section 3, the two data transformations we consider are somewhat complicated. Even checking that they are indeed admissible requires some basic stochastic process theory, but that can be done. A good way to think of these data transformations is that they reorder and modify the original data, so that deviations from the null hypothesis are emphasized, producing larger values of the KS statistic in (2). For example, the data transformations tend to move small interarrival times toward the left end of the interval. It is not difficult to see that the Durbin data transformation in the Lewis test moves all 0-valued interarrival times to the left end of the interval. However, it remains to develop an improved understanding. That effort presumably should take place in the following mathematical framework.

2.1. The Mathematical Framework

The starting point is a KS test (but other statistical tests could be substituted) and its null hypothesis, which we often think of as a continuous cdf, but it should also include the i.i.d. property. Let the sample space be \mathcal{S} , which we require to be a subinterval of the real line.

DEFINITION 2.1: A **KS test** (of sample size n) uses the KS test statistic D_n in (2) to test whether or not an n -dimensional random vector is distributed as the **null hypothesis**, which can be characterized by a random vector $X \equiv (X_1, \dots, X_n)$ of i.i.d. random variables on \mathcal{S} with continuous cdf F .

An admissible data transformation maps one KS test into another, while leaving the distribution of the KS test statistic D_n in (2) under the null hypothesis unchanged. Thus, an admissible data transformation involves two KS null hypotheses indexed by i , which in turn involves two subintervals of the real line, \mathcal{S}_1 and \mathcal{S}_2 . For $i = 1, 2$ and any specified sample size n , let $X^{(i)} \equiv (X_1^{(i)}, \dots, X_n^{(i)})$ be n i.i.d. random variables with continuous cdf F_i on \mathcal{S}_i . Let \mathcal{S}_i^n be the associated n -fold product space. The two associated KS tests involve more general random vectors on \mathcal{S}_i^n .

DEFINITION 2.2: An **admissible data transformation for a KS test** with null hypothesis $X^{(1)}$ with cdf F_1 and sample size n , defined as above, is a measurable map

$$T : \mathcal{S}_1^n \rightarrow \mathcal{S}_2^n,$$

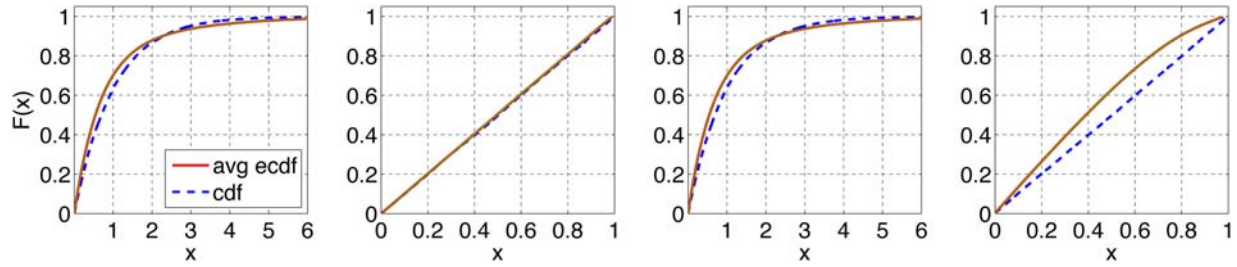


Figure 1. Comparison of the average ecdf based on 10^4 replications for four KS tests of a rate-1 Poisson process applied to a renewal process over $[0, 200]$ with mean-1 hyperexponential (H_2) interarrival times having $c^2 = 2$: standard, conditional-uniform, log and Lewis KS tests (from left to right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

such that $X^{(2)} \equiv \mathcal{T}(X^{(1)})$ is itself a bonafide KS null hypothesis with cdf F_2 and sample size n , i.e., $(X_1^{(2)}, \dots, X_n^{(2)})$ are n i.i.d. random variables on \mathcal{S}_2 with cdf F_2 , and the distribution of the KS statistic D_n in (2) under the null hypothesis remains unchanged.

The two data transformations we consider are admissible and both have domain \mathcal{S}_1^n for $\mathcal{S}_1 = [0, 1]$ and $F(t) = t$, the uniform cdf, because they are applied to test a PC NHPP following the CU transformation. The logarithmic data transformation in [8] has $\mathcal{S}_2 = [0, \infty)$ and $G(t) = 1 - e^{-t}$, the mean-1 exponential cdf. The Durbin [10] data transformation we propose here has $\mathcal{S}_2 = [0, 1]$ and $G(t) = t$, the uniform cdf.

2.2. The Power of Alternative Admissible Data Transformations

What we want to achieve with these admissible data transformations is to increase the power for various alternative hypotheses. Recall that an alternative hypothesis for given sample size n corresponds to an alternative joint probability distribution for the random vector $X \equiv (X_1, \dots, X_n)$.

DEFINITION 2.3: An admissible data transformation for a KS test with sample size n and significance level α (a) **increases power for a given alternative hypothesis X** if $P(D_n(X) > y_\alpha)$ increases after the transformation and (b) **is more powerful than another admissible data transformation for that given alternative hypothesis** if $P(D_n(X) > y_\alpha)$ is greater with the first than the second, where y_α is the P -value for the specified significance level α , that is, $P(D_n > y_\alpha) = \alpha$ under all null hypotheses.

We should point out that there is a large class of admissible data transformations that do not change the power at all. They apply identical transformations to each coordinate. We formalize by stating without proof the following well known proposition.

PROPOSITION 1: Admissible data transformations of KS tests can be defined by coordinate transformations

$$\mathcal{T} = \prod_{j=1}^n \mathcal{T}_j, \quad \text{where } \mathcal{T}_j : \mathcal{S}_1 \rightarrow \mathcal{S}_2$$

with

$$\mathcal{T}_j(X_j) = F_1(X_j) \text{ or } \mathcal{T}_j(X_j) = -\log(1 - F_1(X_j)) \text{ for all } j.$$

The first transformation makes F_2 a uniform cdf on $[0, 1]$, whereas the second makes F_2 a mean-1 exponential cdf. However, both admissible data transformations leave the distribution of the KS test statistic unchanged for alternative hypotheses as well as the null hypotheses.

This framework leads to many natural optimization problems, which evidently all remain unsolved. For given null hypothesis and alternative hypothesis, is there a most powerful admissible data transformation and, if so, what is it? Are there admissible data transformations that are uniformly most powerful for large classes of null and alternative hypotheses? The admissibility condition makes these constrained nonlinear optimization problems.

2.3. Seeing If The Data Transformations Are Effective

Although it remains to develop systematic methods to construct admissible data transformations and establish theoretical results about their power for various sets of alternatives, we conclude that there are good ways to see if these admissible data transformations are effective. For this purpose, simulation is very useful, because it provides controlled experiments. For simulation experiments involving a range of alternative hypotheses, we have found that plots comparing the average ecdf $F_n(x)$ in (1) from many independent replications to the theoretical cdf F are very revealing.

Figures 1 and 2 illustrate for 10,000 replications of a test of a PP applied to two rate-1 renewal process alternative

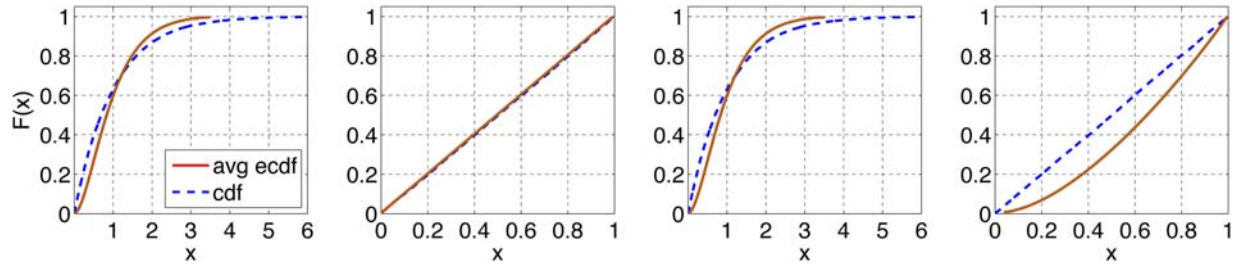


Figure 2. Comparison of the average ecdf based on 10^4 replications for four KS tests of a rate-1 Poisson process applied to a renewal process over $[0, 200]$ with mean-1 Erlang (E_2) interarrival times: standard, conditional-uniform, log and Lewis KS tests (from left to right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

hypotheses, the first with interarrival times having a hyperexponential (H_2) cdf with squared coefficient of variation (scv, variance divided by the square of the mean) $c^2 = 2$ and the second having an Erlang (E_2) cdf, both over the time interval $[0, 200]$. (See section 4.1 for more details.)

Figures 1 and 2 compare (from left to right) the standard, CU, Log, and Lewis KS tests of a PP. For the standard and Log KS tests, the null hypothesis is a mean-1 exponential cdf, while for the CU and Lewis KS tests, the null hypothesis is the uniform cdf on $[0, 1]$. We could apply Proposition 1 to make all four null hypotheses identical, but we leave them in their natural form. These figures dramatically show that the transformation in the Lewis KS test provides much greater separation between the average ecdf and the null hypothesis cdf than the other KS tests, while the CU test provides essentially no separation at all.

Many alternatives to the KS statistical test are also based on the ecdf $F_n(x)$ in (1), using different measures of the distance from the cdf F than the uniform distance in (2). From the plots of the ecdf and cdf such as in Figures 1 and 2, we can see how these alternative tests will perform. For example, we expect that the correct one-sided Lewis KS test will have even greater power than the Lewis KS test itself, whereas the wrong one-sided Lewis KS test will have essentially no power at all, like the CU KS test. That is born out in simulation experiments in the appendix [19]. However, since different one-sided KS tests are needed for these two different alternative hypotheses and since there is not likely to be such a narrow range of alternative hypotheses in applications, the standard KS test or possibly the (also symmetric) AD test seems more appropriate.

3. THE ALTERNATIVE KS STATISTICAL TESTS

As a basis for comparison in our study of the CU, Log, and Lewis KS tests, we also consider variants of the standard KS test as described in section 1.1. The standard KS test applies to a fixed number n of observations with a fully specified cdf, including the mean of the exponential interarrival times. However, in our application to a PP or a NHPP, both

these requirements are violated. First, the number of arrivals in each interval is actually random; second, in an application we would not know the rate of the PP over each subinterval, and thus we do not know the mean of the exponential interarrival times in the PP of the null hypothesis. Nevertheless, we include variants of the standard KS test for comparison.

3.1. Variants of the Standard KS Test

When we consider variants of the standard KS test applied to the interarrival times directly, we let the required fixed number of interarrival times, n , be the random number observed in that sample, and we use the known mean 1. We find that the Lewis test, without using information about the rate of the PP, is usually superior, but the difference is not great.

To understand the implications of the invalid standard KS test, we also study variants of this base standard KS test. First, we consider alternative standard KS tests with the same random sample size but (i) using the estimated mean and (ii) the associated Lilliefors [23] test for the exponential distribution with unknown mean, using the Matlab program *lillietest*. Second, we consider alternative experiments based on a fixed number n of interarrival times. Then the standard KS test of the PP is valid, provided that the mean is known. In this setting with fixed n , we also consider the consequence of using the estimated mean and the Lilliefors [23] test.

We find that all these variants of the standard KS test perform similarly in our experiments. Thus, we conclude that the (invalid) standard KS test with random sample size and known mean provides a reasonable basis of comparison with the CU, Log, and Lewis tests, which are of primary interest for a NHPP. The invalid standard KS test with estimated mean or the invalid Lilliefors [23] test with random sample size can be considered as alternative (approximate) KS tests to use to test a PP. Nevertheless, the valid Lewis test seems to be superior.

3.2. The Four Base KS Tests

We now specify the four main KS tests that we consider.

3.2.1. Standard Test

We use the standard KS test described above to test whether the random number of observations in the interval $[0, t]$ is consistent with a rate-1 PP. We act as if the observed random number is a fixed number and we use the known mean 1 of an exponential interarrival time in the rate-1 PP.

3.2.2. CU Test

In this test, we exploit the basic conditioning property of a PP. Given an arrival process over an interval $[0, t]$, we observe the number n of arrival in this interval and their arrival times $T_j, 1 \leq j \leq n$. Under the null PP hypothesis, these random variables are distributed as the order statistics of i.i.d. random variables uniformly distributed over $[0, t]$. Thus, the random variables $T_j/t, 1 \leq j \leq n$, are distributed as the order statistics of i.i.d. random variables uniformly distributed over $[0, 1]$. Thus, the ecdf can be computed via

$$F_n(x) \equiv n^{-1} \sum_{k=1}^n 1_{\{T_k/t \leq x\}}, \quad 0 \leq x \leq 1,$$

and the KS statistic can be computed as in (2) with uniform cdf $F(x) = x, 0 \leq x \leq 1$.

3.2.3. Log Test

As observed in [8] (and is shown in the appendix), given the n observed arrival times $\{T_j : 1 \leq j \leq n\}$ during the interval $[0, t]$,

$$X_j^{\text{Log}} \equiv -(n+1-j) \log_e \left(\frac{t-T_j}{t-T_{j-1}} \right), \quad 1 \leq j \leq n,$$

are n i.i.d. mean-1 exponential random variables. The KS test in (2) can then be applied using the exponential cdf $F(x) \equiv 1 - e^{-x}$. A variant of the Log test applies to a fixed sample of size n . With T_j again denoting the time of the j^{th} arrival,

$$X_j^{\text{Log},n} \equiv -j \log_e \left(\frac{T_j}{T_{j+1}} \right), \quad 1 \leq j \leq n-1,$$

are again i.i.d. rate-1 exponential random variables.

3.2.4. Lewis Test

Lewis [22] proposed using a different modification of the CU test, exploiting a transformation due to Durbin [10]. Following [10], we start with a sample $U_j, 1 \leq j \leq n$, hypothesized to be uniformly distributed on $[0, 1]$. Then let $U_{(j)}$ be the j^{th} smallest of these, $1 \leq j \leq n$, so that

$U_{(1)} < \dots < U_{(n)}$. This is applied in [22] with $U_{(j)} = T_j/t$ from the CU test. Next we look at the successive intervals between these ordered observations:

$$C_1 = U_{(1)}, C_j = U_{(j)} - U_{(j-1)} \quad 2 \leq j \leq n, \\ \text{and} \quad C_{n+1} = 1 - U_{(n)}. \quad (3)$$

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

$$Z_j = (n+2-j)(C_{(j)} - C_{(j-1)}), \quad 1 \leq j \leq n+1, \\ (\text{with } C_{(0)} \equiv 0). \quad (4)$$

Remarkably, Durbin [10] showed in a simple direct argument (by giving explicit expressions for the joint density functions, exploiting the transformation of random vectors by a function) that, under the PP null hypothesis, the random vector (Z_1, \dots, Z_n) is distributed the same as the random vector (C_1, \dots, C_n) . Hence, again under the PP null hypothesis, the vector of associated partial sums (S_1, \dots, S_n) , where

$$S_k \equiv Z_1 + \dots + Z_k, \quad 1 \leq k \leq n, \quad (5)$$

has the same distribution as the original random vector $(U_{(1)}, \dots, U_{(n)})$ of ordered uniform random variables. Hence, we can apply the KS test with the ecdf

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

for S_k in (5) and (4), comparing it to the uniform cdf $F(x) \equiv x, 0 \leq x \leq 1$. In [10], it is shown that by doing this transformation starting from a sequence of i.i.d. uniform random variables, we should gain an increase in power. In [22], it is shown this transformation increases power after the CU transformation, which is a different setting than in [10].

4. THE EXPERIMENT

In this section, we specify the stochastic processes to which we apply the different KS tests of a PP. We also indicate how we performed the simulation.

4.1. Study Cases

We consider nine cases, each with one-to-five subcases, yielding a total of 26 cases in all. We specify these cases in terms of the sequence $\{X_n : n \geq 1\}$ of interarrival times, each distributed as a random variable X . In all cases, the sequence is assumed to be stationary with $E[X] = 1$. The first five cases are renewal arrival processes, with i.i.d. interarrival times.

The first i.i.d. case is our PP null hypothesis with exponential interarrival times. The other i.i.d. cases have nonexponential interarrival times. Cases 2 and 3 contain Erlang and hyperexponential interarrival times, which are, respectively, stochastically less variable and stochastically more variable than the exponential distribution in convex stochastic order, as in section 9.5 of [30]. Thus, they have scv, respectively, $c^2 < 1$ and $c^2 > 1$. Cases 4 and 5 contain nonexponential cdf's with $c_X^2 = 1$ as well as $E[X] = 1$, just like the exponential cdf.

Case 1, Exponential: The PP null hypothesis with exponential interarrival times (Base Case).

Case 2, Erlang, E_k : Erlang- k (E_k) interarrival times, 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, H_2 : Hyperexponential-2 (H_2) interarrival times, a mixture of 2 exponential cdf's with $c_X^2 = 1.25, 1.5, 2, 4$, and 10. The cdf is $P(X \leq x) \equiv 1 - p_1 e^{-\lambda_1 x} - p_2 e^{-\lambda_2 x}$. We further assume balanced means ($p_1 \lambda_1^{-1} = p_2 \lambda_2^{-1}$) as in (3.7) of [36] 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 $c_X^2 = 1$: 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, LN(1, 1) with $c_X^2 = 1$: Lognormal distribution with mean and variance both equal to 1, so that $c_X^2 = 1$.

Cases 6 and 7 are stationary point processes that deviate from a PP only through dependence among successive interarrival times, each exponentially distributed with mean 1:

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 [16, 17]). 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 [15]. 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 X_n is exponentially distributed with mean m , while

$$\begin{aligned} P(U_n = 0) &= 1 - P(U_n = 1) = \beta \text{ and} \\ P(V_n = 0) &= 1 - P(V_n = 1) = \rho, \end{aligned} \quad (6)$$

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

$$\begin{aligned} S_n &= \beta X_n + U_n Y_{n-1}, \\ Y_n &= \rho Y_{n-1} + V_n X_n, \quad n \geq 1. \end{aligned} \quad (7)$$

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), (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 [28]. We specify these cases by these cumulative correlations.

The final two cases are stationary point processes that have both nonexponential interarrival times and dependence among successive interarrival times:

Case 8, mH_2 , superposition of m i.i.d. H_2 renewal processes: 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 (H_2), dependent H_2 interarrival times with $c^2 = 4$: 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 above have short-range dependence, whereas Case 8 for large m tends to have nearly exponential interarrival times, but longer-range dependence. (but not strong dependence in the sense of section 4.6 of [38]), and so is more complicated; see section 9.8 and section 9.9.3 of [38] and references therein for more discussion. 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.

4.2. Corresponding Random Rate Models

As noted in section 1.1, we may prefer to think of the test being for a doubly stochastic NHPP, with both time-varying and stochastic rate function. We now want to show that our experiments studying the power of alternative statistical tests are still relevant in such a more general setting. We do so by showing that we can use each of the rate-1 stationary-process study cases in section 4.1 to construct corresponding alternative hypotheses to a doubly stochastic PP with a random rate function. (For more on this construction, see section 7 of [26].)

Let $\{\Lambda(t) : t \geq 0\}$ be an arbitrary cumulative rate stochastic process with nondecreasing sample paths, so that $\Lambda(t)$ is the random total rate over $[0, t]$. Let $\{A(t) : t \geq 0\}$ be any rate-1 arrival counting process with stationary increments, such as one of the ones in section 4.1, which for simplicity we take to be independent of $\{\Lambda(t) : t \geq 0\}$. Then, we obtain an alternative hypothesis to a doubly stochastic PP with a random rate function $\{\Lambda(t) : t \geq 0\}$ by letting

$$N(t) \equiv A(\Lambda(t)), \quad t \geq 0. \quad (8)$$

Since A has rate 1, $E[A(s+t) - A(s)] = t$ for all $s, t > 0$ and

$$\begin{aligned} E[N(s+t) - N(s)|\Lambda(u), \\ 0 \leq u \leq t] &= E[A(\Lambda(s+t)) - A(\Lambda(s))|\Lambda(u), \\ 0 \leq u \leq t] &= \Lambda(s+t) - \Lambda(s), \end{aligned}$$

so that

$$\begin{aligned} E[N(s+t) - N(s)] &= E[E[N(s+t) - N(s)|\Lambda(u), \\ 0 \leq u \leq t]] &= E[\Lambda(s+t) - \Lambda(s)]. \end{aligned}$$

If we can approximate the random rate function by a PC rate function, so that the random rate is constant over each subinterval, then the tests we consider are tests of the Poisson property, that is, to test if N is a doubly stochastic NHPP. The construction here makes all our study cases correspond to legitimate alternative hypotheses.

4.3. Simulation Design

For each arrival process, we simulate 10^4 replications of 10^4 interarrival times. We generate much more data than needed to get rid of any initial effects. We are supposing that we observe stationary point processes, which are achieved

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 arrival data for both time intervals of a fixed length t and sample sizes of a fixed size n . Our main results are for time intervals of fixed length, but we also consider the other scenario. For the first scenario with specified intervals $[0, t]$ with $t = 200$, in each replication we transform the 10^4 interarrival times to 10^4 arrival times starting at $t = 0$ by taking cumulative sums and then consider the arrival process in the interval $[10^3, 10^3 + 200]$. We treat this as observations from a stationary point process over the interval $[0, 200]$.

To observe the effect of longer intervals, we subsequently consider the arrival process in the interval $[10^3, 10^3 + 2000]$; we treat that interval as $[0, 2000]$. To examine the impact of introducing subintervals, we use the same arrival process in the interval $[10^3, 10^3 + 200]$, again treated as $[0, 200]$, and divide it into 10 disjoint contiguous subintervals, each of length 20. In forming subintervals, we necessarily break up the subintervals crossing the boundary points. For these sample sizes, that boundary effect matters; about 5% of the subintervals are altered when the subintervals are of length 20.

In the second scenario with fixed sample size $n = 200$, in each replication of the 10^4 simulated interarrival times we use interarrival times from the 10^3 th interarrival time to the $10^3 + 200$ th interarrival time. We then consider the interarrival times from the 10^3 th interarrival time to the $10^3 + 2000$ th interarrival time to observe the effect of larger sample size. In addition, the interarrival times from the 10^3 th interarrival time to the $10^3 + 200$ th interarrival time are used to examine the impact of introducing ten equally sized subsamples, each with sample size 20, but now no intervals are split up.

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 ecdf's appear in [19].

4.4. Performance Implications of the Alternative Hypotheses

In our experiments, we will see how well the different KS tests can detect the deviation of each of the alternative hypotheses from the PP null hypothesis. To show how different these alternative processes are from a PP from a performance perspective, we now show steady-state performance measures for a $G/M/s + M$ queueing model in which each alternative process serves, as the arrival process. These models have i.i.d. exponential service times, s servers, and customer abandonment with i.i.d. exponential patience times. We let the arrival, service, and abandonment rates be $\lambda = 25$, $\mu = 1$, and $\theta = 1$, respectively. Corresponding results for

Table 1. Simulation estimates of steady-state performance measures in the $G/M/s + M$ model ($\lambda = 25, \mu = 1, \theta = 1$) with different staffing levels based on 100 replications of 10^5 customers (first 10^3 customers removed to avoid initial transient behavior).

s	Arrival	$E[W All]$	$E[W Serve]$	$E[W Abandon]$	%Wait	%Abandon
28 ($\beta = 0.5$ for M)	M	0.0324 ± 0.0003	0.0348 ± 0.0003	0.1004 ± 0.0006	29.96 ± 0.15	3.49 ± 0.03
	$H_2/c^2 = 2$	0.0461 ± 0.0003	0.0496 ± 0.0004	0.1165 ± 0.0006	35.93 ± 0.16	4.96 ± 0.03
	$H_2/c^2 = 4$	0.0680 ± 0.0005	0.0731 ± 0.0005	0.1384 ± 0.0005	43.23 ± 0.18	7.32 ± 0.05
	E_2	0.0241 ± 0.0002	0.0258 ± 0.0002	0.0881 ± 0.0006	25.72 ± 0.16	2.58 ± 0.02
	E_4	0.0197 ± 0.0002	0.0210 ± 0.0002	0.0807 ± 0.0005	23.03 ± 0.16	2.10 ± 0.02
	$RRI(p = 0.5)$	0.0562 ± 0.0004	0.0608 ± 0.0005	0.1316 ± 0.0007	38.79 ± 0.19	6.07 ± 0.05
	$RRI(p = 0.9)$	0.1401 ± 0.0009	0.1552 ± 0.0009	0.2371 ± 0.0009	53.58 ± 0.22	15.52 ± 0.09
	$EARMA(1)$	0.0547 ± 0.0005	0.0590 ± 0.0005	0.1271 ± 0.0006	38.81 ± 0.20	5.91 ± 0.05
	$EARMA(5)$	0.1108 ± 0.0008	0.1218 ± 0.0008	0.2007 ± 0.0008	49.29 ± 0.23	12.19 ± 0.08
30 ($\beta = 1$ for M)	M	0.0168 ± 0.0002	0.0181 ± 0.0002	0.0878 ± 0.0006	18.23 ± 0.13	1.81 ± 0.02
	$H_2/c^2 = 2$	0.0272 ± 0.0003	0.0295 ± 0.0003	0.1031 ± 0.0006	24.75 ± 0.15	2.94 ± 0.03
	$H_2/c^2 = 4$	0.0454 ± 0.0004	0.0492 ± 0.0004	0.1233 ± 0.0006	33.33 ± 0.18	4.93 ± 0.04
	E_2	0.0108 ± 0.0001	0.0115 ± 0.0001	0.0762 ± 0.0007	13.69 ± 0.12	1.15 ± 0.01
	E_4	0.0079 ± 0.0001	0.0084 ± 0.0001	0.0690 ± 0.0006	11.01 ± 0.11	0.84 ± 0.01
	$RRI(p = 0.5)$	0.0360 ± 0.0003	0.0392 ± 0.0004	0.1181 ± 0.0007	28.47 ± 0.17	3.91 ± 0.04
	$RRI(p = 0.9)$	0.1151 ± 0.0008	0.1290 ± 0.0009	0.2233 ± 0.0009	47.74 ± 0.22	12.90 ± 0.09
	$EARMA(1)$	0.0345 ± 0.0003	0.0374 ± 0.0004	0.1132 ± 0.0007	28.29 ± 0.19	3.75 ± 0.04
	$EARMA(5)$	0.0869 ± 0.0007	0.0964 ± 0.0007	0.1860 ± 0.0008	42.51 ± 0.22	9.66 ± 0.07
35 ($\beta = 2$ for M)	M	0.0022 ± 0.0001	0.0024 ± 0.0001	0.0647 ± 0.0012	3.41 ± 0.06	0.24 ± 0.01
	$H_2/c^2 = 2$	0.0057 ± 0.0001	0.0061 ± 0.0001	0.0776 ± 0.0009	7.11 ± 0.09	0.61 ± 0.01
	$H_2/c^2 = 4$	0.0140 ± 0.0002	0.0152 ± 0.0002	0.0939 ± 0.0008	14.11 ± 0.13	1.53 ± 0.02
	E_2	0.0008 ± 0.0000	0.0009 ± 0.0000	0.0562 ± 0.0014	1.50 ± 0.04	0.09 ± 0.00
	E_4	0.0004 ± 0.0000	0.0004 ± 0.0000	0.0477 ± 0.0018	0.80 ± 0.02	0.04 ± 0.00
	$RRI(p = 0.5)$	0.0101 ± 0.0002	0.0110 ± 0.0002	0.0924 ± 0.0008	10.61 ± 0.12	1.10 ± 0.02
	$RRI(p = 0.9)$	0.0696 ± 0.0006	0.0796 ± 0.0007	0.1949 ± 0.0010	34.50 ± 0.21	7.96 ± 0.07
	$EARMA(1)$	0.0089 ± 0.0001	0.0097 ± 0.0002	0.0869 ± 0.0008	9.96 ± 0.12	0.96 ± 0.02
	$EARMA(5)$	0.0458 ± 0.0005	0.0515 ± 0.0005	0.1558 ± 0.0008	27.70 ± 0.21	5.15 ± 0.05

The 95% confidence intervals are shown.

models with arrival rate $\lambda = 10$ and without customer abandonment are given in the appendix [19]. These simulation estimates are based on 100 replications of 10^5 customers, with the first 10^3 removed to allow the system to approach steady state.

Table 1 shows five performance measures for three levels of staffing $s = 28, 30,$ and 35 . The five performance measures are: the expected waiting time of (i) all customers, (ii) served customers, and (iii) abandoning customers, and the percentage of customers that (i) must wait before entering service and (ii) who abandon. For example, we see that the abandonment probability of 0.0515 for an EARMA (5) arrival process with staffing $s = 35$ and 0.0493 for an H_2 arrival process with $c^2 = 4$ and staffing $s = 30$ are higher than the abandonment probability of 0.0349 for a PP arrival process with staffing $s = 28$. Hence, those alternative arrival processes require 7 and 2 extra servers than for a PP to achieve comparable performance, respectively.

5. RESULTS

We now present the results of applying the four alternative KS tests in section 3 to the original and transformed arrival

process data in all 26 study cases specified in section 4, with the goal of comparing the power of the alternative tests of a PP. We assume that an initial NHPP has been regarded as a PC NHPP, so that we are looking at a single subinterval, yielding a PP. Of course, a PP also could be considered directly, as in [22].

5.1. The Four Alternatives for the Interval [0, 200]

We report the number of KS tests passed out of 10,000 replications as well as the average P -value with associated 95% confidence intervals. Thus, the estimated power is $1 - (\text{number passed}/10,000)$. The P -value is the significance level below which the hypothesis would be rejected. Thus, low P -values indicate greater power.

The difference is most striking for the middle H_2 alternative with $c^2 = 2$, as shown in Table 2. For this case, the Lewis test has far greater power than all the alternatives.

The results for all 26 cases are given in Table 3. The first “exponential” case is the PP null hypothesis. The results show that all tests behave properly for the PP null hypothesis. The results also show that the tests perform quite differently for the non-PP alternative hypotheses. Table 3 shows that the

Table 2. Summary of the performance of alternative KS tests of a rate-1 Poisson process for the time interval $[0, 200]$ with significance level $\alpha = 0.05$: the case of a renewal process with H_2 interarrival times having $c_x^2 = 2$.

KS test	Lewis	Standard	Log	CU
Power	0.94	0.63	0.51	0.28
Average P -value	0.01	0.10	0.13	0.23

standard, Log, and Lewis tests all perform reasonably well for the cases with nonexponential interarrival-time cdf's, in marked contrast to the CU test. Table 3 also shows that the Lewis test is consistently most powerful for these cases. For H_2 cdf's with lower scv, the power of all methods is less than in Table 2, but the ordering remains; for H_2 cdf's with higher scv, the power of all methods is greater, but the ordering remains. This same conclusion also applies to Case 9 with the $RRI(H_2)$ process.

However, Table 3 tells very different story for Cases 6 and 7 with dependent exponential interarrival times. For the most part, none of the KS tests has much power, except in the cases with extreme dependence, for example, RRI with $P = 0.9$

and EARMA with the larger cumulative correlations (3.0 and 5.25). For RRI, the Lewis and standard tests are somewhat better than the others; for the EARMA subcases 3.0 and 5.25, the CU test is best. For those two cases, Table 3 shows the power ordering of $CU > Standard > Log > Lewis$, which is nearly the opposite of the order for the renewal processes.

The two cases RRI and $RRI(H_2)$ are somewhat different from the others because the dependence involves successive identical exponential interarrival times. Since the successive interarrival times that are copies are exactly the same length as the others, they will be kept together after the Durbin reordering. Thus, the effectiveness of the Lewis KS test should not be lost.

We see that the advantage of the CU KS test lies in the other cases, EARMA and mH_2 with higher m . Moreover, the performance in Table 3 seems to be consistent with this observation.

The lack of uniformity suggests composite statistical tests, which we consider in the appendix [19]. In retrospect, the lack of uniformity should perhaps be expected, because the data transformations emphasize different features of the underlying model. Nevertheless, the differences seem to concentrate on the two dimensions (2D): (i) whether the interarrival-time

Table 3. Performance of alternative KS tests of a rate-1 Poisson process for the time interval $[0, 200]$: 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.

Case	Subcase	Standard		CU		Log		Lewis	
		# P	$E[P\text{-value}]$	# P	$E[P\text{-value}]$	# P	$E[P\text{-value}]$	# P	$E[P\text{-value}]$
Exp	—	9479	0.50 ± 0.006	9503	0.50 ± 0.006	9490	0.50 ± 0.006	9502	0.50 ± 0.006
E_k	$k = 2$	11	0.00 ± 0.000	9991	0.78 ± 0.005	24	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 4$	0	0.00 ± 0.000	10000	0.94 ± 0.002	0	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 6$	0	0.00 ± 0.000	10000	0.98 ± 0.001	0	0.00 ± 0.000	0	0.00 ± 0.000
H_2	$c^2 = 1.25$	8757	0.41 ± 0.006	8988	0.41 ± 0.006	8685	0.40 ± 0.006	7558	0.30 ± 0.006
	$c^2 = 1.5$	7049	0.27 ± 0.005	8428	0.33 ± 0.005	7081	0.28 ± 0.006	3990	0.11 ± 0.004
	$c^2 = 2$	3695	0.10 ± 0.003	7186	0.23 ± 0.005	4059	0.13 ± 0.004	584	0.01 ± 0.001
	$c^2 = 4$	231	0.01 ± 0.001	3551	0.08 ± 0.003	934	0.03 ± 0.002	3	0.00 ± 0.000
	$c^2 = 10$	13	0.00 ± 0.000	650	0.01 ± 0.001	608	0.02 ± 0.002	2	0.00 ± 0.000
Z	—	1200	0.02 ± 0.001	9412	0.57 ± 0.006	1308	0.02 ± 0.001	243	0.01 ± 0.000
LN	—	88	0.01 ± 0.000	9525	0.52 ± 0.006	332	0.01 ± 0.000	69	0.00 ± 0.000
RRI	$p = 0.1$	9093	0.41 ± 0.006	9044	0.42 ± 0.006	9037	0.42 ± 0.006	9080	0.41 ± 0.006
	$p = 0.5$	4631	0.11 ± 0.003	5516	0.15 ± 0.004	5204	0.13 ± 0.003	4633	0.11 ± 0.003
	$p = 0.9$	14	0.00 ± 0.000	826	0.02 ± 0.002	95	0.00 ± 0.000	12	0.00 ± 0.000
EARMA	0.25	9260	0.47 ± 0.006	8536	0.36 ± 0.005	9236	0.46 ± 0.006	9477	0.50 ± 0.006
	0.5	8848	0.42 ± 0.006	7433	0.26 ± 0.005	8870	0.43 ± 0.006	9406	0.49 ± 0.006
	1	8244	0.37 ± 0.006	5964	0.18 ± 0.004	8280	0.37 ± 0.006	8994	0.44 ± 0.006
	3	5196	0.21 ± 0.005	1977	0.04 ± 0.002	5725	0.22 ± 0.005	6736	0.29 ± 0.006
	5.25	4088	0.14 ± 0.004	1594	0.04 ± 0.002	4512	0.15 ± 0.004	5770	0.22 ± 0.005
mH_2	$m = 2$	4531	0.16 ± 0.005	4231	0.10 ± 0.003	4901	0.18 ± 0.005	1237	0.03 ± 0.002
	$m = 5$	7474	0.32 ± 0.006	5282	0.16 ± 0.004	7706	0.34 ± 0.006	7142	0.29 ± 0.006
	$m = 10$	7812	0.35 ± 0.006	6496	0.23 ± 0.005	8334	0.38 ± 0.006	8980	0.44 ± 0.006
	$m = 20$	8034	0.36 ± 0.006	7751	0.32 ± 0.006	8831	0.43 ± 0.006	9366	0.49 ± 0.006
$RRI(H_2)$	$p = 0.1$	322	0.01 ± 0.001	2841	0.06 ± 0.002	1012	0.03 ± 0.002	4	0.00 ± 0.000
	$p = 0.5$	478	0.01 ± 0.001	884	0.02 ± 0.001	1209	0.03 ± 0.002	69	0.00 ± 0.000
	$p = 0.9$	9	0.00 ± 0.000	772	0.06 ± 0.005	92	0.00 ± 0.000	6	0.00 ± 0.000

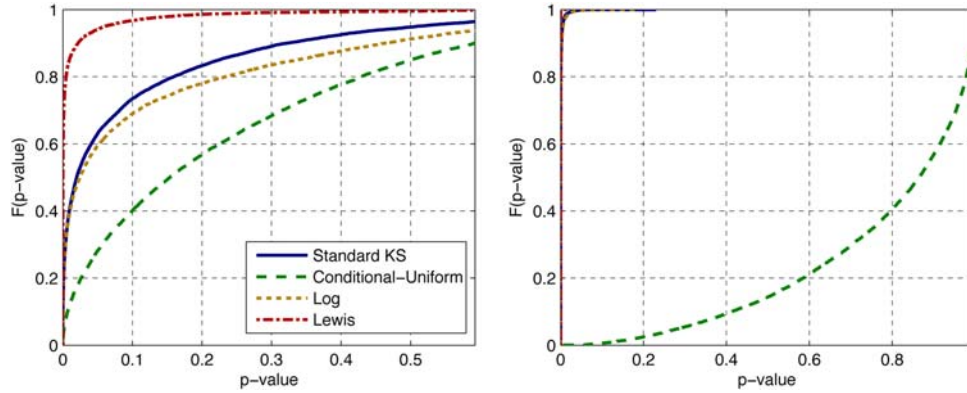


Figure 3. Empirical distribution function of the P -values - H_2 ($c^2 = 2$) (left) and E_2 (right) Based on 10^4 replications for the time interval $[0, 200]$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

distribution is nearly exponential, and (ii) whether nearly exponential interarrival times are nearly independent. For the first issue, the Lewis test is the clear winner; for the second issue, none of the tests is especially effective, but the CU test evidently does capture forms of dependence that are lost by reordering better than the alternative tests.

We also display the analog of Table 3 for the AD test of a PP for these same alternative hypotheses in the appendix [19]. (Both start with the CU transformations and thus the same ecdf.) The results for the AD test are very similar to the results for the KS test. In fact, the power of the AD test against non-exponential interarrival times tends to be slightly higher than that of the KS test in Table 3 for all except the powerless CU test. The power of the AD log test is almost equivalent to the power of the AD Lewis test for these cases. On the other hand, the power of the CU AD test of a PP against dependent exponential interarrival times is slightly lower than for the CU KS test.

5.2. Insightful Plots

In section 2.3, we displayed plots comparing the average of the ecdf's over many independent replications to the cdf associated with the null hypothesis (which depends on the transformation). Figures 1 and 2 illustrate for the renewal process cases with interarrival times having cdf's H_2 with $c^2 = 2$ and E_2 over the time interval $[0, 200]$. These figures show that the transformation in the Lewis KS test provides much greater separation between the average ecdf and the cdf. Indeed, for the Lewis test, the ecdf and cdf appear to be stochastically ordered, whereas the ecdf and cdf cross for the other KS tests. From these plots, it is evident that the CU test should perform poorly. For any new contemplated alternative, we suggest conducting simulations and comparing the plots.

These plots lead to a conjecture about the consequence of convex stochastic order of the underlying cdf's, denoted by \leq_c ; let \leq_{st} denote ordinary stochastic order, as in section 9.1

of [30]. Let $E[F_n]$ be the expected value of the ecdf, which coincides with the limiting average of the sample ecdf's. Since we have convex stochastic order ($E_k \leq_c Exp \leq_c H_k$) among random variables with a given mean, we are led to make the following conjecture:

CONJECTURE 5.1: Consider a renewal process with interarrival times distributed as X and let Exp be an exponential random variable. If $X \leq_c (\geq_c) Exp$, then $E[F_n] \geq_{st} (\leq_{st}) F$ for the Lewis test.

Verifying the conjecture seems challenging in view of the complicated construction of the Lewis test.

Also insightful are plots of the empirical distribution function of the P -values. These plots show that the relative power tends to remain across all P -values, not just for our type I error of $\alpha = 0.05$. Paralleling the plots above, we display the plots for H_2 with $c^2 = 2$ and E_2 in Fig. 3. The advantage of the Lewis KS test for the H_2 ($c^2 = 2$) alternative is even more important in this figure.

5.3. Alternative Standard KS Tests

The standard KS test used to obtain the values reported in Table 3 is actually invalid because the sample size was not specified in advance, but instead was the random number observed in $[0, t]$. Moreover, it used the true Mean 1, which would not be known in application. Hence, to put the results for the standard test Table 3 in perspective, we now give the alternative results for the valid standard KS test with fixed sample size $n = 200$ in the first columns of Table 4. For the case of a specified interval $[0, t]$, we also used the known mean in the standard KS test, which would not in fact be known. Hence, in the later columns of Table 4 we also give the corresponding standard KS results using the estimated mean and the Lilliefors [23] test for the exponential cdf with unknown mean.

Table 4. Performance of alternative standard KS tests for untransformed interarrival times: number of K-S tests passed (denoted by #P) at significance level 0.05 out of 10,000 replications and average P -values (denoted by $E[P\text{-value}]$) with associated 95% confidence intervals.

Case	Subcase	Fixed number $n = 200$		Fixed interval $[0, t]$ for $t = 200$			
		KS with known mean		KS with estimated mean		[23] test	
		#P	$E[P\text{-value}]$	#P	$E[P\text{-value}]$	#P	$E[P\text{-value}]$
Exp	–	9487	0.50 ± 0.006	9942	0.65 ± 0.005	9482	0.37 ± 0.003
E_k	$k = 2$	28	0.00 ± 0.000	8	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 4$	0	0.00 ± 0.000	0	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 6$	0	0.00 ± 0.000	0	0.00 ± 0.000	0	0.00 ± 0.000
H_2	$c^2 = 1.25$	8843	0.42 ± 0.006	9564	0.50 ± 0.006	8239	0.28 ± 0.004
	$c^2 = 1.5$	7204	0.27 ± 0.005	8063	0.29 ± 0.005	5379	0.14 ± 0.003
	$c^2 = 2$	3603	0.09 ± 0.003	3752	0.08 ± 0.003	1398	0.03 ± 0.002
	$c^2 = 4$	90	0.00 ± 0.000	157	0.00 ± 0.000	23	0.00 ± 0.000
	$c^2 = 10$	0	0.00 ± 0.000	76	0.00 ± 0.000	26	0.00 ± 0.000
Z	–	1200	0.02 ± 0.001	1230	0.02 ± 0.001	210	0.01 ± 0.000
LN	–	98	0.00 ± 0.000	281	0.01 ± 0.000	9	0.00 ± 0.000
RRI	$p = 0.1$	9048	0.41 ± 0.006	9845	0.56 ± 0.005	8971	0.32 ± 0.003
	$p = 0.5$	4659	0.11 ± 0.003	6873	0.18 ± 0.004	3775	0.08 ± 0.002
	$p = 0.9$	16	0.00 ± 0.000	44	0.00 ± 0.000	2	0.00 ± 0.000
EARMA	0.25	9284	0.47 ± 0.006	9942	0.65 ± 0.005	9460	0.37 ± 0.003
	0.5	8865	0.43 ± 0.006	9925	0.65 ± 0.005	9420	0.37 ± 0.003
	1	8178	0.37 ± 0.005	9845	0.60 ± 0.005	9102	0.34 ± 0.003
	3	5209	0.21 ± 0.005	8341	0.43 ± 0.006	6702	0.23 ± 0.004
	5.25	4100	0.14 ± 0.004	8015	0.35 ± 0.006	5987	0.19 ± 0.004
mH_2	$m = 2$	4398	0.14 ± 0.004	4949	0.14 ± 0.004	2440	0.06 ± 0.002
	$m = 5$	7514	0.32 ± 0.006	9414	0.49 ± 0.006	7944	0.27 ± 0.004
	$m = 10$	7818	0.35 ± 0.006	9874	0.61 ± 0.005	9177	0.35 ± 0.003
	$m = 20$	7996	0.37 ± 0.006	9935	0.65 ± 0.005	9421	0.37 ± 0.003
$RRI(H_2)$	$p = 0.1$	104	0.00 ± 0.000	202	0.00 ± 0.000	41	0.00 ± 0.000
	$p = 0.5$	253	0.00 ± 0.001	311	0.01 ± 0.001	94	0.00 ± 0.000
	$p = 0.9$	4	0.00 ± 0.000	36	0.00 ± 0.000	1	0.00 ± 0.000

Notice that the simulation results in Table 4 for the valid KS test with known mean (which is unlikely to be so useful) and the valid Lilliefors [23] test are consistent with the fixed significance level of $\alpha = 0.05$, whereas the invalid standard KS test with estimated mean is not, because the null hypothesis is not rejected in over 99% of the cases. Overall, we see that using the estimated mean in this setting tends to make rejection less likely. To obtain a valid test of this form, we can conduct additional simulation experiments with increased α until the null hypothesis is not rejected in only 95% of the cases, but that is just what has been done by Lilliefors [23]. Additional simulation experiments (in the appendix) show that a realized significance level of $\alpha = 0.05$ is achieved when we estimate the mean if we increase the nominal significance level to $\alpha = 0.18$; then the results are very close to the results for the Lilliefors [23] test in Table 4.

Since these performance results do not differ greatly, one might consider the Lilliefors [23] test as a realizable alternative to the standard KS test. However, this is evidently inferior to the Lewis KS test in the cases with nonexponential interarrival times.

5.4. A Longer Time Interval

Table 3 clearly shows how the power decreases as the alternative gets closer to the PP null hypothesis. For the renewal processes, 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 interval $[0, 200]$ with expected sample size of 200. The power also increases as we increase the sample size, as we now illustrate by considering the interval $[0, 2000]$. In Table 5, we now see a simple summary story similar to Table 2 when the H_2 scv is reduced to $c_X^2 = 1.25$.

Table 5. Summary of the performance of alternative KS tests of a rate-1 Poisson process for the longer time interval $[0, 2000]$ with significance level $\alpha = 0.05$: the case of a renewal process with H_2 interarrival times having $c_X^2 = 1.25$.

KS test	Lewis	Standard	Log	CU
Power	0.974	0.660	0.637	0.105
Average P -value	0.01	0.08	0.10	0.40

Table 6. Performance of alternative KS tests of a rate-1 Poisson process for the longer time interval [0, 2000]: 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.

Case	Subcase	Standard		CU		Log		Lewis	
		#P	E[P-value]	#P	E[P-value]	#P	E[P-value]	#P	E[P-value]
Exp	—	9525	0.50 ± 0.006	9491	0.50 ± 0.006	9513	0.50 ± 0.006	9476	0.50 ± 0.006
E_k	$k = 2$	0	0.00 ± 0.000	9983	0.79 ± 0.004	0	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 4$	0	0.00 ± 0.000	10000	0.95 ± 0.002	0	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 6$	0	0.00 ± 0.000	10000	0.98 ± 0.001	0	0.00 ± 0.000	0	0.00 ± 0.000
	$c^2 = 1.25$	3402	0.08 ± 0.003	8945	0.40 ± 0.005	3627	0.10 ± 0.003	263	0.01 ± 0.001
H_2	$c^2 = 1.5$	86	0.00 ± 0.000	8311	0.32 ± 0.005	200	0.00 ± 0.000	0	0.00 ± 0.000
	$c^2 = 2$	0	0.00 ± 0.000	6878	0.21 ± 0.004	0	0.00 ± 0.000	0	0.00 ± 0.000
	$c^2 = 4$	0	0.00 ± 0.000	2804	0.06 ± 0.002	0	0.00 ± 0.000	0	0.00 ± 0.000
	$c^2 = 10$	0	0.00 ± 0.000	174	0.00 ± 0.000	0	0.00 ± 0.000	0	0.00 ± 0.000
	Z	—	0	0.00 ± 0.000	9491	0.52 ± 0.006	0	0.00 ± 0.000	0
LN	—	0	0.00 ± 0.000	9519	0.51 ± 0.006	0	0.00 ± 0.000	0	0.00 ± 0.000
RRI	$p = 0.1$	9010	0.41 ± 0.005	9119	0.41 ± 0.005	9027	0.41 ± 0.005	9009	0.40 ± 0.005
	$p = 0.5$	4449	0.10 ± 0.003	4667	0.11 ± 0.003	4693	0.11 ± 0.003	4502	0.10 ± 0.003
	$p = 0.9$	0	0.00 ± 0.000	21	0.00 ± 0.000	6	0.00 ± 0.000	1	0.00 ± 0.000
EARMA	0.25	9338	0.47 ± 0.006	8318	0.33 ± 0.005	9152	0.46 ± 0.006	9429	0.49 ± 0.006
	0.5	8808	0.42 ± 0.006	7029	0.22 ± 0.004	8870	0.42 ± 0.006	9392	0.49 ± 0.006
	1	8182	0.37 ± 0.006	4749	0.12 ± 0.003	8017	0.36 ± 0.006	8894	0.43 ± 0.006
	3	5233	0.20 ± 0.005	805	0.01 ± 0.001	5465	0.21 ± 0.005	6660	0.29 ± 0.006
	5.25	4111	0.14 ± 0.004	179	0.00 ± 0.000	4308	0.14 ± 0.004	5783	0.21 ± 0.005
mH_2	$m = 2$	2	0.00 ± 0.000	3001	0.06 ± 0.002	21	0.00 ± 0.000	0	0.00 ± 0.000
	$m = 5$	3217	0.09 ± 0.003	3376	0.07 ± 0.002	3403	0.10 ± 0.004	139	0.00 ± 0.000
	$m = 10$	6413	0.25 ± 0.005	3752	0.09 ± 0.003	6571	0.26 ± 0.005	4406	0.13 ± 0.004
	$m = 20$	7360	0.31 ± 0.006	4343	0.11 ± 0.003	7655	0.33 ± 0.006	8147	0.35 ± 0.006
$RRI(H_2)$	$p = 0.1$	0	0.00 ± 0.000	1960	0.04 ± 0.001	0	0.00 ± 0.000	0	0.00 ± 0.000
	$p = 0.5$	0	0.00 ± 0.000	154	0.00 ± 0.000	0	0.00 ± 0.000	0	0.00 ± 0.000
	$p = 0.9$	0	0.00 ± 0.000	1	0.00 ± 0.000	0	0.00 ± 0.000	0	0.00 ± 0.000

All the results for the longer interval [0, 2000] are shown in Table 6. Now, for many cases the three main candidates - standard, Log, and Lewis—reject the alternatives for all 10,000 replications and the estimated P-value is 0.00.

5.5. Subintervals as Occur in a PC NHPP

When we apply these KS tests to a NHPP, we regard it as PC and combine data over subintervals where the rate is constant. Even though we can combine all the data, this process causes the impact of longer-time behavior (occurring over multiple subintervals) to be lost. To gain insight into the consequence of having to combine data over subintervals, we repeat the test for the time interval [0, 200] after dividing that interval into 10 contiguous disjoint intervals of length 20. Table 7 summarizes the results of that experiment.

Comparing Table 7 to Table 3, we see that the use of subintervals degrades the power for the cases with nonexponential interarrival times somewhat for all tests, with the possible exception of the standard test. Consistent with intuition, we see a clear impact for the more highly variable H_2 interarrival times, but little impact for the less variable Erlang

interarrival times. For the CU test, it loses any advantage it had for the dependent cases after using subintervals; now the CU test is uniformly dominated by the Lewis test, even for the dependent cases. Curiously, we see that the power of the Lewis tests improves dramatically, whereas the power of the CU test degrades dramatically when we work with the subintervals in the EARMA(5.25) case with highest cumulative correlations. Overall, the power of the Lewis test and standard KS test are similar in this scenario, but recall that the standard test is invalid and exploits the true mean, even though it would not be known. Table 8 below shows that the relatively good power in the standard KS test is lost if we use the estimated mean or the Lilliefors [23] test. Hence, we conclude that the Lewis test dominates the others. (Also note that, just as in Table 4, in Table 8 the invalid standard KS test with estimated mean fails to reject the null hypothesis too often, in over 99% of the cases.)

5.6. Fixed Sample Sizes

To place the results above in perspective, we also consider experiments with a fixed sample size of $n = 200$ in Tables 9

Table 7. Performance of alternative KS tests for the time interval $[0, 200]$ when we combine the results from 10 disjoint subintervals of length 20: 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.

Case	Subcase	Standard		CU		Log		Lewis	
		$\#P$	$E[P\text{-value}]$	$\#P$	$E[P\text{-value}]$	$\#P$	$E[P\text{-value}]$	$\#P$	$E[P\text{-value}]$
Exp	—	9206	0.45 ± 0.006	9544	0.50 ± 0.006	9533	0.50 ± 0.006	9495	0.50 ± 0.006
E_k	$k = 2$	87	0.00 ± 0.000	9979	0.74 ± 0.005	79	0.00 ± 0.000	2	0.00 ± 0.000
	$k = 4$	0	0.00 ± 0.000	10000	0.90 ± 0.003	0	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 6$	0	0.00 ± 0.000	10000	0.95 ± 0.002	0	0.00 ± 0.000	0	0.00 ± 0.000
H_2	$c^2 = 1.25$	7687	0.32 ± 0.006	9147	0.45 ± 0.006	8993	0.43 ± 0.006	8160	0.35 ± 0.006
	$c^2 = 1.5$	5546	0.18 ± 0.005	8848	0.40 ± 0.006	7848	0.33 ± 0.006	5459	0.18 ± 0.005
	$c^2 = 2$	2441	0.06 ± 0.003	8214	0.34 ± 0.006	5537	0.19 ± 0.005	1783	0.04 ± 0.002
	$c^2 = 4$	118	0.00 ± 0.000	6790	0.25 ± 0.005	2178	0.06 ± 0.003	138	0.00 ± 0.000
	$c^2 = 10$	11	0.00 ± 0.000	6323	0.23 ± 0.005	2177	0.06 ± 0.003	260	0.01 ± 0.001
Z	—	2274	0.04 ± 0.001	9882	0.64 ± 0.005	2014	0.04 ± 0.002	499	0.01 ± 0.001
LN	—	328	0.01 ± 0.000	9730	0.59 ± 0.006	1057	0.02 ± 0.001	175	0.00 ± 0.000
RRI	$p = 0.1$	8768	0.38 ± 0.006	9311	0.46 ± 0.006	9232	0.44 ± 0.006	9054	0.42 ± 0.006
	$p = 0.5$	4824	0.12 ± 0.003	7992	0.33 ± 0.006	6076	0.18 ± 0.004	3199	0.07 ± 0.002
	$p = 0.9$	21	0.00 ± 0.000	5890	0.25 ± 0.006	93	0.00 ± 0.000	1	0.00 ± 0.000
EARMA	0.25	9013	0.44 ± 0.006	9314	0.48 ± 0.006	9453	0.49 ± 0.006	9487	0.50 ± 0.006
	0.5	8589	0.41 ± 0.006	8923	0.44 ± 0.006	9295	0.47 ± 0.006	9244	0.47 ± 0.006
	1	8084	0.36 ± 0.006	9238	0.48 ± 0.006	8926	0.42 ± 0.006	8117	0.34 ± 0.006
	3	5227	0.21 ± 0.005	5869	0.20 ± 0.005	4729	0.14 ± 0.004	4556	0.17 ± 0.005
	5.25	4345	0.15 ± 0.005	9154	0.49 ± 0.006	2562	0.06 ± 0.002	352	0.01 ± 0.001
mH_2	$m = 2$	3401	0.11 ± 0.004	8249	0.35 ± 0.006	7250	0.29 ± 0.006	4611	0.15 ± 0.004
	$m = 5$	6747	0.28 ± 0.006	9291	0.46 ± 0.006	9336	0.48 ± 0.006	9182	0.46 ± 0.006
	$m = 10$	7394	0.32 ± 0.006	9465	0.49 ± 0.006	9462	0.50 ± 0.006	9453	0.49 ± 0.006
	$m = 20$	7774	0.34 ± 0.006	9459	0.49 ± 0.006	9505	0.50 ± 0.006	9491	0.50 ± 0.006
$RRI(H_2)$	$p = 0.1$	191	0.00 ± 0.000	6712	0.24 ± 0.005	2500	0.07 ± 0.003	267	0.01 ± 0.001
	$p = 0.5$	435	0.01 ± 0.001	5764	0.19 ± 0.005	4070	0.11 ± 0.004	2151	0.05 ± 0.002
	$p = 0.9$	15	0.00 ± 0.000	4431	0.19 ± 0.006	849	0.04 ± 0.003	497	0.03 ± 0.003

and 10. These experiments take us out of the original setting of a PC NHPP, because the sample sizes then are necessarily random over each subinterval. However, the standard KS test is valid in this setting. The CU, Log, and Lewis tests still apply because, under the PP null hypothesis, the variables T_k/T_n , $1 \leq k \leq n$, obtained by dividing the arrival times (sum of the interarrival times) by the largest arrival time are again distributed as the order statistics of i.i.d. random variables uniformly distributed on $[0, 1]$. These results are consistent with the results for fixed interval $[0, t]$ with $t = 200$, again supporting the Lewis test. And again the invalid standard KS test with estimated mean fails to reject the null hypothesis too often, in over 99% of the cases.

6. A DEPARTURE PROCESS EXAMPLE

In this section, we consider a new example of a departure process from a many-server queue, which is interesting because it may then become an arrival process to another subsequent queue. In particular, we consider the departure process from an $M_t/H_2/s_t + M$ model with an NHPP arrival process (the M_t), i.i.d. service times with the H_2 distribution

having mean $\mu^{-1} = 1$, $c^2 = 4$ and balanced means, a time-varying staffing function (the s_t) and customer abandonment with i.i.d. exponential patience times having mean $\theta^{-1} = 2$ (the $+M$). The staffing function s_t is set to stabilize the abandonment probability at a target α using the DIS-MOL approximation in [24]. We consider a sinusoidal arrival rate function of the form $\lambda(t) = 100(1 + r \sin t)$ for three relative amplitudes r : $r = 0$ (PP), $r = 0.2$, and $r = 0.6$ (large fluctuations). Since the mean service time is 1 and the average arrival rate is 100, the staffing fluctuates around 100. The approximate departure rate function for this model is $(1 - \alpha)E[\lambda(t - w - S)]$, as shown in (12) of [24]; it has been found to be very accurate; for example, see the appendix of [25].

The departure process for this model is relatively well-understood in the two extreme cases: (i) light load, when α is low, and (ii) heavy load, when α is high. When α is low, the servers are only rarely all busy, so that the departure process is similar to the departure process from the associated $M_t/H_2/\infty$ model, which is known to be exactly NHPP; see Theorem 1 of [11]; when α is high, all servers tend to be busy most of the time, so that the departure process is similar to the superposition of s_t i.i.d. renewal processes with the

Table 8. Performance of alternative tests for untransformed interarrival times on [0, 200] with 10 equally sized subintervals of length 20: 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.

Case	Subcase	KS with estimated mean		Lilliefors test	
		#P	E[P-value]	#P	E[P-value]
Exp	—	9949	0.65 ± 0.005	9493	0.37 ± 0.003
E _k	k = 2	20	0.00 ± 0.000	1	0.00 ± 0.000
	k = 4	0	0.00 ± 0.000	0	0.00 ± 0.000
	k = 6	0	0.00 ± 0.000	0	0.00 ± 0.000
H ₂	c ² = 1.25	9681	0.54 ± 0.006	8550	0.30 ± 0.004
	c ² = 1.5	8601	0.35 ± 0.005	6257	0.18 ± 0.004
	c ² = 2	5290	0.13 ± 0.004	2397	0.05 ± 0.002
	c ² = 4	908	0.02 ± 0.001	183	0.01 ± 0.001
	c ² = 10	971	0.02 ± 0.001	324	0.01 ± 0.001
Z	—	1276	0.02 ± 0.001	233	0.01 ± 0.000
LN	—	297	0.01 ± 0.000	15	0.00 ± 0.000
RRI	p = 0.1	9847	0.57 ± 0.005	9031	0.32 ± 0.003
	p = 0.5	7235	0.20 ± 0.004	4224	0.09 ± 0.002
	p = 0.9	69	0.00 ± 0.000	7	0.00 ± 0.000
EARMA	0.25	9949	0.66 ± 0.005	9485	0.38 ± 0.003
	0.5	9930	0.65 ± 0.005	9439	0.37 ± 0.003
	1	9878	0.61 ± 0.005	9142	0.35 ± 0.003
	3	8432	0.44 ± 0.006	6795	0.24 ± 0.004
	5.25	8324	0.37 ± 0.006	6294	0.20 ± 0.004
mH ₂	m = 2	6883	0.22 ± 0.005	4023	0.10 ± 0.003
	m = 5	9645	0.53 ± 0.006	8462	0.30 ± 0.004
	m = 10	9905	0.63 ± 0.005	9284	0.36 ± 0.003
	m = 20	9940	0.65 ± 0.005	9446	0.37 ± 0.003
RRI(H ₂)	p = 0.1	873	0.02 ± 0.001	220	0.01 ± 0.001
	p = 0.5	715	0.01 ± 0.001	200	0.00 ± 0.001
	p = 0.9	503	0.02 ± 0.002	314	0.01 ± 0.001

H₂ service times as the interrenewal times. Thus, the high-α case with r = 0 is similar to our m-H₂ superposition process experiment, but here with even larger m, namely m = 100, for which it is known that the interarrival times are approximately exponential and the deviation from a PP is due to the cumulative impact of many small correlations; see [12, 34] and references therein. In [25], it is also shown that the histogram of the interarrival times of the departure process for r = 0 matches an exponential pdf perfectly.

This example is interesting for us because the process we are looking at locally looks like a PP, but it has two-forms of dependence: (i) stochastic dependence from the many small correlations over many interdeparture times and (ii) time dependence due to the sinusoidal arrival rate function when r > 0.

Table 11 shows the results of the CU and Lewis KS tests of an NHPP applied to the departure process data over [6, 20] obtained from independent replications of a simulation of the M_t/H₂/s_t + M model starting out empty for the two targets: α = 0.3 (heavily loaded) and α = 0.01 (lightly loaded) for three forms of evenly divided subintervals: L = 0.5, L = 2, and L = 14 (no subintervals). For small (large) L, the impact of both the stochastic and time dependence is reduced

(increased). As in previous tables, Table 11 shows the average P-value and the number of the 50 KS tests that fails to reject the PP null hypothesis. Even though there are random fluctuations, Table 11 strongly confirms our conclusions about stochastic and time dependence. Both forms of dependence increase as L increases and as r increases. We see the impact of stochastic dependence alone for the cases with r = 0. We see that this dependence is only detected by the CU KS test and then only when L = 14.

On the other hand, when the short subintervals of length L = 0.5 are used, it is evident that these tests conclude that both forms of dependence are eliminated completely; the results are essentially the same as for a PP. As L and r increase, these KS tests begin to detect the two forms of dependence, with the CU KS test doing so sooner and more decisively. For the extreme case of r = 0.6 and L = 14, both KS tests decisively reject the PP null hypothesis. For r = 0.2 and L = 14, the CU KS test is evidently substantially more powerful than the Lewis test. The overall power is not great here, but it would be more decisive with a much larger sample size.

The impact of this departure process on the performance of a subsequent queue for which it is the arrival process is

Table 9. Performance of the alternative KS tests for a fixed sample size $n = 200$: 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.

Case	Subcase	Standard		CU		Log		Lewis	
		# P	$E[P\text{-value}]$	# P	$E[P\text{-value}]$	# P	$E[P\text{-value}]$	# P	$E[P\text{-value}]$
Exp	—	9487	0.50 ± 0.006	9511	0.50 ± 0.006	9478	0.50 ± 0.006	9493	0.50 ± 0.006
E_k	$k = 2$	28	0.00 ± 0.000	9985	0.78 ± 0.004	21	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 4$	0	0.00 ± 0.000	10000	0.94 ± 0.002	0	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 6$	0	0.00 ± 0.000	10000	0.98 ± 0.001	0	0.00 ± 0.000	0	0.00 ± 0.000
	$c^2 = 1.25$	8843	0.42 ± 0.006	9169	0.45 ± 0.006	9015	0.43 ± 0.006	8138	0.35 ± 0.006
H_2	$c^2 = 1.5$	7204	0.27 ± 0.005	8811	0.40 ± 0.006	7940	0.33 ± 0.006	5441	0.18 ± 0.005
	$c^2 = 2$	3603	0.09 ± 0.003	7186	0.24 ± 0.005	4447	0.15 ± 0.004	695	0.02 ± 0.001
	$c^2 = 4$	90	0.00 ± 0.000	3648	0.08 ± 0.003	1323	0.04 ± 0.002	22	0.00 ± 0.000
	$c^2 = 10$	0	0.00 ± 0.000	935	0.02 ± 0.001	1205	0.04 ± 0.003	67	0.00 ± 0.001
Z	—	1200	0.02 ± 0.001	9438	0.57 ± 0.006	1228	0.02 ± 0.001	187	0.00 ± 0.000
LN	—	98	0.00 ± 0.000	9517	0.53 ± 0.006	219	0.01 ± 0.000	24	0.00 ± 0.000
RRI	$p = 0.1$	9048	0.41 ± 0.006	9044	0.42 ± 0.006	9056	0.42 ± 0.006	9121	0.41 ± 0.005
	$p = 0.5$	4659	0.11 ± 0.003	5587	0.16 ± 0.004	5118	0.13 ± 0.003	4624	0.11 ± 0.003
	$p = 0.9$	16	0.00 ± 0.000	701	0.01 ± 0.001	83	0.00 ± 0.000	13	0.00 ± 0.000
EARMA	0.25	9284	0.47 ± 0.006	8564	0.36 ± 0.005	9266	0.47 ± 0.006	9498	0.50 ± 0.006
	0.5	8865	0.43 ± 0.006	7519	0.27 ± 0.005	8908	0.43 ± 0.006	9393	0.49 ± 0.006
	1	8178	0.37 ± 0.006	9238	0.48 ± 0.006	8918	0.42 ± 0.006	8115	0.34 ± 0.006
	3	5209	0.21 ± 0.005	5671	0.19 ± 0.005	4534	0.13 ± 0.004	4686	0.17 ± 0.005
	5.25	4100	0.14 ± 0.004	1598	0.03 ± 0.002	4216	0.14 ± 0.004	5680	0.21 ± 0.005
mH_2	$m = 2$	4398	0.14 ± 0.004	4355	0.11 ± 0.003	5332	0.21 ± 0.005	1546	0.04 ± 0.002
	$m = 5$	7514	0.32 ± 0.006	5400	0.17 ± 0.004	7922	0.35 ± 0.006	7228	0.29 ± 0.006
	$m = 10$	7818	0.35 ± 0.006	9446	0.48 ± 0.006	9475	0.49 ± 0.006	9457	0.50 ± 0.006
	$m = 20$	7996	0.37 ± 0.006	9494	0.50 ± 0.006	9517	0.50 ± 0.006	9489	0.50 ± 0.006
$RRI(H_2)$	$p = 0.1$	104	0.00 ± 0.000	2986	0.07 ± 0.002	1432	0.04 ± 0.002	37	0.00 ± 0.000
	$p = 0.5$	253	0.00 ± 0.001	1105	0.02 ± 0.001	1738	0.04 ± 0.002	215	0.00 ± 0.001
	$p = 0.9$	4	0.00 ± 0.000	229	0.00 ± 0.001	52	0.00 ± 0.000	5	0.00 ± 0.000

examined in [25]. The performance impact on the second queue of the departure process from the first queue was found to be roughly consistent with the cumulative impact of the stochastic dependence over longer time intervals, which in turn is well-captured by the index of dispersion for counts (IDC), that is, the ratio of the variance to the mean of the departure counting process. For the case $\alpha = 0.3$, the estimated IDC is increasing in t and approximately 2.1 for time 4. That is consistent with what the CU KS test shows for $L = 14$. Both substantiate that the stochastic dependence is present over longer time intervals.

7. ASYMPTOTICS OF THE KS TESTS

To understand the power of the different KS tests, it is useful to consider the functional law of large numbers (FWLLN) and CLT (FCLT) for the empirical process (the empirical cdf as a function of its argument x) and the associated LLN and CLT for the KS statistic in (2) applied to alternative stochastic processes; see [31, 38] for background. In addition to the PP associated with the null hypothesis, we want to cover a wide range of alternative hypotheses, including all the ones we analyzed. In particular, we want to include renewal processes

with nonexponential distributions, but also more general stationary point processes with dependence among successive interarrival times. We want to see how these two distinct departures from the PP affect the asymptotic behavior.

The desired results apply to all those alternative arrival processes provided that the interarrival times are only weakly dependent, which is the common case in service system applications. For this purpose, it suffices to apply the FCLT for empirical processes as in [31]. We use the version stated in Theorem 2.1 of [29], where explicit conditions for weak dependence are stated and references are given, from which one can find the extensive literature.

Let $\{X_k : k \geq 1\}$ be the sequence of interarrival times, and let $F(x) \equiv P(X_k \leq x)$ be the marginal cdf, which we assume is continuous and strictly increasing, and so has a well-defined inverse F^{-1} . Let $F_n(x)$ be the ecdf as in (1) and let D_n be the KS statistic as in (2).

As is customary with ecdf's, we work with an associated sequence $\{U_k : k \geq 1\}$ of random variables that individually are uniformly distributed on $[0, 1]$. In particular, we let

$$U_k \equiv F(X_k), \quad k \geq 1, \tag{9}$$

Table 10. Performance of alternative standard Tests $n = 200$: 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.

Case	Subcase	KS with estimated mean		Lilliefors test	
		#P	$E[P\text{-value}]$	#P	$E[P\text{-value}]$
<i>Exp</i>	—	9945	0.65 ± 0.005	9495	0.37 ± 0.003
E_k	$k = 2$	2	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 4$	0	0.00 ± 0.000	0	0.00 ± 0.000
	$k = 6$	0	0.00 ± 0.000	0	0.00 ± 0.000
H_2	$c^2 = 1.25$	9516	0.50 ± 0.006	8161	0.28 ± 0.004
	$c^2 = 1.5$	8030	0.29 ± 0.005	5260	0.14 ± 0.003
	$c^2 = 2$	3818	0.09 ± 0.003	1507	0.03 ± 0.002
	$c^2 = 4$	348	0.01 ± 0.001	90	0.00 ± 0.000
	$c^2 = 10$	360	0.01 ± 0.001	153	0.01 ± 0.001
Z	—	1042	0.02 ± 0.001	161	0.00 ± 0.000
LN	—	147	0.01 ± 0.000	4	0.00 ± 0.000
RRI	$P = 0.1$	9842	0.56 ± 0.005	9020	0.32 ± 0.003
	$P = 0.5$	6857	0.19 ± 0.004	3820	0.08 ± 0.002
	$P = 0.9$	50	0.00 ± 0.000	4	0.00 ± 0.000
EARMA	0.25	9947	0.65 ± 0.005	9521	0.37 ± 0.003
	0.5	9918	0.65 ± 0.005	9415	0.37 ± 0.003
	1	9829	0.60 ± 0.005	9073	0.34 ± 0.003
	3	8387	0.43 ± 0.007	6744	0.23 ± 0.004
	5.25	8002	0.34 ± 0.006	5891	0.18 ± 0.004
mH_2	$m = 2$	4908	0.15 ± 0.004	2628	0.07 ± 0.003
	$m = 5$	9421	0.49 ± 0.006	7946	0.27 ± 0.004
	$m = 10$	9877	0.62 ± 0.005	9183	0.35 ± 0.003
	$m = 20$	9929	0.65 ± 0.005	9487	0.37 ± 0.003
$RRI(H_2)$	$P = 0.1$	432	0.01 ± 0.001	122	0.00 ± 0.000
	$P = 0.5$	674	0.01 ± 0.001	216	0.01 ± 0.001
	$P = 0.9$	19	0.00 ± 0.000	0	0.00 ± 0.000

so that $P(U_k \leq x) = P(X_k \leq F^{-1}(x)) = F(F^{-1}(x)) = x$, $0 < x < \infty$. and $F^{-1}(U_k) = X_k, k \geq 1$. Hence,

$$1_{\{U_k \leq F(x_k): 1 \leq k \leq k_0\}} = 1_{\{X_k \leq x_k: 1 \leq k \leq k_0\}}. \tag{10}$$

Hence, we can apply a FCLT for uniform random variables and then transform by the cdf F when we are done. For this purpose, let

$$Y_k(x) \equiv 1_{\{U_k \leq x\}} - x, \quad k \geq 1, \quad 0 \leq x \leq 1, \tag{11}$$

Table 11. The CU and Lewis KS tests applied to the departure processes over [6, 20] from the $M_t/H_2/s_t + M$ model with the sinusoidal arrival rate function in 18 cases: three relative amplitudes [$r = 0$ (constant), 0.2, and 0.6], 2 abandonment probability targets [$\alpha = 0.3$, heavily loaded, and $\alpha = 0.01$, lightly loaded] and three subinterval lengths [$L: 0.5, 2$, and 14].

Arrival rate Rel.	Aband. Prob.	Sample Size	Result	$L = 0.5$		$L = 2$		$L = 14$	
				CU	Lewis	CU	Lewis	CU	Lewis
$r = 0.0$	$\alpha = 0.3$	9717	$P\text{-val}$	0.50	0.48	0.54	0.53	0.20	0.58
			# pass	44	47	46	49	30	50
	$\alpha = 0.01$	13,709	$P\text{-val}$	0.46	0.45	0.54	0.46	0.40	0.48
			# pass	47	47	47	45	41	44
$r = 0.2$	$\alpha = 0.3$	9586	$P\text{-val}$	0.56	0.54	0.46	0.55	0.00	0.27
			# pass	47	47	38	50	0	47
	$\alpha = 0.01$	13,690	$P\text{-val}$	0.46	0.51	0.27	0.56	0.00	0.52
			# pass	47	47	38	50	0	47
$r = 0.6$	$\alpha = 0.3$	9344	$P\text{-val}$	0.47	0.40	0.18	0.12	0.00	0.00
			# pass	46	46	27	16	0	0
	$\alpha = 0.01$	13,672	$P\text{-val}$	0.47	0.48	0.04	0.02	0.00	0.00
			# pass	46	48	10	6	0	0

The KS tests are applied 50 times, once for each 10 replications.

be the process constructed from mean-0 centered indicator variables associated with U_k , let

$$\hat{U}_n(x) \equiv \frac{1}{\sqrt{n}} \sum_{j=1}^n Y_k(x), \quad 0 \leq x \leq 1, \quad (12)$$

be the FCLT-scaled uniform empirical process on $[0, 1]$ and let

$$\begin{aligned} \Gamma(x, y) \equiv & E[Y_1(x)Y_1(y)] + \sum_{j=2}^{\infty} (E[Y_1(x)Y_j(y)] \\ & + E[Y_1(y)Y_j(x)]) \end{aligned} \quad (13)$$

for $0 \leq x \leq 1$ and $0 \leq y \leq 1$. The elements $E[Y_1(x)Y_k(y)]$ in the sum in the second term of (13) are covariance terms, capturing the dependence. Let $D[0, 1]$ be the usual function space of right-continuous real-valued functions on $[0, 1]$, for example, see [38].

In this framework, here is the FCLT for the uniform empirical process, without giving a full description of weak dependence:

THEOREM 7.1 (FCLT for the uniform empirical process): Let $\{U_k : k \geq 1\}$ be a weakly dependent stationary sequence of random variables uniformly distributed on $[0, 1]$, either (i) ϕ -mixing or (ii) S-mixing. Assume that

$$\sum_{i=1}^{\infty} \|E[U_{i+k}|\mathcal{F}_k]\|_{L_2} = \sum_{i=1}^{\infty} (E[(E[U_{i+k}|\mathcal{F}_k])^2])^{\frac{1}{2}} < \infty, \quad (14)$$

where $\mathcal{F}_k \equiv \sigma\{U_i : 1 \leq i \leq k\}$ for each $k \geq 1$. Then, the series $\Gamma(x, y)$ in (13) converges absolutely, and

$$\hat{U}_n \Rightarrow \hat{U} \text{ in } D[0, 1] \quad \text{as } n \rightarrow \infty, \quad (15)$$

where $\hat{U} \equiv \{\hat{U}(x) : 0 \leq x \leq 1\}$ is a mean-0 Gaussian process with covariance function $\Gamma(x, y)$ in (13).

The final results can be obtained by transforming back from $[0, 1]$ to $[0, \infty)$ using the marginal cdf F . Let

$$\hat{F}_n(x) = \sqrt{n}(F_n(x) - F(x)), \quad x \geq 0, \quad (16)$$

be the FCLT-scaled empirical process on $[0, \infty)$. Since $\hat{F}_n(x) = \hat{U}_n(F(x))$, $x \geq 0$, we have the following corollary.

COROLLARY 7.1 (FCLT for the general empirical process): Under the weak dependence condition of Theorem 7.1,

$$\hat{F}_n \Rightarrow \hat{F} \text{ in } D[0, \infty) \quad \text{as } n \rightarrow \infty, \quad (17)$$

where $\hat{F}(x) \equiv \hat{U}(F(x))$, $x \geq 0$, so that $\hat{F} \equiv \{\hat{F}(x) : x \geq 0\}$ is a mean-0 Gaussian process with covariance function $E[\hat{F}(x)\hat{F}(y)] = \Gamma(F(x), F(y))$, $x \geq 0$ and $y \geq 0$, for Γ in (13).

As in Theorem 7.1, the dependence in the limit in Corollary 7.1 is captured by the second term in (13), but now transformed by the cdf F . As a consequence of Corollary 7.1, we have the FWLLN:

COROLLARY 7.2 (FWLLN for the general empirical process): Under the weak dependence condition of Theorem 7.1,

$$F_n \Rightarrow F \text{ in } D[0, \infty) \quad \text{as } n \rightarrow \infty. \quad (18)$$

These results are generalizations of the classical Glivenko–Cantelli theorem and Kolmogorov CLT for i.i.d random variables briefly reviewed in §2.3 of [38].

By the continuous mapping theorem (applied in the case with bounded domain $[0, 1]$), we obtain the associated LLN and CLT for the KS statistic:

COROLLARY 7.3 (CLT and LLN for the KS statistic): Under the weak dependence condition of Theorem 7.1,

$$D_n \Rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (19)$$

and

$$\hat{D}_n \equiv \sqrt{n}D_n \Rightarrow \hat{D} \stackrel{d}{=} \sup_{x \geq 0} \{|\hat{F}(x)|\} \quad \text{as } n \rightarrow \infty. \quad (20)$$

where \hat{F} is the Gaussian limit process in Corollary 7.1.

Without dependence, the limiting random variable has a known distribution, originally found by Kolmogorov, namely,

$$\hat{D} \stackrel{d}{=} \sup_{0 \leq x \leq 1} \{|B_0(x)|\}, \quad (21)$$

where B_0 is the Brownian bridge stochastic process and

$$P(\hat{D} \leq x) = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-\left(\frac{\pi^2(2k-1)^2}{8x^2}\right)}, \quad x > 0. \quad (22)$$

However, the dependence yields an extra term. More generally, we have the limit \hat{D} being the supremum of the absolute value of a mean-0 Gaussian process with covariance function Γ in (13). An explicit analytical expression is evidently difficult to obtain, but there is a well defined limit that can be identified using simulation.

There are important implications for applications. The first-order [order $O(1)$] asymptotic behavior of the ecdf

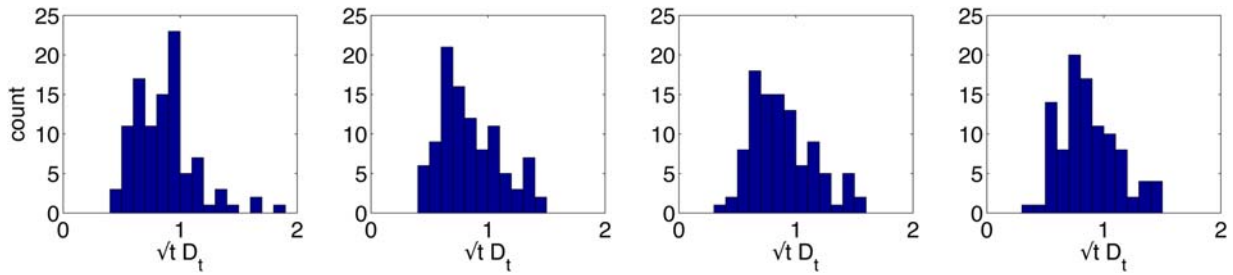


Figure 4. *Exp* - Histogram of $\sqrt{t}D_t$ with $t = 2 \times 10^5$ from 100 replications: standard KS, conditional-uniform, log, and Lewis tests (from left to right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$F_n(x)$ in (1) as $n \rightarrow \infty$ is covered by the LLN in (19). First, the LLN in (19) holds for all processes, with the same limit as if the sequence $\{X_k\}$ is i.i.d. From this perspective, to distinguish an alternative hypothesis from the null hypothesis, we want the marginal cdf's to be significantly different. The dependence does not affect the LLN, so that it should be relatively more difficult to distinguish an alternative arrival process with a dependent exponential distribution from a PP.

The second-order [order $O(1/\sqrt{n})$] asymptotic behavior of the ecdf $F_n(x)$ is covered by the CLT in (20). If the alternative hypothesis has the same marginal cdf as the null hypothesis, or nearly the same, then we will be relying on the second-order behavior to reject the alternative. In that case, we should expect the power to be much less. Moreover, the dependence is captured by the CLT via the covariance function $\Gamma(x, y)$, but the supremum makes the case with dependence quite complicated. Thus to see such second-order effects, it is natural to look at the scaled random variable \hat{D}_n in (20) as well as D_n and the empirical process F_n .

To illustrate the asymptotics described above, we now plot $\sqrt{t}\hat{D}_t$ as in (20) for the rate-1 process over the interval $[0, t]$ for much larger t . The limit in (20) extends immediately from the discrete argument n to the continuous argument t .] To make the large-sample story clear, these figures are histograms based on 100 replications for $t = 2 \times 10^5$, which is 1000 times longer than our base case with $t = 200$. In each case, plots for the four tests are shown, in the order of the standard, CU, Lewis, and Log tests from left to right.

First, Fig. 4 shows the results for the PP null hypothesis. Now all four plots are consistent with the classical limit theorem in (21) and (22). Second, Fig. 5 shows the corresponding plots for the renewal process with H_2 interarrival times having $c^2 = 2$. Consistent with our previous observations, Fig. 5 shows the pathetically weak performance of the CU test and the superiority of the Lewis test. Finally, Fig. 6 shows the corresponding plots for the *EARMA*(5.25) sequence of dependent exponential interarrival times with the largest (5.25) cumulative correlations. Fig. 6 shows that all four KS tests do detect the dependence, but the CU has highest power for this example. These conclusions are confirmed by the sample means displayed in Table 12. The relatively good performance of the CU test for dependent exponential interarrival times is evident, but it is yet to be explained.

The analysis in this section is complicated by considering the entire ecdf $F_n(x)$ as a function of x and thus the FCLT's in Theorem 7.1 and Corollary 7.1. After taking the supremum, that leads to the 1D KS statistic in (2) and its associated CLT in Corollary 7.3. We can see the essential idea in this section if we consider the ecdf $F_n(x)$ for one fixed argument x . This limited view is easy to understand and analyze, because the ecdf $F_n(x)$ is simply the sum of Bernoulli random variables. However, these Bernoulli random variables will be dependent if the X_n are. Nevertheless, analysis is relatively easy.

We illustrate this more elementary perspective for the case of *EARMA*(5.25) for the argument $x = 0.2$ in Fig. 7. It shows a plot of $\sqrt{t}|F_t(x) - F(x)|$ for $x = 0.2$ as a function

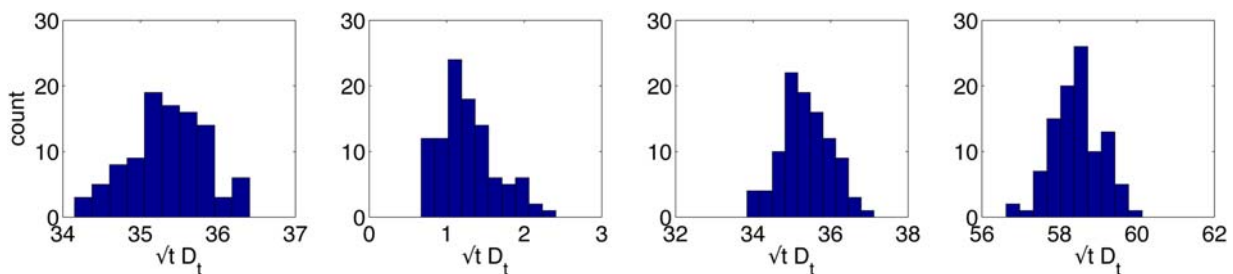


Figure 5. H_2 ($c^2 = 2$) - Histogram of $\sqrt{t}D_t$ with $t = 2 \times 10^5$ from 100 replications: standard KS, conditional-uniform, log, and Lewis tests (from left to right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

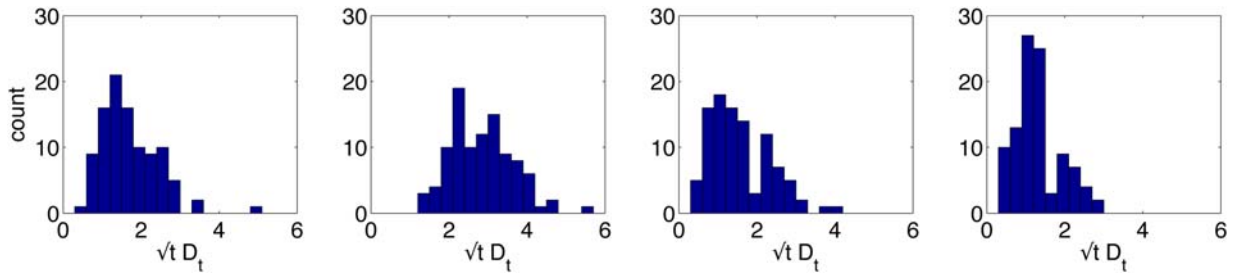


Figure 6. EARMA (5.25) - Histogram of $\sqrt{t}D_t$ with $t = 2 \times 10^5$ from 100 replications: standard KS, conditional-uniform, log, and Lewis tests (from left to right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 12. Average $\sqrt{t}D_t$ with associated 95% confidence intervals from 100 replications and $t = 2 \times 10^5$.

Case	Standard	CU	Log	Lewis
Exp	0.86 ± 0.02	0.85 ± 0.02	0.88 ± 0.02	0.87 ± 0.02
E_2	62.50 ± 0.02	0.63 ± 0.01	62.48 ± 0.02	82.32 ± 0.02
H_2	35.35 ± 0.03	1.28 ± 0.02	35.36 ± 0.04	58.46 ± 0.04
EARMA (0.25)	0.94 ± 0.02	1.06 ± 0.02	0.98 ± 0.02	0.91 ± 0.02
EARMA (0.5)	0.91 ± 0.02	1.20 ± 0.02	0.96 ± 0.02	0.87 ± 0.02
EARMA (1)	1.02 ± 0.02	1.45 ± 0.03	1.05 ± 0.02	0.94 ± 0.02
EARMA (3)	1.48 ± 0.03	2.20 ± 0.04	1.41 ± 0.04	1.25 ± 0.03
EARMA (5.25)	1.70 ± 0.05	2.84 ± 0.05	1.58 ± 0.05	1.31 ± 0.04

of t up to $t = 2 \times 10^5$ as above, with t expressed in log scale. For each KS test, the 95% confidence interval lines are also shown. From this more elementary plot, we also see the greater power of the CU test in this case, but only for $x = 0.2$.

8. WHY THE CONDITIONAL UNIFORM TEST HAS LOW POWER

The purpose of this section is to provide theoretical support for our explanation in section 1.3 why the CU KS test has so little power. We have indicated that it is because the CU KS test focuses on the arrival time instead of the interarrival times. As a consequence, we can show that the

ecdf in (1) converges to the uniform cdf associated with the null hypothesis as the sample size $n \rightarrow \infty$. Thus the long-run average value of the ecdf in (1) for the CU KS test will coincide with the uniform cdf, just as we saw in Fig. 1 and 2. However, the CU KS test can detect dependence among the interarrival times in the alternative hypothesis. The FCLT in section 7 shows that it will produce a difference of order $O(1/\sqrt{n})$, which is the order of the KS statistic in (2) under the null hypothesis. The tables and plots in section 7 demonstrate that the CU KS test can detect the dependence.

To establish the supporting result, it suffices to assume that the sequence of interarrival times $\{X_n : n \geq 1\}$ obeys a strong LLN (SLLN), as occurs (by definition) when the sequence of interarrival times is an ergodic stationary sequence. For an i.i.d. sequence that occurs if an interarrival time has a finite

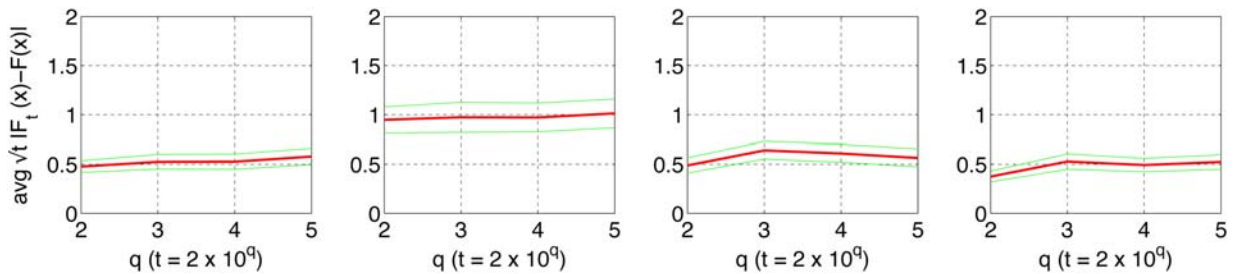


Figure 7. EARMA (5.25) - Average $\sqrt{t}|F_t(x) - F(x)|$ with $F(x) = 0.2$ over 100 replications: standard KS, conditional-uniform, log, and Lewis tests (from left to right). The x-axis is scaled such that $t = 2 \times 10^q$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

mean. That covers all our examples of possible alternative hypotheses and many more. Then

$$n^{-1}S_n \rightarrow E[X] \quad \text{w.p.1} \quad \text{as } n \rightarrow \infty, \quad (23)$$

where $S_n \equiv X_1 + \cdots + X_n, n \geq 1$, with $S_0 \equiv 0$. As a consequence, the interarrival times necessarily also obey the seemingly much stronger functional strong law of large numbers (FSLLN), that is,

$$\sup_{0 \leq t \leq 1} \{|n^{-1}S_{[nt]} - tE[X]|\} \rightarrow 0 \quad \text{w.p.1} \quad \text{as } n \rightarrow \infty, \quad (24)$$

for example, see section 3.2 of the Internet Supplement to [38]. However, the CU test applied to an interval $[0, t]$ applies to the associated counting process. Nevertheless, the FSLLN in (24) implies a corresponding FSLLN for the associated counting process,

$$N(t) \equiv \max \{k \geq 0 : S_k \leq t\}, \quad t \geq 0. \quad (25)$$

As a consequence, of the FSLLN for the partial sums in (24), we automatically have the associated FSLLN for the counting process:

$$\sup_{0 \leq t \leq 1} \{|n^{-1}N(nt) - t/E[X]|\} \rightarrow 0 \quad \text{w.p.1} \quad \text{as } n \rightarrow \infty; \quad (26)$$

for example, see section 13.6 and 13.8 of [38]. That implies that the CU test asymptotically yields the linear plots we see in the CU cases of Figs. 1 and 2. Indeed, all CU test plots looked just like these (see the appendix).

9. CONCLUSIONS AND FUTURE RESEARCH

9.1. Conclusions

We have focused on statistical tests of a homogeneous PP, which we apply to statistical tests of a NHPP based on a approximation for the rate function and the CU transformation, as proposed by Brown et al. [8]. This approach is appealing because the CU transformation allows us to combine the data into a single sequence of i.i.d. random variables uniformly distributed on $[0, 1]$. It thus eliminates all nuisance parameters, that is, there are no parameters to estimate. We can then directly apply a standard statistical test, such as the CU KS test, but we have shown that an extra data transformation, as proposed in [8, 10, 22], should be helpful, because the i.i.d. random variables produced by the CU transformation correspond to the arrival times instead of the interarrival times. As a consequence, for the CU KS test the ecdf in (1) converges to the cdf of the uniform cdf as the sample size n

increases, as we showed in section 8. That explains why we see no separation at all for the CU KS test in Figs. 1 and 2.

The Durbin [10] transformation used by Lewis [22] after the CU transformation starts by ordering the interarrival times in ascending order. As a consequence, it focuses more on differences in the interarrival-time distribution from the exponential interarrival-time distribution of the PP null hypothesis. We have conducted extensive simulation experiments to study the power of the alternative KS tests of a PP. Our analysis supports the approach proposed by Brown et al. [8], but finds that the Lewis [22] KS test consistently has even greater power against alternative hypotheses with different interarrival-time cdf's. The Lewis test also has greater power than the standard KS tests of a PP via its interarrival times, even if the mean is assumed to be known in advance.

Our experiments show that different conclusions hold for the KS tests of a PP for two classes of alternative hypotheses: (i) those with nonexponential interarrival times and (ii) those with dependent exponential interarrival times. For the first class, the Lewis test consistently has most power.

For the second class, none of the procedures has much power, but the CU KS test evidently has the most power. We show that the CU KS test can detect the dependence with enough data. Since the deviation from a PP is due entirely to the dependence in these examples, it evidently is better to keep the original order of the interarrival times. (We have noted the exception for the RRI cases.) Since the data transformations tend to reorder the interarrival times, the data transformations may actually dissipate the impact of the dependence. The final example testing the departure process from a many-server queue with NHPP arrival process illustrates the potential advantages of the CU KS test. Comparing Tables 7 and 3 show that the use of subintervals may significantly weaken the ability of the CU KS test to detect the dependence, as the dependence extending beyond the subinterval boundaries is lost.

We have shown that simulation can be very useful for determining how much power these KS tests have in any given context. Especially, revealing are the plots such as in Figs. 1 and 2 comparing the sample average of the ecdf over many independent replications with the cdf of the null hypothesis. For dependent alternative hypotheses, it may be helpful to display the estimated CLT-scaled ecdf and KS statistic, as in Figs. 4–7.

9.2. Remaining Issues

We also have investigated other important issues arising in tests of a NHPP with service system data. First, we studied the consequence of approximating a general NHPP by a PC NHPP, as required for the CU, Log, and Lewis tests. Second, we studied the consequence of data rounding, which was found to be a serious issue by [8]. Third, we studied the

issue of over-dispersion when we attempt to obtain more data by combining data from multiple days. We report our results for these three issues in [20], where we also apply these KS tests to call center and hospital data.

We also realized that the two data transformations studied here can also be applied to other KS tests. Hence, we studied alternative KS tests to the standard KS test of i.i.d. random variables with a cdf F . We found that the power of the standard test can often be improved by first transforming the null hypothesis to i.i.d. exponential random variables (by applying Proposition 1 in section 2.2) and then applying the Lewis KS test of the resulting PP. We found that this procedure is consistently superior to the direct Durbin transformation (applied after converting the null hypothesis to a uniform cdf on $[0, 1]$ (again by applying Proposition 1 in §2.2); we report the details in [21].

There remain many unanswered questions: can we find even better admissible data transformations? For individual alternative hypotheses or classes of them, can we find the best admissible data transformation, as formulated in section 2? Can we verify Conjecture 5.1 about stochastic order of the expected ecdf's in section 5.2? Can we supplement the FCLT in section 7 by quantifying the asymptotic distribution of the KS test statistic in the general case with dependence?

It remains to decide how to interpret statistical tests in applications. With enough data, any null hypothesis will be rejected by reasonable statistical tests if there is the slightest difference in the hypotheses. Thus, failing the best KS statistical test does not imply that the NHPP arrival process model cannot be useful. Indeed, the model might nevertheless be most useful, depending on how it can be used. This is especially relevant for performance analysis, because queueing models with PP and NHPP arrival processes are much easier to analyze than with all known alternatives. We have begun to study how to analyze performance in queueing models when the KS test indicates that an NHPP is actually inappropriate. We plan to discuss this last issue in subsequent papers.

Whenever we have two or more candidate models, all of which can be applied, and only one passes the statistical tests, then there is quite a good case for using the successful model. However, if we only know how to proceed with one model, then the implications of test failure are unclear. It remains to consider what to do if a NHPP fails the KS test, which we intend to address in a forthcoming paper. We do see that the NHPP is less justified as the estimated P -value decreases; that should provide caution. A more powerful test reduces the likelihood of being overconfident. Applications suggest considering alternative performance-based statistical tests of a NHPP, in the spirit of [7, 37].

ACKNOWLEDGMENTS

The authors thank the Samsung Foundation and NSF for support (NSF grants CMMI 1066372 and 1265070).

REFERENCES

- [1] O.Z. Aksin, M. Armony, and V. Mehrotra, The modern call center: a multi-disciplinary perspective on operations management research, *Prod Oper Manag* 16 (2007), 665–688.
- [2] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov, Patient flow in hospitals: A data-based queueing-science perspective, New York University, 2011. Available at: <http://www.stern.nyu.edu/om/faculty/armony/>.
- [3] A.N. Avramidis, A. Deslauriers, and P. L'Ecuyer, Modeling daily arrivals to a telephone call center, *Manag Sci* 50 (2004), 896–908.
- [4] A.D. Barbour, L. Holst, and S. Janson, Poisson approximation, Oxford University Press, Oxford, U.K., 1992.
- [5] G.A. Barnard, Time intervals between accidents: A note on Maguire, Pearson & Wynn's paper, *Biometrika* 40 (1953), 212–213.
- [6] A. Bassamboo and A. Zeevi, On a data-driven method for staffing large call centers, *Oper Res* 57 (2009), 714–726.
- [7] O. Besbes, R. Philliips, and A. Zeevi, Testing the validity of a demand model: An operations perspective, *Manuf Serv Oper Manag* 12 (2010), 162–183.
- [8] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, Statistical analysis of a telephone call center: A queueing-science perspective, *J Am Stat Assoc* 100 (2005), 36–50.
- [9] D.J. Daley and D. Vere-Jones, An introduction to the theory of point processes, volume II, 2nd ed., Springer, Oxford, U. K., 2008.
- [10] J. Durbin, Some methods for constructing exact tests, *Biometrika* 48 (1961), 41–55.
- [11] S.G. Eick, W.A. Massey, and W. Whitt, The physics of the $M_t/G/\infty$ queue, *Oper Res* 41 (1993), 731–742.
- [12] K.W. Fendick and W. Whitt, Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue, *Proc IEEE* 71 (1989), 171–194.
- [13] S. Holm, A simple sequentially rejective multiple test procedure, *Scan J Stat* 6 (1979), 65–70.
- [14] R. Ibrahim, P. L'Ecuyer, N. Regnard, and H. Shen, On the modeling and forecasting of call center arrivals, Proceedings of the 2012 Winter Simulation Conference, Berlin, Germany, 2012, pp. 256–267.
- [15] P.A. Jacobs and P.A.W. Lewis, A mixed autoregressive-moving average exponential sequence and point process (EARMA 1,1), *Adv Appl Probab* 9 (1977), 87–104.
- [16] P.A. Jacobs and P.A.W. Lewis, Discrete time series generated by mixtures. i: Correlational and runs properties, *J R Stat Soc B* 40 (1978), 94–105.
- [17] P.A. Jacobs and P.A.W. Lewis, Stationary autoregressive discrete moving average time series generated by mixtures, *J R Stat Soc B* 4 (1983), 19–36.
- [18] G. Jongbloed and G. Koole, Managing uncertainty in call centers using Poisson mixtures, *Appl Stoch Model Bus Ind* 17 (2001), 307–318.
- [19] S.-H. Kim and W. Whitt, Appendix to choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process, Columbia University, 2013. Available at: <http://www.columbia.edu/~ww2040/allpapers.html>.
- [20] S.-H. Kim and W. Whitt, Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes?, Columbia University, 2013. Available at: <http://www.columbia.edu/~ww2040/allpapers.html>.

- [21] S.-H. Kim and W. Whitt, The power of alternative Kolmogorov-Smirnov tests based on transformations of the data, Columbia University, 2013. <http://www.columbia.edu/~ww2040/allpapers.html>.
- [22] P.A.W. Lewis, Some results on tests for Poisson processes, *Biometrika* 52 (1965), 67–77.
- [23] H.W. Lilliefors, On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown, *J Am Stat Assoc* 64 (1969), 387–389.
- [24] Y. Liu and W. Whitt, Stabilizing customer abandonment in many-server queues with time-varying arrivals, *Oper Res* 60 (2012), 1551–1564.
- [25] Y. Liu and W. Whitt, Stabilizing performance in feed-forward networks of many-server queues with time-varying arrivals, *Prob Engr Inf Sci*, forthcoming. Available at <http://www.columbia.edu/~ww2040>.
- [26] W.A. Massey and W. Whitt, Unstable asymptotics for nonstationary queues, *Math Oper Res* 19 (1994), 267–291.
- [27] R.G. Miller, *Simultaneous statistical inference*, 2nd ed., Springer, New York, 1981.
- [28] G. Pang and W. Whitt, The impact of dependent service times on large-scale service systems, *Manuf Serv Oper Manag* 14 (2012), 262–278.
- [29] G. Pang and W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues with dependent service times, *Queueing Sys* 73 (2013), 119–146.
- [30] S.M. Ross, *Stochastic processes*, 2nd ed., Wiley, New York, 1996.
- [31] G.R. Shorack and J.A. Wellner, *Empirical Processes with applications in statistics*, SIAM, Philadelphia, SIAM classics in applied mathematics edition, 2009.
- [32] R. Simard and P. L'Ecuyer, Computing the two-sided Kolmogorov-Smirnov distribution, *J Stat Soft* 39 (2011), 1–18.
- [33] R.J. Simes, An improved Bonferroni procedure for multiple tests of significance, *Biometrika* 73 (1986), 751–754.
- [34] K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, *IEEE J Sel Area Commun SAC-4* (1986), 833–846.
- [35] M.A. Stephens, Edf statistics for goodness of fit and some comparisons, *J Am Stat Assoc* 69 (1974), 730–737.
- [36] W. Whitt, Approximating a point process by a renewal process: Two basic methods, *Oper Res* 30 (1982), 125–147.
- [37] W. Whitt, Queue tests for renewal processes, *Oper Res Lett* 2 (1983), 7–12.
- [38] W. Whitt, *Stochastic-process limits*, Springer, New York, 2002.