Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

USING SIMULATION TO STUDY STATISTICAL TESTS FOR ARRIVAL PROCESS AND SERVICE TIME MODELS FOR SERVICE SYSTEMS

Song-Hee Kim
Ward Whitt

Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

ABSTRACT

When fitting queueing models to service system data, it can be helpful to perform statistical tests to confirm that the candidate model is appropriate. The Kolmogorov-Smirnov (KS) test can be used to test whether a sample of interarrival times or service times can be regarded as a sequence of i.i.d. random variables with a continuous cdf, and also to test a nonhomogeneous Poisson Process (NHPP). Using extensive simulation experiments, we study the power of various alternative KS tests based on data transformations. Among available alternative tests, we find the one with the greatest power in testing a NHPP. Furthermore, we devise a new method to test a sequence of i.i.d. random variables with a specified continuous cdf; it first transforms a given sequence to a rate-1 Poisson process (PP) and then applies the existing KS test of a PP. We show that it has greater power than direct KS tests.

1 INTRODUCTION

The most widely used stochastic queueing model of a service system is the Erlang-A ($M/M/s+M$) model, with s servers and i.i.d. exponentially distributed interarrival, service, and patience times. However, it is important to perform statistical tests with service system data to determine if such a model is actually appropriate. To demonstrate the importance of using the correct model, Table 1 compares the simulated performance of two models where one has exponentially distributed interarrival times and the other has hyper-exponentially distributed interarrival times with *squared coefficient of variation* (scv, variance divided by the square of the mean) $c^2 = 2$. It shows that if we choose staffing assuming that interarrival times are exponentially distributed when they are actually hyper-exponentially distributed with $c^2 = 2$, then we would observe an average of 35% increase in the percentage of the customers that wait (here we consider $s = 30$, but the increase is similar in other cases).

Brown et al. (2005) emphasized the importance of performing statistical tests with service system data. Realizing that testing only for a Poisson process (PP) would not be sufficient because service systems, such as call centers and hospital emergency rooms, typically have strongly time-varying arrival rates, they suggested a specific statistical test to test whether the arrival process is a nonhomogeneous Poisson process (NHPP). Their test is based on, first, looking at subintervals where the NHPP can be regarding as a PP, second, applying the conditional-uniform property of a PP to obtain a sample of i.i.d. random variables uniformly distributed on $[0,1]$ and, third, applying the Kolmogorov-Smirnov (KS) statistic after transforming the data. We found there are actually more KS tests exploiting data transformations, and conduct extensive simulation experiments to study the power of various alternatives. We conclude that the general approach of Brown et al. (2005) is excellent, but an alternative KS test proposed by Lewis (1965), exploiting a different transformation due to Durbin (1961), consistently has greater power in testing a NHPP.

Table 1: Simulation estimates of steady-state performance in the $GI/M/s+M$ model with $\lambda = 25$, $\mu = 1$, $\theta = 1$ for two interarrival time cdf's: exponential and hyper-exponential ($c^2 = 2$). The staffing level, s, is chosen using the square root staffing formula assuming exponentially distributed interarrival times, $s \equiv m + \beta\sqrt{m}$, where m is the offered load $\lambda/\mu = 25$. We consider three cases for the quality-of-service parameter β: 0.5, 1, and 2 (yielding s =28, 30, and 35). The results are based on 100 replications of $10^3 + 10^5$ customers (first 10^3 customers removed to get rid of the initial effect). Associated 95% confidence intervals are also shown.

Model	s	$E[W\|All]$	$E[W\|Served]$	$E[W\|Abandoned]$	*%Wait*	*%Abandon*
$M/M/s+M$	28	0.0324 ± 0.0003	0.0348 ± 0.0003	0.1004 ± 0.0006	29.96 ± 0.15	3.49 ± 0.03
	30	0.0168 ± 0.0002	0.0181 ± 0.0002	0.0878 ± 0.0006	18.23 ± 0.13	1.81 ± 0.02
	35	0.0022 ± 0.0001	0.0024 ± 0.0001	0.0647 ± 0.0012	3.41 ± 0.06	0.24 ± 0.01
$H_2/M/s+M$	28	0.0461 ± 0.0003	0.0496 ± 0.0004	0.1165 ± 0.0006	35.93 ± 0.16	4.96 ± 0.03
	30	0.0272 ± 0.0003	0.0295 ± 0.0003	0.1031 ± 0.0006	24.75 ± 0.15	2.94 ± 0.03
	35	0.0057 ± 0.0001	0.0061 ± 0.0001	0.0776 ± 0.0009	7.11 ± 0.09	0.61 ± 0.01

Furthermore, we found that the KS tests used to test a PP can be applied to test whether n observations can be regarded as a sample of size n from an i.i.d. sequence with arbitrary specified continuous cdf F. These KS tests are directly applicable to service systems, because the standard model for the service times is an i.i.d. sequence. The most convenient cdf for analysis is the exponential cdf, but data analysis often suggests a lognormal cdf instead, as in Brown et al. (2005). The idea is to first transform the given sequence to a sequence of mean-1 exponential random variables, which is equivalent to a rate-1 PP, and then apply the alternative KS tests of a PP. Durbin (1961) suggested transforming the data to increase the power of the KS test for an i.i.d. sequence with arbitrary specified continuous cdf F, but a KS test that first transforms a given sequence to a Poisson process has not been considered before. We use simulation experiments to show that the power can often be more consistently and substantially increased by including this step.

2 ALTERNATIVE KOLMOGOROV-SMIRNOV TESTS

The KS statistical test is commonly used to determine if data can be regarded as a sample from a sequence of independent and identically distributed (i.i.d.) random variables $\{X_n : n \geq 1\}$, each distributed as a random variable X with a specified continuous cumulative distribution function (cdf) $F(x) \equiv P(X \leq x)$, $x \in \mathbb{R}$. In this section, we first briefly describe the standard KS test, and introduce three alternative KS tests of a PP (equivalently, whether a sequence of i.i.d. random variables is exponentially distributed) based on transformations of the data. Then, we introduce three additional tests that can be used to test whether a sequence of i.i.d. random variable is from a continuous cdf F.

2.1 The Standard Test

The KS test is based on the maximum difference between the *empirical cdf* (ecdf) $F_n(x) \equiv n^{-1}\sum_{k=1}^{n} 1_{\{X_k \leq x\}}$ where $x \in \mathbb{R}$ and the underlying cdf F, where 1_A is an indicator function that is equal to 1 if the event A occurs, and equal to 0 otherwise, i.e.,

$$D_n \equiv \sup_x \{|F_n(x) - F(x)|\}, \tag{1}$$

which has a distribution that is independent of the continuous cdf F. 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, i.e., for specified probability of rejecting the null hypothesis when it is in fact correct (type I error), which we take to be $\alpha = 0.05$.

In this study, we are mostly interested in the power of alternative KS tests. 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 accepting the null hypothesis (type II error) when it is false, which of course depends on the alternative.

2.2 Alternative Tests for Testing for a NHPP

A key property exploited in alternative tests that will be described in this subsection is the basic conditioning property of a PP. For a PP on an interval, it is well known that, conditional on the total number of arrivals in that interval, the arrival times divided by the length of the interval are distributed as the order statistics of i.i.d. random variables uniformly distributed on $[0,1]$; e.g., see §2.3 of Ross (1996). Since the arrival rate in a service system typically changes relatively slowly compared to the overall arrival rate, it is often reasonable to assume that the arrival rate is *piecewise-constant* (PC); Brown et al. (2005) assume that can be done. A PC NHPP can be regarded as a homogeneous PP over each subinterval.

With that classical *conditional-uniform* (CU) transformation, the data from all the subintervals can be combined into obtain a single sequence of i.i.d. random variables uniformly distributed on $[0,1]$, to which the KS test can be applied directly; we call that the CU test. The CU transformation and the CU test has the important property that it eliminates the nuisance parameter; the method is independent of the rate of the PP. However, Brown et al. (2005) did not stop with the CU KS test, but instead proposed a (scaled) logarithmic transformation into a single sequence of i.i.d. exponential random variables for their Log KS test. We also found that there is relevant history in the statistical literature. In particular, Barnard (1953) first proposed the CU test of a PP. Then Lewis (1965) made a significant contribution for testing a PP, recognizing that a transformation proposed by Durbin (1961) could be effectively applied after the CU transformation to obtain a new KS test; we call that the Lewis test. Here are formal definitions of the **three tests of a PP**:

Conditional-Uniform (CU) Test. We start with the CU test proposed by Barnard (1953). Given an arrival process over an interval $[0,t]$, we observe the number n of arrivals in this interval and their arrival times T_j, $1 \le j \le n$. 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, $1 \le j \le n$, are distributed as the order statistics of i.i.d. random variables uniformly distributed over $[0,1]$. Thus the ecdf can be computed via $F_n(x) \equiv n^{-1}\sum_{k=1}^{n} 1_{\{T_k/t \le x\}}, 0 \le x \le 1$, and the KS statistic can be computed as in (1) with uniform cdf $F(x) = x$, $0 \le x \le 1$.

Log Test. Brown et al. (2005) observed that, given the n observed arrival times $\{T_j : 1 \le j \le n\}$ during the interval $[0,t]$, $X_j^{Log} \equiv -(n+1-j)\log_e[(t-T_j)/(t-T_{j-1})]$ with $1 \le j \le n$ are n i.i.d. mean-1 exponential random variables. The KS test in (1) can then be applied using the exponential cdf $F(x) \equiv 1-e^{-x}$.

We note that 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, $X_j^{Log,n} \equiv -j\log_e(T_j/T_{j+1})$ where $1 \le j \le n-1$ are again i.i.d. rate-1 exponential random variables.

Lewis Test. Lewis (1965) proposed using a different modification of the CU test, exploiting a transformation due to Durbin (1961). Durbin (1961) started with a sample U_j with $1 \le j \le n$ that is hypothesized to be uniformly distributed on $[0,1]$. Then let $U_{(j)}$ be the j^{th} smallest of these, $1 \le j \le n$, so that $U_{(1)} < \cdots < U_{(n)}$. Lewis (1965) applies this with $U_{(j)} = T_j/t$ from the CU test. Next the successive *intervals* between these ordered observations are considered: $C_j = U_{(j)} - U_{(j-1)}$ with $2 \le j \le n$, where $C_1 = U_{(1)}$ and $C_{n+1} = 1 - U_{(n)}$. Then let $C_{(j)}$ be the j^{th} smallest of these intervals, $1 \le j \le n$, so that $0 < C_{(1)} < \cdots < C_{(n+1)} < 1$. Now let Z_j be scaled versions of the intervals between these new variables, i.e., $Z_j = (n+2-j)(C_{(j)} - C_{(j-1)})$ with $1 \le j \le n+1$ where $C_{(0)} \equiv 0$. Remarkably, Durbin (1961) showed that, under the PP null hypothesis, the random vector $(Z_1,\ldots,Z_n)$ is distributed the same as the random vector $(C_1,\ldots,C_n)$. Hence, again under the PP

null hypothesis, the vector of associated partial sums $(S_1,\ldots,S_n)$, where $S_k \equiv Z_1+\cdots+Z_k$ with $1\le k\le n$, has the same distribution as the original random vector $(U_{(1)},\ldots,U_{(n)})$ of ordered uniform random variables. Hence, we can apply the KS test with the ecdf $F_n(x)\equiv n^{-1}\sum_{k=1}^{n} 1_{\{S_k\le x\}}$ with $0\le x\le 1$ for S_k, comparing it to the uniform cdf $F(x)\equiv x$, $0\le x\le 1$. Durbin (1961) showed that by doing this transformation starting from a sequence of i.i.d. uniform random variables, we should gain an increase in power. Lewis (1965) showed that this transformation increases power after the CU transformation, which is a different setting than considered by Durbin (1961).

2.3 Additional Tests for Testing for a General Distribution

In this subsection, we introduce three additional KS tests of an i.i.d. sequence $\{X_n\}$ with continuous cdf F based on the idea that $F(X_n)$ has a uniform cdf on $[0,1]$ and $-\log\{1-F(X_n)\}$ has a mean-1 exponential cdf. It is important to note that the KS statistic is unchanged by these transformations.

The additional KS tests are: (i) Exp+CU, applying the CU transformation to a PP constructed using $-\log\{1-F(X_n)\}$, (ii) Exp+CU+Log, the CU transformation of a PP plus the Log transformation as in Brown et al. (2005), and (iii) Exp+CU+Durbin, the CU transformation of a PP plus the Durbin (1961) transformation, as in Lewis (1965).

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

Exp+CU+Log Test. We start with the partial sums T_k, $1\le k\le n$, used in the Exp+CU test, which are the arrivals times of a rate-1 PP under the null hypothesis. We again use the conditional-uniform property for fixed sample size to conclude that, under the null hypothesis, T_k/T_n, $1\le k\le n-1$, are distributed as $U_{(k)}$, the order statistics of $n-1$ random variables, with $U_{(1)}<\cdots<U_{(n-1)}$. Hence, $Y_j^{(L)}\equiv -\log_e(T_j/T_{j+1}), 1\le j\le n-1$ should be $n-1$ i.i.d. rate-1 exponential random variables, to which we can apply the KS test.

Exp+CU+Durbin Test. We again start with the partial sums T_k, $1\le k\le n$, used in the Exp+CU test, which are the arrivals times of a rate-1 PP under the null hypothesis. We again use the conditional-uniform property for fixed sample size to conclude that, under the null hypothesis, T_k/T_n, $1\le k\le n-1$, are distributed as $U_{(k)}$, the order statistics of $n-1$ random variables uniformly distributed on $[0,1]$, with $U_{(1)}<\cdots<U_{(n-1)}$. From this point, we apply the Durbin (1961) test with n replaced by $n-1$, just as Lewis (1965) did in his test of a PP (and explained in §2.2).

3 THE SIMULATION EXPERIMENT

In this section we specify the alternative hypotheses to the PP and the service time distribution that we will consider, and show how we perform the simulation experiment.

3.1 Study Cases

We consider five cases, each with one or three subcases, yielding a total of nine cases in all. We specify these cases in terms of the sequence $\{X_n : n\ge 1\}$ of interarriva/service times, each distributed as a random variable X. In all cases, the sequence is assumed to be stationary with $E[X]=1$.

For the purpose of testing for our PP null hypothesis, all of the five cases are renewal arrival processes, with i.i.d. interarrival times. The first i.i.d. case is the null hypothesis with exponential interarrival times. The other i.i.d. cases have non-exponential 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 §9.5 of Ross (1996). Thus, they have scv $c_X^2 < 1$ and $c_X^2 > 1$. Cases 4 and 5 contain non-exponential cdf's with $c_X^2 = 1$ as well as $E[X] = 1$, just like the exponential cdf. Similarly, for testing service times, all of the cases are designed to be i.i.d. mean-1 random variables with the specified cdf.

1. **Exponential**: Exponential interarrival/service times.
2. **Erlang, $\mathbf{E_k}$**: Erlang-k (E_k) interarrival/service times, a sum of k i.i.d. exponentials for $k = 2, 4, 6$ with $c_X^2 \equiv c_k^2 = 1/k$.
3. **Hyperexponetial, $\mathbf{H_2}$**: Hyperexponential-2 (H_2) interarrival/service times, a mixture of 2 exponential cdf's with $c_X^2 = 1.25$, 2, and 10. The cdf is $P(X \le x) \equiv 1 - p_1 e^{-\lambda_1 x} - p_2 e^{-\lambda_2 x}$. We further assume balanced means ($p_1 \lambda_1^{-1} = p_2 \lambda_2^{-1}$) as in (3.7) of Whitt (1982) so that given the value of c_X^2, $p_i = [1 \pm \sqrt{(c_X^2 - 1)/(c_X^2 + 1)}]/2$ and $\lambda_i = 2p_i$.
4. **Mixture with $\mathbf{c_X^2 = 1}$**: A mixture of a more variable cdf and a less variable cdf so that the $c_X^2 = 1$; $P(X = Y) = p = 1 - P(X = Z)$, where Y is H_2 with $c_Y^2 = 4$, Z is E_2 with $c_Z^2 = 1/2$ and $p = 1/7$.
5. **Lognormal**: Lognormal interarrival/service times with mean and variance both equal to 1 ($LN(1,1)$), so that $c_X^2 = 1$.

3.2 Simulation Design

For each study case, we simulate 10^4 replications of 10^4 interarrival/service times. We generate much more data than needed in order to get rid of any initial effects. We are supposing that we observe stationary point processes, which is achieved by having the system operate for some time before collecting data.

We use this simulation output to generate arrival for time intervals of a fixed length t and service times for sample sizes of a fixed size n. For the first scenario of testing a PP in interval $[0,t]$ 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 $[0, 2000]$.

In the second scenario for testing service time models with fixed sample size n, in each replication of the 10^4 simulated service times we use $n = 200$ and $n = 2000$.

4 RESULTS

We now present the results of applying the alternative KS tests in §2 to the original and transformed arrival process and service time data in all nine study cases specified in §3, with the goal of comparing the power of the alternative KS tests. We first show the results of tests of a PP in §4.1; 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. In §4.2, we show the results of tests for an i.i.d. sequence with cdf F. In §4.3, we discuss the effect of longer length of intervals (equivalently, larger sample sizes) on the power of the tests.

4.1 Tests of a Nonhomogeneous Poisson Process

In Table 2, we report the power of each test based on 10,000 replications, where the power is estimated as (number rejected/10,000). We also report the average p-value; the p-value is the significance level below which the hypothesis would be rejected. Thus low p-values indicate greater power. The first "Exp" case is the PP null hypothesis, and 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. The Lewis performs the best, but the Log test also performs reasonably well, in marked contrast to the CU test. The difference is most striking for the middle H_2 alternative with $c^2 = 2$. For H_2 cdfs with lower scv, the power

of all methods is less, but the ordering remains; for H_2 cdf's with higher scv, the power of all methods is greater, but the ordering remains.

We also include a "standard" test for comparison. The standard KS test used to obtain the values reported in Table 2 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. In order to check the consequence of this, we also tried the valid standard KS test with fixed sample size $n = 200$, the standard KS test in the interval $[0,200]$ but with estimated mean, and the Lilliefors (1969) test for the exponential cdf with unknown mean. Interestingly, none of the performance results differed greatly, so one might consider the standard KS test with estimated mean or the Lilliefors (1969) test as realizable alternatives to the standard KS test. However, all of these were inferior to the Lewis KS test.

Table 2: Simulation estimates of the power of alternative KS tests of a rate-1 Poisson process for rate-1 renewal processes, with observations over the time interval $[0,200]$ and significance level $\alpha = 0.05$ based on 10,000 replications. The estimated power is computed as (number rejected / 10,000). The average p-value, the significance level below which the hypothesis would be rejected, is shown in parenthesis.

KS test	Lewis	Standard	Log	CU
Exp	0.05(0.50)	0.05(0.50)	0.05(0.50)	0.05(0.50)
E_2	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.78)
E_4	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.94)
E_6	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.98)
$H_2(c^2 = 1.25)$	0.24(0.30)	0.12(0.41)	0.13(0.40)	0.10(0.41)
$H_2(c^2 = 2)$	0.94(0.01)	0.63(0.10)	0.59(0.13)	0.28(0.23)
$H_2(c^2 = 10)$	1.00(0.00)	1.00(0.00)	0.94(0.02)	0.94(0.01)
Z	0.98(0.01)	0.88(0.02)	0.87(0.02)	0.06(0.57)
$LN(1,1)$	0.99(0.00)	0.99(0.01)	0.97(0.01)	0.05(0.52)

We find that useful insight is provided by appropriate plots. Especially revealing are plots comparing the average of the ecdf's over all 10,000 replications to the cdf associated with the null hypothesis (which depends on the transformation). Figure 1 illustrates for the case of a renewal process with H_2 interarrival times having $c_X^2 = 2$ over the time interval $[0,200]$. It shows 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 this plot, it is evident that the CU test should perform poorly. For any new contemplated alternative, we suggest conducting simulations and comparing the plots.

Also insightful are plots of the empirical distribution function of the p values; we display the plots for H_2 with $c_X^2 = 2$ and Z in Figure 2. These plots show that the relative power tends to remain across all p values, not just for our type I error of $\alpha = 0.05$.

4.2 Test for i.i.d. Sequence with cdf F

In this section, we present the results of four KS tests: (i) the standard test, using the variables $U_k \equiv F(X_k)$, (ii) the Durbin (1961) test as described under the Lewis test in §2.2, (iii) the EXP+CU test described in §2.3, and (iv) the EXP+CU+Durbin test described in §2.3. Under the null hypotheses, the cdf in all four cases is uniform on $[0,1]$. We found that the EXP+CU+Log tests were consistently dominated by the Durbin (1961) test or the EXP+CU+Durbin test, so we do not present detailed results for that test here.

As in §4.1, we report the estimated power of each test based on 10,000 replications as well as the average p-value. We use the i.i.d. LN(1,1) null hypothesis in Table 3, since lognormal hypotheses are especially interesting for service systems, e.g., Brown et al. (2005). We first observe that all tests perform

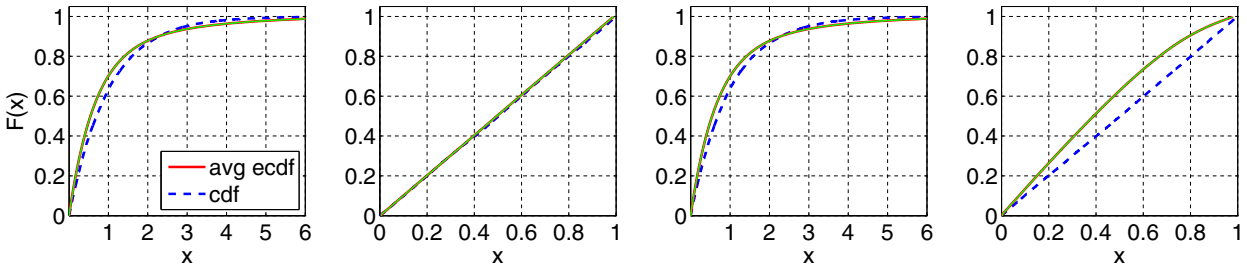

Figure 1: Comparison of the average ecdf for an H_2 renewal process with $c_X^2 = 2$ based on 10^4 replications for $[0,200]$ with the cdf of the null hypothesis: Standard, Conditional-Uniform, Log, Lewis KS Tests (from left to right).

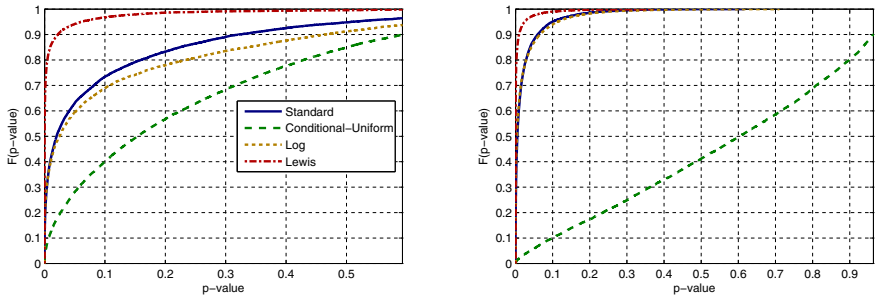

Figure 2: Empirical distribution function of the p-values for two cases of the simulation experiment in §4.1 - H_2 ($c^2 = 2$) (left) and Z (right) based on 10^4 replications for the time interval $[0,200]$.

properly for the null hypotheses. In order to observe the effect of the same distribution but different parameters, we add $LN(1,0.25)$, $LN(1,4)$, and $LN(1,10)$ to the alternative distributions. The results for the EXP+CU+Durbin, standard, and EXP+CU tests are similar to those for the corresponding KS tests of a PP in Table 2, but the results for the Durbin (1961) test are new, and surprisingly bad in some cases. Table 3 also shows that the EXP+CU+Durbin test is consistently most powerful for all study cases. We have tested other null hypotheses (i.e., all study cases given in §3), and found that the conclusions from the i.i.d. LN(1,1) null hypothesis extend to i.i.d. null hypotheses with other marginal cdfs.

We again found that useful insight is provided by plots comparing the average of the ecdf's over all $10,000$ replications to the cdf associated with the null hypothesis. Figure 3 illustrates for the i.i.d. variables having cdf Exp with $n = 200$, tested for the i.i.d. $LN(1,1)$ null hypothesis. It shows that the transformation in the EXP+CU+Durbin KS test provides greater separation (slightly better than the standard KS test) between the average ecdf and the cdf.

The poor results for the Durbin (1961) test for the i.i.d. cases in Table 3 seem inconsistent with the results in Durbin (1961) and the enthusiastic endorsement by Lewis (1965), so we decided to repeat some of the experiments actually performed by Durbin (1961). We now consider the same four KS tests applied to the i.i.d. standard normal, $N(0,1)$, null hypothesis. To keep the same mean for all alternatives, we consider all the previous cases after subtracting 1 to make them all have mean 0. Indeed, the first alternative considered by Durbin (1961) was an i.i.d. sequence of random variables distributed as $Y-1$, where Y is a mean 1 exponential variable; it has the same mean and variance as $N(0,1)$. Since those alternatives have quite a different shape from the symmetric $N(0,1)$ distributions, we also considered i.i.d. sequences of random variables distributed as $Z_k - 1 + \sqrt{1-(1/k)}N(0,1)$, where Z_k has an E_k cdf, for $k = 2,4,6$. These have the same first two moments and approximately the same shape. We summarize the results for this alternative with the sample size $n = 50$ used by Durbin (1961) in Table 4. The new base case is the i.i.d.

Table 3: Simulation estimates of the power of alternative KS tests of i.i.d. $LN(1,1)$ variables for alternative distributions with sample size $n = 200$ and significance level $\alpha = 0.05$ based on 10,000 replications. The estimated power is computed as (number rejected / 10,000). The average p-value, the significance level below which the hypothesis would be rejected, is shown in parenthesis.

KS test	Exp+CU+Durbin	Standard	Durbin Direct	Exp+CU
Exp	0.97(0.01)	0.95(0.01)	0.64(0.09)	0.07(0.45)
E_2	0.93(0.01)	0.69(0.08)	0.15(0.38)	0.00(0.70)
E_4	1.00(0.00)	1.00(0.00)	0.99(0.00)	0.00(0.89)
E_6	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.95)
$H_2(c^2 = 1.25)$	1.00(0.00)	0.99(0.00)	0.74(0.06)	0.12(0.39)
$H_2(c^2 = 2)$	1.00(0.00)	1.00(0.00)	0.92(0.02)	0.24(0.26)
$H_2(c^2 = 10)$	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.71(0.06)
Z	0.47(0.14)	0.31(0.22)	0.15(0.37)	0.03(0.57)
$LN(1,0.25)$	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.90)
$LN(1,1)$	0.05(0.50)	0.05(0.50)	0.05(0.50)	0.05(0.50)
$LN(1,4)$	1.00(0.00)	1.00(0.00)	0.99(0.00)	0.39(0.17)
$LN(1,10)$	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.64(0.08)

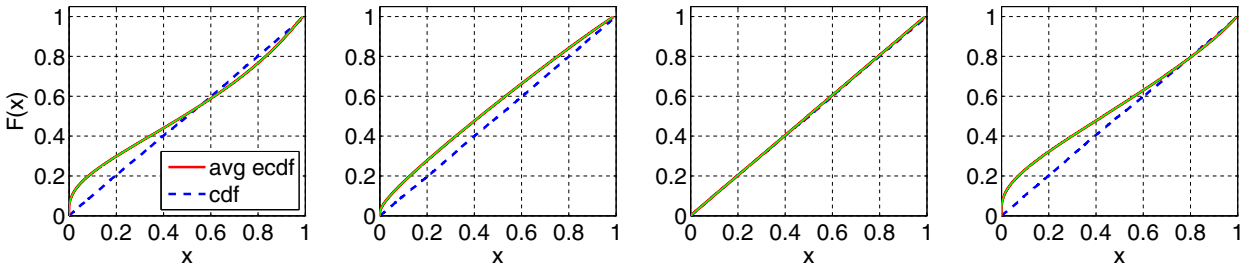

Figure 3: Comparison of the average ecdf from observations of $n = 200$ i.i.d. mean-1 exponential random variables based on 10^4 replications with the cdf of the null hypothesis ($LN(1,1)$): Standard, Direct Durbin, EXP+CU, and EXP+CU+Durbin KS Tests (from left to right).

standard normal null hypothesis. Just as in previous results, the results show that all tests behave properly for the standard normal null hypothesis. In addition, Table 4 shows that the Durbin (1961) test performs much better now, just as originally reported. In this case *both* the Durbin (1961) and EXP+CU+Durbin KS tests perform much better than the standard and EXP+CU alternatives.

4.3 Longer Time Intervals and Larger Sample Sizes

In §4.1 and §4.2, we have observed that the power decreases as the alternative gets closer to the null hypothesis. On the other hand, the power increases as we increase the sample size, as we now illustrate by considering tests of a PP for a longer 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 $c_X^2 = 1.25$ from 2. For all the other cases, the three main candidates - Lewis, standard, and Log - have estimated power of 1.00 with average p-value 0.00. We note that similar results hold when we increase the sample size, n, in tests for an i.i.d. sequence with arbitrary specified continuous cdf F.

Table 4: Simulation estimates of the power of alternative KS tests of i.i.d. $N(0,1)$ variables for alternative distributions with sample size $n = 50$ and significance level $\alpha = 0.05$ based on 10,000 replications. The estimated power is computed as (number rejected / 10,000). The average p-value, the significance level below which the hypothesis would be rejected, is shown in parenthesis. All alternatives have mean 0.

KS test	Exp+CU+Durbin	Standard	Durbin Direct	Exp+CU
$Exp-1$	0.88(0.02)	0.44(0.07)	0.81(0.04)	0.33(0.23)
E_2-1	0.96(0.01)	0.62(0.04)	0.70(0.05)	0.06(0.52)
E_4-1	1.00(0.00)	1.00(0.01)	0.97(0.01)	0.00(0.81)
E_6-1	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.92)
$H_2-1(c^2=1.25)$	0.96(0.01)	0.58(0.05)	0.90(0.02)	0.49(0.16)
$H_2-1(c^2=2)$	1.00(0.00)	0.83(0.02)	0.98(0.00)	0.74(0.07)
$H_2-1(c^2=10)$	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.81(0.09)
$Z-1$	0.95(0.01)	0.52(0.05)	0.73(0.05)	0.27(0.37)
$LN(1,1)-1$	0.99(0.00)	0.83(0.03)	0.93(0.01)	0.40(0.24)
$N(0,1)$	0.05(0.50)	0.06(0.50)	0.05(0.50)	0.05(0.50)
$E_2-1+\sqrt{1/2}N(0,1)$	0.16(0.38)	0.07(0.47)	0.05(0.49)	0.12(0.40)
$E_4-1+\sqrt{3/4}N(0,1)$	0.06(0.48)	0.05(0.51)	0.05(0.50)	0.07(0.47)
$E_6-1+\sqrt{5/6}N(0,1)$	0.06(0.49)	0.05(0.50)	0.05(0.50)	0.06(0.49)

5 CONCLUSION

We have conducted extensive simulation experiments to study the power of alternative Kolmogorov-Smirnov (KS) statistical tests that can be used to test components of service system models: the arrival process and service time distributions. Specifically, when one needs to test for a nonhomogeneous Poisson Process (NHPP), our analysis strongly supports the approach proposed by Brown et al. (2005), but finds another related KS test proposed by Lewis (1965) consistently has even greater power. Since both tests exploit the same conditional-uniform transformation, they both apply directly to piecewise-constant NHPP's as well as Poisson processes (PP's). Moreover, the CU transformation eliminates the nuisance parameter; these KS tests do not depend on the rate of the PP. However, the CU test itself has very low power for non-exponential interarrival times, so should not be used. In agreement with Brown et al. (2005), we find that the banking call center arrival data passed all KS tests for a NHPP, provided that data rounding is addressed.

For testing for the distribution of service times (more generally, testing for an i.i.d. sequence with arbitrary specified continuous cdf F), our analysis strongly supports the data-transformation approach proposed by Durbin (1961), but we find that applying that data transformation after transforming the original sequence to a rate-1 PP produces a KS test with greater power.

As usual with statistical tests, the power increases with the length of the interval or the sample size, so that some sample sizes may be too small to have any power, whereas other sample sizes may be too large to accept even the slightest deviation from a null hypothesis. Thus, as many have discovered before, judgment is required in the use of statistical tests.

Table 5: Simulation estimates of the power of alternative KS tests of a rate-1 Poisson process for rate-1 renewal processes, with observations over the time interval $[0, 2000]$ and significance level $\alpha = 0.05$ based on 10,000 replications. The estimated power is computed as (number rejected / 10,000). The average p-value, the significance level below which the hypothesis would be rejected, is shown in parenthesis.

KS test	Lewis	Standard	Log	CU
Exp	0.05(0.50)	0.05(0.50)	0.05(0.50)	0.05(0.50)
E_2	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.79)
E_4	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.95)
E_6	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.98)
$H_2(c^2 = 1.25)$	0.97(0.01)	0.66(0.08)	0.64(0.10)	0.11(0.40)
$H_2(c^2 = 2)$	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.31(0.21)
$H_2(c^2 = 10)$	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.98(0.00)
Z	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.05(0.52)
$LN(1,1)$	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.05(0.51)

ACKNOWLEDGMENTS

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

REFERENCES

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

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective". *Journal of the American Statistical Association* 100:36–50.

Durbin, J. 1961. "Some Methods for Constructing Exact Tests". *Biometrika* 48 (1): 41–55.

Lewis, P. A. W. 1965. "Some Results on Tests for Poisson Processes". *Biometrika* 52 (1): 67–77.

Lilliefors, H. W. 1969. "On the Kolmogorov-Smimov Test for the Exponential Distribution with Mean Unknown". *Journal of the American Statistical Association* 64:387–389.

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

Whitt, W. 1982. "Approximating a Point Process by a Renewal Process: Two Basic Methods". *Operations Resesarch* 30:125–147.

AUTHOR BIOGRAPHIES

SONG-HEE KIM is a doctoral student in the Department of Industrial Engineering and Operations Research at Columbia University. Her primary research focus is on operations management in service systems with emphasis on problems related to healthcare delivery, using empirical/statistical analysis, simulation, stochastic modeling, and queueing theory. Her e-mail and web addresses are sk3116@columbia.edu and http://www.columbia.edu/~sk3116, respectively.

WARD WHITT is a professor in the Department of Industrial Engineering and Operations Research at Columbia University. He joined the faculty there in 2002 after spending 25 years in research at AT&T. He received his Ph.D. from Cornell University in 1969. His recent research has focused on stochastic models of service systems, using both queueing theory and simulation. His email address is ww2040@columbia.edu and his web page is http://www.columbia.edu/~ww2040.