2010

Parametric bootstrap and approximate tests for two Poisson variates

Sung Nok Chiu
Hong Kong Baptist University, snchiu@hkbu.edu.hk

Parametric bootstrap and approximate tests for two Poisson variates

SUNG NOK CHIU*

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Abstract

The parametric bootstrap tests and the asymptotic or approximate tests for detecting difference of two Poisson means are compared. The test statistics used are the Wald statistics with and without log-transformation, the Cox F statistic and the likelihood ratio statistic. It is found that the type I error rate of an asymptotic/approximate test may deviate too much from the nominal significance level α under some situations. It is recommended that we should use the parametric bootstrap tests, under which the four test statistics are similarly powerful and their type I error rates are all close to α. We apply the tests to breast cancer data and injurious motor vehicle crashes data.

Keywords: Asymptotic tests; Monte–Carlo tests; Parametric bootstrap; Poisson process; Rate ratio.

Running title: Tests for two Poisson variates

*E-mail: snchiu@hkbu.edu.hk

1 Introduction

Count data on the numbers of occurrences often arise in retrospective public health studies in such a way that the cumulative total number is observed at one time point but no information is available on the exact time points of the individual occurrences. In spatial epidemiology [10], we may have count data on the numbers of cases in a politically defined administrative region but the exact locations of the individual cases have not been recorded. With complete information on the exact time points or exact spatial locations, the data could be modelled by a counting process [1] or a spatial point process [5], respectively. Without the detailed temporal or spatial information, such count data are typically modelled by the Poisson distribution. One question we often encounter in the latter scenario is that given two independent counts observed from two time intervals or spatial regions with fixed but not necessarily equal length or size, whether the rate or the intensity, i.e. the number of occurrences per unit time or area, of the two underlying temporal or spatial processes are the same or not.

For example, in a study of the risk of motor vehicle crashes in elderly drivers, Ray *et al.* [16] found that in a sample of 16,262 drivers aged 65–84 years, there were 175 injurious crashes in 17.3 thousand person-years at risk among women whilst there were 320 in 21.4 thousand person-years at risk among men. Another example can be found in the study by Boice and Monson [2], who compared the breast cancer rate in women with tuberculosis after repeated fluoroscopic examinations of the chest with a control group; they observed 41 cases of breast cancer in 28,010 person-years at risk among women repeatedly exposed to multiple

X-ray fluoroscopies and 15 cases in 19,017 person-years at risk among unexposed women. In each of the above examples we have two count data (175 and 320; 41 and 15) coming from two time intervals of unequal length (17.3 and 21.4; 28,010 and 19,017, respectively).

To compare two independent Poisson rates for the breast cancer data in Boice and Monson [2], Greenland and Rothman [7] used the large-sample Wald confidence limits for the logarithm of the rate ratio, whilst Graham *et al.* [6] proposed the use of the likelihood scores to construct large-sample confidence limits for the rate ratio. Note that by saying asymptotic or large-sample we consider the limiting scenario that the means of the Poisson distributions go to infinity, because in our context a large sample (of the point process) comes from a long observation period, leading to a large Poisson mean; if we consider fixed length observation periods, the limiting scenario is then equivalent to the one in which the rates go to infinity. Liu *et al.* [11] compared the coverage of confidence intervals constructed by four different methods. Ng and Tang [14] and Ng *et al.* [13] carried out extensive simulation studies to compare the type I error rates and the powers of Wald, likelihood ratio and score statistics using the asymptotic normality and the numerical approximation for the p-value. However, these comparisons have overlooked a popular approximate test developed by Cox [3], which has been cited, up to the end of August 2008, over one hundred and forty times, mostly in medical research articles. These papers also have not mentioned the possibility of using parametric bootstrap. Krishnamoorthy and Thomson [9] remarked that their ad hoc approach to estimate the p-value for the test statistic T_1 (to be introduced in Section 2) is equivalent to the parametric bootstrap approach in an exact manner and they found that

their approach is better than the conditional test introduced by Przyborowski and Wilenski [15].

This paper reports a power comparison for the statistics recommended in Cox [3], Ng *et al.* [13] and Ng and Tang [14], using the asymptotic/approximate distributions as well as the parametric bootstrap tests [4]. Section 2 introduces the test statistics for detecting difference between the rates of two Poisson variates. Section 3 explains how the parametric bootstrap tests would be carried out in our context and then Monte–Carlo simulation results are reported and discussed in Section 4. Finally, in Section 5 we apply the tests to the above two examples.

2 Test statistics and their asymptotic distributions

Suppose X_1 and X_2 are two independent random variables coming from two Poisson distributions with means $\lambda_1 t_1$ and $\lambda_2 t_2$, respectively. That is to say, X_i is the observed number of occurrences of a temporal or spatial Poisson process with rate or intensity λ_i in a sampling frame of length or size t_i, $i = 1$ and 2. Denote by ρ the rate ratio λ_2/λ_1, and let $t_2/t_1 = d$. Let us consider a one-sided test here, and so the hypotheses of interest are

$$\mathrm{H}_0\colon\ \rho = 1 \quad \text{against} \quad \mathrm{H}_\mathrm{A}\colon\ \rho > 1.$$

Cox [3] argued that approximately

$$F = \frac{t_1(X_2 + \frac{1}{2})\lambda_1}{t_2(X_1 + \frac{1}{2})\lambda_2}$$

has an F-distribution with $(2X_1+1, 2X_2+1)$ degrees of freedom, which, he said, "may lead to accurate results even in very small samples".

Ng and Tang [14] considered four Wald statistics, two of which are obtained after taking logarithmic transformation for skewness correction and variance stabilization. They are:

$$\begin{aligned} W_1 &= \frac{X_2 - dX_1}{(d^2X_1 + X_2)^{1/2}}, \\ W_2 &= \frac{X_2 - dX_1}{\{d(X_1 + X_2)\}^{1/2}}, \\ W_3 &= \frac{\ln(X_2/X_1) - \ln(d)}{(1/X_1 + 1/X_2)^{1/2}}, \\ W_4 &= \frac{\ln(X_2/X_1) - \ln(d)}{\{(2 + 1/d + d)/(X_1 + X_2)\}^{1/2}}, \end{aligned}$$

where for W_1 and W_2, we use the convention that $0/0 = 0$, whilst for W_3 and W_4, we set $X_i = 0.5$ whenever $X_i = 0$, $i = 1, 2$. Under the null hypothesis, W_j follows the standard normal distribution asymptotically, $j = 1, 2, 3, 4$. The simulation results in Ng and Tang [14] suggest that we should use either W_2 and W_3, and so in this paper we do not consider W_1 and W_4.

Ng *et al.* [13] considered the difference, instead of the ratio, of the two rates:

$$\mathrm{H_0}\colon\ \lambda_2 - \lambda_1 = \delta \quad \text{against} \quad \mathrm{H_A}\colon\ \lambda_2 - \lambda_1 > \delta$$

and their test statistics include

$$\begin{aligned} T_1 &= \frac{X_2/t_2 - X_1/t_1 - \delta}{(X_1/t_1^2 + X_2/t_2^2)^{1/2}}, \\ T_2 &= \frac{X_2/t_2 - X_1/t_1 - \delta}{\{(X_1 + X_2)/(t_1t_2) + \delta(t_2 - t_1)/(t_1t_2)\}^{1/2}}, \end{aligned}$$

which are equivalent to W_1 and W_2, respectively, if $\delta = 0$. Note also that the statistic T_1 with its asymptotic normality is the one recommended by Liu *et al.* [11] to construct confidence intervals. Ng *et al.* [13]'s another statistic T_3

$$T_3 = \frac{X_2/t_2 - X_1/t_1 - \delta}{\{\lambda_1^*/t_1 + \lambda_2^*/t_2\}^{1/2}},$$

where λ_i^* are the constrained maximum likelihood estimates of λ_i under the null hypothesis that $\lambda_2 - \lambda_1 = \delta$:

$$\lambda_i^* = \frac{(-1)^i\delta}{2} + \frac{X_1 + X_2}{2(t_1 + t_2)} + \sqrt{\left(\frac{\delta}{2} - \frac{X_1 + X_2}{2(t_1 + t_2)}\right)^2 + \frac{X_1\delta}{t_1 + t_2}},$$

is also equivalent to W_2 when $\delta = 0$. Because the null hypothesis of interest here is that $\rho = 1$ or $\delta = 0$, we do not have to consider T_j, $j = 1, 2, 3$, for non-zero δ in this paper.

They also had the one-sided likelihood ratio statistic:

$$L = \begin{cases} 2\ln \dfrac{(X_1/t_1)^{X_1}(X_2/t_2)^{X_2}}{\{(X_1 + X_2)/(t_1 + t_2)\}^{X_1+X_2}}, & X_2/t_2 - X_1/t_1 > \delta, \\ 0, & X_2/t_2 - X_1/t_1 \le \delta, \end{cases}$$

with the convention that $0^0 = 1$. Under the null hypothesis, asymptotically L is zero with probability 0.5 and follows a χ^2-distribution with one degree of freedom with probability 0.5.

Generically, denote by τ any one of the statistics above. For the one-sided alternative

$$\mathrm{H_A}\colon\ \rho > 1 \quad \text{or} \quad \mathrm{H_A}\colon\ \lambda_2 - \lambda_1 > 0$$

the critical regions of these test statistics are all in the form $\{\tau \ge \tau_0\}$ for some τ_0.

3 Parametric bootstrap tests

The normal, χ^2- and F-distributions of W_j's (and T_j's), L and F, respectively, are large-sample approximation only. Ng *et al.* [13] expressed the p-value of each T_j as a double infinite sum (see also Krishnamoorthy and Thomson [9]) but their approach cannot be applied to L and F. In this paper, we estimate the p-values via parametric bootstrapping [4, pp. 140–148].

More precisely, under the null hypothesis that $\rho = 0$ or $\delta = 0$, the maximum likelihood estimator of $\lambda_1 = \lambda_2 = \lambda$ is

$$\hat{\lambda} = \frac{X_1 + X_2}{t_1 + t_2}.$$

We then generate R pairs of independent Poisson variates with means $\hat{\lambda}t_1$ and $\hat{\lambda}t_2$, respectively. Thus, in total we have $R+1$ pairs of data, namely, the observed counts (X_1, X_2) and the counts (X_{1i}^*, X_{2i}^*), $i = 1, \ldots, R$, simulated under the null hypothesis with the estimated common rate $\hat{\lambda}$. Denote by τ the value of any one of the statistics F, W_2, W_3 or L calculated from the observation (X_1, X_2) and by τ_i^* the value of the same statistic obtained from (X_{1i}^*, X_{2i}^*), $i = 1, \ldots, R$. The sequence $\{\tau_1^*, \ldots, \tau_R^*\}$ forms a random sample of the parametric bootstrap distribution of the chosen test statistic under the null hypothesis. Thus, if exactly k simulated τ_i^* are greater than τ and none is equal to it, the one-sided p-value can be estimated by the sample proportion

$$p_{\text{boot}} = \frac{k+1}{R+1},$$

because under the null hypothesis, τ is another independent realization of the distribution of the chosen statistic and the random sample $\{\tau, \tau_1^*, \ldots, \tau_R^*\}$ of size $R+1$ has $k+1$ members greater than or equal to the value τ.

The counts (X_{1i}^*, X_{2i}^*) are, however, generated under a Poisson distribution, which is discrete; one or more τ_i^* may be equal to the value τ. If exactly m of $\{\tau_1^*, \ldots, \tau_R^*\}$ are equal to τ, then

$$\frac{k+1}{R+1} \leq p_{\text{boot}} \leq \frac{k+m+1}{R+1}.$$

To be conservative, we take the upper bound

$$p_{\text{boot}} = \frac{k+m+1}{R+1} = \frac{\#\{\tau_i^* \geq \tau\}+1}{R+1}.$$

The null hypothesis will be rejected if p_{boot} is less than or equal to the significance level α.

Such a parametric bootstrap test is a straightforward generalization of a Monte–Carlo test, which is the same procedure as above except that for a Monte–Carlo test we do not have any nuisance parameters, such as λ, to estimate. Hope [8] showed that the power loss, compared with the corresponding uniformly most powerful test, resulting from using Monte–Carlo tests is slight and so R is not necessary to be large. Marriott [12] suggested that for $\alpha = 0.05$, $R = 99$ is adequate, whilst Davison and Hinkley [4], p. 156, suggested, for $\alpha \geq 0.05$, that the loss of power with $R = 99$ is not serious and $R = 999$ should generally be safe. Note that in either a Monte–Carlo test or a parametric bootstrap test, it is the rank of τ, and not the value of τ itself, which determines that p-value.

4 Simulation results

In this section, $\alpha = 0.05$ and $R = 999$ will be used. Without loss of generality, we set $t_1 = 1$ and so $t_2 = d$. Tables 1 and 2 show the rejection rates estimated by 10,000 simulations for

$\lambda_1 = 1$ and 20, representing small and large λ_1 scenarios. Nevertheless, conclusions below were drawn from an extensive series of simulation in which different values between 1 and 20 were used for λ_1. By definition, $\lambda_2 = \rho\lambda_1$ and so if $\rho = 1$, the rejection rate is the type I error rate, otherwise the rejection rate is the power.

Because t_1 is fixed and $t_2 = d$, the larger the value of d, the longer the observation interval for the point process corresponding to the count X_2. However, the power of any one of the considered tests is not an increasing function of d, especially when ρ is not much larger than unity. The reason is that the variance of X_2 also increases as d increases. On the other hand, if, for example, we double the value of λ_1, we can rescale the time axis so that the rate of the process corresponding to X_1 remains the same and the length of the observation period is doubled. That is, increasing λ_1 with a fixed t_1 is effectively the same as increasing the observation period whilst the rate is kept fixed. By the same argument above, the power does not necessarily increase when λ_1 increases, because the variances of X_1 and X_2 also increase. Nevertheless, when we vary the values of λ_1 and d, the performance of some tests may become more desirable whilst some others less desirable; this will be discussed in details in the following.

First, note that for small λ_1 or small d, some columns may be identical or almost identical for different statistics because the Poisson variates with small means, and hence small variances, do not vary too much.

We can draw from our simulation the same conclusion as in Ng and Tang [14] that the type I error rate of the asymptotic W_3 is usually smaller than or equal to that of the

asymptotic W_2. However, for large λ_1 and small d, the type I error rates of the approximate Cox F statistic and the asymptotic likelihood ratio statistic L could be smaller than that of the asymptotic W_3, and the latter may be higher than the nominal significance level α under such situations.

Naturally, the price for a low type I error rate is the loss in power; a more powerful test is accompanied by a higher type I error rate, but the gain in power is not remarkable here. Thus, the main issue in this comparison is the type I error rate.

Although the Monte–Carlo tests with continuously distributed statistic are exact in their own right, the parametric bootstrap tests adopted here are not, because the test statistics are discrete and a nuisance parameter $\lambda_1 = \lambda_2 = \lambda$ has to be estimated. Since the upper bound of p_{boot} was used, the discreteness of the statistics would lead to conservative tests. On the other hand, the estimation of the nuisance parameter might lead to a type I error rate that is higher or lower than the nominal significance level α. From our simulation (including those not reported here), in 102 out of the 120 cases considered, the simulated type I error rates of the parametric bootstrap tests are less than α. For $\lambda_1 \geq 5$, the Cox F statistic always gives the lowest simulated type I error rates among the four statistics, using the parametric bootstrap. However, even though it is desirable to control the type I error rate to be below the nominal significance level α, it is unnecessarily conservative to choose a statistic that would give the lowest type I error rate; this rate should be as close to α as possible.

The clear message revealed from our simulation is that no test statistic is uniformly better than the others; no test statistic could really have a completely controllable type I error rate,

and no test statistic would always give the least deviation from the nominal significance level for all different combinations of λ_1 and d.

Nevertheless, we recommend the parametric bootstrap tests, because the type I error rates seldom exceed α by more than 10% of α (in our simulation study, among the 18 cases where the estimated type I error rates exceed 0.05, the maximum is only 0.0535), whilst the asymptotic/approximate tests in 64 out of the 120 cases considered have type I error rates higher than 0.05, in which 17 cases are 0.065 or above, i.e. 30% or more higher than α. Using the parametric bootstrap setting, unlike Ng *et al.* [13] and Ng and Tang [14], we do not see W_3 or any one of these four statistics is universally superior to the others; any one of them is sometimes better than the others under different combinations of λ_1 and d. Thus, generally speaking, the parametric bootstrap tests using these four statistics are more or less equally trustworthy.

If only a pocket calculator is available so that parametric bootstrapping is not feasible, then for small d and small λ_1, we recommend the asymptotic W_3 for its success in conservativeness; for large d or large λ_1, we recommend the approximate Cox F statistic for the smaller deviation from the nominal significance level.

5 Real data

As we mentioned in Section 1, Boice and Monson [2] observed 41 cases of breast cancer in 28,010 person-years at risk among women repeatedly exposed to multiple X-ray fluoroscopies and 15 cases in 19,017 person-years at risk among unexposed women, and Ray *et al.* [16]

reported 175 injurious crashes in 17.3 thousand person-years at risk among women whilst there were 320 in 21.4 thousand person-years at risk among men.

The p-values of the four statistics calculated from the asymptotic/approximate distributions and estimated by using parametric bootstrapping are shown in Table 3; note that when $R = 999$, the value of p_{boot} is at least 0.001. We have strong evidence to reject the equal rate null hypothesis at the 0.05 significance level and conclude that (1) the incidence rate of breast cancer for women who had been exposed repeatedly to X-ray fluoroscopy is higher than that for those who had not and (2) the incidence rate of injurious crashes for men drivers is higher than that for women.

Since the differences are significant, the powers of the tests are irrelevant and the concern is the possibility of committing a type I error. If we rescale the time axis so that the length of the sampling frame of the first sample t_1 is one unit time in each example, then the maximum likelihood estimator $\hat{\lambda}_1 = X_1$ and the constrained maximum likelihood estimator under the null hypothesis $\hat{\lambda}_1 = \hat{\lambda_2} = (X_1 + X_2)/(1 + d)$ are large, suggesting that the true λ_1 is likely to be large in each example. Moreover, d's are greater than 1 (1.47 and 1.27). An inspection of Table 2 suggests that the type I error rates would not be substantially greater than the nominal level $\alpha = 0.05$ for large λ_1 and $d \approx 1.5$ and so none of these tests is overly liberal.

Acknowledgements

This research was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Numbers HKBU200503 and

HKBU200605) and an FRG grant of the Hong Kong Baptist University. I thank the referee for helpful comments.

References

[1] Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N., 1993, *Statistical Models Based on Counting Processes* (New York: Springer-Verlag).

[2] Boice, J. D. and Monson, R. R., 1977, Breast cancer in women after repeated fluoroscopic examinations of the chest. *Journal of the National Cancer Institute*, **59**, 823–832.

[3] Cox, D. R., 1953, Some simple approximate tests for Poisson variates. *Biometrika*, **40**, 354–360.

[4] Davison, A. C. and Hinkley, D. V., 1997, *Bootstrap Methods and their Application* (Cambridge: Cambridge University Press).

[5] Diggle, P., 2003, *Statistical Analysis of Spatial Point Patterns* (2nd edn) (New York: Arnold).

[6] Graham, P.L., Mengersen, K. and Morton, A. P., 2003, Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. *Statistics in Medicine*, **22**, 2071–2083.

[7] Greenland, S. and Rothman, K. J., 1998, Introduction to categorical statistics. In:

K. J. Rothman and S. Greenland *Modern Epidemiology* (2nd edn) (Philadelphia: Lippincott-Raven), pp. 231–252.

[8] Hope, A. C. A., 1968, A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society Series B (Methodological)*, **30**, 582–598.

[9] Krishnamoorthy, K. and Thomson, J., 2004, A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*, **119**, 23–35.

[10] Lawson, A. B., 2006, *Statistical Methods in Spatial Epidemiology* (2nd edn) (Chichester: Wiley).

[11] Liu, G. K., Wang, J., Liu, K. and Snavely, D. B., 2006, Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. *Statistics in Medicine*, **25**, 1275–1286.

[12] Marriott, F. H. C., 1979, Barnard's Monte Carlo tests: how many simulations? *Applied Statistics*, **28**, 75–77.

[13] Ng, H. K. T., Gu, K. and Tang, M. L., 2007, A comparative study of tests for the difference of two Poisson means. *Computational Statistics and Data Analysis*, **51**, 3085–3099.

[14] Ng, H. K. T. and Tang, M. L., 2005, Testing the equality of two Poisson means using the rate ratio. *Statistics in Medicine*, **24**, 955–965.

[15] Przyborowski, J. and Wilenski, H., 1940, Homogeneity of results in testing samples from

Poisson series: with an application to testing clover seed for dodder. *Biometrika*, **31**, 313–323.

[16] Ray, W. A., Fought, R. L. and Decker, M. D., 1992, Psychoactive drugs and the risk of injurious motor vehicle crashes in elderly drivers. *Americal Journal of Epidemiology*, **136**, 873–883.

Table 1: Estimated rejection rate by simulation, $\lambda_1 = 1$ (Asym = Asymptotic; Approx = Approximate; Boots = Bootstrap).

		W_2		W_3		F		L	
d	ρ	Asym	Boots	Asym	Boots	Approx	Boots	Asym	Boots
0.1	1.0	0.0761	0.0413	0.0422	0.0407	0.0761	0.0415	0.0422	0.0413
	1.1	0.0749	0.0408	0.0418	0.0407	0.0749	0.0403	0.0418	0.0408
	1.2	0.0882	0.0449	0.0455	0.0448	0.0882	0.0447	0.0455	0.0449
	1.5	0.1062	0.0567	0.0571	0.0565	0.1062	0.0567	0.0571	0.0567
	2.0	0.1403	0.0746	0.0756	0.0744	0.1402	0.0740	0.0755	0.0746
	2.5	0.1631	0.0912	0.0929	0.0905	0.1631	0.0917	0.0929	0.0912
	4.0	0.2604	0.1628	0.1655	0.1621	0.2604	0.1596	0.1655	0.1628
0.5	1.0	0.0393	0.0328	0.0012	0.0211	0.0393	0.0338	0.0393	0.0330
	1.1	0.0497	0.0419	0.0026	0.0269	0.0497	0.0418	0.0497	0.0420
	1.2	0.0538	0.0451	0.0021	0.0302	0.0538	0.0452	0.0538	0.0452
	1.5	0.0790	0.0642	0.0063	0.0442	0.0790	0.0637	0.0790	0.0643
	2.0	0.1313	0.1019	0.0168	0.0802	0.1313	0.1013	0.1313	0.1022
	2.5	0.1824	0.1408	0.0303	0.1155	0.1824	0.1387	0.1824	0.1408
	4.0	0.3600	0.2761	0.1160	0.2538	0.3600	0.2678	0.3600	0.2763
1.0	1.0	0.0300	0.0260	0.0008	0.0091	0.0310	0.0268	0.0972	0.0265
	1.1	0.0421	0.0372	0.0009	0.0129	0.0441	0.0388	0.1139	0.0386
	1.2	0.0450	0.0384	0.0015	0.0149	0.0477	0.0396	0.1345	0.0392
	1.5	0.0749	0.0637	0.0028	0.0309	0.0798	0.0669	0.1687	0.0664
	2.0	0.1256	0.1135	0.0126	0.0709	0.1389	0.1165	0.2402	0.1160
	2.5	0.1900	0.1743	0.0333	0.1325	0.2150	0.1775	0.3095	0.1781
	4.0	0.3818	0.3628	0.1686	0.3530	0.4420	0.3594	0.4989	0.3657
1.5	1.0	0.0072	0.0185	0.0000	0.0074	0.0238	0.0199	0.0712	0.0198
	1.1	0.0096	0.0260	0.0000	0.0099	0.0327	0.0274	0.0852	0.0267
	1.2	0.0146	0.0312	0.0001	0.0127	0.0399	0.0328	0.0995	0.0318
	1.5	0.0303	0.0589	0.0005	0.0273	0.0692	0.0613	0.1386	0.0607
	2.0	0.0712	0.1157	0.0025	0.0711	0.1402	0.1177	0.2264	0.1167
	2.5	0.1372	0.1855	0.0127	0.1404	0.2229	0.1885	0.2971	0.1872
	4.0	0.3725	0.4072	0.1186	0.4021	0.4747	0.4023	0.5081	0.4037
2.0	1.0	0.0071	0.0168	0.0000	0.0069	0.0215	0.0175	0.0551	0.0172
	1.1	0.0102	0.0237	0.0000	0.0086	0.0298	0.0244	0.0692	0.0241
	1.2	0.0130	0.0286	0.0000	0.0125	0.0362	0.0293	0.0825	0.0290
	1.5	0.0321	0.0580	0.0000	0.0302	0.0701	0.0593	0.1316	0.0591
	2.0	0.0793	0.1159	0.0012	0.0781	0.1411	0.1180	0.2124	0.1174
	2.5	0.1502	0.1948	0.0063	0.1517	0.2289	0.1962	0.2958	0.1957
	4.0	0.4181	0.4464	0.1192	0.4537	0.5047	0.4412	0.5261	0.4408
4.0	1.0	0.0009	0.0049	0.0000	0.0010	0.0073	0.0051	0.0392	0.0051
	1.1	0.0015	0.0088	0.0000	0.0020	0.0132	0.0091	0.0557	0.0090
	1.2	0.0040	0.0156	0.0000	0.0039	0.0208	0.0161	0.0761	0.0161
	1.5	0.0148	0.0408	0.0000	0.0145	0.0534	0.0415	0.1424	0.0414
	2.0	0.0591	0.1180	0.0000	0.0602	0.1406	0.1192	0.2440	0.1192
	2.5	0.1582	0.2280	0.0005	0.1610	0.2544	0.2297	0.3378	0.2294
	4.0	0.4362	0.4850	0.0652	0.5066	0.5239	0.4840	0.5687	0.4809

Table 2: Estimated rejection rate by simulation, $\lambda_1 = 20$ (Asym = Asymptotic; Approx = Approximate; Boots = Bootstrap).

		W_2		W_3		F		L	
d	ρ	Asym	Boots	Asym	Boots	Approx	Boots	Asym	Boots
0.1	1.0	0.0674	0.0511	0.0621	0.0509	0.0582	0.0455	0.0497	0.0509
	1.1	0.0828	0.0630	0.0762	0.0638	0.0711	0.0586	0.0621	0.0635
	1.2	0.1098	0.0857	0.1015	0.0858	0.0959	0.0794	0.0835	0.0857
	1.5	0.1929	0.1543	0.1814	0.1546	0.1719	0.1434	0.1526	0.1546
	2.0	0.3716	0.3193	0.3570	0.3205	0.3421	0.3061	0.3166	0.3205
	2.5	0.5511	0.4990	0.5377	0.5008	0.5216	0.4865	0.4955	0.5008
	4.0	0.8883	0.8573	0.8825	0.8581	0.8702	0.8489	0.8546	0.8580
0.5	1.0	0.0514	0.0480	0.0505	0.0481	0.0505	0.0436	0.0505	0.0480
	1.1	0.0855	0.0801	0.0836	0.0800	0.0836	0.0752	0.0836	0.0800
	1.2	0.1256	0.1197	0.1247	0.1197	0.1247	0.1129	0.1247	0.1197
	1.5	0.3205	0.3088	0.3189	0.3092	0.3187	0.2989	0.3187	0.3092
	2.0	0.6988	0.6855	0.6960	0.6865	0.6949	0.6813	0.6948	0.6857
	2.5	0.9166	0.9099	0.9145	0.9103	0.9143	0.9080	0.9143	0.9102
	4.0	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
1.0	1.0	0.0498	0.0488	0.0478	0.0494	0.0514	0.0472	0.0514	0.0488
	1.1	0.0887	0.0876	0.0864	0.0880	0.0916	0.0836	0.0918	0.0875
	1.2	0.1503	0.1518	0.1470	0.1522	0.1552	0.1485	0.1553	0.1518
	1.5	0.4023	0.3989	0.4014	0.3994	0.4071	0.3933	0.4071	0.3989
	2.0	0.8353	0.8330	0.8347	0.8329	0.8360	0.8298	0.8360	0.8328
	2.5	0.9787	0.9779	0.9782	0.9779	0.9788	0.9776	0.9788	0.9779
	4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.5	1.0	0.0479	0.0503	0.0463	0.0504	0.0528	0.0480	0.0556	0.0499
	1.1	0.0963	0.0993	0.0931	0.0994	0.1024	0.0955	0.1042	0.0989
	1.2	0.1556	0.1604	0.1519	0.1606	0.1649	0.1551	0.1672	0.1604
	1.5	0.4513	0.4536	0.4462	0.4535	0.4581	0.4482	0.4621	0.4537
	2.0	0.8912	0.8917	0.8889	0.8917	0.8929	0.8889	0.8961	0.8917
	2.5	0.9931	0.9927	0.9927	0.9927	0.9934	0.9924	0.9935	0.9927
	4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2.0	1.0	0.0483	0.0497	0.0469	0.0496	0.0520	0.0491	0.0546	0.0497
	1.1	0.0933	0.0980	0.0918	0.0985	0.0989	0.0946	0.1053	0.0980
	1.2	0.1538	0.1623	0.1519	0.1625	0.1622	0.1592	0.1714	0.1623
	1.5	0.4917	0.4979	0.4874	0.4983	0.5039	0.4937	0.5115	0.4981
	2.0	0.9134	0.9157	0.9128	0.9157	0.9192	0.9144	0.9205	0.9158
	2.5	0.9954	0.9951	0.9953	0.9951	0.9956	0.9948	0.9957	0.9951
	4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0	0.0423	0.0468	0.0423	0.0472	0.0473	0.0465	0.0508	0.0467
	1.1	0.0940	0.1001	0.0940	0.1003	0.1004	0.0996	0.1080	0.0996
	1.2	0.1692	0.1822	0.1681	0.1821	0.1825	0.1784	0.1911	0.1816
	1.5	0.5400	0.5523	0.5341	0.5524	0.5572	0.5498	0.5670	0.5522
	2.0	0.9459	0.9499	0.9457	0.9499	0.9506	0.9497	0.9529	0.9499
	2.5	0.9991	0.9990	0.9990	0.9990	0.9992	0.9991	0.9992	0.9990
	4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3: Estimated p-value for the real data sets (Asym = Asymptotic; Approx = Approximate; Boots = Bootstrap).

	W_2		W_3		F		L	
Source of data	Asym	Boots	Asym	Boots	Approx	Boots	Asym	Boots
Boice and Monson [2]	0.019	0.011	0.020	0.010	0.017	0.012	0.016	0.012
Ray *et al.* [16]	0.000	0.001	0.000	0.001	0.000	0.001	0.000	0.001