2009

Homogeneity tests for several Poisson populations

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

Ling Wang

Follow this and additional works at: https://repository.hkbu.edu.hk/math JA

Part of the Mathematics Commons

This document is the authors' final version of the published article.
Link to published article: http://dx.doi.org/10.1016/j.csda.2009.05.017

APA Citation

This Journal Article is brought to you for free and open access by the Department of Mathematics at HKBU Institutional Repository. It has been accepted for inclusion in Department of Mathematics Journal Articles by an authorized administrator of HKBU Institutional Repository. For more information, please contact repository@hkbu.edu.hk.
Homogeneity tests for several Poisson populations

Sung Nok Chiu*, Ling Wang

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

Abstract

In this paper we compare the size distortions and powers for Pearson’s $\chi^2$-statistic, likelihood ratio statistic $LR$, score statistic $SC$ and two statistics, which we call $UT$ and $VT$ here, proposed by Potthoff and Whittinghill (1966) for testing the equality of the rates of $K$ Poisson processes. Asymptotic tests and parametric bootstrap tests are considered. It is found that the asymptotic $UT$ test is too conservative to be recommended, while the other four asymptotic tests perform similarly and their powers are close to those of their parametric bootstrap counterparts when the observed counts are large enough. When the observed counts are not large, Monte Carlo simulation suggested that the asymptotic test using $SC$, $LR$ and $UT$ statistics are discouraged; none of the parametric bootstrap tests with the five statistics considered here is uniformly best or worst, and the asymptotic tests using Pearson’s $\chi^2$ and $VT$ always have similar powers to their bootstrap counterparts. Thus, the asymptotic Pearson’s $\chi^2$ and $VT$ tests have an advantage over all five parametric bootstrap tests in terms of their computational simplicity and convenience, and over the other four asymptotic tests in terms of their powers and size distortions.

Keywords: Poisson process, homogeneity, asymptotic $\chi^2$-test, parametric bootstrap, count data.

* Corresponding author: Tel: +852 3411 7020; fax: +852 3411 5811.

E-mail address: snchiu@hkbu.edu.hk
1 Introduction

To analyze the patient survival after heart valve replacement operations, Laird and Olivier (1981) observed a sample of 109 patients, who were classified by valve (aortic, mitral) and by age (<55, ≥55). For younger subjects in the sample, 4 deaths in 1259 months of observation occurred with aortic valve replacement, 1 death in 2082 months of observation occurred with mitral valve replacement, while for older subjects, 7 deaths in 1417 months of observation occurred with aortic valve replacement, and 9 deaths in 1647 months of observation occurred with mitral valve replacement. This example, in which each patient was observed (at risk) until he died or the study ended, calls for statistical tests of homogeneity for several count data, each of which is the total number of occurrences in a different group with a not-necessarily-equal observation period. Such homogeneity tests can be applied to the analysis of regional count data that arise in public health studies; because of confidentiality restrictions, official agencies often release disease, census, or other data as summary counts for regions (e.g., counties) that partition the study area. Examples of such data can be found in Lawson (2006, pp. 19–24) and Waller and Gotway (2004, Chapter 7).

Without the detailed temporal or spatial information of the individual occurrences, such count data are typically modeled by the Poisson distributions. If we have $K$ independent Poisson variates, which are counts observed from $K$ time intervals or spatial regions with fixed but not necessarily equal length or area, the very first question of interest is whether the rates or the intensities, which are the average numbers of occurrences per unit time or area, of the $K$ underlying temporal or spatial processes are the same or not.

The most natural way to test the homogeneity for several Poisson count data is to use Pearson’s $\chi^2$-test (Hoel, 1945). Potthoff and Whittinghill (1966) proposed two other statistics having asymptotic $\chi^2$-distributions for testing the homogeneity. In particular, when $K = 2$, there are quite a few other statistics available; see the power comparison studies in Chiu (2009), Ng et al. (2007), and Ng and Tang (2005). If the two counts are observed in equal length intervals or equal area regions, there are even more statistics, e.g., Best (1975), and Detre and White (1970). However, to the best of our knowledge, there is no power comparison
study reported for Pearson’s $\chi^2$-statistic and the two statistics proposed by Potthoff and Whittinghill (1966). The aim of this paper is to perform such a power comparison, in which we also include the likelihood ratio test and the score test derived particularly here for testing homogeneity.

In addition to the asymptotic tests, we would also consider the parametric bootstrap tests for these statistics. Bootstrap, formally introduced in the seminal paper by Efron (1979), is a widely used technique to generate Monte Carlo resamples from the original observed sample so that a functional of a unknown population distribution function can be approximated by the average of the values of the corresponding statistic computed from individual resamples; see the monographs by Davison and Hinkley (1997), Efron and Tibshirani (1993), and Hall (1992). In particular, a bootstrap test is a testing procedure in which the P-value of the observed test statistic is approximated by bootstrap. Such a test is first introduced in Barnard (1963) and then further developed by Beran (1986), Davidson and MacKinnon (1999, 2000, 2006), Hall and Titterington (1989), Hope (1968), Marriott (1979), and Zhu (2005). Chiu (2009) recommended to use parametric bootstrap tests instead of asymptotic tests for testing the equality of the rates of two Poisson processes. Therefore, in this paper, we also report the sizes and powers of the bootstrap tests.

In Section 2, we introduce our five test statistics for detecting inhomogeneity of $K$ Poisson variates. Section 3 explains how the parametric bootstrap tests would be carried out in our context and then Monte Carlo simulation results on size distortion, power and robustness are reported and discussed in Section 4. Finally, in Section 5, we apply the tests to the data of patient survival after heart valve replacement operations mentioned above.

2 Test Statistics and their Asymptotic Distributions

Suppose we observe $K$ independent random variables $x_1, x_2, \ldots, x_K$, and each $x_i$ follows a Poisson distribution with mean $\lambda_i t_i$, in which $t_i$ is known but $\lambda_i$ is not. That is, $x_i$ is the observed number of occurrences of a temporal (or spatial, respectively) Poisson process with
rate (or intensity) $\lambda_i$ in a sampling frame of length (or area) $t_i$, $i = 1, 2, \ldots, K$. Hence, the density function of $x_i$ is,

$$f(x_i) = \frac{e^{-\lambda_i t_i} (\lambda_i t_i)^{x_i}}{x_i!}, \text{ for } i = 1, 2, \ldots, K.$$  

The maximum likelihood estimate (MLE) of $\lambda_i$ is simply the empirical rate $\hat{\lambda}_i = x_i/t_i$. By the homogeneity test of $K$ independent Poisson variates, we mean a test for:

$$H_0: \lambda_1 = \cdots = \lambda_K \text{ against } H_A: \lambda_i \neq \lambda_j \text{ for some } i, j \in \{1, \ldots, K\}.$$  

In the following, we introduce five statistics for this homogeneity test. By saying asymptotic or large-sample we consider the limiting scenario in which the means of the Poisson distributions go to infinity, because in our context a large sample (of a point process) comes from a long observation period, leading to a large Poisson mean; if the sampling intervals are of fixed lengths, the limiting scenario is then equivalent to the one in which the rates go to infinity.

Because of independence, the likelihood function is simply

$$L(\lambda_1, \ldots, \lambda_K|x_1, \ldots, x_K) = e^{-\sum_{i=1}^{K} \lambda_i t_i} \prod_{i=1}^{K} (\lambda_i t_i)^{x_i} / \prod_{i=1}^{K} x_i!.$$  

Under the homogeneity null hypothesis, all $\lambda_i$ are the same. Let $\lambda$ denote the common value and the MLE of $\lambda$ is

$$\hat{\lambda} = \frac{\sum_{i=1}^{K} x_i}{\sum_{i=1}^{K} t_i}.$$  

Pearson’s $\chi^2$-test is the original and most widely used $\chi^2$-test. It tests a null hypothesis that the relative frequencies of occurrence of observed events follow some specified frequency distribution, and the test statistic in our context is (see equation (11) in Potthoff and Whittinghill, 1966)

$$\text{Pearson’s } \chi^2 = \sum_{i=1}^{K} \frac{(x_i - \hat{\lambda} t_i)^2}{\hat{\lambda} t_i} = \sum_{i=1}^{K} t_i \frac{x_i^2}{\sum_{i=1}^{K} t_i} - \sum_{i=1}^{K} x_i,$$

which has an asymptotic $\chi^2$-distribution with $K - 1$ degrees of freedom if the homogeneity hypothesis is true. We take the convention that $0/0 = 0$, in case the sum of $x_i$ equals 0.
Because the likelihood function is explicitly known, we can obtain the likelihood ratio
\[ \Lambda = \frac{L(\hat{\lambda}, \ldots, \hat{\lambda}|x_1, \ldots, x_K)}{L(\hat{\lambda}_1, \ldots, \hat{\lambda}_K|x_1, \ldots, x_K)}. \]

For commonly occurring families of probability distributions, such as Poisson distributions we discussed here, the likelihood ratio test statistic
\[ LR = -2 \ln \Lambda = 2 \left( \sum_{i=1}^{K} x_i \ln \frac{x_i}{t_i} - \sum_{i=1}^{K} x_i \ln \frac{\sum_{i=1}^{K} x_i}{\sum_{i=1}^{K} t_i} \right), \]

with the convention that \(0 \ln 0 = 0\), has an asymptotic \(\chi^2\)-distribution with degrees of freedom \(K - 1\), if the homogeneity hypothesis is true. The particular case that \(t_1 = \cdots = t_K = 1\) has already been considered in Rao and Chakravarti (1956), but we are not aware if the above general form in (1) that allows different \(t_i\) has ever been discussed in the literature.

Another test that utilizes the likelihood function is the score test, which is often simpler than the likelihood ratio test because the score statistic requires parameter estimators to be obtained only under the null hypothesis. More precisely, the score test statistic is
\[ SC = u(\hat{\lambda}, \ldots, \hat{\lambda})' I(\hat{\lambda}, \ldots, \hat{\lambda})^{-1} u(\hat{\lambda}, \ldots, \hat{\lambda}) = \left( \sum_{i=1}^{K} \frac{x_i}{\sum_{i=1}^{K} t_i} \right)^2 \sum_{i=1}^{K} \frac{t_i^2}{x_i} - \sum_{i=1}^{K} x_i, \]

where
\[
\begin{align*}
u(\lambda_1, \ldots, \lambda_K) &= \left( \frac{\partial \ln L(\lambda_1, \ldots, \lambda_K|x_1, \ldots, x_K)}{\partial \lambda_1}, \ldots, \frac{\partial \ln L(\lambda_1, \ldots, \lambda_K|x_1, \ldots, x_K)}{\partial \lambda_K} \right)', \\
I(\lambda_1, \ldots, \lambda_K) &= \left[ -\frac{\partial^2 \ln L(\lambda_1, \ldots, \lambda_K|x_1, \ldots, x_K)}{\partial \lambda_i \partial \lambda_j} \right]_{i,j=1,\ldots,K}.
\end{align*}
\]

To avoid the problem of division by zero, we add 0.5 to every \(x_i\) whenever any one of them is zero. Under the homogeneity hypothesis, \(SC\) has an asymptotic \(\chi^2\)-distribution with \(K - 1\) degrees of freedom. Ng and Cook (1999) investigated score tests of homogeneity of Poisson processes in which, unlike in the present application, the times of occurrence of the recurrent event are available, and we believe that applying the above score statistic in (2) to test the homogeneity of \(K \geq 3\) Poisson count data with different \(t_i\) has not been considered in the literature.
The above three tests are in fact applications of general classes of tests to a particular problem. Potthoff and Whittinghill (1966) proposed two test statistics specifically designated for testing the homogeneity of several Poisson count data, and we state the formulae below.

The first one, denoted by $VT$ here, is based on the statistic

$$V = \sum_{i=1}^{K} t_i \sum_{i=1}^{K} x_i (x_i - 1)/t_i,$$

where $[V - (\sum_{i=1}^{K} x_i)(\sum_{i=1}^{K} x_i - 1)]/\sqrt{2(K - 1)(\sum_{i=1}^{K} x_i)(\sum_{i=1}^{K} x_i - 1)}$ follows the standard normal distribution asymptotically under the null hypothesis. By choosing proper constants $e = 2(K - 1)/[\sum_{i=1}^{K} t_i \sum_{i=1}^{K} 1/t_i - 3K + 2 + 2(K - 1)(\sum_{i=1}^{K} x_i - 2)]$ and $f = e[(K - 1)e - 1]/[\sum_{i=1}^{K} x_i)(\sum_{i=1}^{K} x_i - 1)]$, we can have a more refined approximation

$$VT = eV + f \sim \chi^2_{\nu_1},$$

where the degrees of freedom $\nu_1 = e^2(K - 1)(\sum_{i=1}^{K} x_i)(\sum_{i=1}^{K} x_i - 1)$. The constants $e$ and $f$ are determined to let $VT$ and $\chi^2_{\nu_1}$ have the same first three moments.

The other one, denoted by $UT$ here, is also an asymptotic $\chi^2$-test based on another statistic $U = \sum_{i=1}^{K} x_i^2 - \sum_{i=1}^{K} x_i - 2\lambda \sum_{i=1}^{K} t_i x_i$, which is proved to be the locally most powerful test for the null hypothesis against a certain selected alternative if the value $\lambda$ is known. Under the null hypothesis, $(U + \lambda^2 \sum_{i=1}^{K} t_i^2)/\sqrt{2\lambda^2 \sum_{i=1}^{K} t_i^2}$ follows the standard normal distribution asymptotically. Again, by choosing proper constants $g = \sum_{i=1}^{K} t_i^2/[(\sum_{i=1}^{K} t_i^2)/2 + \lambda \sum_{i=1}^{K} t_i^2]$ and $h = g(g + 1)\lambda^2 \sum_{i=1}^{K} t_i^2$, we have a more refined approximation

$$UT = gU + h \sim \chi^2_{\nu_2},$$

where the degrees of freedom $\nu_2 = g^2 \lambda^2 \sum_{i=1}^{K} t_i^2$, and $g$ and $h$ are chosen so that $UT$ and $\chi^2_{\nu_2}$ have the same first three moments. In our context, we do not know $\lambda$ and we follow Potthoff and Whittinghill’s approach, i.e., replace it by $\lambda^*$, where $\lambda^* = \sqrt{\frac{\sum_{i=1}^{K} x_i^2 - \sum_{i=1}^{K} x_i}{\sum_{i=1}^{K} t_i^2}}$, which is the value of $\lambda$ that minimizes $(U + \lambda^2 \sum_{i=1}^{K} t_i^2)/\sqrt{2\lambda^2 \sum_{i=1}^{K} t_i^2}$. Potthoff and Whittinghill (1966) suspected, and we would confirm this in our simulation in Section 4, that $UT$ would be generally conservative, and might sometimes be too conservative in smaller experiments as to entail a marked loss of power, a handicap not shared by Pearson’s $\chi^2$ and $VT$. 

6


3 Parametric Bootstrap Tests

The $\chi^2$-distributions of the five statistics in the previous section are valid only asymptotically, and hence a test using the P-value calculated by the $\chi^2$-distribution may encounter with size distortion. Here we also consider the parametric bootstrap test (Davison and Hinkley, 1997, pp. 148–149), in which we approximate the P-value from Monte Carlo resamples generated according to the homogeneity hypothesis in the way described briefly as follows.

We first calculate a statistic value based on the observed counts and generically denote it by $\tau$. Then we generate a Monte Carlo resample according to the null hypothesis using the MLE $\hat{\lambda}$, i.e., draw $K$ independent Poisson variates with mean $\hat{\lambda}t_i$, $i = 1, 2, \ldots, K$, respectively and calculate the value of the statistic; denote the value by $\tau^*_1$. This sample generation process will be done independently $R$ times in total, so that we have $\tau^*_1, \tau^*_2, \ldots, \tau^*_R$. Then, the bootstrap P-value is given by (Davison and Hinkley, 1997, p. 148)

$$
\text{bootstrap P-value} = \frac{\#\{i : \tau^*_i \geq \tau\} + 1}{R + 1}.
$$

(3)

The null hypothesis will be rejected if the bootstrap P-value is less than or equal to the nominal significance level $\alpha$. Some authors (e.g. Davidson and MacKinnon, 1998, 2007) do not add 1 to the numerator and the denominator of the ratio in (3), but here we do so because it will then correspond to the original testing procedure suggested in Barnard (1963), which admits exact significance levels when there is no nuisance parameter (see e.g. Davison and Hinkley, 1997, pp. 140-141, Hall and Titterington, 1989, and Zhu, 2005, pp. 1-2).

Hope (1968) showed that when there is no nuisance parameter, the power loss, compared with the corresponding uniformly most powerful test, resulting from using the parametric bootstrap tests is slight and so $R$ is not necessarily large. Marriott (1979) suggested that for $\alpha = 0.05$, $R = 99$ is adequate, whilst Davison and Hinkley (1997, 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.
4 Simulation

To compare the powers of these statistics, using P-values approximated by their asymptotic distributions or by bootstrap, we carry out a Monte Carlo study and estimate the rejection rates by 10,000 simulations, and so for each sample proportion of rejection the standard error is at most 0.005. The rejection rate is just the size if the null hypothesis is true and is the power otherwise. As mentioned in the previous section, when $\alpha = 0.05$, choosing $R = 999$ in the parametric bootstrap tests should be quite sufficient in view of the loss of power and so has been adopted in this study.

There are quite a few parameters in this Monte Carlo study, the number of Poisson variates $K$, the lengths (or areas, respectively) of the sampling intervals (or regions) $t_1, \ldots, t_K$, and the rates $\lambda_1, \ldots, \lambda_K$. We have considered many different values for $K$, but report here only the cases $K = 5, 10$ and $50$, because, according to our simulation, these three can represent the small, intermediate and large values cases, respectively. Without loss of generality we can always let one of the $t_i$ equal 1.

For $K > 3$, we consider $K - N$ processes with the same rate $\lambda$, whilst the remaining $N$ processes have rates different from $\lambda$. In our study we found that the ranks of the tests, ranked according to their powers, depend mostly on the maximum difference between $\lambda_i$. Therefore, we believe that it is fair enough to restrict ourselves to $N = 2$ in this study. (This restriction, as it turns out below, does not matter at all because our major conclusion is that none of these test statistics is uniformly best nor worst among the cases we considered; including more cases would by no means refute this conclusion.) We report the cases $\lambda = 10, 100$ and 1000, representing small, intermediate and large sample cases.

Moreover, we consider three scenarios for the lengths of the sampling intervals. In the first scenario we have equal $t_i$, and all are 1. In the second scenario, the lengths $t_i$ of sampling intervals are all different, and the two processes with different rates are observed in the longest two intervals. In contrast, in the last scenario, the two count data with different rates are observed in the shortest two intervals, and all $t_i$ are still all different.

A major difficulty in this comparison study is that, even under the above restrictions,
we still have a huge number of possible combinations of parameter values. We chose representative cases and reported the results by the P-value plots suggested by Davidson and MacKinnon (1998) in the following sections. Nevertheless, the observations below, from which the conclusions are drawn, are applicable to not only these few cases but also other cases in our much more extensive simulation.

4.1 Size distortion

In Figure 1, all the $K$ variates have the same $\lambda$, so the rejection rate here is just the actual size, where $t_i = 1 + \frac{i-1}{K-1}$. Following Ng and Tang (2005), we call a test conservative and liberal, respectively, if the actual size is 20% lower and higher than the nominal level, and call it robust otherwise.

We can see that the actual sizes of the asymptotic tests using $UT$ are always smaller than the others and, unless $K$ is large, much smaller than the nominal levels; the price to pay for conservativeness is a loss in power and so such a conservative test might be a better choice only if it also had higher powers than other tests at the same nominal level, but it is not the case for $UT$ as we can see in the next section.

On the other hand, the actual sizes of the asymptotic score test would be much higher than the nominal levels when $\lambda$ is not large, leading to liberalness, the problem of which is that a rejection conclusion by a liberal test should be taken with caution because such a testing procedure would reject a correct null hypothesis much more frequently than we expect to see at the nominal level.

For small $\lambda$ and large $K$, the asymptotic likelihood ratio test, though may be still robust, also has actual sizes larger than the nominal levels.

The actual sizes of Pearson’s $\chi^2$ and $VT$ are always very close to the nominal levels in the cases considered.

Though we have not shown the P-value plots here, as expected, the sizes of the parametric bootstrap tests with all these five statistics are pretty close to the nominal levels, no matter what the values of $K$ and $\lambda$ are.
Thus, in view of their size distortions, we would not recommend the asymptotic $\chi^2$-tests using $UT$ and $SC$, unless $\lambda$ and $K$ are large. The same conclusion can be drawn when all $t_i = 1$.

4.2 Power

Davidson and MacKinnon (1998) suggested that the power can be plotted against the actual size by performing two experiments using the same sequence of random numbers from which the data are generated according to an alternative model and the null model that minimizes the Kullback–Leibler divergence from the alternative model (see also Davidson and MacKinnon, 2006). However, as we mentioned in the previous section, we would not recommend tests that have serious size distortion, and so in this paper we simply compare the powers of the test at the same nominal levels, which unlike actual sizes, can be controlled.

Figures 2–7 are plots of the power against the nominal level and are arranged in the following way. Figures 2 and 3 are the powers in the first scenario described above, while Figures 4 and 5 consider the second scenario, and Figures 6 and 7 the third one. From these figures and numerical values (not reported here) we have the following observations.

First, we should discard the asymptotic test with $UT$. We have already seen in the previous section that this test is conservative and the price to pay is a substantial loss in power for $K = 5$ or $10$, as shown in Figures 2(a), 3(a), 6(a) and 7(a). This agrees with the comment by Potthoff and Whittinghill (1966). However, it is interesting to see that even it is conservative, its power is not much lower and sometimes is even higher than other asymptotic tests for large $K$ in Figures 4(a) and 5(a), corresponding to the second scenario, in which the two processes with different rates are observed in the longest two intervals. However, in practice $K$ is typically not as large as 50, and more importantly, the heterogeneity alternative hypothesis in general does not specify which processes are of different rates, and so whether processes with different rates, if exist, are observed in longer or shorter intervals are unknown. Therefore, overall speaking we would not encourage the use of the asymptotic test using $UT$. 
Second, consider Pearson’s $\chi^2$ and $VT$. Even when $\lambda$ is not large, the powers of the asymptotic test using Pearson’s $\chi^2$-statistic are always close to those of the corresponding bootstrap test, and it is also the case when the test statistic is $VT$. Moreover, the powers of Pearson’s $\chi^2$ and $VT$ are very close, no matter what the values of $K$ and $\lambda$ are.

Third, consider the likelihood ratio $LR$ and score $SC$ test statistics. When $\lambda$ is small, the sizes of the asymptotic test using $SC$ is much higher than the others, and much higher than the nominal level $\alpha$. Such a liberal test may commit the type I error much more frequently than acceptable and so would not be recommended. The sizes of the asymptotic $LR$ test are often higher than the asymptotic Pearson’s $\chi^2$-test. However, numerical values corresponding to Figure 2(a) show that the powers of the former are always not larger than the powers of the latter. Therefore, even though in other figures the asymptotic $LR$ test sometimes (but not always) has a higher power, at the price of larger size, than the asymptotic Pearson’s $\chi^2$-test, we do not consider it as appealing as Pearson’s $\chi^2$ in general.

Fourth, consider large $\lambda$. When all $t_i$ are the same (see Figures 2 and 3), except the asymptotic $UT$ test, the rejection rates of the other four asymptotic tests as well as all five parametric bootstrap tests are often very close to each other. However, this is not the case when $t_i$ are not the same. For the second scenario, i.e., the two processes with different rates are observed in the longest intervals (see Figures 4 and 5), the bootstrap test using $UT$ is more powerful than all the other tests (and this is also true for small and intermediate values of $\lambda$), whilst in the third scenario (Figures 6 and 7), the $UT$ statistic is the least powerful among these five statistics. Moreover, the bootstrap test with $UT$, as well as other four statistics, has very small size distortion. What is subtler is that if in the shortest two intervals the observed counts came from processes with larger rates (Figure 6), Pearson’s $\chi^2$ and $VT$ (asymptotic and bootstrap tests of them are more or less equally powerful) are the most powerful, and $SC$ is only better than the poorest $UT$ statistic, but if in the shortest intervals the observed counts came from processes with smaller rates (Figure 7), the score statistic $SC$, either asymptotic or bootstrap, is always the most powerful.

Fifth, consider small and intermediate $\lambda$. We already disapproved the overly liberal
asymptotic score test and the less appealing likelihood ratio test in our third observation above. We also discarded the overly conservative asymptotic UT test in our first observation. Moreover, in our second observation we concluded that the asymptotic test and the bootstrap test have very close powers if the test statistic is Pearson’s $\chi^2$ or $VT$. Therefore, for $\lambda$ that is not large, we may simply compare the powers of these five statistics in the parametric bootstrap test setting. In Figures 2(b), 4(b) and 6(b), i.e., when there are two processes with larger rates than the other $K - 2$ processes, the score test is usually the least powerful; in contrast, in Figures 3(b) and 7(b), where there are two processes with smaller rates than the other $K - 2$ processes, it is the most powerful. However, we cannot conclude that the score test is more powerful in detecting the existence of some smaller rates because in Figure 5(b), where there are also two processes with smaller rates than the other $K - 2$ processes, the most powerful statistic is UT and the score test is, surprisingly, the least powerful if $K = 5$, but its performance improves as $K$ increases, and becomes the second powerful when $K = 50$. On the other hand, the performance of Pearson’s $\chi^2$ and $VT$ behaves exactly in the opposite manner. They are the most powerful in Figures 2(b) and 6(b), but the least powerful in Figure 3(b) and the second least powerful in Figure 7(b); in Figure 5(b) they are the least powerful if $K$ is large but becomes more powerful when $K$ decreases. The LR statistic performs similarly to the score statistic $SC$; when $SC$ is the best, the LR statistic is the second best, and when $SC$ is the worst, LR statistic is the second worst. Unlike the other four, the performance of UT does not depend on whether the rates of the two different processes are larger or smaller than the others but depends on whether the processes of the different rates are observed in the longest or shortest intervals. The UT statistic is the best or the second best if the different processes are observed in the longest intervals and is the worst or the second worst if observed in the shortest intervals. However, in practice the actual scenario will not be specified in the heterogeneity hypothesis, and hence practitioners would not be able to tell which statistic would be the most appropriate choice.
4.3 Robustness

Saha and Bilisoly (2009) compared several homogeneity test for count data sampled from negative binomial distributions, which are usually used to model data that exhibit more variation than they should under the Poisson assumption. In this section we report the robustness of the above five statistics without any adjustment for overdispersion when the true data generating process for \( x_i \) is the negative binomial distribution with mean \( \lambda_i \) and variance \((1 + \psi)\lambda_i\). The bootstrap samples would still be generated according to the Poisson model.

It is intuitively clear that the extra variation among the negative binomial count data will increase the rejection rate of each test statistic and our simulation confirmed this.

For the size distortion, our concern is that whether the test statistics which are neither conservative nor liberal (i.e. Pearson’s \( \chi^2 \), LR and VT for the asymptotic test and all five statistics for the bootstrap test) under the Poisson model would deviate more than 20% from the nominal level if \( \psi > 0 \). When \( \psi = 0.01 \), the sizes of all five statistics, using asymptotic or bootstrap test, are pretty much the same as the corresponding sizes under the Poisson model. However, when \( \psi = 0.05 \), the asymptotic test using Pearson’s \( \chi^2 \), LR or VT statistic becomes liberal in all nine combinations of the values of \( K \) and \( \lambda \) considered, and so does the bootstrap test using any one of the five statistic, except in the case that \( K = 5 \) and \( \lambda = 10 \); when \( \psi = 0.1 \), even the bootstrap test becomes liberal in all, including \( K = 5 \) and \( \lambda = 10 \), cases.

For the power, we found that for \( \psi = 0.01 \), 0.05 and 0.1, the pattern of the power curves remains the same as that for the Poisson model. Consequently, the observations and hence the conclusions stated in the previous section are still valid.

Thus, the five statistics considered here have the same level of robustness with respect to overdispersion.
4.4 Conclusion

In summary, if $\lambda$ is large, i.e., when the observed counts are large, we may use the asymptotic test with any one of Pearson’s $\chi^2$, $LR$, $SC$ and $VT$ statistics but not the overly conservative asymptotic $UT$ test. If $\lambda$ is not large, i.e., the observed counts are not large, none of the five test statistics, using the parametric bootstrap test setting, is uniformly best nor worst. However, Pearson’s $\chi^2$ and $VT$ statistics have a desirable advantage, namely, their computationally simpler and more convenient asymptotic tests, because their size distortions are mild, have powers similar to their bootstrap counterparts even when $\lambda$ is not large, while the asymptotic test using $SC$, $LR$ and $UT$ statistics are disapproved. All statistics considered are robust to mild overdispersion.

5 Real Data

Consider the survival data described in Section 1. Here, $K = 4$ is small, and all counts are less than 10, which means we are considering a small $K$ and small $\lambda$ case. The P-values of the ten tests considered in this paper are reported in Table 1, where all P-values are less than 0.05. The overly liberal asymptotic score test gives the smallest P-value 0.00, whilst the overly conservative asymptotic $UT$ test gives the largest P-value 0.045, which is still less than 0.05. Thus, we have strong evidence against the homogeneity hypothesis at the 0.05 significance level and conclude that the rates are affected by the value and/or the age.

Acknowledgements

We thank the referees for their helpful suggestions, especially for suggesting the use of graphical instead of tabular presentations. This research was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region (Project Numbers HKBU200605 and HKBU200807) and an FRG grant of the Hong Kong Baptist University.
References


Table 1: Estimated P-value for the data in Laird and Olivier (1981) (Asym = Asymptotic; Boot = Bootstrap, $R = 999$).

<table>
<thead>
<tr>
<th></th>
<th>Pearson’s $\chi^2$</th>
<th>$LR$</th>
<th>$SC$</th>
<th>$VT$</th>
<th>$UT$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asym</td>
<td>Boot</td>
<td>Asym</td>
<td>Boot</td>
<td>Asym</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>0.036</td>
<td>0.021</td>
<td>0.013</td>
<td>0.017</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figure 1
Figure 2(a)
Figure 2(b)
Figure 3(a)
Figure 3(b)
Figure 4(a)
Figure 4(b)
Figure 5(a)
Figure 5(b)
Figure 6(a)
Figure 6(b)
Figure 7(a)
Figure 7(b)
Figure 1: Actual size of the asymptotic test, estimated by 10,000 simulations, vs nominal level, where $t_i = 1 + (i - 1)/(K - 1)$ for all $i = 1, 2, \ldots, K$. The solid lines are reference lines with slopes 0.8, 1 and 1.2, indicating whether the tests are conservative, liberal or robust.

Figure 2: Power, estimated by 10,000 simulations, vs nominal level, where $t_i = 1$ for all $i = 1, 2, \ldots, K$ and $(\lambda_1, \ldots, \lambda_K) = (\lambda, \ldots, \lambda, \lambda(1 + \rho_1), \lambda(1 + \rho_2))$, in which $(\rho_1, \rho_2) = (0.7, 1.0)$ when $\lambda = 10$, (0.15, 0.30) when $\lambda = 100$, and (0.05, 0.07) when $\lambda = 1000$, for (a) asymptotic tests, and (b) bootstrap tests with $R = 999$.

Figure 3: Power, estimated by 10,000 simulations, vs nominal level, where $t_i = 1$ for all $i = 1, 2, \ldots, K$ and $(\lambda_1, \ldots, \lambda_K) = (\lambda, \ldots, \lambda, \lambda(1 + \rho_1), \lambda(1 + \rho_2))$, in which $(\rho_1, \rho_2) = (0.7, 1.0)$ when $\lambda = 10$, (0.15, 0.30) when $\lambda = 100$, and (0.05, 0.07) when $\lambda = 1000$, for (a) asymptotic tests, and (b) bootstrap tests with $R = 999$.

Figure 4: Power, estimated by 10,000 simulations, vs nominal level, where $t_i = 1 + (i - 1)/(K - 1)$ for $i = 1, 2, \ldots, K$ and $(\lambda_1, \ldots, \lambda_K) = (\lambda, \ldots, \lambda, \lambda(1 + \rho_1), \lambda(1 + \rho_2))$, in which $(\rho_1, \rho_2) = (0.7, 1.0)$ when $\lambda = 10$, (0.15, 0.30) when $\lambda = 100$, and (0.05, 0.07) when $\lambda = 1000$, for (a) asymptotic tests, and (b) bootstrap tests with $R = 999$. 
Figure 5: Power, estimated by 10,000 simulations, vs nominal level, where \( t_i = 1 + (i - 1)/(K - 1) \) for \( i = 1, 2, \ldots, K \) and \((\lambda_1, \ldots, \lambda_K) = (\lambda, \ldots, \lambda, \lambda/(1 + \rho_1), \lambda/(1 + \rho_2))\), in which \((\rho_1, \rho_2) = (0.7, 1.0)\) when \( \lambda = 10 \), \((0.15, 0.30)\) when \( \lambda = 100 \), and \((0.05, 0.07)\) when \( \lambda = 1000 \), for (a) asymptotic tests, and (b) bootstrap tests with \( R = 999 \).

Figure 6: Power, estimated by 10,000 simulations, vs nominal level, where \( t_i = 2 - (i - 1)/(K - 1) \) for \( i = 1, 2, \ldots, K \) and \((\lambda_1, \ldots, \lambda_K) = (\lambda, \ldots, \lambda, \lambda/(1 + \rho_1), \lambda/(1 + \rho_2))\), in which \((\rho_1, \rho_2) = (0.7, 1.0)\) when \( \lambda = 10 \), \((0.15, 0.30)\) when \( \lambda = 100 \), and \((0.05, 0.07)\) when \( \lambda = 1000 \), for (a) asymptotic tests, and (b) bootstrap tests with \( R = 999 \).

Figure 7: Power, estimated by 10,000 simulations, vs nominal level, where \( t_i = 2 - (i - 1)/(K - 1) \) for \( i = 1, 2, \ldots, K \) and \((\lambda_1, \ldots, \lambda_K) = (\lambda, \ldots, \lambda, \lambda/(1 + \rho_1), \lambda/(1 + \rho_2))\), in which \((\rho_1, \rho_2) = (0.7, 1.0)\) when \( \lambda = 10 \), \((0.15, 0.30)\) when \( \lambda = 100 \), and \((0.05, 0.07)\) when \( \lambda = 1000 \), for (a) asymptotic tests, and (b) bootstrap tests with \( R = 999 \).