ORIGINAL RESEARCH PAPER

Testing Poissonity of a large number of populations

M. D. Jiménez-Gamero[1] · **J. de Uña-Álvarez**[2]

Abstract

This paper studies the problem of simultaneously testing that each of k samples, coming from k count variables, were all generated by Poisson laws. The means of those populations may differ. The proposed procedure is designed for large k, which can be bigger than the sample sizes. First, a test is proposed for the case of independent samples, and then the obtained results are extended to dependent data. In each case, the asymptotic distribution of the test statistic is stated under the null hypothesis as well as under alternatives, which allows to study the consistency of the test. Specifically, it is shown that the test statistic is asymptotically free distributed under the null hypothesis. The finite sample performance of the test is studied via simulation. A real data set application is included.

Keywords Goodness-of-fit · Count data · High-dimensional data

Mathematics Subject Classification 62G10 · 62G09

1 Introduction

Univariate count data appear in many real life situations and the Poisson distribution is frequently used to model this kind of data. Testing the goodness-of-fit of given observations with a probabilistic model is a crucial aspect of data analysis. Because of these reasons, a number of authors have proposed tests for the Poisson law. Most papers on this issue deal with testing Poissonity for a sample, and the properties of the proposed procedures are studied as the sample size increases. This paper studies the problem of simultaneously testing Poissonity of k samples, where k can be larger than the sample sizes.

✉ M. D. Jiménez-Gamero
   dolores@us.es

[1] Department of Statistics and Operations Research, Universidad de Sevilla, Seville, Spain

[2] CINBIO, Universidade de Vigo, SiDOR Research Group, Vigo, Spain

Let $\mathbf{X}_1 = \{X_{1,1}, \ldots, X_{1,n_1}\}, \ldots, \mathbf{X}_k = \{X_{k,1}, \ldots, X_{k,n_k}\}$ be k independent samples with sizes $n_1, \ldots, n_k$, which may be different, where $X_{i,1}, \ldots, X_{i,n_i}$ are independent and identically distributed (i.i.d.) coming from $X_i \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$, $1 \leq i \leq k$. In this setting, we deal with the problem of testing

$$H_{0,k} : X_i \sim Pois(\lambda_i), \text{ for some } \lambda_i > 0, \ 1 \leq i \leq k,$$

were $X \sim Pois(\lambda)$ means that the random variable X has a Poisson law with mean λ, against general alternatives,

$$H_{1,k} : X_i \nsim Pois(\lambda), \ \forall \lambda > 0, \text{ for some } 1 \leq j \leq k,$$

where k is allowed to be large (the precise meaning of "large" will be stated in the following sections). Notice that, under $H_{0,k}$, the parameter of each Poisson law may vary across the k populations.

Testing $H_{0,k}$ is relevant in many settings. Just to mention an important example, sequencing experiments performed in biological sciences usually report long sequences of read counts, corresponding to a large number of genes or DNA bases; see for instance Jiménez-Otero et al. (2019) and references theirein. Such counts are typically analyzed under the umbrella of the Poisson model, which has been found particularly well-suited for certain sequencing experiments (see Lander and Waterman 1988). The validity of these analyses critically depends on the goodness-of-fit of the Poisson distribution to the count data at hand, and here is where our testing problem comes into play.

The problem of simultaneously testing goodness-of-fit for k populations have been studied in Gaigall (2021) by using test statistics based on comparing the empirical distribution function of each sample with a parametric estimator derived under the null hypothesis. The asymptotic properties studied in Gaigall (2021) are for continuous distribution functions, fixed k and increasing sample sizes. The tests in Gaigall (2021) are sums of goodness-of-fit test statistics for each of the k populations. For the problem of simultaneously testing normality of a large number of populations, Jiménez-Gamero (2023) considered several test statistics, obtaining best overall results for one based on a sum of statistics for testing normality in each population. Here we will also study a test based on a statistic of this type.

The paper is organized as follows. Section 2 briefly reviews Poissonity tests for one sample. The review is restricted to consistent tests, with special emphasis on the test in Baringhaus and Henze (1992). Section 3 studies a test statistic that is based on sum of statistics for one-sample Poissonity testing. Specifically, the one-sample statistic considered is a modification of the one proposed in Baringhaus and Henze (1992). We start by assuming that the samples are independent. It is shown that the test statistic is asymptotically free distributed, not relying on resampling or Monte-Carlo methods to obtain critical values. Here by asymptotic we mean as $k \to \infty$, no assumption is made on the sample sizes, which may remain bounded or increase with k. Section 4 derives the asymptotic power. As the asymptotic approximation to the null distribution is rather conservative for small to moderate values of k, Sect. 5 studies a bootstrap estimator of the null distribution that gives good results in such

cases. The finite sample performance of the test was numerically assessed by means of an extensive simulation experiment, whose results are reported in Sect. 6. Section 7 shows that, by assuming suitable mixing conditions, the results in sections 3 and 4 can be extended to the case of dependent data. A real data illustration is provided in Sect. 8. Section 9 summarizes the paper and comments on extensions and further research. All proofs and some simulation results can be found in the Supplementary Material (SM). The computations in this paper have been carried out using programs written in the R language [39]. The R code for the calculation of the modification of the test statistic in Baringhaus and Henze (1992) is also available in the SM.

Throughout the paper we will make use of the following standard notation: all random variables and random elements will be defined on a sufficiently rich probability space $(\Omega, \mathcal{A}, \mathbb{P})$; the symbols $\mathbb{E}$ and $\mathbb{V}$ denote expectation and variance, respectively; $\mathbb{P}_0$, $\mathbb{E}_0$ and $\mathbb{V}_0$ denote probability, expectation and variance under the null hypothesis, respectively; $\overset{\mathcal{D}}{=}$ and $\overset{\mathcal{D}}{\to}$ mean equality in distribution and convergence in distribution, respectively; $\overset{\mathbb{P}}{\to}$ denotes convergence in probability; unless otherwise specified, all limits are taken as $k \to \infty$.

2 Poissonity tests for one sample

2.1 A brief review

This subsection gives a brief, non-exhaustive review of Poissonity tests for one sample, i.e. tests of the null hypothesis $H_{0,1}$ v.s. $H_{1,1}$. The review will be centered on tests that are consistent against fixed alternatives. Most of them are discussed in Mijburgh and Visagie (2020), that gives an overview of Poissonity tests, including some non-universally consistent ones.

Henze (1996) studied tests based on the empirical distribution function for discrete models, that can be applied to test $H_{0,1}$. The test statistics are functions of $F_n(x) - F(x; \hat{\lambda})$, where F_n denotes the empirical distribution function, $F(\cdot; \lambda)$ stands for the cumulative distribution function of the law $Pois(\lambda)$ and $\hat{\lambda}$ is an estimator of λ, which is taken equal to the sample mean in Sect. 4 of that paper. Specifically, the Kolmogorov-Smirnov and Cramér-von Mises statistics are studied. A Kolmogorov-Smirnov type statistic was also proposed by Klar (1999), based on the difference between the integrated distribution function and its empirical counterpart. Székely and Rizzo (2004) proposed a Cramér-von Mises type test statistic where the empirical distribution function is replaced by an estimator which is obtained by using a characterization of the distribution of count variables in terms of mean distances.

When dealing with count data, Nakamura and Pérez-Abreu (1993a) argue in favor of using inferential methods based on the empirical probability generating function (EPGF). A number of tests use this tool. Rueda and O'Reilly (1999) (see also Jiménez-Gamero and Batsidis 2017) proposed a test of $H_{0,1}$ based on an L^2 norm of the difference between the probability generating function (PGF) of the Poisson law, with λ replaced with the sample mean, and the EPGF. That test was latter extended for the bivariate (and d-variate) Poisson law in Novoa-Muñoz and Jiménez-Gamero (2014).

The PFG of the Poisson law is the unique PGF g of a law with finite mean, satisfying the differential equation,

$$g'(t) - \mu g(t) = 0, \quad \forall t \in [0, 1], \tag{1}$$

where $g'(t) = \frac{\mathrm{d}}{\mathrm{d}t} g(t)$ and μ is a constant, which is necessarily the mean of the distribution. Baringhaus and Henze (1992) consider an L^2 norm of an empirical version of that equation as a test statistic for $H_{0,1}$. A related test was proposed by Nakamura and Pérez-Abreu (1993b), that was later generalized in Jiménez-Gamero and Alba-Fernández (2019) to testing for a family of distributions that contains the Poisson law as a special case, and in Novoa-Muñoz and Jiménez-Gamero (2016) to the bivariate Poisson. Puig and Weiss (2020) give a characterization of the Poisson distribution, based on an identity involving the binomial thinning operator, that is used to propose some tests of $H_{0,1}$, whose test statistics replace the PGF in the characterization by the EPGF.

The last decade has witnessed an increasing interest in the developing of inferential methods based on Stein's method (see Anastasiou et al. (2023) for a recent overview). In this line, Betsch et al. (2022) give a characterization of count variables that is used to propose a test of $H_{0,1}$.

As expected from Janssen (2000), in the simulations carried out in Gürtler and Henze (2000), Székely and Rizzo (2004), Puig and Weiss (2020) and Betsch et al. (2022), there is no test having the largest power against all considered alternatives. Nevertheless, the test in Baringhaus and Henze (1992), that will be called the BH test, is one of those having best overall performance. Next subsection deals in some detail with its test statistic.

2.2 The BH test statistic

In order to simplify notation, for the one-sample case we will write $Y, Y_1, \ldots, Y_n$ and μ instead of $X_1, X_{1,1}, \ldots, X_{1,n_1}$ and λ_1, respectively. Along this subsection, it will be assumed that $\mu = \mathbb{E}(Y) < \infty$. Let g_n denote the EPGF associated to the sample,

$$g_n(t) = \frac{1}{n} \sum_{i=1}^{n} t^{Y_i}.$$

Since the PFG of the Poisson law is the unique satisfying Eq. (1) it follows that, calling

$$\theta = \theta(Y) = \int_0^1 \{g'(t) - \mu g(t)\}^2 \mathrm{d}t, \tag{2}$$

where $g(t)$ is the PGF of Y, we have that $\theta = 0$ if and only if Y has a Poisson law. Baringhaus and Henze (1992), estimate θ by means of $\tilde{\theta}$, obtained by replacing g, g' and μ with g_n, g'_n and $\overline{Y} = (1/n) \sum_{i=1}^{n} Y_i$, respectively, in the expression of θ, where

$$g'_n(t) = \frac{1}{n} \sum_{i=1}^{n} Y_i \mathbf{1}(Y_i \geq 1) t^{Y_i - 1},$$

and $\mathbf{1}(\cdot)$ stands for the indicator function. The proposed test rejects the null for large values of $n\tilde{\theta}$.

$\tilde{\theta}$ is a degree-4 V-statistic with symmetric kernel H_{sym}

$$\tilde{\theta} = \frac{1}{n^4} \sum_{i_1,i_2,i_3,i_4=1}^{n} H_{sym}(Y_{i_1}, Y_{i_2}, Y_{i_3}, Y_{i_4}),$$

where

$$H_{sym}(Y_{i_1}, Y_{i_2}, Y_{i_3}, Y_{i_4}) = \frac{1}{24} \sum_{\sigma=(\sigma(1),\sigma(2),\sigma(3),\sigma(4))\in\mathcal{P}_4} H(Y_{\sigma(1)}, Y_{\sigma(2)}, Y_{\sigma(3)}, Y_{\sigma(4)}),$$

$\mathcal{P}_4$ is the set of all permutations of (i_1, i_2, i_3, i_4) (the four elements are treated as different, even if there are ties between them), and

$$H(x, y, u, v) = \int_0^1 h(x, y; t)h(u, v; t)\mathrm{d}t, \quad h(x, y; t) = x\mathbf{1}(x \geq 1)t^{x-1} - yt^x.$$

(3)

Clearly, $\tilde{\theta}$ is not an unbiased estimator of θ. In our developments, we will work with the associated U-statistic, denoted as $\hat{\theta}$,

$$\hat{\theta} = \hat{\theta}(Y_1, \ldots, Y_n) = \frac{1}{n(4)} \sum_{1\leq i_1\neq i_2\neq i_3\neq i_4\leq n} H_{sym}(Y_{i_1}, Y_{i_2}, Y_{i_3}, Y_{i_4}), \quad (4)$$

where $n(4) = n(n-1)(n-2)(n-3)$, which is unbiased for θ, and thus $\mathbb{E}_0(\hat{\theta}) = 0$. In the expression (4), the notation $1 \leq i_1 \neq i_2 \neq i_3 \neq i_4 \leq n$ means that the four indices i_1, i_2, i_3, i_4 are all of them different and that their values range from 1 to n. Notice that the calculation of $\hat{\theta}$ requires $n \geq 4$.

For the practical calculation of $n\hat{\theta}$, it is convenient to use the following expression

$$n\hat{\theta} = \frac{1}{n-1} \sum_{i_1\neq i_2} \frac{Y_{i_1}Y_{i_2}}{Y_{i_1} + Y_{i_2} - 1}\mathbf{1}(Y_{i_1} \geq 1)\mathbf{1}(Y_{i_2} \geq 1)$$

$$-2\frac{1}{(n-1)(n-2)} \sum_{i_1\neq i_2\neq i_3} \frac{Y_{i_1}Y_{i_2}}{Y_{i_1} + Y_{i_3}}\mathbf{1}(Y_{i_1} \geq 1)$$

$$+\frac{1}{(n-1)(n-2)(n-3)} \sum_{i_1\neq i_2\neq i_3\neq i_4} \frac{Y_{i_1}Y_{i_2}}{Y_{i_3} + Y_{i_4} + 1}.$$

The R code we used in the numerical results for the calculation of $n\hat{\theta}$ is provided in the SM.

3 Sum of BH test statistics

This section studies a test of $H_{0,k}$ vs $H_{1,k}$ whose associate statistic is a sum of one-sample modified BH test statistics. As seen in Subsection 2.2, for $n\hat{\theta}$ to be well-defined we must assume that the population generating the data has finite expectation and that the sample size is larger than or equal to 4. Accordingly, the setting is as follows:

Let $\mathbf{X}_1 = \{X_{1,1}, \ldots, X_{1,n_1}\}, \ldots, \mathbf{X}_k = \{X_{k,1}, \ldots, X_{k,n_k}\}$ be k
independent samples with sizes $n_1, \ldots, n_k$, which may be different,
where $X_{i,1}, \ldots, X_{i,n_i}$ are i.i.d. coming from $X_i \in \mathbb{N}_0$, with
$\lambda_i = \mathbb{E}(X_i) < \infty$, and $n_i \geq 4, 1 \leq i \leq k$. (5)

Let

$$\mathcal{T}_k = \sum_{i=1}^{k} n_i \hat{\theta}_i,$$

where $\hat{\theta}_i = \hat{\theta}(X_{i,1}, \ldots, X_{i,n_i})$, $1 \leq i \leq k$, and $\hat{\theta}$ is as defined in (4). Let $\theta_i = \theta(X_i)$, where θ is as defined in (2), $1 \leq i \leq k$. By construction, $\mathbb{E}(\mathcal{T}_k) = \sum_{i=1}^{k} n_i \theta_i \geq 0$ with $\mathbb{E}(\mathcal{T}_k) = 0$ if and only if $H_{0,k}$ is true. Thus, it seems reasonable to reject $H_{0,k}$ for large values of $\mathcal{T}_k$.

Remark 1 The reason for multiplying $\hat{\theta}_i$ by n_i in the definition of $\mathcal{T}_k$ is that, under $H_{0,k}$ and for large n_i, we have that $n_i\hat{\theta}_i = O_{\mathbb{P}}(1)$. Specifically, it converges in law to a linear combination of χ^2 variables (see, for example, Chapter 5 of Serfling 2009). For small to moderate n_i we still have that $n_i\hat{\theta}_i = O_{\mathbb{P}}(1)$. Therefore, in all cases (small to large sample sizes), under $H_{0,k}$, $\mathcal{T}_k$ is a sum of zero-mean, non-degenerate random variables.

To test $H_{0,k}$ we must calculate upper percentiles of the null distribution of $\mathcal{T}_k$, that is unknown because it depends on $\lambda_1, \ldots, \lambda_k$, which are unknown quantities. Moreover, even if they were known, it is rather difficult to determine the exact null distribution of $\mathcal{T}_k$. Because of these reasons, we try to approximate it. The next result shows that, under $H_{0,k}$ and conveniently normalized, $\mathcal{T}_k$ converges in law to a standard normal law, no matter how large (or small) are the sample sizes $n_1, \ldots, n_k$. To derive it we only assume that the means of the populations take values in a compact interval bounded away from 0. This is mainly done to avoid the case in which the variance goes to infinity or to 0. Let $\tau_{0,i}^2 = \mathbb{V}_0(n_i\hat{\theta}_i) = \mathbb{E}_0(n_i^2\hat{\theta}_i^2)$, $1 \leq i \leq k$, and $\sigma_{0k}^2 = (1/k)\sum_{i=1}^{k} \tau_{0,i}^2$. Let Φ denote the cumulative distribution function of a standard normal distribution.

Theorem 1 *Suppose that* (5) *holds and that* $H_{0,k}$ *is true with* $\lambda_i \in [L_1, L_2]$, $\forall i$, *for some fixed* $0 < L_1 < L_2 < \infty$, *then*

$$\frac{1}{\sqrt{k}} \frac{\mathcal{T}_k}{\sigma_{0k}} \xrightarrow{\mathcal{D}} Z \sim N(0, 1),$$

moreover

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}_0 \left(\frac{1}{\sqrt{k}} \frac{\mathcal{T}_k}{\sigma_{0k}} \le x \right) - \Phi(x) \right| \le C \frac{1}{\sqrt{k}}, \tag{6}$$

for each k, where $C = C(L_1, L_2) > 0$ is a finite constant.

Recall that the test is one-sided, rejecting the null hypothesis for large values of $\mathcal{T}_k$. If σ_{0k} were a known quantity, in view of Theorem 1, the test that rejects $H_{0,k}$ when $\mathcal{T}_k/(\sqrt{k}\sigma_{0k}) \ge z_{1-\alpha}$, for some $\alpha \in (0, 1)$, where $\Phi(z_{1-\alpha}) = 1 - \alpha$, would have (asymptotic) level α. The assertion remains true if σ_{0k} is replaced by a consistent estimator. We will consider as estimator of σ_{0k}^2 the sample variance of $n_1\hat{\theta}_1, \ldots, n_k\hat{\theta}_k$,

$$S_{1k}^2 = \frac{1}{k} \sum_{i=1}^{k} (n_i\hat{\theta}_i - \overline{n\hat{\theta}})^2, \quad \overline{n\hat{\theta}} = \frac{1}{k} \sum_{i=1}^{k} n_i\hat{\theta}_i.$$

Since $\mathbb{E}_0(n_i\hat{\theta}_i) = 0$, $\forall i$, we can also consider the following estimator of σ_{0k}^2,

$$S_{2k}^2 = \frac{1}{k} \sum_{i=1}^{k} n_i^2\hat{\theta}_i^2.$$

The next proposition shows that, under $H_{0,k}$, S_{1k}^2 and S_{2k}^2 are both of them ratio consistent estimators of σ_{0k}^2.

Proposition 1 *Suppose that the assumptions in Theorem 1 hold, then $S_{ik}^2/\sigma_{0k}^2 \xrightarrow{a.s.} 1$, $i = 1, 2$.*

Corollary 1 *Suppose that the assumptions in Theorem 1 hold, then $\mathcal{T}_k/(\sqrt{k}S_{ik}) \xrightarrow{\mathcal{D}} Z$, $i = 1, 2$.*

Let $\alpha \in (0, 1)$. For testing $H_{0,k}$ vs $H_{1,k}$, we consider the test that rejects the null when

$$T_i = \mathcal{T}_k/(\sqrt{k}S_{ik}) \ge z_{1-\alpha}. \tag{7}$$

From Corollary 1, it has asymptotic level α, $i = 1, 2$.

Remark 2 Notice that to derive the above results no assumption has been done on the sample sizes (apart from $n_i \ge 4$); hence they are valid whether the sample sizes remain bounded or increase with k at any rate.

Remark 3 To estimate the variance of $\mathcal{T}_k$ we have considered nonparametric estimators. Parametric estimators, that is, estimators obtained by calculating the exact expression of $\tau_{0,i}^2$ and then replacing λ_i with the sample mean of the sample from X_i, or any other estimator (see, e.g. Jiménez-Gamero and Batsidis 2017, Jiménez-Gamero et al.

2016), are more involved. In addition, the study of the behavior of such estimators under alternatives (necessary to study the power of the proposed test) may become rather complicated.

Remark 4 In simulations we observed that the tests in (7), based on the normal approximation,

$$\hat{\mathbb{P}}_1(x) := \hat{\mathbb{P}}_1(T_i \leq x) = \Phi(x), \quad i = 1, 2,$$

are rather conservative, for small to moderate k (the results will be shown in Sect. 6). In order to improve the normal approximation one could try a one-term Edgeworth expansion which, if it exists, should be (see, e.g., display (2.6) in Liu 1988),

$$\hat{\mathbb{P}}_2(x) := \hat{\mathbb{P}}_2(T_i \leq x) = \Phi(x) + \frac{\varsigma}{6\sqrt{k}}\phi(x)(2x^2 + 1),$$

$$\varsigma = \frac{\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_0(n_i^3\hat{\theta}_i^3)}{\sigma_{0,k}^3}, \quad i = 1, 2, \tag{8}$$

where ϕ denotes the probability density function of a standard normal distribution. For practical application, the term ς must be replaced by some consistent estimator, such as $\hat{\varsigma}_1 = \frac{1}{k}\sum_{i=1}^{k}(n_i\hat{\theta}_i - \overline{n\hat{\theta}})^3/S_{1,k}^3$ or $\hat{\varsigma}_2 = \frac{1}{k}\sum_{i=1}^{k}n_i^3\hat{\theta}_i^3/S_{2,k}^3$ (the proof of the consistency of these two estimators of ς can be easily derived using the results in Subsection 1.1 of the SM). Notice that since $\phi(x)(2x^2 + 1)$ is a bounded function and that, under assumptions in Theorem 1, ς is a bounded quantity (this fact is consequence of Lemma 2 in Subsection 1.1 of the SM), it follows that,

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}_0(T_i \leq x) - \hat{\mathbb{P}}_j(x)\right| \leq M\frac{1}{\sqrt{k}}, \quad i = 1, 2, \ j = 1, 2,$$

where M is a positive constant. Thus, one can try both approximations to the null distribution of the test statistic T_i, $i = 1, 2$. Typically, one-term Edgeworth expansions give approximations with error of order $o(k^{-1/2})$, whenever the involved populations satisfy certain assumptions related to the existence of certain moments and their characteristic functions (see, e.g. Bhattacharya and Ranga Rao 1976 and Hall 1992). According to the results in Subsection 1.1 of the SM, such moment assumptions are satisfied in our setting. The situation is not so clear for the conditions on the characteristic function. Perhaps, the most popular sufficient condition on the characteristic function is Cramér condition. See display (20.55) in Bhattacharya and Ranga Rao (1976) for the case of independent random vectors. We did not succeed to prove it in our setting. However, if all $n_i\hat{\theta}_i$ are replaced by their limit in law, as $n_i \rightarrow \infty$ (which are linear combinations of centered independent χ^2 variables, as observed in Remark 1), in the expression of $\mathcal{T}_k$, then it can be checked that under the assumptions in Theorem 1, the approximation error of $\hat{\mathbb{P}}_2(x)$ is of order $o(k^{-1/2})$. This suggests that the order $o(k^{-1/2})$ could be valid for large sample sizes.

Nevertheless, in simulations we observed that when using the approximation (8), the tests become a bit liberal, specially for small k. Section 5 investigates a bootstrap approximation, that gave really good practical results.

4 Power

This section studies the asymptotic power of the tests in (7). With this aim, it will be assumed that the sample sizes are comparable in the following sense:

$$n_i = c_i m, \quad m \geq 1, \quad 0 < c \leq c_i \leq C < \infty, \quad \text{for some constants } c \text{ and } C.$$

$$(9)$$

To be precise, m should be denoted as m_k, since it can vary with k. To simplify notation, the subindex k will be omitted. No restriction is assumed on m (except that $m \geq 1$), so if (9) holds, the sample sizes can either remain bounded or increase arbitrarily. It will be also assumed w.l.o.g. that $X_1, \ldots, X_r$ have Poisson laws, while the remaining populations have alternative distributions, for some $0 \leq r < k$. Here r is also allowed to vary with k, $r = r_k$, but such dependence on k will be skipped. If $r = 0$, then $1 \leq i \leq r$ is understood to be the empty set.

The next result, which is similar to Theorem 1, shows that under $H_{1,k}$ and conveniently normalized, $\mathcal{T}_k$ also converges in law to a standard normal law.

Theorem 2 *Suppose that* (5) *holds, that* $X_1, \ldots, X_r$ *have Poisson laws with* $\lambda_i \in [L_1, L_2]$, $1 \leq i \leq r$, *for some* $0 \leq r < k$, *and for some fixed* $0 < L_1 < L_2 < \infty$, *that* $X_{r+1}, \ldots, X_k$ *have alternative distributions satisfying* $0 < a \leq \mathbb{V}(\sqrt{n_i}\hat{\theta}_i)$ *and* $\mathbb{E}\{|\sqrt{n_i}(\hat{\theta}_i - \theta_i)|^{2+\delta}\} \leq A$, $r+1 \leq i \leq k$, *for some positive constants* a, δ *and* A, *and that* $n_1, \ldots, n_k$ *satisfy* (9). *Let* $p_k = (k - r)/k$. *Assume that either* $p_k \to p \in (0, 1]$, *or* $p_k \to 0$ *and* $k - r \to \infty$, *or* $p_k \to 0$, $k - r$ *remain bounded and* $m/k \to 0$. *Then*

$$\frac{1}{\sqrt{k}} \frac{\mathcal{T}_k - \sum_{i=r+1}^k n_i \theta_i}{\sqrt{\frac{1}{k} \sum_{i=1}^k \mathbb{V}(n_i \hat{\theta}_i)}} \xrightarrow{\mathcal{D}} Z.$$

Remark 5 If X_i has an alternative distribution, then $\sqrt{n_i}(\hat{\theta}_i - \theta_i)$ converges in law, as $n_i \to \infty$, to a normal law; so, it makes sense to impose restrictions on $\sqrt{n_i}(\hat{\theta}_i - \theta_i)$ rather than on $n_i(\hat{\theta}_i - \theta_i)$, $r + 1 \leq i \leq k$.

Remark 6 Notice that to derive the above result it has been assumed that the sample sizes satisfy $n_i \geq 4$, and (9) with $m/k \to 0$ if $p_k \to 0$ and $k - r$ remain bounded. In any case, the result is valid whether the sample sizes remain bounded or increase with k.

Remark 7 Theorem 2 assumes that $0 < a \leq \mathbb{V}(\sqrt{n_i}\hat{\theta}_i)$, $r + 1 \leq i \leq k$, for some fixed constant a. From the proof, we see that it suffices that the variance of $\sqrt{n_i}\hat{\theta}_i$ is bounded away from 0 for a positive proportion of alternatives.

Under alternatives,

$$\mathbb{E}(S_{1k}^2) = \frac{1}{k}\left(1 - \frac{1}{k}\right)\sum_{i=1}^k \mathbb{V}(n_i\hat{\theta}_i) + \frac{1}{k}\sum_{i=1}^k (n_i\theta_i - \overline{n\theta})^2, \quad \overline{n\theta} = \frac{1}{k}\sum_{i=1}^k n_i\theta_i,$$

$$\mathbb{E}(S_{2k}^2) = \frac{1}{k}\sum_{i=1}^k \mathbb{V}(n_i\hat{\theta}_i) + \frac{1}{k}\sum_{i=r+1}^k n_i^2\theta_i^2,$$

and so S_{1k}^2 and S_{2k}^2 are both of them biased estimators of $(1/k)\sum_{i=1}^k \mathbb{V}(n_i\hat{\theta}_i)$. The next result shows that S_{ik}^2 is a ratio (strongly) consistent estimator of its expectation, $i = 1, 2$.

Proposition 2 *Suppose that the assumptions in Theorem* 2 *hold. Then* $S_{ik}^2/\mathbb{E}(S_{ik}^2) \xrightarrow{a.s.} 1$, $i = 1, 2$.

Let $\mathrm{pwd}_{ik} = \mathbb{P}(\mathcal{T}_k/\sqrt{k}S_{ik} \geq z_{1-\alpha})$, $i = 1, 2$. From Theorem 2 and Proposition 2,

$$\mathrm{pwd}_{ik} - \Phi\left(\mathrm{coc}_{i1}^{1/2}z_\alpha + \mathrm{coc}_2\right) \to 0, \tag{10}$$

$i = 1, 2$, where

$$\mathrm{coc}_{i1} = \frac{\mathbb{E}(S_{ik}^2)}{\frac{1}{k}\sum_{i=1}^k \mathbb{V}(n_i\hat{\theta}_i)}, \quad \mathrm{coc}_2 = \frac{1}{\sqrt{k}}\frac{\sum_{i=r+1}^k n_i\theta_i}{\sqrt{\frac{1}{k}\sum_{i=1}^k \mathbb{V}(n_i\hat{\theta}_i)}}, \tag{11}$$

$i = 1, 2$. The next result gives the asymptotic power of the tests (7), that is, the limit of $\Phi\left(\mathrm{coc}_{i1}^{1/2}z_\alpha + \mathrm{coc}_2\right)$ in some special cases.

Theorem 3 *Suppose that the assumptions in Theorem* 2 *hold, that* $kp_k \to \infty$, *that* $\theta_i \leq \Theta$, $\forall i$, *for some positive constant* Θ *and that* $\bar{\gamma} = \frac{1}{k-r}\sum_{i=r+1}^k c_i\theta_i > \epsilon$, *for some positive constant* ϵ *(at least for large k).*

1. *If* $p_k \to p \in (0, 1]$, *then* $\mathrm{pwd}_{ik} \to 1$, $i = 1, 2$.
2. *If* $p_k \to p = 0$ *and* $mp_k \to a > 0$, *then* $\mathrm{pwd}_{ik} \to 1$, $i = 1, 2$.
3. *If* $p_k \to p = 0$, $mp_k \to 0$ *and* $\sqrt{k}mp_k \to \infty$, *then* $\mathrm{pwd}_{ik} \to 1$, $i = 1, 2$.
4. *If* $p_k \to p = 0$, $mp_k \to 0$ *and* $\sqrt{k}mp_k \to \kappa \in [0, \infty)$, *then*

$$\mathrm{pwd}_{ik} = \Phi\left(z_\alpha + \frac{\kappa\bar{\gamma}}{\sqrt{v_1}}\right) + r_{ik}, \tag{12}$$

with $r_{ik} \to 0$, $i = 1, 2$, *where* $v_1 = (1/r)\sum_{i=1}^r \mathbb{E}(n_i^2\hat{\theta}_i^2)$.

Theorem 3 assumes that all θ_i are bounded. If $\mathbb{E}(X_i) \leq M$, $\forall i$, for some positive constant M, then, from the definition of θ in (2), it follows that $\theta_i \leq \Theta$, $\forall i$, for some positive $\Theta = \Theta(M)$. Recall that for the populations obeying the null we are assuming

that $\lambda_i = \mathbb{E}(X_i) \leq L_2$, so it is not very restrictive to assume that a similar bound is true for the expectations of the alternative populations.

Theorem 3 reveals that the tests in (7) are consistent in a variety of scenarios. To be specific: if the proportion of alternative distributions is positive, then the proposed tests are consistent (this is case 1 in Theorem 3); if the proportion of alternative distributions goes to 0, then the proposed tests are still consistent provided that the sampling effort m compensates for the vanishing p_k (cases 2 and 3 in Theorem 3); otherwise, the tests may be still able to detect alternatives (power greater than α), but the power may not reach 1 (case 4 in Theorem 3). Finally, we point out that there are more cases not considered in that theorem, which could be studied in a similar way.

5 A bootstrap approximation to the null distribution

We observed in simulations that neither the normal nor the one-term Edgeworth approximations to the null distribution of the test statistics T_1 and T_2 are satisfactory. Because of this reason, this section studies another approximation based on bootstrap. Two bootstrap approximations are conceivable. First, since $H_{0,k}$ is a parametric hypothesis, a parametric bootstrap could be applied: estimate λ_i through $\hat{\lambda}_i$ (which can be taken the sample mean of the sample $\mathbf{X}_i$), generate a sample of size n_i from $X_i^{\star} \sim Pois(\hat{\lambda}_i)$, say $X_{i,1}^{\star}, \ldots, X_{i,n_i}^{\star}$, $1 \leq i \leq k$, evaluate the test statistic at the bootstrap data $X_{1,1}^{\star}, \ldots, X_{1,n_1}^{\star}, \ldots, X_{k,1}^{\star}, \ldots, X_{k,n_k}^{\star}$, obtaining say $T_i^{\star}$, and finally, approximate the null distribution of T_i through its bootstrap distribution, which is the conditional distribution of $T_i^{\star}$, given the original data, $X_{1,1}, \ldots, X_{1,n_1}, \ldots, X_{k,1}, \ldots, X_{k,n_k}$. In practise, the approximation is carried out by simulation, that is, one repeats B times the above procedure, that way obtaining B bootstrap replicates of T_i, say $T_i^{\star 1}, \ldots, T_i^{\star B}$, and then the bootstrap distribution of T_i is estimated through the empirical distribution of the B bootstrap replicates. Although in simulations we got excellent results (in terms of obtaining a really good approximation to the null distribution), the main issue with this bootstrap approximation is that it is very time-consuming, due to the calculation of $\hat{\theta}_i^{\star}$, since it involves a sum with $O(n_i^2)$ terms, $1 \leq i \leq k$.

Alternatively, a non-parametric bootstrap estimator can be obtained as follows. First, define

$$Y_i = n_i \hat{\theta}_i - \overline{n\hat{\theta}}, \quad 1 \leq i \leq k.$$

By construction, $\bar{Y} = (1/k) \sum_{i=1}^{k} Y_i = 0$. Let $Y_1^*, \ldots, Y_k^*$ be a random sample from the empirical distribution of $Y_1, \ldots, Y_k$. Approximate the null distribution of T_i, $i = 1, 2$, through the conditional distribution of $T_i^* = \sum_{i=1}^{k} Y_i^* / (\sqrt{k} S_{ik}^*)$, $i = 1, 2$, given the original data, denoted as $\mathbb{P}_*(T_i^* \leq x)$, where $S_{1k}^{*2} = (1/k) \sum_{i=1}^{k} (Y_i^* - \overline{Y^*})^2$, $\overline{Y^*} = (1/k) \sum_{i=1}^{k} Y_i^*$ and $S_{2k}^{*2} = (1/k) \sum_{i=1}^{k} Y_i^{*2}$. As before, in practise the approximation is carried out by simulation, that is, one repeats B times the above procedure, obtaining $T_i^{*1}, \ldots, T_i^{*B}$, and then the bootstrap distribution of T_i is estimated through the empirical distribution of the B bootstrap replicates, $i = 1, 2$. As this bootstrap

avoids the calculation (at the bootstrap level) of the most time-consuming part (the $\hat{\theta}_i^*$, $1 \leq i \leq k$), its implementation is really fast, even for large k. Because of this reason, we restrict ourselves to study the consistency of this non-parametric bootstrap approximation.

Theorem 4 *Suppose that either $H_{0,k}$ is true or it is not true and the assumptions in Theorem 2 hold with $m^2/k \to 0$, and that $\theta_i \leq \Theta$, $\forall i$, for some positive constant Θ. Then* $\sup_x \left| \mathbb{P}_*(T_i^* \leq x) - \Phi(x) \right| \xrightarrow{\mathbb{P}} 0$, $i = 1, 2$.

To get the consistency of the bootstrap null distribution estimator when the data do not obey $H_{0,k}$, Theorem 4 requires $m^2/k \to 0$, which is used to cope with the bias of S_{ik}^2 as an estimator of the variance of $(1/\sqrt{k})\mathcal{T}_k$.

The result in Theorem 4 holds whether $H_{0,k}$ is true or not. In particular, the test that rejects $H_{0,k}$ when $p_* = \mathbb{P}_*\{T_i^* > T_i\} \leq \alpha$, is asymptotically correct in the sense that when $H_{0,k}$ is true, $p_* \to \alpha$, in probability. Under alternatives, the results in Theorem 3 also apply to that bootstrap test.

6 Finite sample results

Two tests of $H_{0,k}$ have been studied and three approximations to the null distribution of their test statistics have been proposed. We conducted two simulation experiments, in order to assess the finite sample performance of the proposed tests, using each of the three null distribution approximations, under the null hypothesis (level) and under alternatives (power). Next we describe those experiments and summarize their outputs. As slightly better results (in the sense of closeness of the empirical levels to the nominal values under the null, and greater power under alternatives) were obtained for the test based on the variance estimator S_{1k}^2, we will only display the results obtained for such test.

Experiment to study the level To study the level of the tests for a finite value of k, we carried out the following simulation experiment. We generated a sample from each of k populations, for $k = 10, 20, 100, 200, 1000, 2000$. Population i has a Poisson law with mean λ_i, $1 \leq i \leq k$, where the values of $\lambda_1, \ldots, \lambda_k$ were randomly generated from a continuous uniform distribution in the interval (a, b), $U(a, b)$, for $(a, b) = (1, 5)$, $(5, 10)$, $(10, 15)$, $(15, 20)$. The size of each sample has been randomly generated from a discrete uniform random law $UD\{N, N+1, \ldots, M\}$, with $(N, M) = (6, 10)$, $(11, 15)$, $(16, 20)$, $(21, 25)$, $(26, 30)$. Each case was run $R = 10,000$ times and to calculate the bootstrap approximation, $B = 1000$ bootstrap samples were generated. Table 1 displays the percentages of p-values less than or equal to 0.05 and 0.10, which are the estimated type I error probabilities for nominal significance level α=0.05 and 0.10, respectively. Looking at this table we observe that: for small to moderate k, in general, the normal approximation gives conservative tests and the one-term Edgeworth approximation gives liberal tests, better results are observed for the bootstrap approximation; the sample sizes seems to have no effect on the closeness of these approximations to the null distribution (for the sample sizes tried); as expected from the theory, as k increases all approximations yield levels closer to the nominal values.

Table 1 Empirical levels, expressed in percentages, at the nominal levels $\alpha = 5\%$, 10%

$k = 10$

λ_i	U(1, 5) 5%	U(1, 5) 10%	U(5, 10) 5%	U(5, 10) 10%	U(10, 15) 5%	U(10, 15) 10%	U(15, 20) 5%	U(15, 20) 10%
n_i								
(a)	3.8	8.2	4.8	10.1	4.9	9.5	4.8	9.5
	4.6	10.1	5.1	12.0	5.0	11.1	4.9	11.2
	7.6	12.5	9.2	14.6	8.6	13.8	8.4	13.9
(b)	3.2	7.6	3.8	8.9	4.1	8.5	3.7	8.1
	4.0	9.1	4.5	11.2	4.3	10.1	4.0	9.8
	6.8	11.8	8.3	13.7	7.7	12.8	7.2	12.5
(c)	2.7	6.8	3.3	8.0	3.4	7.7	3.1	7.4
	3.2	8.6	3.8	10.0	3.8	9.4	3.6	9.0
	6.1	11.0	7.5	12.7	6.9	12.1	6.6	11.6
(d)	2.9	7.1	3.4	8.1	2.8	7.3	2.8	7.1
	3.4	8.5	3.9	10.3	3.4	9.1	3.3	8.6
	6.4	11.1	7.3	12.9	6.6	11.6	6.2	11.1
(e)	3.0	7.0	3.1	7.3	3.2	7.1	2.5	6.6
	3.2	8.4	3.8	9.7	3.6	8.9	3.0	8.2
	6.1	11.1	7.0	11.9	6.6	11.3	5.9	10.9

$k = 20$

λ_i	U(1, 5) 5%	U(1, 5) 10%	U(5, 10) 5%	U(5, 10) 10%	U(10, 15) 5%	U(10, 15) 10%	U(15, 20) 5%	U(15, 20) 10%
n_i								
(a)	2.7	7.3	3.4	8.3	3.3	8.2	3.5	8.3
	4.8	10.2	5.6	11.5	5.5	11.3	5.3	11.1
	6.9	12.1	8.2	13.0	7.8	12.6	7.7	12.7
(b)	2.5	6.8	3.2	7.8	3.2	7.3	3.3	7.8
	4.3	9.5	5.2	10.9	5.0	10.1	5.1	10.4
	6.3	11.0	7.3	12.6	7.0	11.8	7.1	12.2
(c)	2.6	6.7	3.0	7.4	2.5	7.0	2.9	7.1
	4.2	9.3	4.9	10.6	4.2	10.1	4.7	10.1
	5.9	11.0	7.1	12.3	6.3	11.8	6.7	11.7
(d)	2.6	6.7	3.0	7.4	2.5	7.0	2.9	7.1
	4.2	9.3	4.9	10.6	4.2	10.1	4.7	10.1
	5.9	11.0	7.1	12.3	6.3	11.8	6.7	11.7
(e)	2.5	6.4	2.8	7.0	2.3	6.1	2.6	6.7
	4.0	9.0	4.8	10.3	4.0	8.9	4.0	9.3
	5.7	10.8	7.2	12.0	5.7	10.8	5.8	11.0

$k = 100$

λ_i	U(1, 5) 5%	U(1, 5) 10%	U(5, 10) 5%	U(5, 10) 10%	U(10, 15) 5%	U(10, 15) 10%	U(15, 20) 5%	U(15, 20) 10%
n_i								
(a)	2.1	6.4	2.5	6.7	3.0	7.3	2.7	7.4
	4.7	9.2	5.1	10.8	5.4	10.6	5.2	10.7
	5.5	10.0	6.7	11.8	6.4	11.3	6.3	11.5

$k = 200$

λ_i	U(1, 5) 5%	U(1, 5) 10%	U(5, 10) 5%	U(5, 10) 10%	U(10, 15) 5%	U(10, 15) 10%	U(15, 20) 5%	U(15, 20) 10%
n_i								
(a)	2.9	7.5	2.5	6.7	2.5	6.9	2.9	7.8
	5.2	10.1	5.1	9.8	4.9	9.9	5.6	10.4
	6.0	10.7	6.3	10.7	5.8	10.5	6.4	11.0

Table 1 continued

	$k = 100$								$k = 200$							
(b)	2.6	7.3	2.5	6.8	2.9	7.6	2.8	7.2	2.9	7.0	3.1	7.3	3.2	7.8	3.4	7.8
	4.8	10.0	5.0	10.2	5.4	10.7	5.0	9.9	5.4	10.4	5.0	9.7	5.3	10.0	5.2	10.1
	5.7	10.6	6.2	11.0	6.4	11.4	5.9	10.8	6.3	11.2	5.8	10.4	5.9	10.5	5.7	10.6
(c)	2.9	7.7	2.8	7.3	2.9	7.1	2.7	7.4	3.0	7.7	3.4	8.0	3.4	7.5	2.9	7.7
	5.1	10.3	5.6	10.4	5.4	9.9	5.0	10.2	5.9	10.7	5.5	10.4	5.3	10.0	4.7	9.9
	5.7	10.9	6.7	11.2	6.0	10.7	5.7	10.8	6.9	11.5	6.1	11.1	5.7	10.4	5.1	10.2
(d)	3.0	7.5	2.3	6.7	2.8	7.3	2.9	7.6	2.5	7.2	2.9	7.6	3.3	7.8	3.2	7.7
	5.0	10.0	5.0	10.2	5.1	10.1	5.3	10.1	5.0	10.2	5.0	10.1	5.1	10.2	5.1	10.0
	5.8	10.4	6.0	10.8	5.9	10.7	6.0	10.8	5.9	10.7	5.0	10.1	5.5	10.6	5.6	10.2
(e)	2.9	7.4	2.5	7.2	2.9	7.6	3.1	7.8	2.8	7.1	3.4	7.8	3.2	8.2	3.4	8.1
	5.1	9.8	5.2	10.5	5.2	10.0	5.3	10.5	5.2	10.2	5.2	10.2	3.2	8.2	5.3	10.0
	5.8	10.4	6.3	11.3	6.0	10.8	6.1	11.2	6.0	10.9	5.8	10.7	5.7	10.7	5.6	10.6

	$k = 1000$								$k = 2000$							
(a)	3.7	8.6	3.0	7.8	3.7	8.6	3.9	8.5	3.1	7.7	3.7	8.2	4.2	8.9	3.9	9.2
	5.1	10.1	5.6	10.5	5.4	10.6	5.2	10.1	5.4	10.4	5.0	9.8	5.5	10.3	5.2	10.4
	5.4	10.2	6.4	11.0	5.6	10.9	5.4	10.4	6.0	10.7	5.1	9.8	5.5	10.3	5.2	10.6
(b)	3.9	8.6	2.9	7.5	3.5	8.2	3.7	8.4	3.2	7.9	3.8	8.8	4.3	9.1	4.3	9.3
	5.0	9.9	5.0	10.0	5.0	10.0	5.1	10.0	5.2	10.1	5.0	10.0	5.3	10.3	5.0	10.4

Table 1 continued

	k = 1000								k = 2000							
	5.1	9.9	5.6	10.4	5.2	10.3	5.2	10.1	5.5	10.3	5.5	10.3	5.1	10.2	5.3	10.4
(c)	3.6	8.6	3.6	8.6	4.3	9.3	3.9	8.9	3.9	8.8	3.4	8.3	4.0	8.9	4.5	9.2
	4.7	9.7	5.9	10.8	5.7	10.8	5.2	10.4	4.7	9.7	5.1	10.2	5.1	10.0	5.5	10.3
	4.7	9.9	6.2	11.1	5.9	11.0	5.4	10.5	4.8	9.8	5.4	10.4	5.1	10.2	5.5	10.2
(d)	2.3	6.6	2.5	6.4	2.4	6.7	2.7	7.3	3.0	7.5	2.3	6.7	2.8	7.3	2.9	7.6
	4.5	9.6	4.6	10.1	4.8	9.8	4.8	10.4	5.0	10.0	5.0	10.2	5.1	10.1	5.3	10.1
	5.6	10.3	6.1	11.1	6.0	10.6	6.0	11.3	5.8	10.4	6.0	10.8	5.9	10.7	6.0	10.8
(e)	4.0	8.8	3.6	8.0	3.9	8.5	3.7	8.5	4.4	9.2	3.7	8.5	4.2	9.1	4.1	9.0
	5.2	10.0	5.1	10.1	5.1	10.0	4.9	9.9	5.2	10.1	5.2	10.1	5.1	10.0	5.0	9.8
	5.1	10.0	5.6	10.2	5.1	10.1	4.9	10.1	5.2	10.0	5.4	10.1	5.1	10.1	5.0	9.9

In each case, the first line gives the results for the normal approximation, the second line for the bootstrap approximation, and the third line for the one-term Edgeworth approximation. The case (a) for n_i means that $n_i \sim UD\{6,\ldots,10\}$, (b) $n_i \sim UD\{11,\ldots,15\}$, (c) $n_i \sim UD\{16,\ldots,20\}$, (d) $n_i \sim UD\{21,\ldots,25\}$ and (e) $n_i \sim UD\{26,\ldots,30\}$

Experiment to study the power To study the power, we repeated the above experiment with $100p\%$ of the samples coming from an alternative distribution with mean μ and the remaining $100(1-p)\%$ of the samples were generated from a Poisson law with the same mean μ, for $p = 0.1, 0.2$. The considered alternative distributions are

- ALT1: a discrete uniform distribution, $UD\{0, \ldots, 5\}$, $\mu = 2.5$.
- ALT2: a mixture of two Poisson distributions, $0.5 \times Pois(2) + 0.5 \times Pois(5)$, $\mu = 3.5$.
- ALT3: a zero inflated Poisson distribution, $0.2 \times 0 + 0.8 \times Pois(2)$, $\mu = 1.6$.
- ALT4: a binomial distribution, $Bin(10, 0.5)$, $\mu = 5$.
- ALT5: a negative binomial distribution, $BN(2, 2/3)$, $\mu = 1$.
- ALT6: a negative binomial distribution, $BN(10, 2/3)$, $\mu = 5$.

Table 2 displays the power results for $k = 100, 200$. Looking at this table we see that power increases with the sample sizes, with k and with p.

7 Dependent data

This section aims to relax the assumption that $X_1, \ldots, X_k$ are independent. For example, $X_1, \ldots, X_k$ can be seen as a segment of a process (observed at discrete times) whose random variables take values in $\mathbb{N}_0$. Along this section, instead of (5), the setting will be the following:

> The available data consist of n independent copies of
> $(X_1, \ldots, X_k)$, say $(X_{11}, \ldots, X_{k1}), \ldots,$
> $(X_{1n}, \ldots, X_{kn})$, $X_i \in \mathbb{N}_0$, with $\mathbb{E}(X_i) < \infty$, and $n \geq 4, 1 \leq i \leq k$. (13)

Notice that, if the independence assumption is dropped, the expression of the variance of $\mathcal{T}_k$ given in Sect. 3 is not valid; moreover, to derive a result similar to that in Theorem 1, some mixing conditions have to be added; and finally, adequate variance estimators must be considered.

As before, let $\hat{\theta}_i = \hat{\theta}(X_{i1}, \ldots, X_{in})$, $1 \leq i \leq k$. Under the setting (13), we again have that $\mathbb{E}(\mathcal{T}_k) = \sum_{i=1}^{k} n\theta_i \geq 0$ with $\mathbb{E}(\mathcal{T}_k) = 0$ if and only if $H_{0,k}$ is true. Therefore, it seems reasonable to reject $H_{0,k}$ for large values of $\mathcal{T}_k$. Because of the same reasons argued in Sect. 3 in the *independent setting* (5), we will build a test of $H_{0,k}$ in the current *dependent setting* (13), based on the asymptotic null distribution of $\mathcal{T}_k$. As observed before, some mixing conditions must be assumed with this aim. The definition of the mixing coefficients in Theorem 5 below can be found in Definition 3.42 of White (2001). Let

$$\bar{\sigma}_{0k}^2 = \mathbb{V}_0(\mathcal{T}_k/\sqrt{k}) = \frac{1}{k} \sum_{i,j=1}^{k} \mathbb{E}_0(n\hat{\theta}_i n\hat{\theta}_j).$$

Theorem 5 *Suppose that* (13) *holds, that* $\{X_i\}_{i \in \mathbb{Z}}$ *has ϕ-mixing coefficients of size* $-v/2(v-1)$ *or α-mixing coefficients of size* $-v/(v-2)$, *for some* $v > 2$, *that*

Table 2 Empirical powers, expressed in percentages at the nominal level $\alpha = 5\%$

k	p	ALT1					ALT2				
		(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
100	0.1	8.0	14.1	22.7	31.2	41.4	5.1	11.3	17.8	25.6	34.0
		12.6	21.6	33.0	43.2	55.1	12.3	21.1	30.3	40.6	50.2
		14.1	23.6	36.0	46.7	58.2	15.4	24.4	33.9	45.2	54.7
	0.2	16.9	37.2	58.5	77.0	88.3	9.9	28.4	48.4	67.6	81.7
		25.0	48.6	70.0	85.7	93.8	21.7	43.8	64.7	80.4	90.4
		27.5	51.6	72.7	87.5	95.0	25.8	48.4	68.8	83.2	91.8
200	0.1	12.9	26.5	43.2	60.3	74.5	8.2	20.5	35.7	51.3	66.9
		18.4	34.2	52.0	69.9	82.6	16.5	31.6	48.4	64.5	77.9
		19.7	36.0	54.2	71.6	83.9	19.2	34.3	51.3	67.1	80.1
	0.2	29.7	65.4	88.6	97.4	99.6	21.5	54.4	81.5	94.2	98.6
		38.5	73.2	92.4	98.5	99.8	35.1	67.2	89.0	96.9	99.3
		40.4	74.5	93.0	98.6	99.8	38.4	69.6	90.1	97.5	99.4

k	p	ALT3					ALT4				
100	0.1	9.1	16.6	24.6	36.0	45.8	6.7	10.9	17.2	24.6	33.7
		13.1	23.5	33.6	46.9	58.4	11.3	17.1	23.8	31.4	40.5
		14.2	25.4	36.1	49.8	61.4	13.4	19.3	26.3	33.4	42.2
	0.2	20.3	44.5	68.6	86.2	94.8	16.0	35.7	55.7	73.8	86.7
		27.6	55.4	78.0	91.5	97.3	21.4	41.4	61.0	77.1	88.6
		30.0	58.1	80.0	92.9	97.8	23.6	43.4	62.2	78.0	89.3
200	0.1	14.8	29.2	48.3	65.8	79.9	8.5	17.3	30.6	44.4	58.7
		19.1	36.4	56.6	73.4	86.3	14.4	25.3	38.1	51.6	64.2
		20.0	37.9	58.0	74.8	87.2	16.4	27.2	39.9	52.7	65.4
	0.2	36.3	75.3	94.8	99.2	100	26.3	58.9	84.3	95.5	99.2
		43.6	81.0	96.7	99.5	100	34.5	64.1	86.5	96.0	99.3
		45.0	82.0	97.0	99.5	100	36.8	65.5	87.3	96.3	99.3

k	p	ALT5					ALT6				
100	0.1	4.8	7.6	9.8	13.1	16.1	3.4	5.4	8.5	11.4	15.2
		7.8	11.6	15.7	20.3	24.7	8.4	13.3	18.0	23.0	28.1
		8.7	13.0	17.4	22.4	27.1	11.2	16.2	21.7	26.3	32.1
	0.2	8.1	13.7	23.2	32.4	44.2	5.1	11.1	19.9	31.9	43.8
		12.6	20.8	32.4	43.8	56.5	13.1	23.7	35.7	48.8	60.8
		14.0	23.3	35.2	46.9	58.9	16.9	28.1	40.2	53.1	64.6
200	0.1	7.0	11.0	16.6	23.9	30.4	4.5	8.4	15.3	22.9	32.4
		10.0	15.8	22.7	31.7	39.9	11.3	17.4	27.0	35.7	46.1
		10.8	16.8	24.1	33.2	41.5	13.7	20.2	30.0	38.8	49.2

Table 2 continued

k	p	ALT5					ALT6				
	0.2	11.8	25.6	42.7	60.2	75.9	8.3	22.0	41.6	61.3	76.6
		16.9	33.5	51.4	69.2	82.4	18.7	37.0	56.8	74.2	85.5
		18.1	34.9	53.0	70.6	83.4	22.2	40.8	60.8	76.4	87.2

In each case, the first line gives the results for the normal approximation, the second line for the bootstrap approximation, and the third line for the one-term Edgeworth approximation. The case (a) means that $n_i \sim UD\{6, \ldots, 10\}$, (b) $n_i \sim UD\{11, \ldots, 15\}$, (c) $n_i \sim UD\{16, \ldots, 20\}$, (d) $n_i \sim UD\{21, \ldots, 25\}$ and (e) $n_i \sim UD\{26, \ldots, 30\}$

$\bar{\sigma}_{0k} > 0$ for all k sufficiently large, and that $H_{0,k}$ is true with $\lambda_i \in [L_1, L_2]$, $\forall i$, for some fixed $0 < L_1 < L_2 < \infty$. Then $\mathcal{T}_k/(\sqrt{k}\bar{\sigma}_{0k}) \xrightarrow{\mathcal{D}} Z$.

For the result in Theorem 5 to be useful to construct a test of $H_{0,k}$, a ratio consistent estimator of $\bar{\sigma}_{0k}^2$ is needed. With this aim we consider the following estimator, called in the literature the HAC estimator,

$$\hat{\bar{\sigma}}_{0k}^2 = \hat{c}_0 + 2\sum_{i=1}^{\ell_k} \omega(i, \ell_k)\hat{c}_i, \quad \hat{c}_i = \frac{n^2}{k}\sum_{j=1}^{k-i}(\hat{\theta}_j - \bar{\hat{\theta}})(\hat{\theta}_{j+i} - \bar{\hat{\theta}}), \quad \bar{\hat{\theta}} = \frac{1}{k}\sum_{j=1}^{k}\hat{\theta}_j.$$

Several papers have dealt with the consistency of the HAC estimator for non-stationary sequences (see e.g. Andrews 1988, Andrews 1991, Newey and West 1987, White 2001). The next proposition shows that, for adequate choices of the weights $\omega(i, \ell_k)$ and of ℓ_k, and under some mixing conditions, $\hat{\bar{\sigma}}_{0k}^2$ is ratio consistent when $\bar{\sigma}_{0k} > 0$.

Proposition 3 *Suppose that* (13) *holds, that* $\{X_i\}_{i\in\mathbb{Z}}$ *has ϕ-mixing coefficients of size* $-2v/(2v-1)$ *or α-mixing coefficients of size $-2v/(v-1)$, for some $v > 1$, that $\bar{\sigma}_{0k} > 0$ for all k sufficiently large, that $H_{0,k}$ is true with $\lambda_i \in [L_1, L_2]$, $\forall i$, for some fixed $0 < L_1 < L_2 < \infty$, that the weights are bounded, $|\omega(i, \ell_k)| \leq \Delta$, for some $\Delta > 0$, that $\ell_k \to \infty$, that $\omega(i, \ell_k) \to 1$ for each i, and that $\ell_k/k^{1/4} \to 0$. Then, $\hat{\bar{\sigma}}_{0k}^2/\bar{\sigma}_{0k}^2 \xrightarrow{\mathbb{P}} 1$.*

Corollary 2 *Suppose that the assumptions in Proposition 3 hold. Then $\mathcal{T}_k/(\sqrt{k}\hat{\bar{\sigma}}_{0k}) \xrightarrow{\mathcal{D}} Z$.*

Remark 8 Notice that to derive the above results no assumption has been done on the sample size n (apart from $n \geq 4$); hence they are valid whether n remains bounded or increases with k at any rate.

Let $\alpha \in (0, 1)$. For testing $H_{0,k}$ vs $H_{1,k}$, in the setting (13), we consider the test that rejects the null when

$$\mathcal{T}_k/(\sqrt{k}\hat{\bar{\sigma}}_{0k}) \geq z_{1-\alpha}. \tag{14}$$

From Corollary 2, it has asymptotic level α.

Remark 9 The condition $\ell_k/k^{1/4} \to 0$ can be relaxed in the sense that ℓ_k can be chosen larger (see Andrews (1988), Andrews (1991)). Nevertheless, in our simulations we found no practical advantage when taking larger values of ℓ_k.

Remark 10 Theorem 1 gives a Berry-Esseen type bound for the normal approximation in the *independent setting* (5) and Remark 4 gives the expression of the one-term Edgeworth expansion. We are only aware of Berry-Essen type bounds for stationary sequences of weakly dependent data (see, e.g. Rio 1996, Bentkus et al. 1997, Jirak 2016), and the same apply to Edgeworth expansions (see, e.g. Lahiri 2010), which are more involved than in the independent setting. Recall that in our setting, the means of the components of the sequence $X_1, \ldots, X_k$ may differ, therefore the sequences we are dealing with are not stationary, and hence the aforementioned bounds and expansions do not apply.

To study the power, as in the independence setting (5), we will assume that $X_1, \ldots, X_r$ have Poisson laws, while the rest of the data have alternative distributions. Let

$$\bar{\sigma}_k^2 = \frac{n}{k} \sum_{i,j=1}^{k} \mathbb{E}\left\{(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)\right\}.$$

Theorem 6 *Suppose that* (13) *holds, that* $\{X_i\}_{i\in\mathbb{Z}}$ *has* ϕ*-mixing coefficients of size* $-v/2(v-1)$ *or* α*-mixing coefficients of size* $-v/(v-2)$, *for some* $v > 2$, *that* $X_1, \ldots, X_r$ *have Poisson laws with* $\lambda_i \in [L_1, L_2]$, $1 \leq i \leq r$, *for some* $r < k$, *that* $X_{r+1}, \ldots, X_k$ *have alternative distributions satisfying* $\mathbb{E}(|\sqrt{n}(\hat{\theta}_i - \theta_i)|^{2+\delta}) \leq A < \infty$, $r+1 \leq i \leq k$, *for some constants* A, $\delta > 0$, *and that* $\bar{\sigma}_k^2 > 0$ *for all* k *sufficiently large. Then,*

$$\frac{1}{\sqrt{k}} \frac{\mathcal{T}_k - n\sum_{i=r+1}^{k} \theta_i}{\sqrt{n}\bar{\sigma}_k} \xrightarrow{\mathcal{D}} Z.$$

Remark 11 The observation in Remark 5, made for the setting (5), also applies here in the setting (13).

The next proposition gives the expectation and the limit in probability of $\hat{\bar{\sigma}}_{0k}^2$ under alternatives.

Proposition 4 *Suppose that* (13) *holds, that* $\{X_i\}_{i\in\mathbb{Z}}$ *has* ϕ*-mixing coefficients of size* $-v/(v-1)$ *or* α*-missing coefficients of size* $-2v/(v-2)$, *for some* $v > 2$, *that* $X_1, \ldots, X_r$ *have Poisson laws with* $\lambda_i \in [L_1, L_2]$, $1 \leq i \leq r$, *for some* $r < k$, *that* $X_{r+1}, \ldots, X_k$ *have alternative distributions satisfying* $\mathbb{E}(|\sqrt{n}(\hat{\theta}_i - \theta_i)|^{2+\delta}) \leq A < \infty$, $r+1 \leq i \leq k$, *for some positive constants* A, δ, *that that* $\theta_i \leq \Theta$, $\forall i$, *for some positive constant* Θ, *that* $|\omega(i, \ell_k)| \leq \Delta$, *for some* $\Delta > 0$, *that* $\ell_k \to \infty$, *that* $\omega(i, \ell_k) \to 1$ *for each* i, *that* $\ell_k/k^{1/4} \to 0$, *that* $n/\sqrt{k}$ *is a bounded quantity, at least*

for large k, and that $\bar{\sigma}_k > 0$ for all k sufficiently large. Then,

$$\frac{1}{n}\mathbb{E}\left(\hat{\bar{\sigma}}_{0k}^2\right) = \bar{\sigma}_k^2 + \frac{n}{k}\sum_{j=1}^{k}(\theta_j - \bar{\theta})^2 + 2\sum_{i=1}^{\ell_k}\omega(i,\ell_k)\frac{n}{k}\sum_{j=1}^{k-i}(\theta_j - \bar{\theta})(\theta_{j+i} - \bar{\theta}) + r_k$$

(15)

where $\bar{\theta} = (1/k)\sum_{i=1}^{k}\theta_i$ and $r_k \to 0$ as $k \to \infty$. Assume also that $(1/n)\mathbb{E}\left(\hat{\bar{\sigma}}_{0k}^2\right) \geq a > 0$, at least for large k. Then,

$$\hat{\bar{\sigma}}_{0k}^2/\mathbb{E}\left(\hat{\bar{\sigma}}_{0k}^2\right) \xrightarrow{\mathbb{P}} 1.$$

(16)

Finally, from Theorem 6 and Proposition 4, the power of the test (14) can be approximated by means of

$$\Phi\left(\mathrm{coc.d}_1^{1/2}z_\alpha + \mathrm{coc.d}_2\right),$$

where

$$\mathrm{coc.d}_1 = \frac{1}{n}\frac{\mathbb{E}\left(\hat{\bar{\sigma}}_{0k}^2\right)}{\bar{\sigma}_k^2}, \quad \mathrm{coc.d}_2 = \frac{\sqrt{n}\sum_{i=r+1}^{k}\theta_i}{\sqrt{k}\bar{\sigma}_k}.$$

In particular, the test is consistent against alternatives for which $\mathrm{coc.d}_1$ remains bounded and $\mathrm{coc.d}_2 \to \infty$.

7.1 Finite sample results for simulated data

We conducted two simulations experiments, in order to assess the finite sample performance of the test in (14) under the null hypothesis (level) and under alternatives (power). Next we describe those experiments and summarize their outputs.

Experiment to study the level To generate a sequence of k dependent Poisson data, we first generated $k + 2m$ independent Poisson data, and then calculated the moving sums of each $2m + 1$ consecutive elements. To generate the independent Poisson data, we considered the following three cases:

- Case inc: the means of the data form an increasing sequence from 0.25 to 1
- Case ran: the means of the data are a random sample from a uniform law in the interval (0.25, 1).
- Case alt: the means of the data form a sequence with the first and third quarters equal to 0.25, and the remaining equal to 1.

We tried $m = 2, 3$, the case $m = 3$ inducing a stronger dependence compared to $m = 2$. As for ℓ_k, we took $\xi\lceil\log(k)\rceil$, with $\xi = 1, 2, 3$, where $\lceil x \rceil$ is the smaller integer greater than or equal to x. This ensures in particular $\ell_k = o(k^{1/4})$, as requested in Proposition 3. We considered several weights $\omega(i, \ell_k)$. Specifically we considered the weights generated by the following kernels (see, for example, Andrews 1991)

- Truncated kernel, which gives $\omega_T(i, \ell_k) = 1$, $1 \le i \le \ell_k$,
- Bartlett kernel, which gives $\omega_B(i, \ell_k) = 1 - i/(\ell_k + 1)$, $1 \le i \le \ell_k$,
- Tukey-Hanning kernel, which gives $\omega_{TH}(i, \ell_k) = 0.5\{1 + \cos(\pi i/\ell_k)\}$, $1 \le i \le \ell_k$,
- Parzen kernel, which gives $\omega_P(i, \ell_k) = 1 - 6(i/\ell_k)^2 + 6(i/\ell_k)^3$, for $1 \le i \le \ell_k/2$, and $\omega(i, \ell_k) = 2(1 - i/\ell_k)^3$, for $\ell_k/2 < i \le \ell_k$.

The weights ω_T and ω_P do not always generate positive variance estimators in finite samples. In our simulations we got negative variance estimators in many instances when using ω_T, and in less cases with ω_P (only for $k = 50$). As for the other weights, with the use of ω_B we got, in general, level results closer to the nominal values than those given by ω_{TH}. Because of these reasons, we will only show up the results for ω_B. Table 1 in the SM displays the percentages of p-values less than or equal to 0.05 and 0.10, which are the estimated type I error probabilities for nominal significance level α=0.05 and 0.10, respectively. Each of those percentages were calculated over $R = 10,000$ runs. Looking at this table we see that, in contrast to the independent case, here the sample size has an effect on the closeness of the actual level to the nominal value; in general, better results (in the sense of getting closer values) are obtained for $n = 10, 15$, than for $n = 5$. As expected from the theory, as k increases, the observed levels stick to the nominal values.

Experiment to study the power To generate a sequence of dependent data, we proceed as before: first generate $k + 2m$ independent data, and then calculate the moving sums of each $2m + 1$ consecutive elements; the difference lies in the way of generating the independent data: the first $100p\%$ come from an alternative law, and the remaining $100(1 - p)\%$ come form a Poisson law with the same mean μ as the alternative, for $p = 0.2, 0.4$. The considered alternatives were

- ALTD1: a zero inflated Poisson distribution, $0.2 \times 0 + 0.8 \times Pois(2)$, $\mu = 1.6$,
- ALTD2: a binomial distribution, $Bin(5, 0.3)$, $\mu = 1.5$,
- ALTD3: a negative binomial distribution, $BN(2, 5/7)$, $\mu = 0.8$.

Because of the same reasons given for the level, we will only show up the results for ω_B. Table 2 in the SM displays them. Looking at this table, the same conclusions given for the independent case apply here.

8 Real data illustration

In order to illustrate the proposed test, we consider the results of the MAQC-2 mRNA-Seq experiment in Bullard et al. (2010). In this experiment, two types of biological samples (Brain and UHR) were analyzed, each using $n = 7$ lanes. As a result, seven sequences of read counts were obtained for each group along $k = 52,580$ genes. In sequencing experiments, the Poisson has been suggested as a suitable model for the read counts; see for instance Lander and Waterman (1988). In this setting, the Poisson distribution is often used to calculate p-values from the read counts, so genetic disorders or other features can be detected (Jiménez-Otero et al. 2019). Therefore, verification of this assumption is important in practice in order to validate the Poisson-based inference.
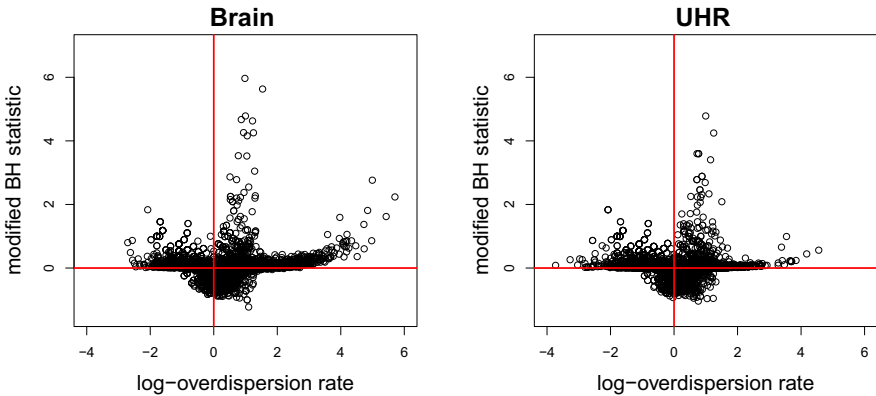
Fig. 1 BH local test statistics against log-overdispersion rates along the genes for the MAQC-2 experiment: Brain group (left) and UHR group (right). Horizontal and vertical lines are located at zero

The null hypothesis of Poissonity was tested separately for the two biological samples by applying criterion (14). We used the weigths $\omega_B(i, \ell_k)$ with $\ell_k = 11, 22$ and 33, corresponding to the choice $\ell_k = \lceil \xi \log(k) \rceil$, with $\xi = 1, 2, 3$ respectively. The p-values were very small for the Brain group ($2.36 \times 10^{-9}, 3.63 \times 10^{-9}$ and 5.29×10^{-9}), leading to the rejection of the Poisson model. The p-value obtained for the UHR type was 0.076 for the three choices of ℓ_k. The very small p-values obtained for the Brain group can be explained from the overdispersion of the read counts. In order to see this, in Fig. 1, left, the modified BH test statistics $n\hat{\theta}_i$, $1 \leq i \leq k$, are plotted against the log-ratio of sample variance and mean. Note that this log-ratio is zero for the Poisson model. From Fig. 1, left, it is seen that the overdispersion is clearly present for many of the genes, which report a relatively large value of the modified BH test statistic too. This figure also reveals that the modified BH statistic may be large even when mean and variance are in well agreement; this is not surprising, since the absence of overdispersion does not guarantee Poissonity. The situation is less critical for the UHR group, see Fig. 1, right; this is why the p-value is several orders of magnitude greater compared to the Brain group. In this case, the log-overdispersion rate roughly ranges between -2 and 2, with only few genes reporting values larger than 2. The modified BH statistics are smaller for the UHR group too.

For comparison purposes, the p-values were re-calculated by using the tests based on the independence assumption; the results were very similar to those reported by the test for dependent data. Specifically, the p-values reported by T_1 were 1.17×10^{-9} and 0, obtained with the normal and the non-parametric bootstrap approximations, respectively, for the Brain group; and similarly 0.078 and 0.015 for the UHR group. The similarity among the results obtained with the normal approximation in each group suggests that the dependence among the read counts is not strong.

9 Concluding remarks and further research

This paper proposes and studies procedures for simultaneously testing that k samples from a count variable come from Poisson laws, which can have different means. The samples can be independent or dependent. The proposed tests are based on the BH test and allow k to increase. The normal approximation to the null distribution of the test statistics works reasonably well for large k. The tests have been shown to be consistent against a wide range of possible configurations for the alternatives, including cases where the proportion of alternative distributions goes to 0.

The methods in this paper are based on the test in Baringhaus and Henze (1992) for one sample. As seen in Subsection 2.1, there are other Poissonity tests designed for the one-sample setting. Each of them could be used to build similar procedures to those developed here. Moreover, parallel strategies could be tried for simultaneously testing that k samples come from other parametric families.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Anastasiou A, Barp A, Briol FX, Ebner B, Gaunt RE, Ghaderinezhad F, Gorham J, Gretton A, Ley C, Liu Q, Mackey L, Oates CJ, Reinert G, Swan Y (2023) Stein's Method Meets Computational Statistics: A Review of Some Recent Developments. Stat Sci 38(1):120–139. https://doi.org/10.1214/22-STS863

Andrews DWK (1988) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Cowles Foundation Discussion Paper No. 877, Yale University

Andrews DWK (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59(3):817–858. https://doi.org/10.2307/2938229

Baringhaus L, Henze N (1992) A goodness of fit test for the Poisson distribution based on the empirical generating function. Statist Probab Lett 13(4):269–274. https://doi.org/10.1016/0167-7152(92)90033-2

Bentkus V, Götze F, Tikhomirov A (1997) Berry-Esseen bounds for statistics of weakly dependent samples. Bernoulli 3(3):329–349. https://doi.org/10.2307/3318596

Betsch S, Ebner B, Nestmann F (2022) Characterizations of non-normalized discrete probability distributions and their application in statistics. Electron J Stat 16(1):1303–1329. https://doi.org/10.1214/22-ejs1983

Bhattacharya RN, Ranga Rao R (1976) Normal approximation and asymptotic expansions. John Wiley & Sons

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in MRNA-seq experiments. BMC Bioinformatics 11(94):1–13

Gaigall D (2021) On a new approach to the multi-sample goodness-of-fit problem. Commun Stat Simul Comput 50(10):2971–2989. https://doi.org/10.1080/03610918.2019.1618472

Gürtler N, Henze N (2000) Recent and classical goodness-of-fit tests for the Poisson distribution. J Statist Plan Inference 90(2):207–225. https://doi.org/10.1016/S0378-3758(00)00114-2

Hall P (1992) The bootstrap and Edgeworth expansion. Springer-Verlag. https://doi.org/10.1007/978-1-4612-4384-7

Henze N (1996) Empirical-distribution-function goodness-of-fit tests for discrete models. Canad J Statist 24(1):81–93. https://doi.org/10.2307/3315691

Janssen A (2000) Global power functions of goodness of fit tests. Ann Statist 28(1):239–253. https://doi.org/10.1214/aos/1016120371

Jiménez-Gamero MD (2023) Testing normality of a large number of populations. Statist Pap. https://doi.org/10.1007/s00362-022-01384-y

Jiménez-Gamero MD, Alba-Fernández MV (2019) Testing for the Poisson-Tweedie distribution. Math Comput Simul 164:146–162. https://doi.org/10.1016/j.matcom.2018.08.001

Jiménez-Gamero MD, Batsidis A (2017) Minimum distance estimators for count data based on the probability generating function with applications. Metrika 80(5):503–545. https://doi.org/10.1007/s00184-017-0614-3

Jiménez-Gamero MD, Batsidis A, Alba-Fernández MV (2016) Fourier methods for model selection. Ann Inst Statist Math 68(1):105–133. https://doi.org/10.1007/s10463-014-0491-8

Jiménez-Otero N, de Uña-Álvarez J, Pardo-Fernández JC (2019) Goodness-of-fit tests for disorder detection in NGS experiments. Biom J 61(2):424–441. https://doi.org/10.1002/bimj.201700284

Jirak M (2016) Berry-Esseen theorems under weak dependence. Ann Probab 44(3):2024–2063. https://doi.org/10.1214/15-AOP1017

Klar B (1999) Goodness-of-fit tests for discrete models based on the integrated distribution function. Metrika 49(1):53–69. https://doi.org/10.1007/s001840050025

Lahiri SN (2010) Edgeworth expansions for Studentized statistics under weak dependence. Ann Statist 38(1):388–434. https://doi.org/10.1214/09-AOS722

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2(3):231–239. https://doi.org/10.1016/0888-7543(88)90007-9

Liu RY (1988) Bootstrap procedures under some Non-I.I.D. models. Ann Statist 16(4):1696–1708. https://doi.org/10.1214/aos/1176351062

Mijburgh PA, Visagie IJH (2020) An overview of goodness-of-fit tests for the poisson distribution. South Afr Statist J 54(2):207–230

Nakamura M, Pérez-Abreu V (1993) Empirical probability generating function: an overview. Insur Math Econom 12(3):287–295. https://doi.org/10.1016/0167-6687(93)90239-L, https://www.sciencedirect.com/science/article/pii/016766879390239L

Nakamura M, Pérez-Abreu V (1993) Use of an empirical probability generating function for testing a Poisson model. Canad J Statist 21(2):149–156. https://doi.org/10.2307/3315808

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55(3):703–708. https://doi.org/10.2307/1913610

Novoa-Muñoz F, Jiménez-Gamero MD (2016) A goodness-of-fit test for the multivariate Poisson distribution. SORT 40(1):113–138, https://raco.cat/index.php/SORT/article/view/310071,

Novoa-Muñoz F, Jiménez-Gamero MD (2014) Testing for the bivariate Poisson distribution. Metrika 77(6):771–793. https://doi.org/10.1007/s00184-013-0464-6

Puig P, Weiss CH (2020) Some goodness-of-fit tests for the Poisson distribution with applications in biodosimetry. Comput Statist Data Anal 144(106878):12. https://doi.org/10.1016/j.csda.2019.106878

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria, https://www.R-project.org/

Rio E (1996) Sur le théorème de Berry-Esseen pour les suites faiblement dépendantes. Probab Theory Related Fields 104(2):255–282. https://doi.org/10.1007/BF01247840

Rueda R, O'Reilly F (1999) Tests of fit for discrete distributions based on the probability generating function. Comm Statist Simul Comput 28(1):259–274. https://doi.org/10.1080/03610919908813547

Serfling RJ (2009) Approximation theorems of mathematical statistics, vol 162. John Wiley & Sons, New York

Székely GJ, Rizzo ML (2004) Mean distance test of Poisson distribution. Statist Probab Lett 67(3):241–247. https://doi.org/10.1016/j.spl.2004.01.005

White H (2001) Asymptotic theory for econometricians. Academic press