Comparison of conditional tests on Poisson data

Comparison of conditional tests on Poisson data

Un confronto di test condizionati su dati di Poisson

Francesca Romana Crucinio and Roberto Fontana

Abstract We compare four conditional tests for Poisson data through a simulation study: the exact binomial test, its asymptotic approximation, a Markov Chain Monte Carlo test and the standard permutation test. Despite being non-parametric, we observe that permutation tests are as effective as the others. From a theoretical point of view we justify this result by observing that the orbits of permutations form a *good* partition of the conditional space.

Abstract *Si confrontano quattro test condizionati per dati di Poisson: il test binomiale esatto, la sua approssimazione asintotica, un test Markov Chain Monte Carlo e un test di permutazione standard. Si osserva che il test di permutazione, pur non parametrico, ha un comportamento simile agli altri. Una giustificazione teorica di questo risultato sta nell'osservare che le orbite di permutazione costituiscono una* buona *partizione dello spazio condizionato.*

Key words: Algebraic statistics, Conditional test, Permutation test, Poisson data

1 Introduction

We address the problem of comparing the means of two Poisson distributions with unknown parameter λ_i, $i = 1,2$. We consider two independent samples, $\mathbf{Y}_1^{(n_1)} = (Y_1, \ldots, Y_{n_1})$ of size n_1 from Poisson(λ_1) and $\mathbf{Y}_2^{(n_2)} = (Y_{n_1+1}, \ldots, Y_{n_1+n_2})$ of size n_2

Francesca Romana Crucinio
Politecnico di Torino, Dipartimento di Scienze Matematiche
e-mail: francesca.crucinio@gmail.com

Roberto Fontana
Politecnico di Torino, Dipartimento di Scienze Matematiche
e-mail: roberto.fontana@polito.it

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online),

from Poisson(λ_2). Then we use the joint sample $\mathbf{Y} = (\mathbf{Y}_1^{(n_1)}, \mathbf{Y}_2^{(n_2)})$ to perform the test $H_0 : \lambda_1 = \lambda_2$ against $H_1 : \lambda_1 \neq \lambda_2$.

The problem has been extensively studied in the literature. Among the several testing procedures available to researchers, we consider *conditional* tests, i.e. tests that are performed considering only samples $\mathbf{Y}$ such that the sum $\mathbf{Y}_+$ of their elements is equal to the sum $\mathbf{y}_{obs,+}$ of the elements of the observed sample $\mathbf{y}_{obs}$

$$\mathbf{Y}_+ = \sum_{i=1}^{n_1+n_2} Y_i = \sum_{i=1}^{n_1+n_2} y_{i,obs} = \mathbf{y}_{obs,+}. \tag{1}$$

A justification for this choice is that, if we assume that the model for the means of the two distributions is the standard one-way ANOVA model, which according to [6] is $\log(\lambda_i) = \beta_0 + \beta_1 x_i$ with $x_i = 1$ if $1 \leq i \leq n_1$ and $x_i = -1$ if $n_1 + 1 \leq i \leq n_1 + n_2$, the statistic $T = \mathbf{Y}_+ = \sum_{i=1}^{n_1+n_2} Y_i$ is sufficient for the population constant β_0, which is the nuisance parameter of the test.

For the sake of simplicity we denote the sum of the observed sample $\mathbf{y}_{obs,+}$ by t and the set of the samples $\mathbf{Y}$ which satisfy (1) by $\mathscr{F}_t$. We refer to $\mathscr{F}_t$ as the *fiber* corresponding to t. We focus on four conditional tests:

1. the exact binomial test by Przyborowski and Wilenski [8];
2. an asymptotic version of the exact binomial test [8], which is based on the normal approximation of the binomial distribution [4];
3. a Markov Chain Monte Carlo testing procedure which exploits Markov basis [3] and the Metropolis-Hastings algorithm [9];
4. a standard permutation test [7].

In Section 2 we briefly describe the structure of the tests under study. In Section 3 we compare the effectiveness of the tests through a simulation study and in Section 4 we analyse the link between fibers and permutations from a theoretical perspective. Conclusions are in Section 5.

2 Conditional Tests

Exact and Asymptotic Conditional Binomial Test

It is well-known that the distribution of the sum of n independent Poisson variables of mean λ is a Poisson variable with mean $n\lambda$. Then it can be shown that the distribution of the variable $T_1|T = t$, i.e. of the variable $T_1 = \sum_{i=1}^{n_1} Y_i$ conditioned to $T = \sum_{i=1}^{n_1+n_2} Y_i = t$, is a Binomial distribution with probability of success $\theta = (n_1\lambda_1)/(n_1\lambda_1 + n_2\lambda_2)$ and t trials. It follows that under $H_0 : \lambda_1 = \lambda_2$ the variable $T_1|T = t$ follows a binomial distribution with probability of success $\theta_0 = n_1/(n_1 + n_2)$ and t trials. If t_1 is the observed value of T_1 the p-value is computed as

$$\min\{2\min\{p(T_1 \leq t_1), p(T_1 \geq t_1)\}, 1\} \tag{2}$$

where $p(T_1 \le t_1) = \sum_{k=0}^{t_1} \binom{t}{k} \theta_0^k (1-\theta_0)^{t-k}$ and $p(T_1 \ge t_1) = \sum_{k=t_1}^{t} \binom{t}{k} \theta_0^k (1-\theta_0)^{t-k}$.

The asymptotic version of the conditional binomial test uses the asymptotic test statistic

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} \sim \mathrm{N}(0,1) \qquad \text{where } \hat{\theta} = T_1/n_1.$$

The p-value is computed as $2*(1-\Phi(|z_{obs}|)$ where Φ is the cumulative distribution of the standard normal variable and $z_{obs} = (t_1/n_1 - \theta_0)/\sqrt{\theta_0(1-\theta_0)/n}$.

The Markov Chain Monte Carlo Test

As mentioned above we condition on the sum t of the elements of the observed sample $\mathbf{y}_{obs}$ and we explore the fiber

$$\mathscr{F}_t = \{(Y_1, \ldots, Y_{n_1+n_2}) \in \mathbb{N}^{n_1+n_2} : \sum_{i=1}^{n_1+n_2} Y_i = t\}. \tag{3}$$

To explore the fiber $\mathscr{F}_t$ as defined in (3) we set up a connected Markov chain by means of a Markov basis, i.e. a set $\mathscr{B}$ of moves which have to be added/subtracted to the vectors in $\mathscr{F}_t$ in order to move on the fiber (see [3] for a formal definition of Markov Basis). This basis can be found using the `4ti2` software [10] or, in this specific case, simply by induction on the sample size $N = n_1 + n_2$. We get that $\mathscr{B}$ is made of $N-1$ moves $\mathbf{m}_U = (1, \delta_{1,U}, \ldots, \delta_{N-1,U}), U = 1, \ldots, N-1$ where $\delta_{a,b} = -1$ if $a = b$ and 0 otherwise. $\mathscr{B}$ allows us to build a graph over the fiber, where each pair of vectors $\mathbf{y}, \mathbf{x} \in \mathscr{F}_t$ is linked by an edge if a move $\mathbf{m} \in \mathscr{B}$ exists such that $\mathbf{y} = \mathbf{x} \pm \mathbf{m}$. An example when $t = 6$ and $N = 3$ is shown in Figure 1.

Under $H_0 : \lambda_1 = \lambda_2 = \lambda$ we exploit the Metropolis Hastings algorithm (an accelerated version as in [1], [2]) to modify the transition probabilities and grant convergence to

$$p(\mathbf{y}) = e^{-\lambda} \frac{\lambda^{y_1}}{y_1!} \cdot \ldots \cdot e^{-\lambda} \frac{\lambda^{y_N}}{y_N!} = e^{-N\lambda} \frac{\lambda^t}{\prod_{i=1}^N y_i!} = C \prod_{i=1}^N \frac{1}{y_i!} \propto \prod_{i=1}^N \frac{1}{y_i!} \tag{4}$$

where $C = e^{-N\lambda}\lambda^t$. At each step if we are in state $\mathbf{y}$ we select a random move $\mathbf{m}_U \in \mathscr{B}$ and we consider every possible transition $\mathbf{y} + \gamma \cdot \mathbf{m}_U$ with $\gamma \in \Gamma = \{\gamma \in \mathbb{Z} : \mathbf{y} + \gamma \cdot \mathbf{m}_U \in \mathscr{F}_t\} = [-y_1, y_{U+1}] \cap \mathbb{Z}$. We move to $\mathbf{y} + \gamma^\star \cdot \mathbf{m}_U$ with $\gamma^\star$ randomly drawn from the set above with probability

$$q_{\gamma^\star} = \frac{p(\mathbf{y} + \gamma^\star \cdot \mathbf{m}_U)}{\sum_{\gamma \in \Gamma} p(\mathbf{y} + \gamma \cdot \mathbf{m}_U)} \propto \frac{1}{(y_1 + \gamma^\star)! \cdot (y_{U+1} - \gamma^\star)!}.$$

This walk on $\mathscr{F}_t$ allows us to build an approximation of the distribution, under H_0, of the test statistic $W = \bar{Y}_1 - \bar{Y}_2 = T_1/n_1 - T_2/n_2$. Finally the p-value is computed as

$$\frac{\#(|W| \ge |w_{obs}|)}{M} \tag{5}$$

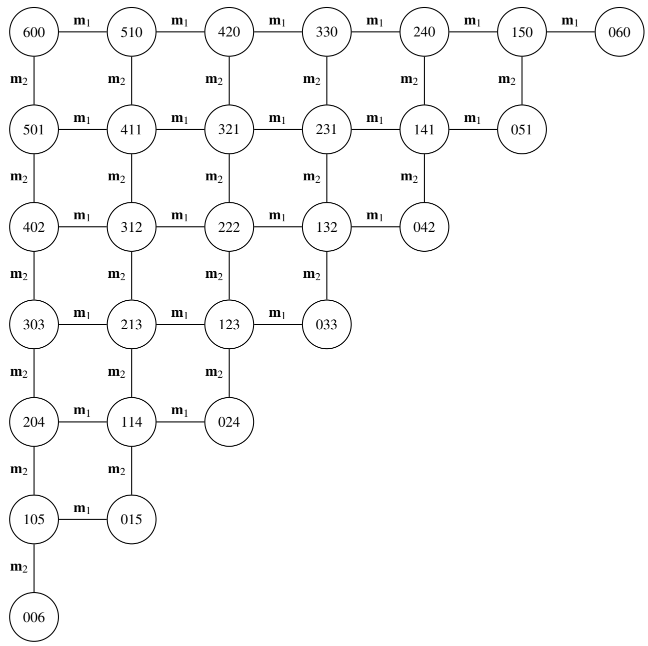

Fig. 1: Graph on the fiber $\mathscr{F}_t$ with $t = 6$ and $N = 3$

where M is the number of transitions and w_{obs} is the observed value of W.

Permutation Test

We perform a standard permutation test [7], randomly selecting M permutations of $\mathbf{y}_{obs}$ (M is at least $1,000$), computing the corresponding values of W and the p-value as in (5).

3 Simulation Study

We consider 27 scenarios that have been built taking three different sample sizes (n_1, n_2) (Table 1a) and, for each sample size, nine different population means (λ_1, λ_2) (Table 1b).

For each scenario $1,000$ samples have been randomly generated. For each sample the corresponding p-values for the four testing procedures under study have been

	1	2	3
n_1	3	8	35
n_2	17	12	15

(a) Sample sizes

	1	2	3	4	5	6	7	8	9
λ_1	0.5	0.5	0.5	1	1	1	5	5	5
λ_2	0.5	0.75	1	1	1.5	2	5	7.5	10

(b) Population means

computed. Specifically for the MCMC test $10,000$ moves after the $1,000$ used for the burn-in step have been used. For the permutation test $2,000$ permutations have been used .

We summarise the most important results:

- the behaviour of the binomial tests (exact and asymptotic) looks different from the behaviour of the Monte Carlo tests (MCMC and permutation). This difference is due to the non-equivalent definitions of p-value ((2) and (5)) and, possibly, to the sampling of the fiber;
- the significance values achieved by the permutation test are almost equivalent to the ones achieved by the MCMC test although this test explores a much wider sample space. We discuss this point in Section 4.

4 Fiber and Permutation Sample Space

The permutation operator does not alter the sum of entries. Hence the *orbits* of permutations $\pi_{\mathbf{y}}$, where $\mathbf{y}$ is the generating vector, are subsets of the fiber. The orbits do not intersect and then we can create a partition of $\mathscr{F}_t$ made up of $\mathrm{part}(t,N)$ orbits $\pi_{\mathbf{y}}$, where $\mathrm{part}(t,N)$ is the partition function defined in [5].

In the same orbit, $p(\mathbf{y})$ is constant and then the probability of taking $\mathbf{y} \in \pi_{\mathbf{y}}$ is $p(\pi_{\mathbf{y}}) = \sum_{\mathbf{y}^* \in \pi_{\mathbf{y}}} p(\mathbf{y}^*) = \#\pi_{\mathbf{y}} \cdot p(\mathbf{y}) = \#\pi_{\mathbf{y}} \cdot C \prod_{i=1}^{N} \frac{1}{y_i!}$, where $\#\pi_{\mathbf{y}}$ is the cardinality of $\pi_{\mathbf{y}}$. It can be proved that C, the normalizing constant defined in (4), can be computed as $C = (\sum_{\pi_{\mathbf{y}} \subseteq \mathscr{F}_t} \#\pi_{\mathbf{y}} \prod_{i=1}^{N} \frac{1}{y_i!})^{-1}$, an expression that does not contain the unknown parameter $\lambda = \lambda_1 = \lambda_2$.

As an example let us consider the fiber in Figure 1. It can be partitioned into $\mathrm{part}(6,3) = 7$ orbits. We get $C = 80/81$ and we can compute the probability of each orbit

$\mathbf{y}$	$p(\mathbf{y})$	$\#\pi_{\mathbf{y}}$	$p(\pi_{\mathbf{y}})$
$(6,0,0)$	$80/(81 \cdot 6!0!0!)$	3	$3/729$
$(5,1,0)$	$80/(81 \cdot 5!1!0!)$	6	$36/729$
$(4,2,0)$	$80/(81 \cdot 4!2!0!)$	6	$90/729$
$(3,3,0)$	$80/(81 \cdot 3!3!0!)$	3	$60/729$
$(3,2,1)$	$80/(81 \cdot 3!2!1!)$	6	$360/729$
$(4,1,1)$	$80/(81 \cdot 4!1!1!)$	3	$90/729$
$(2,2,2)$	$80/(81 \cdot 2!2!2!)$	1	$90/729$

The partition of $\mathscr{F}_t$ into permutation orbits looks somehow *optimal*, because we can approximate well the fiber with one orbit if its probability $p(\pi_{\mathbf{y}})$ is large enough. This result is confirmed in Figure 1. If we select $n_1 = 2$ and $n_2 = 1$ and we compute the exact null cumulative distribution of W over $\mathscr{F}_6$ and its approximation using the orbit $\pi_{(1,2,3)}$ (which has the highest probability), we obtain two distributions which are considerably close, even if the cardinality of the selected orbit is low ($\#\pi_{(1,2,3)} = 6$) compared to the the cardinality of $\mathscr{F}_6$, which is 28.

Table 1: Cumulative distribution of W on $\mathscr{F}_6$ and $\pi_{(1,2,3)}$

w	-6	-4.5	-3	-1.5	0	1.5	3
$\mathscr{F}_6$	0.001	0.018	0.100	0.320	0.649	0.912	1
$\pi_{(1,2,3)}$	0	0	0	0.333	0.667	1	1

5 Conclusion

This study can easily be extended to the non-negative discrete distributions of the exponential family. The convergence of the MCMC to the exact binomial and a mathematical statement on the *optimality* of the partition of the fiber into orbits of permutations are part of our ongoing research.

References

1. Aoki, S., Hara, H., Takemura, A.: Markov Bases in Algebraic Statistics. Springer Series in Statistics. Springer New York (2012)
2. Aoki, S., Takemura, A.: Markov chain monte carlo tests for designed experiments. Journal of Statistical Planning and Inference **140**(3), 817 – 830 (2010)
3. Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Ann. Statist. **26**(1), 363–397 (1998). DOI 10.1214/aos/1030563990
4. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical methods for rates and proportions. John Wiley & Sons (2013)
5. Kunz, M.: Partitions and their lattices. ArXiv Mathematics e-prints (2006)
6. McCullagh, P., Nelder, J.: Generalized Linear Models, Second Edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis (1989)
7. Pesarin, F., Salmaso, L.: Permutation tests for complex data: theory, applications and software. John Wiley & Sons (2010)
8. Przyborowski, J., Wilenski, H.: Homogeneity of results in testing samples from poisson series: With an application to testing clover seed for dodder. Biometrika **31**(3/4), 313–323 (1940)
9. Robert, C., Casella, G.: Monte Carlo Statistical Methods. Springer Texts in Statistics. Springer New York (2013). URL https://books.google.it/books?id=lrvfBwAAQBAJ
10. 4ti2 team: 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at www.4ti2.de