

Some results on the number of coincidences under exchangeability

Alcuni risultati sul numero di coincidenze sotto l'ipotesi di scambiabilità

Sandra Fortini, Annalisa Cerquetti¹

Istituto di Metodi Quantitativi

Università L. Bocconi di Milano

sandra.fortini@uni-bocconi.it, annalisa.cerquetti@uni-bocconi.it

Riassunto: Si presentano alcuni risultati sulla distribuzione asintotica del numero di coincidenze in ambito bayesiano, cioè qualora il meccanismo aleatorio che genera i dati sia specificato mediante una distribuzione a priori. In particolare si determinano condizioni necessarie e sufficienti per la convergenza in distribuzione del numero di coincidenze ad una mistura di Poisson, nel caso di una generica distribuzione a priori, cioè nella sola ipotesi che i dati siano scambiabili. Risultati asintotici in questo ambito erano già noti, in letteratura, per particolari distribuzioni a priori. Le dimostrazioni si basano su recenti risultati sui grafi aleatori colorati, con colori scambiabili.

Keywords: Birthday problem, Chen-Stein methods, Exchangeability, Poisson approximations, Random colored graphs.

1. Introduction

One of the most renowned probability problems is the *birthday problem* (Von Mises, 1939): if n balls are randomly dropped into k boxes, what is the chance of a *match*, that is, that two or more balls fall in the same box? The classical answer is given under the assumption that balls are dropped independently and uniformly into each box. If $k = 365$, the answer gives the probability of two or more coincident birthdays, in a group of n individuals.

In a recent paper Diaconis and Holmes (2002) take up again the birthday problem in a new perspective: "Consider a practical demonstration of the birthday problem on the first day of a class. Are the birthdays of the students reasonably considered as balls dropped uniformly into 365 categories? After all there are well known weekend/weekdays effects, and perhaps seasonal and lunar trends for birth rates. On reflection, the instructor may conclude that the chance p_i of a student being born on day i is not uniform and in fact is not known."

Some results on the birthday problem with non-uniform occurrence probabilities can be found in the literature (see Joag-Dev and Proschan (1992), Mase (1992), Henze (1998), Camarri and Pitman (2000)). The main innovation in Diaconis and Holmes (2002) consists in putting prior probabilities on the p_i 's, in a Bayesian perspective. Exact calculations, for the chance of a match, are presented under uniform prior and symmetric Dirichlet prior. Moreover a Poisson approximation for the law of the number of boxes

¹Indirizzo per corrispondenza: Università L. Bocconi, Istituto di Metodi Quantitativi, Viale Isonzo, 25 20135 - Milano

²Ricerca parzialmente finanziata da: Programma di Ricerca MIUR 2002 *Bayesian nonparametric methods and their applications*.

containing two or more balls, is proved, under general Dirichlet priors. This last result relies on the Chen-Stein approximation methods, applied to negatively associated random variables (see Chen (1975), Barbour *et al.* (1992)). In a remark, Diaconis and Holmes (2002) claim that the limiting law of the number of boxes, containing two or more balls, is a mixture of Poisson distributions, whatever the prior is.

In the present paper, the approximation problem for the number of matches is faced, under a general prior, that is only assuming that the balls are exchangeable. Following a suggestion in Barbour *et al.* (1992), we treat the birthday problem as a counting problem for colored graphs. In this setting, a match consists in a pair of balls falling in the same box. Results for this kind of birthday problem are given in Barbour *et al.* (1992), in the classical i.i.d. setting. Taking up again this point of view, we apply some recent results on random graphs with exchangeable colors (Cerquetti and Fortini (2003)) and find necessary and sufficient conditions for the law of the number of matches to be well approximated by a mixture of Poisson distributions, under a general prior on the p_i 's. This result enables, in a Bayesian setting, to decide whether to assume a Poisson model, for the number of matches, on the base of prior information.

The paper is organized as follows. Section 2 contains the main result and some examples. An application to the surname problem in Japan (Mase (1992)) is presented in Section 3, together with some remarks and suggestions for further developments.

2. The Poisson approximation

For every $n = 1, 2, \dots$, let $(X_{in})_{i=1, \dots, n}$ be a finite sequence of exchangeable random variables taking values in $\{1, \dots, k_n\}$, that can be extended to an infinite exchangeable sequence. According to de Finetti's representation theorem, there exists, for every n , a random vector $Q_n = (Q_{1n}, \dots, Q_{k_n n})$ such that $X_{1n}, \dots, X_{nn}$ are conditionally i.i.d., given Q_n , and

$$P(X_{in} = j | Q) = Q_{jn} \quad P - a.s. \quad (1)$$

for every i, j, n . For every n , let Z_n denote the number of matches in $X_{1n}, \dots, X_{nn}$:

$$Z_n = \sum_{i < l} 1_{(X_{in} = X_{ln})}. \quad (2)$$

The following theorem gives a sufficient condition for the limiting law of $(Z_n)_n$ to be a mixture of Poisson distributions.

Theorem 1. *For every n , let $(X_{in})_{i=1, 2, \dots}$ be a sequence of exchangeable random variables satisfying*

$$E(\max_{1 \leq j \leq k_n} Q_{jn}) = o\left(\frac{1}{n}\right) \quad \text{as } n \rightarrow \infty. \quad (3)$$

a) If $(Z_n)_n$ converges in distribution to a random variable Z , as $n \rightarrow \infty$, then

$$P(Z \in A) = \int_{[0, +\infty)} P_{o_\lambda}(A) d\mu(\lambda), \quad (4)$$

where $P_{o_0}(\{0\}) = 1$ and, for $\lambda > 0$, P_{o_λ} denotes the Poisson distribution with parameter λ . Moreover, the mixing measure μ in (4) is unique.

b) The sequence $(Z_n)_n$ converges, in distribution, if and only if there exists a random variable Λ such that

$$\binom{n}{2} \sum_{j=1}^{k_n} Q_{jn}^2 \xrightarrow{d} \Lambda \quad \text{as } n \rightarrow \infty. \quad (5)$$

Moreover, the mixing measure μ in (4) is the law of Λ .

The proof of Theorem 1 can be obtained by applying a recent result in Cerquetti and Fortini (2003) to the following random colored graph. For every n , let G_n be the complete graph with n vertices and $\binom{n}{2}$ edges and let H be the graph composed of one edge joining two vertices. Consider a sequence of colors, named $1, 2, \dots$. Color the vertices of each graph G_n assigning color X_{jn} to the j -th vertex of graph G_n . Let W_n be the number of edges left after deleting the edges with different colors at the endpoints. Then Z_n , as defined in (2), coincides with W_n . In Cerquetti and Fortini (2003), necessary and sufficient conditions are given for the number of subgraphs of a fixed kind, in a graph with exchangeable colors, to converge, under a negligibility assumption on the frequencies of colors. Moreover it is proved that the limiting law, when it exists, is a mixture of Poisson distributions. The negligibility condition is easily seen to follow from (3) and the necessary and sufficient condition is seen to coincide with (5).

Remark 1. Assumption (3) implies that

$$P(X_{in} = X_{ln}) = o\left(\frac{1}{n}\right), \quad (6)$$

as $n \rightarrow \infty$, for every i, l , with $1 \leq i < l \leq n$. In fact, if (3) holds,

$$nP(X_{in} = X_{ln}) = nE\left(\sum_{j=1}^{k_n} Q_{jn}^2\right) \leq nE\left(\max_{1 \leq j \leq k_n} Q_{jn}\right) \rightarrow 0$$

as $n \rightarrow \infty$.

On the other hand, the probability distribution of Z_n might fail to converge to $E(P_{\circ\Lambda})$, if (3) does not hold, even under (5) and (6). See Example 2.

Example 1. Theorem 1 is consistent with the solution of the classical birthday problem, in which $X_{1n}, \dots, X_{nn}$ are i.i.d. with uniform distribution over $\{1, \dots, k_n\}$, with $\binom{n}{2}/k_n \rightarrow \lambda$, as $n \rightarrow \infty$. In this case $Q_{jn} = 1/k_n$ and, therefore, $E(\max_{1 \leq j \leq k_n} Q_{jn}) = 1/k_n = o(1/n)$ as $n \rightarrow \infty$. Furthermore $\lambda = \lim_{n \rightarrow \infty} \binom{n}{2}/k_n = \lim_{n \rightarrow \infty} \binom{n}{2} \sum_{j=1}^{k_n} Q_{jn}^2$.

Example 2. For every n , let $X_{1n}, \dots, X_{nn}$ be i.i.d. random variables, taking values $1, \dots, n^2$ with probabilities, $1/n, 1/(n^2 + n), 1/(n^2 + n), \dots, 1/(n^2 + n)$, respectively. Since $\binom{n}{2} \sum_{j=1}^{n^2} Q_{jn}^2 \rightarrow 1$, (5) holds with $\Lambda = 1$. On the other hand, the law of Z_n does not converge to the $P_{\circ 1}$ distribution, as $n \rightarrow \infty$. In fact, it can be easily proved that $P(Z_n = 0) \not\rightarrow e^{-1}$ as $n \rightarrow \infty$. Notice that (3) does not hold in this case, since $E(\max_{1 \leq j \leq n^2} Q_{jn}) = 1/n$ fails to be $o(1/n)$, as $n \rightarrow \infty$.

3. Final remarks.

In a Bayesian framework, if assumption (3) is believed to be coherent with the mechanism that generates the data, then, according to Theorem 1, a Poisson model should be fixed for the number of coincidences Z . As far as the prior probability distribution is concerned, notice that the random parameter is the mean of the random probability distribution of Z . Hence, the mixing measure μ in (4) must reflect the statistician's prior opinion on the mean number of coincidences.

Example 3. The birthday problem with unequal probabilities has recently been applied, in a classical i.i.d. setting, to estimate the coincidence probability of surnames in Japan (Mase (1992)). Suppose that you want to face the same estimation problem in a Bayesian framework. Consider a group composed of n individuals and suppose that you are interested in the number of couples of individuals whose surnames coincide. If, on the base of prior information, the probabilities of the different surnames are thought to be uniformly small, compared to n , then (3) can be assumed to be true. Therefore the Poisson model can be chosen for the number Z of matches in the surnames, the random parameter being the mean number of coincidences.

To conclude, notice that, in Theorem 1, $X_{1n}, \dots, X_{nn}$ are assumed to be the first n elements of an infinite exchangeable sequence. It would be interesting to explore the limiting distribution of the number of coincidences, when, for every n , $X_{1n}, \dots, X_{nn}$ is a finite exchangeable sequence.

References

- Barbour A.D., Holst L. and Janson S. (1992) *Poisson Approximation*, Clarendon Press, Oxford.
- Camarri M. and Pitman J. (2000) Limit distributions and random trees derived from the birthday problem with unequal probabilities, *Electronic Journal of Probability*, 5, 1–18.
- Cerquetti A. and Fortini S. (2003) A Poisson approximation for colored graphs under exchangeability, *Studi Statistici, n. 79 Istituto di Metodi Quantitativi - Università "L. Bocconi" - Milano*.
- Chen L. (1975) Poisson approximation for dependent trials, *Annals of Probability*, 3, 534–545.
- Diaconis P. and Holmes S. (2002) A Bayesian peek into Feller volume I, *Sankhyā, Series A*, 64, 820–841.
- Henze N. (1998) A Poisson limit law for a generalized birthday problem, *Statistics & Probability Letters*, 39, 333–336.
- Joag-Dev K. and Proschan F. (1992) Birthday problem with unlike probabilities, *The American Mathematical Monthly*, 99, 10–12.
- Mase S. (1992) Approximations to the birthday problem with unequal occurrence probabilities and their application to the surname problem in Japan, *Annals of the Institute of Statistical Mathematics*, 44, 479–499.
- Von Mises R. (1939) Über aufteilungs und besetzungswahrscheinlichkeiten, *Rev. Fac. Sci. Istanbul*, 4, 145–163.