Z. Wahrscheinlichkeitstheorie verw. Gebiete 33, 69–74 (1975)

An Approximation Theorem for Sums of Certain Randomly Selected Indicators

Louis H. Y. Chen
Department of Mathematics, University of Singapore
Singapore 10, Republic of Singapore

1. Introduction

The Poisson approximation for sums of independent Bernoulli random variables has been of considerable interest in the literature [see Prohorov (1953), Hodges and LeCam (1960), LeCam (1960), Kerstan (1964) and Vervaat (1969)]. The problem was generalised by Chen (1975) to include certain classes of dependent Bernoulli random variables. In this paper a similar approximation theorem is proved for the distribution of $\sum_{i=1}^{n} X_{i\pi(i)}$, where X_{ij}, $i, j=1, 2, \ldots, n$, are independent Bernoulli random variables and $(\pi(1), \pi(2), \ldots, \pi(n))$ a random permutation of $(1, 2, \ldots, n)$ independent of the X_{ij}'s. The nature of the dependence among $X_{1\pi(1)}, \ldots, X_{n\pi(n)}$ differs from that considered in [2].

A number of corollaries are derived from the main theorem. These include results of LeCam (1960), but with larger absolute constants in the bounds. The largeness of the absolute constants can be attributed to the greater generality of the present problem. One of the corollaries is a Poisson counterpart of a theorem of Wald and Wolfowitz (1944), where the latter is actually a limit theorem. [See also Noether (1949), Hoeffding (1951) and Robinson (1972).] Another corollary is an approximation theorem for the hypergeometric distribution.

Throughout this paper, all summations will be from 1 to n unless otherwise stated.

2. The Main Theorem

We first state the theorem as follows:

Theorem 2.1. *Let X_{ij}, $i, j=1, 2, \ldots, n$ be independent Bernoulli random variables with $P(X_{ij}=1)=1-P(X_{ij}=0)=p_{ij}$ and let $(\pi(1), \pi(2), \ldots, \pi(n))$ be a random perpemutation of $(1, 2, \ldots, n)$ independent of the X_{ij}'s. Then for $n \geqq 5$ and every real-valued function h defined on the non-negative integers such that $|h| \leqq 1$, we have*

$$\left|Eh\left(\sum_i X_{i\pi(i)}\right)-\mathscr{P}_\lambda h\right| \leqq 15.75 \min(\lambda^{-\frac{1}{2}}, 1)\left\{\sum_i \bar{p}_{i+}^2+\sum_j \bar{p}_{+j}^2\right\} \tag{2.1}$$

and

$$|Eh(\sum_i X_{i\pi(i)}) - \mathscr{P}_\lambda h| \leqq 45.25\,\lambda^{-1}\{\sum_i \bar{p}_{i+}^2 + \sum_j \bar{p}_{+j}^2\} \tag{2.2}$$

where

$$\mathscr{P}_\lambda h = e^{-\lambda}\sum_{k=0}^{\infty} h(k)\lambda^k/k!\,,$$

$$\bar{p}_{i+} = \sum_j p_{ij}/n, \qquad \bar{p}_{+j} = \sum_i p_{ij}/n,$$

$$\lambda = \sum_i \bar{p}_{i+} = \sum_j \bar{p}_{+j} = \sum_i\sum_j p_{ij}/n.$$

The proof of the theorem is based on the derivation of an identity similar to that in [2] and a few lemmas. (An interesting application of a special case of these identities is in [1].)

Let I, J, K, L, M be random variables, each uniformly distributed on $\{1, 2, \ldots, n\}$ and let $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$, $\tilde{\pi} = (\tilde{\pi}(1), \tilde{\pi}(2), \ldots, \tilde{\pi}(n))$ and $\tilde{\tilde{\pi}} = (\tilde{\tilde{\pi}}(1), \tilde{\tilde{\pi}}(2), \ldots, \tilde{\tilde{\pi}}(n))$ be random permutations of $(1, 2, \ldots, n)$ such that

$$\{I, J, K, L, M, \pi, \tilde{\pi}, \tilde{\tilde{\pi}}\} \text{ is independent of } \{X_{ij}\colon i, j = 1, 2, \ldots, n\}, \tag{2.3}$$

$$(I, K) \text{ and } (L, M) \text{ are uniformly distributed on } \{(i, k)\colon i \neq k;\ i, k = 1, 2, \ldots, n\}, \tag{2.4}$$

$$J, (I, K), (L, M) \text{ and } \tilde{\tilde{\pi}} \text{ are mutually independent,} \tag{2.5}$$

$$J, (I, K) \text{ and } \tilde{\pi} \text{ are mutually independent,} \tag{2.6}$$

$$I \text{ and } \pi \text{ are independent,} \tag{2.7}$$

$$\tilde{\pi}(\alpha) = \begin{cases} \tilde{\tilde{\pi}}(\alpha), & \alpha \neq I, K, \tilde{\tilde{\pi}}^{-1}(L), \tilde{\tilde{\pi}}^{-1}(M) \\ L, & \alpha = I \\ M, & \alpha = K \\ \tilde{\tilde{\pi}}(I), & \alpha = \tilde{\tilde{\pi}}^{-1}(L) \\ \tilde{\tilde{\pi}}(K), & \alpha = \tilde{\tilde{\pi}}^{-1}(M) \end{cases} \tag{2.8}$$

and

$$\pi(\alpha) = \begin{cases} \tilde{\pi}(\alpha), & \alpha \neq I, \tilde{\pi}^{-1}(J) \\ J, & \alpha = I \\ \tilde{\pi}(I), & \alpha = \tilde{\pi}^{-1}(J), \end{cases} \tag{2.9}$$

where $\tilde{\pi}(\tilde{\pi}^{-1}(\alpha)) = \alpha$ and $\tilde{\tilde{\pi}}(\tilde{\tilde{\pi}}^{-1}(\alpha)) = \alpha$.

The consistency of the conditions (2.3)–(2.9) can easily be verified.

Now let $(\Omega, \mathscr{B}, P)$ be a common probability space on which all the above random vectors are defined and let

$\mathscr{F}$ = the σ-algebra generated by π and the X_{ij}'s,

$$W = \sum_i X_{i\pi(i)}, \qquad \tilde{W} = \sum_i X_{i\tilde{\pi}(i)}, \qquad \tilde{\tilde{W}} = \sum_i X_{i\tilde{\tilde{\pi}}(i)},$$

$$W^* = \sum_{i \neq I} X_{i\pi(i)}, \qquad \tilde{W}^* = \sum_{i \neq I} X_{i\tilde{\pi}(i)}, \qquad \tilde{W}^{**} = \sum_{i \neq I, K} X_{i\tilde{\pi}(i)}.$$

Also define the operator Δ by $\Delta f(w)=f(w+1)-f(w)$. Then, using the basic properties of conditional expectations, the fact that each X_{ij} takes on 0 and 1, the conditional independence of $X_{I\pi(I)}$ and W^* given (I,π) and the independence of I and W, we obtain, for every real-valued function f defined on $\{0,1,2,\ldots\}$.

$$\begin{aligned}
E[Wf(W)] &= nE\{[E^{\mathscr{F}} X_{I\pi(I)}]f(W)\} = nE[X_{I\pi(I)}f(W)] = nE[p_{I\pi(I)}f(W^*+1)] \\
&= nE[(p_{I\pi(I)}-\bar{p}_{I+})f(W^*+1)] + nE\{\bar{p}_{I+}[f(W^*+1)-f(W+1)]\} \\
&\quad + \lambda Ef(W+1) \\
&= nE[(p_{IJ}-\bar{p}_{I+})f(W^*+1)] - nE[\bar{p}_{I+}\,p_{IJ}\,\Delta f(W^*+1)] \\
&\quad + \lambda Ef(W+1), \qquad (2.10)
\end{aligned}$$

where $E^{\mathscr{F}}$ denotes conditional expectation given the σ-algebra $\mathscr{F}$. Using again the fact that each X_{ij} takes on 0 and 1, the conditional independence of p_{IJ} and $\tilde{W}^*$ given I and the conditional independence of

$$(X_{\tilde{\pi}^{-1}(J),\tilde{\pi}(I)}, X_{\tilde{\pi}^{-1}(J),J}) \quad \text{and} \quad \sum_{\alpha \neq I, \tilde{\pi}^{-1}(J)} X_{\alpha\tilde{\pi}(\alpha)}$$

given $(I,J,\tilde{\pi})$, we obtain

$$\begin{aligned}
& nE[(p_{IJ}-\bar{p}_{I+})f(W^*+1)] \\
&\quad = nE\{(p_{IJ}-\bar{p}_{I+})[f(W^*+1)-f(\tilde{W}^*+1)]\} \\
&\quad = nE\{(p_{IJ}-\bar{p}_{I+})(X_{\tilde{\pi}^{-1}(J),\tilde{\pi}(I)}-X_{\tilde{\pi}^{-1}(J),J})\,\Delta f(\textstyle\sum_{\alpha\neq I,\tilde{\pi}^{-1}(J)} X_{\alpha\tilde{\pi}(\alpha)}+1)\} \\
&\quad = nE\{(p_{IJ}-\bar{p}_{I+})(p_{\tilde{\pi}^{-1}(J),\tilde{\pi}(I)}-p_{\tilde{\pi}^{-1}(J),J})\,\Delta f(\textstyle\sum_{\alpha\neq I,\tilde{\pi}^{-1}(J)} X_{\alpha\tilde{\pi}(\alpha)}+1)\} \\
&\quad = n(n-1)E\{(p_{IM}-\bar{p}_{I+})(p_{KL}-p_{KM})\,\chi(J=M)\,\Delta f(\tilde{W}^{**}+1)\} \qquad (2.11)
\end{aligned}$$

where $\chi(A)$ is the indicator function of the set A. Now, as in [2], we choose f such that

$$wf(w)-\lambda f(w+1)=h(w)-\mathscr{P}_\lambda h \qquad (2.12)$$

where h is another real-valued and bounded function defined on $\{0,1,2,\ldots\}$, and let $S_\lambda h(w)$ denote the solution of the difference equation (2.12) for $w \geqq 1$ (the solution is unique except at $w=0$). Then (2.10) and (2.11) yield

$$\begin{aligned}
Eh(W) &= \mathscr{P}_\lambda h + n(n-1)E[(p_{IM}-\bar{p}_{I+})(p_{KL}-p_{KM})\,\chi(J=M)\,\Delta S_\lambda h(\tilde{W}^{**}+1) \\
&\quad - nE[\bar{p}_{I+}\,p_{IJ}\,\Delta S_\lambda h(W^*+1)]. \qquad (2.13)
\end{aligned}$$

In order to bound the error terms on the right hand side of (2.13), we need a few lemmas.

Lemma 2.1 [2]. *For* $|h| \leqq 1$ *and* $w \geqq 1$,

$$|\Delta S_\lambda h(w)| \leqq 6 \min(\lambda^{-\frac{1}{2}}, 1).$$

Lemma 2.2. [2] *For* $|h| \leqq 1$ *and* $w \geqq 1$,

$$|\Delta S_\lambda h(w)| \leqq \lambda^{-1}\{2+4|w-\lambda| \min(\lambda^{-\frac{1}{2}}, 1)\}.$$

Lemma 2.3. *For* $Z=\tilde{W}$ *or* $\tilde{\tilde{W}}$,

$$E(Z-\lambda)^2 \leqq \lambda.$$

Proof. Direct computations yield

$$E(Z-\lambda)^2 = \lambda + \lambda^2/(n-1) - n\sum_i \bar{p}_{i+}^2/(n-1) - n\sum_j \bar{p}_{+j}^2/(n-1) + \sum_i\sum_j p_{ij}^2/n(n-1).$$

This together with the inequalities $\lambda^2 \leqq n\sum_i \bar{p}_{i+}^2$ and $\sum_i p_{ij}^2 \leqq (\sum_i p_{ij})^2$ proves the lemma.

Lemma 2.4.

$$n(n-1)\,E|(p_{IM}-\bar{p}_{I+})(p_{KL}-p_{KM})\,\chi(J=M)| \leqq (3n-2)\,\lambda^2/n(n-1) + \sum_j \bar{p}_{+j}^2.$$

Proof. By direct computations.

We now prove Theorem 2.1. In the following, we shall take h in (2.13) to be such that $|h| \leqq 1$. We first bound the second error term on the right hand side of (2.13). By Lemma 2.1, we obtain

$$|nE[\bar{p}_{I+}\,p_{IJ}\,\Delta S_\lambda h(W^*+1)]| \leqq 6\min(\lambda^{-\frac{1}{2}}, 1)\sum_i \bar{p}_{i+}^2. \tag{2.14}$$

Also, by Lemma 2.2, the inequality $|W^*+1-\tilde{W}| \leqq 2$ and the independence of I, J and $\tilde{W}$, we obtain

$$|nE[\bar{p}_{I+}\,p_{IJ}\,\Delta S_\lambda h(W^*+1)]| \leqq \lambda^{-1}[nE\bar{p}_{I+}\,p_{IJ}][10 + 4\min(\lambda^{-\frac{1}{2}}, 1)\,E|\tilde{W}-\lambda|]$$

which by Jensen's inequality and Lemma 2.3

$$\leqq 14\lambda^{-1}\sum_i \bar{p}_{i+}^2. \tag{2.15}$$

Next we consider the first error term on the right hand side of (2.13). By Lemmas 2.1 and 2.4, we obtain

$$\begin{aligned} &|n(n-1)E[(p_{IM}-\bar{p}_{I+})(p_{KL}-p_{KM})\chi(J=M)\,\Delta S_\lambda h(\tilde{W}^{**}+1)]| \\ &\quad \leqq 6\min(\lambda^{-\frac{1}{2}}, 1)\,\{(3n-2)\,\lambda^2/n(n-1) + \sum_j \bar{p}_{+j}^2\}. \end{aligned} \tag{2.16}$$

Also, by Lemma 2.2, the inequality $|\tilde{W}^{**}+1-\tilde{\tilde{W}}| \leqq 3$, the independence of $J, (I, K), (L, M)$ and $\tilde{\tilde{W}}$, and Lemma 2.4, we obtain as in (2.15)

$$\begin{aligned} &|n(n-1)E[(p_{IM}-\bar{p}_{I+})(p_{KL}-p_{KM})\,\chi(J=M)\,\Delta S_\lambda h(\tilde{W}^{**}+1)]| \\ &\quad \leqq 18\lambda^{-1}\{(3n-2)\lambda^2/n(n-1) + \sum_j \bar{p}_{+j}^2\}. \end{aligned} \tag{2.17}$$

Finally, noting that $n \geqq 5$, that $\lambda^2 \leqq n\sum_i \bar{p}_{i+}^2$ and that $\lambda^2 \leqq n\sum_j \bar{p}_{+j}^2$, we obtain (2.1) from (2.13), (2.14) and (2.16), and obtain (2.2) from (2.13), (2.15) and (2.17). This completes the proof of Theorem 2.1.

3. Corollaries

Except for Corollary 3.2, all the corollaries in this section have already been mentioned in the Introduction. Corollaries 3.3 and 3.4 are actually corollaries to

Corollary 3.2. Unless otherwise stated, all notations are the same as in the preceding sections.

Corollary 3.1 (LeCam). *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with $P(X_i=1)=1-P(X_i=0)=p_i$. Then for $n \geqq 5$ and $|h| \leqq 1$, we have*

$$|Eh(\sum_i X_i) - \mathscr{P}_\lambda h| \leqq 31.5 \min(\lambda^{-\frac{1}{2}}, 1) \sum_i p_i^2$$

and

$$|Eh(\sum_i X_i) - \mathscr{P}_\lambda h| \leqq 90.5\, \lambda^{-1} \sum_i p_i^2$$

where

$$\lambda = \sum_i p_i.$$

Proof. Let $p_{ij}=p_j$ for all i and j and observe that $\lambda^2 \leqq n \sum_i p_i^2$.

Corollary 3.2. *Let a_{ij}, $i,j=1,2,\ldots,n$, be an array of 0's and 1's and let $(\pi(1), \pi(2), \ldots, \pi(n))$ be a random permutation of $(1,2,\ldots,n)$. Then for $n \geqq 5$ and $|h| \leqq 1$, we have*

$$|Eh(\sum_i a_{i\pi(i)}) - \mathscr{P}_\lambda h| \leqq 15.75 \min(\lambda^{-\frac{1}{2}}, 1) \{\sum_i p_i^2 + \sum_j q_j^2\}$$

and

$$|Eh(\sum_i a_{i\pi(i)}) - \mathscr{P}_\lambda h| \leqq 45.25\, \lambda^{-1} \{\sum_i p_i^2 + \sum_j q_j^2\}$$

where

$$p_i = \sum_j a_{ij}/n, \qquad q_j = \sum_i a_{ij}/n,$$

$$\lambda = \sum_i p_i = \sum_j q_j = \sum_i \sum_j a_{ij}/n.$$

Proof. Let $p_{ij}=a_{ij}$ for every i and j.

Corollary 3.3. *Let $a_1, \ldots, a_n$, $b_1, \ldots, b_n$ be 0's and 1's and let $(\pi(1), \ldots, \pi(n))$ be a random permutation of $(1, \ldots, n)$. Then for $n \geqq 5$ and $|h| \leqq 1$, we have*

$$|Eh(\sum_i a_i b_{\pi(i)}) - \mathscr{P}_\lambda h| \leqq 45.25(\bar{a}+\bar{b})$$

where

$$\bar{a} = \sum_i a_i/n, \qquad \bar{b} = \sum_i b_i/n, \qquad \lambda = n\bar{a}\bar{b}.$$

Proof. This follows from Corollary 3.2 with $a_{ij}=a_i b_j$.

Corollary 3.4. *Let*

$$h(r; n, a, b) = \begin{cases} \binom{a}{r}\binom{n-a}{b-r} \Big/ \binom{n}{b} & \text{if } \max(a+b-n, 0) \leqq r \leqq a \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\sum_{r=0}^{\infty} |h(r; n, a, b) - e^{-\lambda} \lambda^r / r!| \leqq 45.25 (a+b)/n$$

where

$$\lambda = ab/n.$$

Proof. This follows from Corollary 3.2 with $a_{ij} = 1$ if $1 \leqq i \leqq a$, $1 \leqq j \leqq b$ and $= 0$ otherwise, and with $h(r) = 1$ or -1 according as $h(r; n, a, b) \geqq$ or $< e^{-\lambda} \lambda^r / r!$.

Acknowledgement. This paper is a part of the author's Ph. D. dissertation written at Stanford University under the supervision of Professor Charles Stein. The author wishes to thank Professor Stein for his guidance and suggestions.

References

1. Chen, Louis H. Y.: On the convergence of Poisson binomial to Poisson distributions. Ann. Probability **2**, 178–180 (1974)
2. Chen, Louis H. Y.: Poisson approximation for dependent trials. Ann. Probability **3**, in press (1975)
3. Hodges, J. L., LeCam, L.: The Poisson approximation to the Poisson binomial distribution. Ann. Math. Statist. **31**, 737–740 (1960)
4. Hoeffding, W.: A combinatorial limit theorem. Ann. Math. Statist. **22**, 558–566 (1951)
5. Kerstan, J.: Verallgemeinerung eines Satzes von Prochorow und LeCam. Z. Wahrscheinlichkeitstheorie verw. Gebiete **2**, 173–179 (1964)
6. LeCam, L.: An approximation theorem for the Poisson binomial distribution. Pacific J. Math. **10**, 1181–1197 (1960)
7. Noether, G. E.: On a theorem by Wald and Wolfowitz. Ann. Math. Statist. **20**, 455–458 (1949)
8. Prohorov, J. V.: Asymptotic behavior of the binomial distribution. Uspehi Mat. Nauk. **8**, 135–142 (1953)
9. Robinson, J.: A converse to a combinatorial limit theorem. Ann. Math. Statist. **43**, 2053–2057 (1972)
10. Vervaat, W.: Upper bounds for the distance in total variation between the binomial or negative binomial and the Poisson distribution. Statistica Neerlandica **23**, 79–86 (1969)
11. Wald, A., Wolfowitz, J.: Statistical tests on permutations of observations. Ann. Math. Statist. **15**, 358–372 (1944)

Received October 1, 1973