International Journal of Mathematics and
Computer Science, **16**(2021), no. 4, 1771–1781

Error estimation for nonoverlapping success runs with length k via Stein-Chen method

C. Sahatsathatsana[1], S. Sahatsathatsana[2], W. Pimpasalee[3]

[1,3]Department of Science and Mathematics
Faculty of Science and Health Technology
Kalasin University
Kalasin, Thailand

[2]Department of Foreign Language
Faculty of Liberal Arts
Kalasin University
Kalasin, Thailand

email: chanokgan.na@ksu.ac.th

(Received April 28, 2021, Accepted June 14, 2021)

Abstract

Run statistics and patterns in a sequence of Bernoulli trials, and multistate trials have broadly been used for various purposes in the areas of statistics and applied probability. A class of enumeration schemes for success runs of a specified length including both nonoverlapping and overlapping success runs. Consequently, we are interested in studying the problem of the number of nonoverlapping success runs of length k ($1 \leq k \leq n$) to provide an error estimation of this problem through the use of the Stein-Chen coupling method.

1 Introduction

The concept of runs has been used in a variety of applications such as hypothesis testing (run-test), (Wald and Wolfowitz [3] and Walsh [7]), statistical

Key words and phrases: Nonoverlapping success runs, Poisson approximation, Stein-Chen method.
AMS (MOS) Subject Classifications: 60G07.
ISSN 1814-0432, 2021, http://ijmcs.future-in-tech.net

quality control (Mosteller [5] and Wolfowitz [9]), DNA sequencing, psychology, and ecology (Schwager [15]). The theory of distribution of runs seems to have been created at the end of the nineteenth century (Stevens [18]). There were approximately 1940 research works on the distribution of runs theory such as the studies of Wishart and Hirsheld [8], Cochran [17], Mood [2], Wald and Wolfowitz [3], and Wolfowitz [9]. Most authors focused on the study of the conditional distributions of runs given the total number of successes. There was less research focusing on the exact and limiting distributions of runs during the period from 1950 to 1970. However, there were some very interesting studies on the approximate formulae for the distributions of runs and patterns developed such as (Walsh [7] and Gibbons [6]) which recently has been successfully used in many diverse areas of applied research including molecular biology, financial engineering, psychology, ecology, and computer science. A class of enumeration schemes for success runs of a specified length including nonoverlapping success runs and overlapping success runs is rigorously studied. In this special model, the success run is defined as a sequence of consecutive successes (S) preceded and succeeded by failures (F). The number of successes in the run will be referred to as its length.

In this paper, we study the distribution of the number of nonoverlapping occurrence of the success run of length k in n trials. Moreover, We define the most important and frequently used statistics of success runs associated with nonidentical independent Bernoulli trials $X_1, X_2, \ldots, X_n$ with success probabilities p and failure probabilities $q = 1 - p$, for n and k $(1 \leq k \leq n)$, in case when $W_{n,k}$ is the number of nonoverlapping success runs of length k. Furthermore, we give examples 1.1 and 1.2 of enumeration scheme for nonoverlapping success runs:

Example 1.1. *Let $n = 17$ trials be performed, numbered from 1 to 17, where S denotes success, F denotes failure of a specific trial and that we get the following outcomes*

$$SSSSSFFSSFSSSFFSS.$$

By the nonoverlapping way of enumeration, the sequence

$$(SSS)SSFFSSF(SSS)FFSS$$

contains 2 (0-overlapping) success runs of length $k = 3$ and the outcomes corresponding to the trials numbered by

1 2 3 and 11 12 13,

Then the number of nonoverlapping success runs of length 3 *is 2*

Example 1.2. *Let us enumerate the number of success runs of length* $k = 4$ *in the sequence*

$$SSSSSSFFSSSFSSSFFSSS$$

By the nonoverlapping way of enumeration, the sequence

$$(SSSS)SSFF(SSSS)FSSSFFSS$$

contains 2 (0-overlapping) success runs of length $k = 4$ *and the outcomes corresponding to the trials numbered by*

1 2 3 4 and 9 10 11 12

Then the number of nonoverlapping success runs of length 4 *is 2.*

Examples 1.1 and 1.2 suggest investigating n beginning sufficiently large of the probability of the number of success runs of length k. We set up this problem in the framework of Stein-chen method. Let Ω be the space of possible arrangements of n Bernoulli trials with success probabilities p, $(0 < p < 1)$ and failure probabilities $q = 1 - p$. Given n and k, $(1 \leq k \leq n)$, for each $i \in \{1, 2, \ldots, n\}$, we define the indicator random variable X_i, as follows:

$$X_i = \begin{cases} 1 & \text{if nonoverlapping success run of} \\ & \text{length } k \text{ starts at the } i^{th} \text{ trials,} \\ 0 & \text{otherwise.} \end{cases}$$

We can construct random variables to solve the problem as follows:
For each $i \in \{1, 2, \ldots n\}$, let

$$W_{n,k} = \sum_{i=1}^{n-k+1} X_i.$$

Then $W_{n,k}$ is the number of nonoverlapping success runs with length k which appear within the first n Bernoulli trials with success probability p.
Therefore,

$$P(X_i = 1) = p^k.$$

For n sufficiently large, it is logical to approximate the distribution of W_n by a Poisson distribution with mean

$$\lambda = \mathrm{E}(W_{n,k}) = (n - k + 1)p^k.$$

The objective of our research is to study and approximate the distribution of the number of nonoverlapping success runs of length k by using the Stein-Chen coupling method which is introduced in Section 2. The following theorem is our main result.

Theorem 1.1. *Let $W_{n,k}$ be the number of nonoverlapping success runs of length k which appear within the first n Bernoulli trials with success probability p and $\lambda = (n-k+1)p^k$. Then*

1. $|P(W_n \in A) - \mathcal{P}_\lambda(A)| \leq C_{\lambda,A}$.
2. $|P(W_n \in A) - \mathcal{P}_\lambda(A)| \leq 1 - e^{-\lambda}$,

where $C_{\lambda,A} = \min\left\{1, \lambda, \frac{\triangle(\lambda)}{M_A+1}\right\}$,

$$\triangle(\lambda) = \begin{cases} e^\lambda + \lambda - 1 & \text{if } \lambda^{-1}(e^\lambda - 1) \leq M_A, \\ 2(e^\lambda - 1) & \text{if } \lambda^{-1}(e^\lambda - 1) > M_A, \end{cases}$$

and

$$M_A = \begin{cases} \max\{w \mid C_w \subseteq A\} & \text{if } 0 \in A, \\ \min\{w \mid w \in A\} & \text{if } 0 \notin A \end{cases}$$

when $C_w = \{0, 1, \ldots, w-1\}$.

According to Theorem 1.1, for n sufficiently large and $A \subseteq \mathbb{N} \cup \{0\}$, we can approximate the cumulative probability of the the number of nonoverlapping success runs of length k , $P(W_n \in A)$, by the cumulative Poisson probability, $Poi_\lambda(A)$ with $\lambda = (n-k+1)p^k$; i.e.,

$$P(W_n \in A) \approx Poi_\lambda(A), \qquad \text{when } n \to \infty \tag{1.1}$$

2 Poisson approximation via Stein-Chen method

The Stein-Chen method was developed to show that the probabilities of rare events can be approximated by Poisson probabilities. Stein [4] introduced a new powerful technique for obtaining the rate of convergence to standard normal distribution. Moreover, his basic idea has many applications. Chen [13] modified Stein's method so as to obtain approximation results for the Poisson distribution.

Our starting point is the Stein equation for the Poisson distribution which gives:

$$I_A(j) - \mathcal{P}_\lambda(A) = \lambda g_{\lambda,A}(j+1) - j g_{\lambda,A}(j) \tag{2.2}$$

$\lambda > 0$, $j \in \mathbb{N} \cup \{0\}$, $A \subseteq \mathbb{N} \cup \{0\}$ and $I_A : \mathbb{N} \cup \{0\} \to \mathbb{R}$ is defined by

$$I_A(w) = \begin{cases} 1 & ; w \in A, \\ 0 & ; w \notin A. \end{cases}$$

The solution $g_{\lambda,A}$ of (2.2) is of the form

$$g_{\lambda,A}(w) = \begin{cases} (w-1)!\lambda^{-w}e^{\lambda}[\mathcal{P}_\lambda(I_{A\cap C_{w-1}}) - \mathcal{P}_\lambda(I_A)\mathcal{P}_\lambda(I_{C_{w-1}})] & ; w \geq 1, \\ 0 & ; w = 0 \end{cases}$$

where

$$\mathcal{P}_\lambda(I_A) = e^{-\lambda}\sum_{l=0}^{\infty} I_A(l)\frac{\lambda^l}{l!}$$

and

$$C_{w-1} = \{0, 1, \ldots, w-1\}.$$

By replacing j and λ in (2.2) by any integer-valued random variable W_n and $\lambda = \mathrm{E}W_n$, we have

$$P(W_n \in A) - \mathcal{P}_\lambda(A) = E(\lambda g_{\lambda,A}(W_n + 1)) - E(W_n g_{\lambda,A}(W_n)). \tag{2.3}$$

In the case where the dependence among the instances of X_i is global, there is an alternative approach to approximate the distribution of W_n. This approach is referred to as The Coupling Approach, which was initially proposed by Barbour [1], is particularly useful when it is possible to construct a random variable $W_{n,i}$, for each a $i \in \{1, 2, \ldots, n\}$ on a common probability space with W_n such that $W_{n,i}$ is distributed as $W_n - X_i$ conditional on the event $X_i = 1$.

There has been a number of successful applications of this method (see [1],[14] and [11]).

Theorem 2.1. *If W_n and $W_{n,i}$ are defined as above, then*

$$|P(W_n \in A) - Poi_\lambda(A)| \leq \| g_{\lambda,A} \| \sum_{i=1}^{n} p_i E|W_n - W_{n,i}| \tag{2.4}$$

where $\| g_{\lambda,A} \| := \sup_w [g_{\lambda,A}(w+1) - g_{\lambda,A}(w)]$.

Many authors wanted to determine a bound of $\| g_{\lambda,A} \|$. For $A \subseteq \mathbb{N} \cup \{0\}$, in 1975, Chen [13] proved that:

$$\| g_{\lambda,A} \| \leq \min\{1, \lambda^{-1}\}$$

and in 1994, Janson [14 showed that

$$\| g_{\lambda,A} \| \leq \lambda^{-1}(1 - e^{-\lambda}). \tag{2.5}$$

In case of non-uniform bound,in 2003, Neammanee [10] showed that

$$\| g_{\lambda,A} \| \leq \min\left\{\frac{1}{w_0}, \lambda^{-1}\right\}$$

and, in 2005, Teerapabolarn and Neammanee [12] gave a bound of $\| g_{\lambda,A} \|$, where $A = \{0, 1, \ldots, w_0\}$ in terms of:

$$\| g_{\lambda,A} \| \leq \lambda^{-1}(1 - e^{-\lambda}) \min\left\{1, \frac{e^{\lambda}}{w_0 + 1}\right\}.$$

In the general case, for any subset A of $\{0, 1, \ldots, n\}$, in 2006, Santiwipanont and Teerapabolarn [16) gave a bound in the form of

$$\| g_{\lambda,A} \| \leq \lambda^{-1} \min\left\{1, \lambda, \frac{\triangle(\lambda)}{M_A + 1}\right\} \tag{2.6}$$

where

$$\triangle(\lambda) = \begin{cases} e^{\lambda} + \lambda - 1 & \text{if } \lambda^{-1}(e^{\lambda} - 1) \leq M_A, \\ 2(e^{\lambda} - 1) & \text{if } \lambda^{-1}(e^{\lambda} - 1) > M_A, \end{cases}$$

and

$$M_A = \begin{cases} \max\{w \mid C_w \subseteq A\} & \text{if } 0 \in A, \\ \min\{w \mid w \in A\} & \text{if } 0 \notin A. \end{cases}$$

The difficult part in applying Theorem 2.1 is to find $W_{n,i}$ which makes $E|W_n - W_{n,i}|$ small. For the case when $X_1, \ldots, X_n$ are independent, we let $W_{n,i} = W_n - X_i$. Then $E|W_n - W_{n,i}| = p_i$ and from (2.4), we have

$$|P(W_n \in A) - Poi_{\lambda}(A)| \leq \| g_{\lambda,A} \| \sum_{i=1}^{n} p_i^2.$$

The problem of constructing of $W_{n,i}$ is difficult in the case of dependent indicator summand. In the next section, we will use Theorem 2.1 to prove our main result by constructing the random variable $W_{n,i}$ which makes $E|W_n - W_{n,i}|$ small.

3 Proof of Main Results

Proof. (Theorem 1.1). For each $j \in \{1, 2, 3, \ldots, n\}$ such that $j \neq i$, let us define the indicator random variable X_{ij} as follows:

$$X_i = \begin{cases} 1 & \text{if the } j^{th} \text{ trial contained success runs} \\ & \text{length } k \text{ after removing the } i^{th} \text{ trial} \\ & \text{which contained success with length } k \\ 0 & \text{otherwise.} \end{cases}$$

Let $W_{n,k,i}$ be the number of nonoverlapping success runs with length k which appear within the first n Bernoulli trials after we take the i^{th} trials having the nonoverlapping success runs with length k and write $W_{n,k} = \sum_{i=1}^{n-k+1} X_i$. For each $w_0 \in \{1, 2, \ldots, n-k\}$, suppose that $\{j_s | s = 1, 2, \ldots, w_0\}$ is the set of w_0 nonoverlapping success runs with length k. Then

$$P(W_{n,k,i} = w_0) = (\sum_{s=1}^{w_0} p_{j_s})^k$$

and

$$\begin{aligned} P(W_{n,k} - X_i = w_0 \mid X_i = 1) &= \frac{P(W_{n,k} - X_i = w_0, X_i = 1)}{P(X_i = 1)} \\ &= \frac{P(W_{n,k} = w_0 + 1, X_i = 1)}{P(X_i = 1)} \\ &= \frac{(p_i \sum_{s=1}^{w_0} p_{j_s})^k}{{p_i}^k} \\ &= (\sum_{s=1}^{w_0} p_{j_s})^k . \end{aligned}$$

It is clear that $W_{n,k,i}$ so constructed is distributed as $W_{n,k} - 1$ conditional on $X_i = 1$. We observe that:

- In case $X_i = 1$, we have the number of nonoverlapping success runs with length k which appear within the first n Bernoulli trials after we take the i^{th} trials that has nonoverlapping success runs with length k equals to the number of nonoverlapping success runs with length k which appear within the first n Bernoulli trials minus 1; that is,

$$W_{n,k,i} = W_{n,k} - 1. \tag{3.7}$$

- In case $X_i = 0$, the number of nonoverlapping success runs with length k which appear within the first n Bernoulli trials after we take the i^{th} trials that has nonoverlapping success runs with length k and we take them again as defined, equals to the appearances of nonoverlapping success runs with length k which appear within the first n trials minus the sum of number of the j^{th} trials, $i \neq j$, is the number of appearances of nonoverlapping success runs with length k in the first count, and they do not contain success runs with length k after we take them again; that is,

$$W_{n,k,i} = W_{n,k} - \sum_{i,j=1,i\neq j}^{n-k+1} X_{ij}Y_{ij}. \tag{3.8}$$

For each $j \in \{0, 1, 2, \ldots, n\}$, $i \neq j$, we define indicator random variable as follows:

$$Y_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ trial did not contain success runs} \\ & \text{with length } k \text{ after removing the } i^{th} \text{ trial} \\ & \text{which contained success with length } k \\ 0 & \text{otherwise.} \end{cases}$$

So the probability that $Y_{ij} = 1$ is given by

$$\begin{aligned} P(Y_{ij} = 1) &= 1 - (\frac{p_i p_j}{p_i})^k \\ &= 1 - p_j^k. \end{aligned} \tag{3.9}$$

We know that

$$E \mid W_{n,k} - W_{n,k,i} \mid = E(W_{n,k} - W_{n,k,i})^+ + E(W_{n,k} - W_{n,k,i})^-,$$

where

$$(W_{n,k} - W_{n,k,i})^+ = \max\{W_{n,k} - W_{n,k,i}, 0\},$$

and

$$(W_{n,k} - W_{n,k,i})^- = -\min\{W_{n,k} - W_{n,k,i}, 0\}.$$

Form (3.7) and (3.8).

- In case $X_i = 1$, we have $(W_{n,k} - W_{n,k,i})^+ = 1$ and $(W_{n,k} - W_{n,k,i})^- = 0$

- In case $X_i = 0$, we have $(W_{n,k} - W_{n,k,i})^+ = \sum_{i,j=1,i\neq j}^{n-k+1} X_{ij}Y_{ij}$ and $(W_{n,k} - W_{n,k,i})^- = 0$.

Therefore, $(W_{n,k} - W_{n,k,i})^+ = \sum_{i,j=1,i\neq j}^{n-k+1} X_{ij}Y_{ij}$ and $(W_{n,k} - W_{n,k,i})^- = 0$. By the fact that

$$\begin{aligned} E(W_{n,k} - W_{n,k,i}) &= E(W_{n,k} - W_{n,k,i})^+ \qquad (3.10)\\ &= E\{\sum_{i,j=1,i\neq j}^{n-k+1} X_{ij}Y_{ij}\} \\ &= \sum_{i,j=1,i\neq j}^{n-k+1} E\{X_{ij}Y_{ij}\} \\ &= \sum_{i,j=1,i\neq j}^{n-k+1} P(X_{ij}=1, Y_{ij}=1) \\ &= \sum_{i,j=1,i\neq j}^{n-k+1} P(X_{ij}=1)P(Y_{ij}=1) \\ &= \sum_{i,j=1,i\neq j}^{n-k+1} p_j^k(1-p_j^k) \\ &= (n-k)\{p_j^k - p_j^{2k}\} \qquad (3.11) \end{aligned}$$

Hence, by (2.4), (2.6) and (3.10), we have

$$\begin{aligned} |P(W_{n,k}\in A) - \mathcal{P}_\lambda(A)| &\le \lambda^{-1}C_{\lambda,A}\sum_{i=1}^{n} p_i E \mid W_{n,k} - W_{n,k,i} \mid \\ &\le \lambda^{-1}C_{\lambda,A}\sum_{i=1}^{n} p_i(n-k)\{p_j^k - p_j^{2k}\} \\ &\le C_{\lambda,A}, \end{aligned}$$

and, by (2.4), (2.5) and (3.10), we have

$$|P(W_{n,k}\in A) - \mathcal{P}_\lambda(A)| \le (1-e^{-\lambda}),$$

where $C_{\lambda,A} = \min\left\{1, \lambda, \frac{\triangle(\lambda)}{M_A+1}\right\}$,

$$\triangle(\lambda) = \begin{cases} e^\lambda + \lambda - 1 & \text{if } \lambda^{-1}(e^\lambda - 1) \le M_A, \\ 2(e^\lambda - 1) & \text{if } \lambda^{-1}(e^\lambda - 1) > M_A, \end{cases}$$

and

$$M_A = \begin{cases} \max\{w \mid C_w \subseteq A\} & \text{if } 0 \in A, \\ \min\{w \mid w \in A\} & \text{if } 0 \notin A \end{cases}$$

when $C_w = \{0, 1, \ldots, w-1\}$. □

Acknowledgement. We would like to thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

References

[1] A. D. Barbour, Poisson convergence and random graphs, Mathematical Proceedings of the Cambridge Philosophical Society, **92**, no. 2, (1982), 349–359.

[2] A. M. Mood, The distribution theory of runs. Ann. Math. Stat., **11**, (1940), 367–392.

[3] A. Wald, J. Wolfowitz, On a test whether two samples are from the same population, The Annals of Mathematical Statistics, **11**, no. 2, (1940), 147–162.

[4] C. M. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, Proc. Sixth Berkeley Symposium, Math. Stat. and Prob., **3**, (1972), 533–602.

[5] F. Mosteller, Note on an application of runs to quality control charts, Ann. Math. Stat., **12**, (1941), 228–232.

[6] J. D. Gibbons, S. Chakraborti, Nonparametric statistical inference, CRC Press, 2020.

[7] J. E. Walsh, Handbook of Nonparametric Statistics: Results for two and several sample problems, symmetry, and extremes, **2**, Van Nostrand, 1965.

[8] J. Wishart, H. O. Hirschfeld, A theorem concerning the distribution of joins between line segments, Journal of the London Mathematical Society, **1**, no. 3, (1936), 227–235.

[9] J. Wolfowitz, On the theory of runs with some applications to quality control, The Annals of Mathematical Statistics, **14**, no. 3, (1943), 280–288.

[10] K. Neammanee, Pointwise approximation of Poisson binomial by Poisson distribution, Stochastic Modelling and Applications, **6**, (2003), 20–26.

[11] K. Lange, Applied probability, Spinger-Verlag, New York, 2003.

[12] K. Teerapabolarn, K. Neammanee, Poisson approximation for sums of dependent Bernoulli random variables, Acta Mathematica Academiae Paedagogicae Nyiregyhaziensis, **22**, (2006), 87–99.

[13] L. H. Y. Chen, Poisson approximation for dependent trials, The Annals of Probability, **3**, (1975), 534–545.

[14] S. Janson, Coupling and Poisson approximation, Acta Applicandae Mathematica, **34**, (1994), 7–15.

[15] S. J. Schwager, Run probabilities in sequences of Markov-dependent trials. J. Amer. Stat. Assoc., (1983), 168–175.

[16] T. Santiwipanont, K. Teerapabolarn, Two formulas of non-uniform bounds on Poisson approximation for dependent indicators, Thai Journal of Mathematics, **1**, (2006), 15–39.

[17] W. G. Cochran, An extension of Gold's method for examining the apparent persistence of one type of weather, Royal Meteorological Soc. Quart. Journal, **64**, (1938), 631–634.

[18] W. L. Stevens, Distribution of groups in a sequence of alternatives, Annals of Eugenics, **9**, no. 1, (1939), 10–17.