International Journal of Pure and Applied Mathematics
Volume 78 No. 5 2012, 635-640
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu

ijpam.eu

BINOMIAL APPROXIMATION TO THE PÓLYA DISTRIBUTION

K. Teerapabolarn

Department of Mathematics
Faculty of Science
Burapha University
Chonburi, 20131, THAILAND

Abstract: This paper uses Stein's method and the characterization associated with the Pólya random variable to give a bound for the total variation distance between the binomial and Pólya distributions.

AMS Subject Classification: 62E17, 60F05
Key Words: binomial approximation, negative hypergeometric distribution, Pólya distribution, total variation distance, Stein's method

1. Introduction

The Pólya or Pólya-Eggenberger distribution is typically and simply introduced via Pólya urn scheme, discussed in Feller [5]. Start with a single urn containing r red and $N - r$ black balls. A ball is drawn at random, note the color, and return it into the urn together with c additional balls of the same color. Repeat this way for n draws. Let X be the number of red balls taken out in the n drawings, then the distribution of X is a Pólya distribution with parameters N, n, r and c, written by $\mathbb{PY}(N, n, r, c)$. The probability function of X is given

Received: January 16, 2012

by

$$p_X(x) = \frac{\binom{\frac{r}{c}+x-1}{x}\binom{\frac{N-r}{c}+n-x-1}{n-x}}{\binom{\frac{N}{c}+n-1}{n}},\ x = 0, 1, ..., n, \tag{1.1}$$

where $N, n, r, c \in \mathbb{N}$ and the mean and variance of X are $\mu = \frac{nr}{N}$ and $\sigma^2 = \frac{rn(N+cn)(N-r)}{N^2(N+c)}$, respectively. It follows from Brown and Phillips [2] that limiting distribution of X is a negative binomial distribution with parameters $\frac{r}{c}$ and $\frac{1}{1+c\rho}$, where $\rho = \lim_{n,N\to\infty} \frac{n}{N}$ is a constant, and they also gave a bound on the rate of this convergence. Note that, in the case of $c = 1$, Teerapabolarn and Wongkasem [6] gave a bound on the error of the binomial and Pólya probability functions.

Let us consider the probability function in (1.1), it can be expressed as

$$\begin{aligned} p_X(x) &= \binom{n}{x} \frac{\left[\left(\frac{r}{c}+\eta\right)\cdots\frac{r}{c}\right]\left[\left(\frac{N-r}{c}+n-x-1\right)\cdots\frac{N-r}{c}\right]}{\left(\frac{N}{c}+n-1\right)\cdots\frac{N}{c}} \\ &= \binom{n}{x} \frac{\left[\left(\frac{r}{N}+\frac{\eta}{N/c}\right)\cdots\frac{r}{N}\right]\left[\left(\frac{N-r}{N}+\frac{n-x-1}{N/c}\right)\cdots\frac{N-r}{N}\right]}{\left(1+\frac{n-1}{N/c}\right)\cdots 1}, \end{aligned} \tag{1.2}$$

where $\eta = \begin{cases} 0 & \text{if } x = 0, \\ x-1 & \text{if } x = 1, ..., n. \end{cases}$ It is seen that if $r, N \to \infty$ while $\frac{r}{N}$ remains constant, then $p_X(x) \to \binom{n}{x}\left(\frac{r}{N}\right)^x\left(1-\frac{r}{N}\right)^{n-x}$ for every $x = 0, 1, ..., n$, that is, $\mathbb{PY}(N, n, r, c)$ converges to a binomial distribution with parameters n and $\frac{r}{N}$, denoted by $\mathbb{B}(n, r/N)$. Therefore $\mathbb{B}(n, r/N)$ can be used as an estimate of $\mathbb{PY}(N, n, r, c)$ when N is sufficiently large and $\frac{r}{N}$ is a constant.

In this paper, we determine a bound for the total variation distance between $\mathbb{B}(n, r/N)$ and $\mathbb{PY}(N, n, r, c)$. This total variation is defined as follows:

$$d(\mathbb{B}(n,p), \mathbb{PY}(N,n,r,c)) = \sup_{A\subseteq\{0,1,...,n\}} |\mathbb{B}(n,p)\{A\} - \mathbb{PY}(N,n,r,c)\{A\}|, \tag{1.3}$$

where $p = \frac{r}{N}$, $\mathbb{B}(n,p)\{A\} = \sum_{k\in A}\binom{n}{k}p^k(1-p)^{n-k}$ and $\mathbb{PY}(N,n,r,c)\{A\} = \sum_{k\in A} p_X(k)$.

2. Method

The tools for deriving the result for this approximation consist of Stein's method for the binomial distribution and the characterization associated with the Pólya

random variable. We start with Stein's equation in Barbour et al. [1]. Stein's equation for the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1)$ is, for given h, of the form

$$h(x) - \mathcal{B}_{n,p}(h) = (n-x)pg(x+1) - qxg(x), \tag{2.1}$$

where $\mathcal{B}_{n,p}(h) = \sum_{k=0}^{n} h(k)\binom{n}{k}p^k q^{n-k}$ and g and h are bounded real-valued functions defined on $\{0,1,...,n\}$.

For $A \subseteq \{0,1,...,n\}$, let $h_A : \{0,1,...,n\} \to \mathbb{R}$ be defined by

$$h_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases} \tag{2.2}$$

By following Barbour et al. [1] on pp. 189, let $g_A : \mathbb{N} \cup \{0\} \to \mathbb{R}$ satisfy (2.1), where $g_A(0) = g_A(1)$ and $g_A(x) = g_A(n)$ for $x \geq n$. Let $x \in \mathbb{N}$ and $\Delta g_A(x)$=$g_A(x+1) - g_A(x)$, Ehm [4] showed that

$$\sup_A |\Delta g_A(x)| \leq \frac{1 - p^{n+1} - q^{n+1}}{(n+1)pq}. \tag{2.3}$$

Consider an important property of the characterization associated with a non-negative integer-valued random variable Y in Lemma 3.1 of Cacoullos and Papathanasiou [3]. It is stated that if Y has a finite variance, then

$$Cov(Y, f(Y)) = \sum_{y=0}^{\infty} \Delta f(y) \sum_{k=0}^{y} [E(Y) - k] p_Y(k), \tag{2.4}$$

for any function $f : \mathbb{N} \cup \{0\} \to \mathbb{R}$ for which the infinite series is absolutely convergent, where $p_Y(k)$ is the probability function of Y. The following lemma presents the characterization associated with the Pólya random variable.

Lemma 2.1. *Let the Pólya random variable X with $p_X(k) > 0$ for every $k \in \{0,1,...,n\}$ have the associated characterization $a(x) = \frac{\sum_{k=0}^{x}(\mu-k)p_X(k)}{p_X(x)}$, $x = 0,1,...,n$, then the following relations hold:*

$$a(x) = \frac{(n-x)(r+cx)}{N}, \ x = 0,1,...,n \tag{2.5}$$

and for $A \subseteq \{0,1,...,n\}$,

$$\sum_{x=0}^{\infty} \Delta g_A(x) \sum_{k=0}^{x} (\mu - k) p_X(k) = \sum_{x=0}^{n} \Delta g_A(x) \frac{(n-x)(r+cx)}{N} p_X(x). \tag{2.6}$$

Proof. First, we shall show that (2.5) holds.

It is observed that $a(x) = \frac{\sum_{k=0}^{x}(\mu-k)p_X(k)}{p_X(x)}$ can be expressed in the form of recurrence relation

$a(0) = \mu = \frac{nr}{N}$ and $a(x) = a(x-1)\frac{p_X(x-1)}{p_X(x)} + \mu - x$, $x = 1, ..., n$.

Using this relation, we have

$a(1) = \frac{(n-1)(r+c)}{N}, a(2) = \frac{(n-2)(r+2c)}{N}, ..., a(k) = \frac{(n-k)(r+kc)}{N}$ for $2 < k \leq n$.

Therefore, by mathematical induction, (2.5) holds.

For the relation (2.6), it is clear that

$$\sum_{x=0}^{\infty}\Delta g_A(x)\sum_{k=0}^{x}(\mu-k)p_X(k) = \sum_{x=0}^{n}\Delta g_A(x)\sum_{k=0}^{x}(\mu-k)p_X(k)$$
$$= \sum_{x=0}^{n}\Delta g_A(x)a(x)p_X(x). \tag{2.7}$$

Substituting $a(x) = \frac{(n-x)(r+cx)}{N}$ in (2.7), the relation (2.6) is also obtained. □

3. Result

The following theorem shows a bound for the total variation distance between $\mathbb{B}(n, r/N)$ and $\mathbb{PY}(N, n, r, c)$.

Theorem 3.1. *Let X be the Pólya random variable with $p_X(k) > 0$ for every $k \in \{0, 1, ..., n\}$ and $p = 1 - q = \frac{r}{N}$. Then we have*

$$d(\mathbb{B}(n,p), \mathbb{PY}(N,n,r,c)) \leq \frac{(1-p^{n+1}-q^{n+1})c(n-1)n}{(n+1)(N+c)}. \tag{3.1}$$

Proof. For $A \subseteq \{0, 1, ..., n\}$, substituting h, x by h_A, X respectively and taking expectation in (2.1), we obtain

$$\mathbb{B}(n,p)\{A\} - \mathbb{PY}(N,n,r,c)\{A\} = E[(n-X)pg(X+1) - qXg(X)], \tag{3.2}$$

where $g = g_A$ is defined as mentioned above.

Let $\delta(\mathbb{B}, \mathbb{PY}) = \mathbb{B}(n,p)\{A\} - \mathbb{PY}(N,n,r,c)\{A\}$, then we obtain

$$\delta(\mathbb{B}, \mathbb{PY}) = E[npg(X+1) - pX\Delta g(X) - Xg(X)]$$

$$
\begin{aligned}
&= E[npg(X+1)] - pE[X\Delta g(X)] - E[Xg(X)] \\
&= npE[g(X+1)] - pE[X\Delta g(X)] - Cov(X, g(X)) - \mu E[g(X)] \\
&= (nr/N)E[\Delta g(X)] - (r/N)E[X\Delta g(X)] - Cov(X, g(X)).
\end{aligned}
$$

Using Lemma 2.1 and (2.3), we have

$$
\sum_{x=0}^{\infty} \left| \Delta g(x) \sum_{k=0}^{x} (\mu - k) p_X(k) \right| = \sum_{x=0}^{n} |\Delta g(x)| \, \frac{(n-x)(r+cx)}{N} p_X(x) < \infty
$$

and by (2.4) and Lemma 2.1, it follows that

$$
\begin{aligned}
\delta(\mathbb{B}, \mathbb{PY}) &= \sum_{x=0}^{n} \Delta g(x) r \left(\frac{n-x}{N} \right) p_X(x) - \sum_{x=0}^{n} \Delta g(x) \frac{(n-x)(r+cx)}{N} p_X(x) \\
&= -\sum_{x=1}^{n} \Delta g(x) \frac{(n-x)cx}{N} p_X(x).
\end{aligned}
$$

Therefore, it follows from (3.2) and (1.3),

$$
\begin{aligned}
d(\mathbb{B}(n,p), \mathbb{PY}(N,n,r,c)) &\le \sum_{x=1}^{n} |\Delta g(x)| \frac{(n-x)cx}{N} p_X(x) \\
&\le \frac{1 - p^{n+1} - q^{n+1}}{(n+1)pq} \sum_{x=1}^{n} \frac{(n-x)cx}{N} p_X(x) \\
&= \frac{(1 - p^{n+1} - q^{n+1})c(n-1)n}{(n+1)(N+c)}.
\end{aligned}
$$

□

Remark. If $c = 1$ and $N = M + 1$, then $p_X(x) = \frac{\binom{r+x-1}{x}\binom{M-r+n-x}{n-x}}{\binom{M+n}{n}}$, $x = 0, 1, ..., n$, is the negative hypergeometric probability function with parameters M, n and r. Thus, immediately from (3.1), a result on binomial approximation to the negative hypergeometric distribution with parameters M, n and r, denoted by $\mathbb{NH}(M, n, r)$, can also be obtained in the following corollary.

Corollary 3.1. *Let $p = 1 - q = \frac{r}{M+1}$, then we have the following:*

$$
d(\mathbb{B}(n,p), \mathbb{NH}(M,n,r)) \le \frac{(1 - p^{n+1} - q^{n+1})(n-1)n}{(n+1)(M+2)}. \tag{3.3}
$$

4. Conclusion

In this study, a bound for the total variation distance between the binomial and Pólya distributions is obtained using Stein's method and the characterization associated with the Pólya random variable. In view of this bound, it is observed that if $\frac{cn}{N}$ is small, or N is large, then the result in Theorem 3.1 gives a good binomial approximation, that is, the binomial distribution with parameters n and r/N can be used as an approximation of the Pólya with parameters N, n, r and c when N is sufficiently large or $\frac{cn}{N}$ is sufficiently small.

References

[1] A.D. Barbour, L. Holst, S. Janson, *Poisson Approximation*, Oxford Studies in Probability 2, Clarendon Press, Oxford (1992).

[2] T.C. Brown, M.J. Phillips, Negative binomial approximation with Stein's method, *Meth. Comp. Appl. Probab.*, **1** (1999), 407-421.

[3] T. Cacoullos, V. Papathanasion, Characterization of distributions by variance bounds, *Statist. Probab. Lett.*, **7** (1989), 351-356.

[4] W. Ehm, Binomial approximation to the Poisson binomial distribution, *Statist. Probab. Lett.*, **11** (1991), 7-16.

[5] W. Feller, *An Introduction to Probability Theory and its Applications*, Volume 1, Wiley, New York (1968).

[6] K. Teerapabolarn, P. Wongkasem, On pointwise binomial approximation by w-functions, *Int. J. Pure Appl. Math.*, **71** (2011), 57-66.