BULLETIN *of the*
Malaysian Mathematical
Sciences Society
http://math.usm.my/bulletin

Bull. Malays. Math. Sci. Soc. (2) **30**(2) *(2007), 121–128*

Bounds on Random Infinite Urn Model

S. Boonta and K. Neammanee
Department of Mathematics, Faculty of Science,
Chulalongkorn University Bangkok 10330, Thailand
kritsana.n@chula.ac.th, k_neammanee@hotmail.com

Abstract. Let $N(n)$ be a Poisson random variable with parameter n. An infinite urn model is defined as follows: $N(n)$ balls are independently placed in an infinite set of urns and each ball has probability $p_k > 0$ of being assigned to the k-th urn. We assume that $p_k \geq p_{k+1}$ for all k and $\sum_{k=1}^{\infty} p_k = 1$. Let U_n be the number of occupied urns after $N(n)$ balls have been thrown. Dutko showed in 1989 that under the condition $\lim_{n\to\infty} Var(U_n) = \infty$ we have $\frac{U_n - E(U_n)}{\sqrt{Var(U_n)}} \xrightarrow{d} \mathcal{N}(0,1)$ as $n \to \infty$ where $\mathcal{N}(0,1)$ is the standard normal random variable. However, Dutko did not give a bound of his approximation. So in this paper, we give uniform and non-uniform bounds of the approximation.

2000 Mathematics Subject Classification: 60F05, 60G50

Key words and phrases: Infinite urn model, Central limit theorem, Uniform and non-uniform bounds.

1. Introduction and main result

In their paper, Milenkovie and Compton [4] said that there are a lot of application on urn model, since many problems in the area of physics, communication theory, computer science, combinatorial analysis of algorithms can be described in terms of distributing balls (object) into specified urn models in physics are the so called Maxwell-Boltzman, Bose-Einstein and Fermi-Dirac model. In computer science, urn models are used for database performance evaluations and for modeling and analyzing algorithms. Two well known examples of the latter kind are hashing and sorting algorithms. In communication theory, some transmission channels can be described in terms of contagion urn models. There are many problems in the area of network analysis that can be described in terms of urn models (see for more detail on Milenkovie and Compton and references there in).

Let $N(n)$ be a Poisson random variable with parameter n, i.e., $P(N(n) = k) = \frac{e^{-n}n^k}{k!}$ for $k = 0, 1, 2, \ldots$. An infinite urn model is defined as follows: $N(n)$ balls are

Received: July 31, 2006; *Revised:* December 13, 2006.

independently placed in an infinite set of urns and each ball has probability $p_k > 0$ of being assigned to the k-th urn. We assume that $p_k \geq p_{k+1}$ for all k and that $\sum_{k=1}^{\infty} p_k = 1$. Let Z_n be the number of occupied urns after n balls have been thrown. And U_n is the number of accupied urns after $N(n)$ balls have been thrown. Since the number of urns is infinite and the number of throws is random, we cannot apply the usual central limit theorem to U_n. Karlin(1967) gave the condition on (p_k) for the convergence of $\dfrac{Z_n - E(Z_n)}{b_n}$ to $\mathcal{N}(0,1)$ where $\mathcal{N}(0,1)$ is the standard normal random variable and $b_n^2 \sim Var(Z_n)$. Dutko [2] considered in case of the number of balls to be thrown into the urns is Poisson distributed with mean n and showed that

$$Var(U_n) = \sum_{k=1}^{\infty} (e^{-np_k} - e^{-2np_k})$$

and under the condition

$$\lim_{n\to\infty} Var(U_n) = \infty, \tag{1.1}$$

we have

$$\frac{U_n - E(U_n)}{\sqrt{Var(U_n)}} \xrightarrow{d} \mathcal{N}(0,1) \text{ as } n \to \infty. \tag{1.2}$$

However, Dutko did not give a bound of his approximation. So in this work, we give uniform and non-uniform bounds of (1.2) by using Stein's method. In 1972, Stein [5] gave a new technique to find a bound in normal approximation. His technique relied instead on the elementary differential equation. Chen and Shao [1] combined truncation with Stein's method and by taking the concentration inequality approach to find uniform and non-uniform bounds on Berry-Esseen theorem. In this paper, we use the technique in Chen and Shao [1] to obtain bounds on the convergence of (1.2). Here are our main results.

Theorem 1.1. *Let F_n and Φ be the distribution functions of $\dfrac{U_n - E(U_n)}{\sqrt{Var(U_n)}}$ and $\mathcal{N}(0,1)$ respectively. Then*

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{6.66}{\sqrt{Var(U_n)}} \tag{1.3}$$

and

$$|F_n(x) - \Phi(x)| \leq \frac{C}{(1+|x|)^3\sqrt{Var(U_n)}} \tag{1.4}$$

for some $C > 0$.

Furthermore, under the condition (1.1) *we have the bounds in* (1.3) *and* (1.4) *tend to zero.*

Theorem 1.2. *Let F_n and Φ be defined as in Theorem 1.1. Then*

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{3.24}{\sqrt{Var(U_n)}}. \tag{1.5}$$

Dutko [2] gave examples that make $\lim_{n\to\infty} Var(U_n) = \infty$, for examples, $p_k = \frac{C}{k^{\log k}}$ and $p_k = \frac{C}{k^r}$ where C is a normalizing constant and $r > 1$.

2. Proof of main theorems

Let S_{nk} = number of balls in the k-th urn after n throws, and
$T_{n,k}$ = number of balls in the k-th urn after $N(n)$ throws.
The random variables $\{T_{n,k}\},\ k = 1, 2, ...$ are mutually independent Poisson random variables with respective mean $\{np_k\}$([2], p. 1259), so that

$$P(T_{n,k} = r) = \frac{e^{-np_k}(np_k)^r}{r!} \text{ for } r = 0, 1, 2, \dots . \tag{2.1}$$

and

$$E(I(T_{n,k})) = 1 - e^{-np_k}. \tag{2.2}$$

Note that

$$Z_n = \sum_{k=1}^{\infty} I(S_{nk}), \text{ where } I(u) = \begin{cases} 1, & \text{if } u > 0, \\ 0, & \text{if } u = 0, \end{cases}$$

and

$$U_n = \sum_{k=1}^{\infty} I(T_{n,k}).$$

From Dutko [2] we know

$$Var(I(T_{n,k})) = e^{-np_k} - e^{-2np_k},\ E(U_n) = \sum_{k=1}^{\infty}(1 - e^{-np_k}),$$

$$Var(U_n) = \sum_{k=1}^{\infty}(e^{-np_k} - e^{-2np_k})$$

and both of $E(U_n)$ and $Var(U_n)$ are finite.
Let

$$X_{nk} = \frac{I(T_{n,k}) - E(I(T_{n,k}))}{\sqrt{Var(U_n)}}. \tag{2.3}$$

Then

$$\frac{U_n - E(U_n)}{\sqrt{Var(U_n)}} = \sum_{k=1}^{\infty} X_{nk},$$

$$E(\sum_{k=1}^{\infty} X_{nk}) = 0 \text{ and } Var(\sum_{k=1}^{\infty} X_{nk}) = 1.$$

To prove Theorem 1.1, we need the following theorems.

Theorem 2.1. *Let $Y_1, Y_2, \ldots$ be independent random variables with zero means and $\sum_{i=1}^{\infty} EY_i^2 = 1$ and $W = \sum_{i=1}^{\infty} Y_i$. Then*

$$\sup_{x\in\mathbb{R}} |P(W \le x) - \Phi(x)| \le 6.66 \sum_{i=1}^{\infty} \{EY_i^2 I_{\{|Y_i|\ge 1\}} + E|Y_i|^3 I_{\{|Y_i|<1\}}\}.$$

where $I_A : \Omega \to \mathbb{R}$ be defined by

$$I_A(\omega) = \begin{cases} 1 \text{ if } \omega \in A, \\ 0 \text{ if } \omega \notin A. \end{cases}$$

Theorem 2.2. *Under the assumptions of Theorem 2.1, we have*

$$|P(W \le x) - \Phi(x)| \le C \sum_{i=1}^{\infty} \left\{ \frac{EY_i^2 I_{\{|Y_i|\ge 1+|x|\}}}{(1+|x|)^2} + \frac{E|Y_i|^3 I_{\{|Y_i|<1+|x|\}}}{(1+|x|)^3} \right\}$$

for some a constant $C > 0$.

Proofs of Theorem 2.1 and Theorem 2.2 are similar to the arguments in the proof of Chen and Shao Theorems [1].

Proof of Theorem 1.1. It is easy to see that (1.3) and (1.4) follow from Theorem 2.1, Theorem 2.2 and the fact that

$$E|X_{nk}|^2 I_{\{|X_{nk}|\ge 1\}} \le E|X_{nk}|^3 I_{\{|X_{nk}|\ge 1\}} \text{ and}$$
$$E|X_{nk}|^2 I_{\{|X_{nk}|\ge 1+|x|\}} \le \frac{E|X_{nk}|^3 I_{\|X_{nk}|\ge 1+|x|\}}}{1+|x|}.$$

To complete the proof of Theorem 1.1, it suffices to show that

$$\sum_{k=1}^{\infty} E|X_{nk}|^3 \le \frac{1}{\sqrt{Var(U_n)}}. \tag{2.4}$$

By (2.1) and (2.2), we have

$$P\Big(X_{nk} = \frac{e^{-np_k} - 1}{\sqrt{Var(U_n)}}\Big) = P(I(T_{n,k}) = 0) = e^{-np_k} \text{ and}$$
$$P\Big(X_{nk} = \frac{e^{-np_k}}{\sqrt{Var(U_n)}}\Big) = 1 - P(I(T_{n,k}) = 0) = 1 - e^{-np_k}.$$

Hence,

$$\begin{aligned} E\ |X_{nk}|^3 &= \sum_{x\in Im(X_{nk})} |x|^3 P(X_{nk} = x) \\ &= \frac{(1-e^{-np_k})^3}{(Var(U_n))^{\frac{3}{2}}} e^{-np_k} + \frac{e^{-3np_k}}{(Var(U_n))^{\frac{3}{2}}}(1-e^{-np_k}) \\ &= \frac{-2e^{-4np_k} + 4e^{-3np_k} - 3e^{-2np_k} + e^{-np_k}}{(Var(U_n))^{\frac{3}{2}}} \end{aligned}$$

which implies

$$\sum_{k=1}^{\infty} E|X_{nk}|^3 = A_n + B_n + C_n$$

where

$$A_n = \frac{2\sum_{k=1}^{\infty} e^{-2np_k}(e^{-np_k} - e^{-2np_k})}{(Var(U_n))^{\frac{3}{2}}},$$

$$B_n = \frac{-2\sum_{k=1}^{\infty} e^{-np_k}(e^{-np_k} - e^{-2np_k})}{(Var(U_n))^{\frac{3}{2}}}$$

and

$$C_n = \frac{\sum_{k=1}^{\infty}(e^{-np_k} - e^{-2np_k})}{(Var(U_n))^{\frac{3}{2}}}.$$

Since $A_n + B_n < 0$,

$$\sum_{k=1}^{\infty} E|X_{nk}|^3 \le C_n = \frac{1}{\sqrt{Var(U_n)}}. \tag{2.5}$$

∎

Proof of Theorem 1.2. Let $W = \sum_{k=1}^{\infty} X_{nk}$, $W^{(k)} = W - X_{nk}$ and

$$K_k(t) = EX_{nk}\{I_{\{0\le t\le X_{nk}\}} - I_{\{X_{nk}\le t<0\}}\}.$$

Hence

$$\sum_{k=1}^{\infty}\int_{-\infty}^{\infty} K_k(t)dt = \sum_{k=1}^{\infty} EX_{nk}^2 = 1. \tag{2.6}$$

Let f be a real-value, bounded, continuous and piecewise differentiable function defined on the real line. Then

$$\begin{aligned}
&EWf(W)\\
&= E(\sum_{k=1}^{\infty} X_{nk}f(W))\\
&= \sum_{k=1}^{\infty} EX_{nk}f(W)\\
&= \sum_{k=1}^{\infty} E\{X_{nk}f(W^{(k)} + X_{nk}) - X_{nk}f(W^{(k)})\}\\
&= \sum_{k=1}^{\infty} EX_{nk}\int_0^{X_{nk}} f'(W^{(k)} + t)dt
\end{aligned}$$

$$= \sum_{k=1}^{\infty} E \int_{-\infty}^{\infty} f'(W^{(k)} + t) X_{nk} \{ I_{\{0 \le t \le X_{nk}\}} - I_{\{X_{nk} \le t < 0\}} \} dt$$

$$= \sum_{k=1}^{\infty} E \int_{-\infty}^{\infty} f'(W^{(k)} + t) E\{ X_{nk} (I_{\{0 \le t \le X_{nk}\}} - I_{\{X_{nk} \le t < 0\}}) \} dt$$

$$= \sum_{k=1}^{\infty} E \int_{-\infty}^{\infty} f'(W^{(k)} + t) K_k(t) dt \tag{2.7}$$

where we have used the fact that

$$\sum_{k=1}^{\infty} E|X_{nk} f(W)| \le (\sup f) \sum_{k=1}^{\infty} E|X_{nk}| = \frac{2(\sup f)}{\sqrt{Var(U_n)}} \sum_{k=1}^{\infty} (e^{-np_k} - e^{-2np_k})$$
$$= 2\sqrt{Var(U_n)} < \infty$$

and Lebesgue Dominated Convergence Theorem (LDC) in the second equality. Let f in (2.7) be the unique bound solution f_x of the Stein equation

$$f'(\omega) - \omega f(\omega) = I_{\{\omega \le x\}} - \Phi(x)$$

i.e.,

$$f_x(\omega) = \begin{cases} \sqrt{2\pi} e^{\frac{1}{2}\omega^2} \Phi(\omega)[1 - \Phi(x)], & \text{if } \omega \le x; \\ \sqrt{2\pi} e^{\frac{1}{2}\omega^2} \Phi(x)[1 - \Phi(\omega)], & \text{if } \omega > x \end{cases}$$

(see [5], p. 22).
Then

$$EW f_x(W) = \sum_{k=1}^{\infty} E \int_{-\infty}^{\infty} \{ (W^{(k)} + t) f_x(W^{(k)} + t) + I_{\{W^{(k)} + t \le x\}} - \Phi(x) \} K_k(t) dt$$

and

$$\sum_{k=1}^{\infty} \int_{-\infty}^{\infty} P(W^{(k)} + t \le x) K_k(t) dt - \Phi(x)$$

$$= \sum_{k=1}^{\infty} E \int_{-\infty}^{\infty} \{ W f_x(W) - (W^{(k)} + t) f_x(W^{(k)} + t) \} K_k(t) dt. \tag{2.8}$$

From the fact that

$$|(\omega + u) f_x(\omega + u) - (\omega + v) f_x(\omega + v)| \le (|\omega| + \frac{\sqrt{2\pi}}{4})(|u| + |v|) \text{ for all real } \omega,\ u \text{ and } v,$$

([3], p. 247) and (2.3) we have

$$E \sum_{k=1}^{\infty} \int_{-\infty}^{\infty} |W f(W) - (W^{(k)} + t) f(W^{(k)} + t)| K_k(t) dt$$

$$\le \sum_{k=1}^{\infty} \int_{-\infty}^{\infty} E(|W^{(k)}| + \frac{\sqrt{2\pi}}{4})(|X_{nk}| + |t|) K_k(t) dt$$

$$\leq (1+\frac{\sqrt{2\pi}}{4})\sum_{k=1}^{\infty}\int_{-\infty}^{\infty}(E|X_{nk}|+|t|)K_k(t)dt$$

$$\leq (1+\frac{\sqrt{2\pi}}{4})\sum_{k=1}^{\infty}\{E|X_{nk}|EX_{nk}^2+0.5E|X_{nk}|^3\}$$

$$\leq 2.44\sum_{k=1}^{\infty}E|X_{nk}|^3$$

$$\leq \frac{2.44}{\sqrt{Var(U_n)}}(\text{by (2.5)}). \tag{2.9}$$

Since

$$|X_{nk}| = |\frac{I(T_{n,k})-E(I(T_{n,k}))}{\sqrt{Var(U_n)}}| \leq \frac{1}{\sqrt{Var(U_n)}},$$

$$K_k(t) = 0 \text{ for } |t| > \frac{1}{\sqrt{Var(U_n)}}$$

and

$$\sum_{k=1}^{\infty}\int_{-\infty}^{\infty}P(W^{(k)}+t\leq x)K_k(t)dt$$

$$=\sum_{k=1}^{\infty}\int_{|t|\leq\frac{1}{\sqrt{Var(U_n)}}}P(W-X_{nk}+t\leq x)K_k(t)dt$$

$$\geq\sum_{k=1}^{\infty}\int_{|t|\leq\frac{1}{\sqrt{Var(U_n)}}}P(W\leq x-\frac{2}{\sqrt{Var(U_n)}})K_k(t)dt$$

$$= P(W\leq x-\frac{2}{\sqrt{Var(U_n)}}). \tag{2.10}$$

Combining (2.8)–(2.10) we have

$$P(W\leq x-\frac{2}{\sqrt{Var(U_n)}})-\Phi(x-\frac{2}{\sqrt{Var(U_n)}})$$

$$\leq \Phi(x)-\Phi(x-\frac{2}{\sqrt{Var(U_n)}})+\frac{2.44}{\sqrt{Var(U_n)}}$$

$$\leq \frac{2}{\sqrt{2\pi}\sqrt{Var(U_n)}}+\frac{2.44}{\sqrt{Var(U_n)}}$$

$$\leq \frac{3.24}{\sqrt{Var(U_n)}}. \tag{2.11}$$

Since x is arbitrary, by (2.11),

$$P(W\leq x)-\Phi(x)\leq\frac{3.24}{\sqrt{Var(U_n)}}.$$

By using the same argument one can show that

$$P(W \leq x) - \Phi(x) \geq \frac{-3.24}{\sqrt{Var(U_n)}}.$$

Hence

$$|F_n(x) - \Phi(x)| \leq \frac{3.24}{\sqrt{Var(U_n)}}.$$

∎

Remark 2.1. The following remarks can be made:

(i) From (1.3) and (2.4) we see that in case of uniform bound, the result in Theorem 1.2 is better than Theorem 1.1.

(ii) In proving Theorem 2.1 and Theorem 2.2 we have to used the fact that $\sum_{k=1}^{\infty} EX_{nk} = E(\sum_{k=1}^{\infty} X_{nk})$ and $\sum_{k=1}^{\infty} EX_{nk}^2 = E(\sum_{k=1}^{\infty} X_{nk}^2)$ which follow from LDC and the fact that $\sum_{k=1}^{\infty} E|X_{nk}| \leq 2\sqrt{Var(U_n)} < \infty$ and $\sum_{k=1}^{\infty} EX_{nk}^2 = 1$.

Acknowledgement. The idea of the proof of Theorem 1.2 come from the lecture of Prof. Shao in the conference of "Stein's method and Application: A program in honor of Charles Stein" in the case of finite sums and bounded random variable. The authors also would like to thanks the referees for their insightful comments.

References

[1] L.H.Y. Chen and Q.M. Shao, A non-uniform Berry-Esseen bound via Stein's method, *Probab. Theory Relat. Fields.* **120**(2001), 236–254.

[2] M. Dutko, Central limit theorems for infinite urn models, *Ann. Probab.* **17**(3)(1989), 1255–1263.

[3] S. Karlin, Central limit theorems for certain infinite urn schemes, *S. Math. Mech.* **17**(1967), 373–401.

[4] O. Milenkovic and K.J. Compton, Probabilistic transforms for combinatorial urn models, *Combin. Probab. Comput.* **13**(2004), 645–675.

[5] C. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, *Proc, Sixth Berkeley Symp. Math. Stat. Prob.* 2, 583–602, Univ. California Press, Berkeley, Calif., 1972.

[6] C. Stein, *Approximation Computation of Expectations*, Lecture Note 7, Inst. Math. Statist., Hayward, Calif., 1986.