UDC 517.55

# On the estimation of the remainder term in the CLT for the number of occupied cells in a multinomial random allocation scheme

**Mirakhmedov Sh. A.**
V.I. Romanovskiy Institute of Mathematics, Academy of Sciences of Uzbekistan, Tashkent
shmirakhmedov@yahoo.com
**Lazareva V. A.**
V.I. Romanovskiy Institute of Mathematics, Academy of Sciences of Uzbekistan, Tashkent
1748mailbox@mail.ru

**RESUME**

In this paper the best estimate for the remainder term in the central limit theorem for the number of occupied cells in the multinomial random allocation scheme is established.

***Key words:*** Random allocation, Poisson distribution, central limit theorem.

## 1. Introduction

We consider a model with $n$ balls and $N$ cells (urns) numbered $1, 2, \dots, N$, where $N = N(n) \to \infty$ as $n \to \infty$. Balls are allocated at random independently of each other. The probability of a ball falling into the $m$th cell is $p_m > 0$, where $p_1 + p_2 + \cdots + p_N = 1$. Let $\eta_m$ be the number of balls in the $m$th cell after allocation of all balls, where $\eta(\eta_1, \dots, \eta_N)$.

In specific applications, instead of cells, one has types or species of sampling units, and the sample array $\eta$ is of interest because it reveals population frequencies $p_j$. Such species sampling problems arise in ecology, also in database query optimization, where the sampling units maybe entries in columns of a database while the species consist of all of distinct values appearing in the column; in literature, where the sampling units may be words appearing in a given author's known works while the species consist of all words known to that author; in disclosure risk limitation, where the sampling units may be people or firms listed in a microdata file, without names or other overtly identifying information, while the types are unique combinations of values of variables with which the people or firms might be implicitly identified; and in many other areas.

The subject of our interest here is the following important in practice count statistic: the number of occupied i.e.,

$$K_N = \sum_{m=1}^{N} \mathbf{I}\{\eta_m > 0\},$$

where $\mathbf{I}(A) = 1$ if event A occurred, otherwise zero.

The occupancy counts statistic $K_N$ has been given many names, such as the "profile" (in information theory) or the "fingerprint" (in theoretical computer science) of the probability distribution $\{p_m\}$. Our goal in this paper is to establish the best estimate of the remainder term when approximating the distribution function of $K_N$ with a normal distribution.

## 2. Results and Proof.

We adopt the following notation: $\xi_1, \xi_2, \dots$ is a sequence of independent random variables, where $\xi_m$ is a Poisson random variable with parameter $\lambda_m := np_m$, $\Phi(x)$ is standard normal distribution function,

$$\Lambda_N = \sum_{m=1}^{N} \left(1 - e^{-\lambda_m}\right), \quad \gamma_N = n^{-1} \sum_{m=1}^{N} \lambda_m e^{-\lambda_m},$$

$$\sigma_N^2 = \sum_{m=1}^{N} e^{-\lambda_m}(1-e^{-\lambda_m}) - n\gamma_N^2,$$

$$g_m(\xi_m) = \mathbf{I}\{\xi_m > 0\} - (1-e^{-\lambda_m}) - \gamma_N(\xi_m - \lambda_m)$$

We assume that

$$N \max_m p_m \le C_0 \tag{6}$$

where, hereafter, $C_k$ denotes a positive universal constant, which may vary across instances.

**Theorem.** There exists a constant $C > 0$ such that

$$\Delta_N := \sup_{-\infty<x<\infty} |P\{K_N < x\sigma_N + \Lambda_N\} - \Phi(x)| \le \frac{C}{\sigma_N}.$$

**Proof.** Write

$$\sigma_N^2 = \sum_{m=1}^{N} e^{-\lambda_m}\left(1-(1+\lambda_m)e^{-\lambda_m}\right) + \sum_{m=1}^{N} \lambda_m(\gamma_N - e^{-\lambda_m})^2,$$

$$g_m(\xi_m) = -(\mathbf{I}\{\xi_m = 0\} + \xi_m e^{-\lambda_m}) + (1+\lambda_m)e^{-\lambda_m} - (\gamma_N - e^{-\lambda_m})(\xi_m - \lambda_m).$$
$$= -[(\mathbf{I}\{\xi_m = 0\} + \xi_m e^{-\lambda_m} - (1+\lambda_m)e^{-\lambda_m}) + (\gamma_N - e^{-\lambda_m})(\xi_m - \lambda_m)].$$

We have for arbitrary $\nu \ge 3$,

$$\sum_{m=1}^{N} E|g_m(\xi_m)|^\nu \le 2^{\nu-1} \sum_{m=1}^{N} E|\mathbf{I}\{\xi_m = 0\} + \xi_m e^{-\lambda_m} - (1+\lambda_m)e^{-\lambda_m}|^\nu$$

$$+2^{\nu-1} \sum_{m=1}^{N} |\gamma_N - e^{-\lambda_m}|^\nu E|\xi_m - \lambda_m)|^\nu =: J_1 + J_2. \tag{7}$$

Next,

$$J_1 = 2^{\nu-1} \sum_{m=1}^{N} \left((1-(1+\lambda_m)e^{-\lambda_m})^\nu\right) e^{-\lambda_m} + 2^{\nu-1} \sum_{m=1}^{N} \lambda_m^\nu e^{-\nu\lambda_m}$$

$$+2^{\nu-1} \sum_{m=1}^{N} \sum_{j=2}^{\infty} |j-1-\lambda_m)|^\nu e^{-(\nu+1)\lambda_m} \frac{\lambda_m^j}{j!} =: J_{11} + J_{12} + J_{13}. \tag{8}$$

We have

$$J_{11} \le 2^{\nu-1} \sum_{m=1}^{N} \left(1-(1+\lambda_m)e^{-\lambda_m}\right) e^{-\lambda_m} \le 2^{\nu-1}\sigma_N^2. \tag{9}$$

By using inequalities $xe^{-x} \le e^{-1}$ and $2^{-1}x^2e^{-x} \le 1-(1+x)e^{-x}$ we obtain

$$J_{12} \le 2^\nu e^{-(\nu-2)} \sum_{m=1}^{N} \frac{1}{2}\lambda_m^2 e^{-2\lambda_m} \le 2^\nu e^{-(\nu-2)} \sum_{m=1}^{N} \left(1-(1+\lambda_m)e^{-\lambda_m}\right) e^{-\lambda_m} \le 2^\nu e^{-(\nu-2)}\sigma_N^2. \tag{10}$$

Assume $\nu \ge 4$ is an even integer. We have

$$2^{1-\nu} J_{13} = \sum_{m=1}^{N} \sum_{j=2}^{\infty} |j-1-\lambda_m)|^\nu e^{-(\nu+1)\lambda_m} \frac{\lambda_m^j}{j!} = \sum_{m=1}^{N} e^{-\nu\lambda_m} \sum_{j=2}^{\infty} (j-1-\lambda_m)^\nu e^{-\lambda_m} \frac{\lambda_m^j}{j!}$$

$$= \sum_{m=1}^{N} e^{-\nu\lambda_m}[\sum_{j=0}^{\infty}(j-1-\lambda_m)^{\nu}e^{-\lambda_m}\frac{\lambda_m^j}{j!} - (1+\lambda_m)^{\nu}e^{-\lambda_m} - \lambda_m^{\nu+1}e^{-\lambda_m}]$$

$$= \sum_{m=1}^{N} e^{-\nu\lambda_m}[E(\xi_m-\lambda_m-1)^{\nu} - (1+\lambda_m)^{\nu}e^{-\lambda_m} - \lambda_m^{\nu+1}e^{-\lambda_m}]$$

$$= \sum_{m=1}^{N} e^{-(\nu-1)\lambda_m}[E(\xi_m-\lambda_m-1)^{\nu}e^{-\lambda_m} - (1+\lambda_m)^{\nu}e^{-2\lambda_m} - \lambda_m^{\nu+1}e^{-2\lambda_m}]$$

$$\le \sum_{m=1}^{N} e^{-(\nu-1)\lambda_m}[\sum_{k=0}^{\nu}(-1)^{\nu-k}C_{\nu}^{k}E(\xi_m-\lambda_m)^k e^{-\lambda_m} - (1+\lambda_m)^{\nu-1}e^{-\lambda_m} + (1+\lambda_m)^{\nu-1}e^{-\lambda}(1-(1+\lambda)e^{-\lambda})]$$

$$= \sum_{m=1}^{N} e^{-(\nu-1)\lambda_m}[\sum_{k=0}^{\nu}(-1)^{\nu-k}C_{\nu}^{k}E(\xi_m-\lambda_m)^k e^{-\lambda_m} - \sum_{k=0}^{\nu-1}C_{\nu-1}^{k}\lambda_m^k e^{-\lambda_m}]$$

$$+ \sum_{m=1}^{N}[(1+\lambda_m)e^{-\lambda_m}]^{\nu-1}e^{-\lambda}(1-(1+\lambda)e^{-\lambda})] =: J'_{13} + J''_{13}. \tag{11}$$

We have,

$$J'_{13} = \sum_{m=1}^{N} e^{-\nu\lambda_m}[\sum_{k=0}^{\nu}(-1)^{\nu-k}C_{\nu}^{k}E(\xi_m-\lambda_m)^k - \sum_{k=0}^{\nu-1}C_{\nu-1}^{k}\lambda_m^k]$$

$$= \sum_{m=1}^{N} e^{-\nu\lambda_m}[-\frac{1}{6}(\nu-1)(\nu-2)(\nu-3)\lambda_m + \sum_{k=4}^{\nu}(-1)^{\nu-k}C_{\nu}^{k}E(\xi_m-\lambda_m)^k - \sum_{k=2}^{\nu-1}C_{\nu-1}^{k}\lambda_m^k].$$

It is known that $E(\xi_m-\lambda_m)^k = \lambda_m + \sum_{l=2}^{[k/2]} c(l)\lambda_m^l$. Furthermore, $\sum_{k=0}^{3}(-1)^{\nu-k}C_{\nu}^{k} - (\nu-1) = (\nu-1)(\nu-2)(\nu-3)/6$. Applying these facts at the last formula we obtain

$$J'_{13} \le c(\nu)\sum_{m=1}^{N} e^{-\nu\lambda_m}(\lambda_m^{v/2} + \lambda_m^2) \le c_1(\nu)\sum_{m=1}^{N}\lambda_m^2 e^{-2\lambda_m} \le c_2(\nu)\sigma^2(n)$$

since $2^{-1}x^2e^{-x} \le 1-(1+x)e^{-x}$. Also $J''_{13} \le c(\nu)\sigma^2(n)$ since $(1+\lambda_m)e^{-\lambda_m} \le 1$. So, by (6) $J_{13} \le c_3(\nu)\sigma^2(n)$. This together with (4), (5) imply

$$J_1 \le c_3(\nu)\sigma_N^2. \tag{12}$$

Under condition (1) by reasoning similar to those of [2] one can show that $J_2 \le c(\nu)\sigma^2(n)$. Thus by (2), (7)

$$\sum_{m=1}^{N} E|g_m(\xi_m)|^{\nu} \le c(\nu)\sigma_N^2.$$

On using this inequality with $\nu = 4$ and well-known inequality between Liyapunov's ratio we obtain

$$\frac{1}{\sigma_N^3}\sum_{m=1}^{N} E|g_m(\xi_m)|^3 \le \left[\frac{1}{\sigma_N^4}\sum_{m=1}^{N} E|g_m(\xi_m)|^4\right]^{1/2} \le \frac{c}{\sigma_N}. \tag{13}$$

On the other hand it follows from Theorem 1 of [1] that

$$\Delta_N \le c\frac{1}{\sigma_N^3}\sum_{m=1}^{N} E|g_m(\xi_m)|^3.$$

Applying here (8) we complete the proof of Theorem.

## REFERENCES

1. Mirakhmedov S.M.. An approximations of multiple randomized divisible statistics by means of Normal distribution, Theory Probabl. and Appl., 32, 761-771 .

2. Quine, M.P., Robinson, J., A Berry-Esseen bound for an occupancy problem, Ann. Probabl., 10, 663-671, 1982.

## REZYUME

Ushbu maqolada zarrachalarni qutichalarga tasodifiy joylashtirishning multinomial modelida band bo'lgan qutichalar uchun markaziy limit teoremasidagi qoldiq hadga eng yaxshi baho olingan.

***Kalit so'zlar:*** tasodifiy joylash, Puasson taqsimoti, normal taqsimot, markaziy limit teorema.

## РЕЗЮМЕ

В работе устанавливается наилучшая оценка остаточного члена в центральной предельной теореме для числа занятых ячеек в полиномиальной схеме случайного размещения.

***Ключевые слова:*** случайное размещение, распределение Пуассона, нормальное распределение, центральная предельная теорема.