Asymptotic Normality Associated with Allocation Schemes of Particles into Infinity Number of Cells and Applications.

Sherzod M. Mirakhmedov

G I K Institute of Engineering Sciences & Technology, Topi-23460, Swabi ,NW.F.P. Pakistan.
E- mail: shmirakhmedov@yahoo.com

1. Introduction.

Many combinatorial problems in probability and statistics can be formulated and best understood by using appropriate allocation schemes (alternatively known as urn models). Such models naturally arise in statistical mechanics, clinical trials, cryptography etc. The properties of several types of the random allocation scheme have been extensively studied in the probabilistic and statistical literature, see Kotz and Balakrishman (1997).

The classical allocation scheme supposes an equiprobable allocation of *n* particles into a finite number of cells, say *N*, i.e. the probability of a particle falling into any particular cell is $1/N$. There are several generalizations of the classical scheme. Here we consider an allocation scheme defined as follows.

Into an infinite number of cells, numbered by $\aleph=\{1,2,...\}$, we throw particles of s types, where n_k particles are of the *k*-th type, $k=1,2,...,s$. The particles are thrown randomly one at a time and independently of each other. The probability of a particle of *l*-th type falling into *m*-th cell is $p_{lm}=p_{lm}(n_l)$, $p_{l1}+p_{l2}+...=1$, *l*=1,2,…,s. Let $\eta_{lm}=\eta_{lm}(n_l)$ be the number of particles of the *l*-th type in the *m*-th cell after the allocation of all $n_1, n_2,\ldots, n_s$ particles, $\eta_l=(\eta_{l1},\eta_{l2},...)$, $\eta=(\eta_1,...,\eta_s)$, $n=(n_1,\ldots,n_s)$, and $f_m^{(n)}(x_1,...,x_s)=(f_{1m}^{(n)}(x_1,...,x_s),..., f_{km}^{(n)}(x_1,...,x_s))$, $m\in\aleph$, be a sequence of random functions of integer arguments $x_1,...,x_s$, not depending on the random vector η. We assume that the series $R_j^{(n)}(\eta)=\sum_{m=1}^{\infty} f_{jm}^{(n)}(\eta_{1m},...,\eta_{sm})$, *j*=1,2,…,k, are convergent with probability one. We shall consider the *k*-dimensional statistic: $R^{(n)}(\eta)=(R_1^{(n)}(\eta),...,R_k^{(n)}(\eta))$.

2. Main Result. In what follows all asymptotic relations are considered as $\min(n_1,\ldots,n_s)\to\infty$.

Let $\xi_{\nu m}$, ν=1,2,…,s; m= 1,2,…, be a mutually independent Poisson random variables with parameter $n_\nu p_{\nu m}$ respectively. Put $\xi_m^{(s)}=(\xi_{1m},...,\xi_{sm})$, $\eta_m^{(s)}=(\eta_{1m},...,\eta_{sm})$,

$$g_{jm}(x_1,...,x_s)=f_{jm}^{(n)}(x_1,...,x_s)-Ef_{jm}^{(n)}(\xi_m^{(s)})-\sum_{l=1}^{s} n_l^{-1}(x_l-n_l p_{lm})\sum_{m=1}^{\infty}\operatorname{cov}(\xi_{lm}, f_{jm}^{(n)}(\xi_m^{(s)})),$$

$$\sigma_{ij}=\sum_{m=1}^{\infty}\operatorname{cov}(g_{im}(\xi_m^{(s)}), g_{jm}(\xi_m^{(s)})),\quad \hat g_{jm}(x_1,...,x_s)=g_{jm}(x_1,...,x_s)/\sqrt{\sigma_{jj}},$$

$$\hat g_m(x_1,...,x_s)=(\hat g_{1m}(x_1,...,x_s),...,\hat g_{km}(x_1,...,x_s)),\quad \hat R_j(\eta)=\sum_{m=1}^{\infty}[f_{jm}^{(n)}(\eta_m^{(s)})-Ef_{jm}^{(n)}(\xi_m^{(s)})]/\sqrt{\sigma_{jj}}.$$

We assume that $\sigma_{jj} > 0$ for all $j = 1,2,...,k$. Let $\Phi_{0,\sigma}(u_1,...,u_k)$ be the k dimensional normal distribution with zero expectation and matrix of correlations σ, and $\Uparrow\{A\}$ be an indicator of the event A.

Theorem. If

a) There exists constant C such that $\max\limits_{l=1\div s, m\in\aleph} p_{lm} \le C < 1$,

b) $\sum\limits_{j=}^{k}\sum\limits_{m=1}^{\infty} E\left[\hat{g}_{jm}(\xi_m^{(s)})\Uparrow\left\{\left|\hat{g}_{jm}(\xi_m^{(s)})\right| > \varepsilon\right\}\right]^2 \to 0$ for arbitrary $\varepsilon > 0$.

Then $\sup\limits_{u_1,...,u_k}\left|P\left\{\hat{R}_j(\eta) < u_j,\ j = 1,...,k\right\} - \Phi_{0,\sigma}(u_1,...,u_k)\right| \to 0$, where matrix $\sigma = (\sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}})$.

Earlier in Mirakhmedov (1992) this Theorem was proved for the case k=s=1 under Lyapunov's type condition and $p_{11} + p_{12} + ... = o(n^{-1/2})$.

3**. Applications.**

1. We consider a probabilistic set-up without memory, input of which is sequence of i.i.d. discrete random variables $X_1,...,X_n$, on output we have a sequence of independent random variables $Y_1,...,Y_n$ with two possible values 0 and 1 only. Let $P\{X_j = x_m\} = p_m$, $p_1 + p_2 + ... = 1$, and $P\{Y_j = 1/X_j = x_m\} = q_m$, $m\in\aleph$. Also, let η_{mj} be a frequency of the event $(X_l = x_m, Y_l = j)$ and $\eta_m = \eta_{m0} + \eta_{m1}$ be a frequency of the event $(X_l = x_m)$ among the sample $(X_1,Y_1),...,(X_n,Y_n)$. We wish to test a simple hypothesis $H_0 : \{q_m = 1/2, m\in\aleph\}$ versus the sequence of alternatives H_{1n} :

$\sum\limits_{m=1}^{\infty} p_m^2 k_m^2 = K_n^2 > 0$, where $k_m = 2q_m - 1$. We consider a test based on $X^2(n) = \sum\limits_{m=1}^{\infty}(\eta_{m1} - \eta_{m0})^2$, large values of which reject the null hypothesis H_0. Put $P(n) = \sum\limits_{m=1}^{\infty} p_m^2$. From Theorem of Section 2 we get: if

$$n^2 P(n) \to \infty, \text{ and } \max_{m\in\aleph} p_m = o\left(\sqrt{P(n)}\right), \text{ as } n\to\infty, \tag{1}$$

then for the $X^2(n)$ test of size $\alpha\in(0,1)$ we have: (i) the critical point is $c = n(1 + u_\alpha\sqrt{2P(n)})$,

(ii) the asymptotic power is $\Phi(-u_\alpha + \dfrac{nK_n^2}{\sqrt{2P(n)}})$, where $\Phi(u)$ is standard normal distribution function,

and $u_\alpha = \Phi^{-1}(1-\alpha)$. Thus this test detects the alternatives at a "distance" $\sqrt{2P(n)}\big/n$ from H_0.

Condition (1) implies that among all values of the input r.v. X_i, there must be enough values having probabilities comparable to $\max\limits_{m\in\aleph} p_m$.

2. We observe a discrete random variable (population), say X, having probability mass distribution $\mathrm{P} = (p_1, p_2, ...)$, $p_1 + p_2 + ... = 1$. We wish to test the simple hypothesis H_0:

$$p_m = p_{0m},\ \forall\, m = 1,2,... \tag{2}$$

We shall now construct a test of sequential type as follows. Let's at first have a sample of size n_1 from the population *X*. Also let $(\eta_{11}, \eta_{12}, ...)$ be a vector of frequencies of the values of *X* in this sample, $\eta_{11} + \eta_{12} + ... = n_1$. We choose a set of functions $f_{1m}(x), m \in \aleph,$ and define the statistic $R_1 = \sum_{m=1}^{\infty} f_{1m}(\eta_{1m})$. We determine a critical point; say c_1.If $R_1 \le c_1$ then the hypothesis H_0 will be accepted. If $R_1 > c_1$ then we continue the process of testing of H_0: we then take a second sample of size n_2, accounting for a vector of frequencies $(\eta_{21}, \eta_{22}, ...)$ in the second sample, $\eta_{21} + \eta_{22} + ... = n_2$, choose a set of functions $f_{2m}(x)$, $m \in \aleph$, define the statistic $R_2 = \sum_{m=1}^{\infty} f_{2m}(\eta_{1m} + \eta_{2m})$ and a critical point c_2. If $R_2 \le c_2$, then the hypothesis H_0 will be accepted. If $R_2 > c_2$, then we again continue the process of testing of the hypothesis H_0 taking the next sample. This process contains at most *k* steps, where *k* is a pre-selected number. In the *k*-th step we must or accept or reject hypothesis H_0. Thus we have the test of sequential type based on the statistics $(R_1, ..., R_k)$, where $R_j = \sum_{m=1}^{\infty} f_{jm}(\eta_{1m} + \eta_{2m} + ... + \eta_{jm})$ and $(\eta_{j1}, \eta_{j2}, ...)$ is vector of the frequencies of the values of the *X* in the *j*- th sample of size n_j, $j = 1,2,..,k$. Put $\overline{A}_j$ is compliment of $A_j = \{R_1 > c_1, ..., R_{j-1} > c_{j-1}, R_j \le c_j\}$. The test described is of the sequential type containing at most *k* steps: rejects the hypothesis H_0 if and only if event $B = \bigcap_{j=1}^{k} \overline{A}_j$ $= \{R_1 > c_1, ..., R_{k-1} > c_{k-1}, R_k > c_k\}$ is occurs, and H_0 may be accepted in the j-th step if A_j is occurred, *j*=1,2,…,*k*. Note that the kernel functions f_{jm} as well as the probability mass distribution $\mathrm{P} = (p_1, p_2, ...)$ on *j*-th step may depend on sample sizes $n_1, n_2, ..., n_j$. We omit respective suffix for simplicity of notation. The Theorem of Section 2 (with *s*=*k*) allows us to determine the critical points $c_1, ..., c_k$ as follows.

Let H_1 be an alternative hypothesis and P_v, E_v and cov_v stands for the probability, expectation and co-variation respectively counted under H_v, *v*=0,1. We denote

$$\zeta_{jm} = \xi_{1m} + ... + \xi_{jm}, \; N_j = n_1 + ... + n_j, \; \Delta f(x) = f(x+1) - f(x), \; M_{vj} = \sum_{m=1}^{\infty} E_v f_{jm}(\zeta_{jm}),$$

$$\sigma_{vij} = \sum_{m=1}^{\infty} \mathrm{cov}_v(f_{im}(\zeta_{im}), f_{jm}(\zeta_{jm})) - N_{\min(i,j)} \sum_{m=1}^{\infty} p_{0m} E_v \Delta f_{im}(\zeta_{im}) \sum_{m=1}^{\infty} p_{0m} E_v \Delta f_{jm}(\zeta_{jm}),$$

$$g_{jm}(\zeta_{jm}) = f_{jm}(\zeta_{jm}) - E f_{jm}(\zeta_{jm}) - \sum_{l=1}^{j} (\xi_{lm} - n_l p_m) \sum_{m=1}^{\infty} p_{0m} E \Delta f_{jm}(\zeta_{jm}). \tag{3}$$

Let $\alpha \in (0,1)$ and the critical region B is defined as $B = B_\alpha = \{R_j > M_{0j} + c_{j\alpha}\sqrt{\sigma_{0jj}}, \forall j = 1,..., k\}$, where the critical points $c_{1\alpha},...,c_{k\alpha}$ are defined such that $\Phi_{0,\sigma_0}(-c_{1\alpha},...,-c_{k\alpha}) = \alpha$, and σ_0 is a matrix with entries $\sigma_{0ij} / \sqrt{\sigma_{0ii}\sigma_{0jj}}$, $i, j = 1,..., k$. From Theorem we get: If $\max_m p_m \le C < 1$ and condition b) is fulfilled with kernel function $g_{jm}(\zeta_{jm})$ given in (3), then $P_0\{B_\alpha\} \to \alpha$, as $n_1 \to \infty$.

We note that the set of critical points is not defined uniquely.

Let's consider sequential chi-square type test based on above defined statistics $(R_1,..., R_k)$ with kernel functions

$$f_{jm}(x) = \gamma_m p_{0m}^{-2}\left(x(x-1) - 2xN_j p_{0m} + N_j^{\,2} p_{0m}^2\right), \; j = 1,2,..., k; m \in \aleph, \tag{4}$$

We wish to test the hypothesis (2) versus to sequence of alternatives H_{1n}:

$$\sum_{m=1}^{\infty} \gamma_m \left(\frac{p_m}{p_{0m}} - 1\right)^2 = \omega(n) > 0. \tag{5}$$

where $\gamma_m \ge 0, m = 1,2,...,$ and $\gamma_1 + \gamma_2 + ... = 1$. In this case in (3) we have $M_{1j} = N_j^2 \omega(n)$, and

$$\sigma_{1ij} = 2N_{\min(i,j)}^2 \sum_{m=1}^{\infty} \frac{\gamma_m^2}{p_{0m}^4} p_m^2 + 4N_i N_j \sum_{m=1}^{\infty} \frac{\gamma_m^2}{p_{0m}^2} p_m \left(\frac{p_m}{p_{0m}} - 1\right)^2 - 4N_i N_j N_{\min(i,j)} \left(\sum_{m=1}^{\infty} \frac{\gamma_m}{p_{0m}} p_m \left(\frac{p_m}{p_{0m}} - 1\right)\right)^2.$$

Putting $\mu(n) = \sum_{m=1}^{\infty} \gamma_m^2 p_{0m}^{-2}$ we obtain $\sigma_{0ij} = 2N_{\min(i,j)}^2 \mu(n)$, $M_{0j} = 0$, $B_\alpha = \{R_j > c_{j\alpha}\sqrt{2\mu(n)} N_j, j = 1,..., k\}$.

Let $\lim_{n_1 \to \infty} \frac{N_k \omega(n)}{\sqrt{\mu(n)}} = \Delta^2, 0 \le \Delta^2 < \infty$. Using Theorem of Section 2 with s=k we get: if $\max_{m \in \aleph} p_{0m} < c < 1$, and $\max\left(\sum_{m=1}^{\infty} \gamma_m^4 p_{0m}^{-4}, \sum_{m=1}^{\infty} \gamma_m^4 p_{0m}^{-6} n_1^{-2}\right) = o(\mu^2(n))$, then the asymptotic power of the test based on statistic $(R_1,..., R_k)$ with kernel functions given in (4) is $\Phi_{0,\sigma_0}\left(\frac{N_1\omega(n)}{\sqrt{2\mu(n)}} - c_{1\alpha},..., \frac{N_k\omega(n)}{\sqrt{2\mu(n)}} - c_{k\alpha}\right)$, where σ_0 is a matrix with entries $N_{\min(i,j)}^2 / N_i N_j$, i, j =1,…,k. Thus this test discriminates alternatives characterized by the sequence $\gamma_1, \gamma_2,...$ such that the measure $\omega(n)$ of closeness of the alternative (5) to H_0 in (2) has order $\sqrt{\mu(n)} N_k^{-1}$.

Reference.

Kotz, S. and Balakrishman, N., (1997) .Advances in urn models during the past two decades. In (Ed. N. Balakrshnan) Birkhauser, Boston, 203-257.

Mirakhmedov, Sh.A. (1992), Randomized decomposable statistics in the scheme of independent allocation of particles into boxes. Discrete Math. Appl., 2, 91-108.