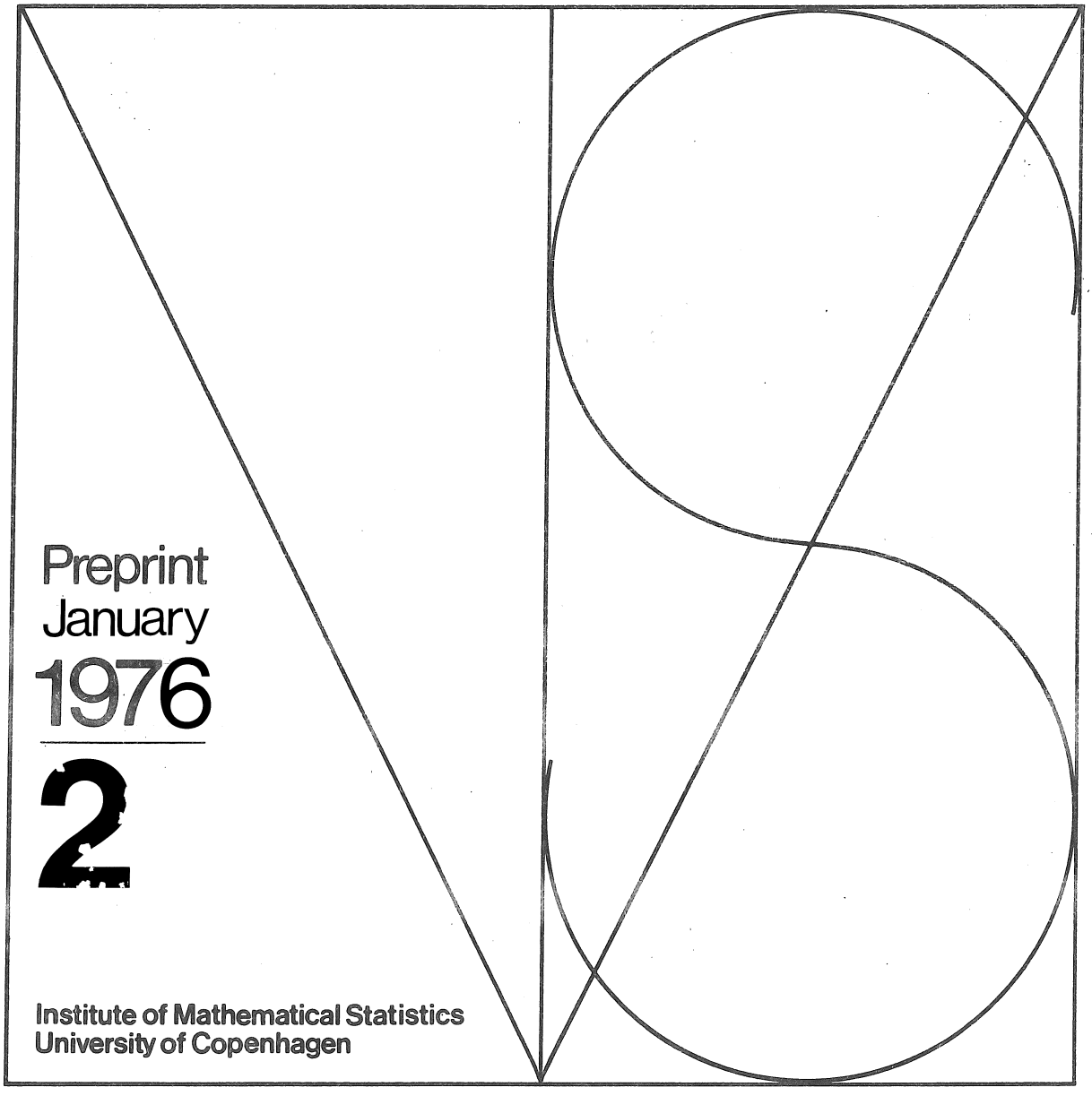
Norman Kaplan

A Generalization of a Result of Erdös and Renyi and a Related Problem



Norman Kaplan*

A GENERALIZATION OF A RESULT OF ERDÖS AND RENYI AND A RELATED PROBLEM

Preprint 1976 No. 2

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

January 1976

*Department of Statistics, University of California, Berkeley.

Abstract

Two problems dealing with urn schemes are studied. The first looks at a finite urn scheme with balls uniformly distributed amongst the urns. The asymptotic behavior of the number of throws necessary to achieve at least r balls in each urn is examined. The second problem deals with an infinite urn scheme and studies the number of urns containing at least r balls. A method of proof is developed which applies equally well in both cases.

A GENERALIZATION OF A RESULT OF ERDÖS AND RENYI AND A RELATED PROBLEM

1. INTRODUCTION

Consider the following problem. Balls are thrown into n urns uniformly and independently. Let $N_m(n)$ equal the number of throws necessary to obtain at least m balls in each urn. Erdös and Renyi [1] proved that for any $-\infty < x < \infty$,

$$(1) \qquad \lim_{n\to\infty} P(N_m(n) \le n(\log n + (m-1)\log\log n + x)) = e^{-\frac{e^{-x}}{(m-1)!}}$$

The purpose of this note is to prove the following generalization of (1). Define for each $1 \le \ell \le n$,

$N_m(n,\ell)$ = number of throws after which ℓ boxes will first contain at least m balls.

Observe that $N_m(n) = N_m(n,n)$. We then have:

Theorem 1. Let $m \ge 1$ *and define for each* $-\infty < x < \infty$,

$$b_n = [n(\log n + (m-1)\log\log n + x)]$$

where [y] *denotes the greatest integer in* y. *Then,*

$$(2) \qquad \max_{0\le\ell\le n} \left| P(N_m(n,\ell) \le b_n) - \sum_{j=0}^{n-\ell} q_j \left(e^{-\frac{x}{(m-1)!}}\right)\right| = o(1)$$

where for any $z > 0$, $j \ge 0$,

$$q_j(z) = \frac{e^{-z} z^j}{j!}.$$

A heuristic proof of Theorem 1 can easily be given. Define:

$X_j(i)$ = number of balls in urn j after i throws, $1 \le j \le n$, $i \ge 1$;

and

$$I_m(y) = \begin{cases} 1 & \text{if } y < m \\ 0 & \text{otherwise} \end{cases}$$

It then follows that

$$P(N_m(n,\ell) \le b_n) = P\Big(\sum_{j=1}^{n} I_m(X_j(b_n)) \le n - \ell\Big).$$

Also note that

$$(3) \qquad a_j = P(I_m(X_j(b_n)) = 1) = \sum_{k=0}^{m-1} \binom{b_n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{b_n - k}$$

$$= \frac{e^{-x}}{n(m-1)!}(1 + o(1)) \qquad 1 \le j \le n.$$

If we were able to treat the $\{X_j(b_n)\}$ as "independent random variables," then Theorem 1 is a consequence of the next result of Hodges and LeCam [5].

Before stating the result we introduce some notation and definitions. For $\theta > 0$, $Y(\theta)$ denotes a Poisson variable with parameter θ. Let X_1 and X_2 be any two nonnegative integer-valued random variables, and define

$$d(X_1, X_2) = \sup_A \left| P(X_1 \in A) - P(X_2 \in B) \right|,$$

where A ranges over all subsets of the nonnegative integers. For a discussion of the metric d, the reader is referred to [2]. The two properties of d that we will need are:

(4) $\quad d(X_1,X_2) \leq P(X_1 \neq X_2),$

and if X_1 and X_2 are Poisson with parameters θ_1 and θ_2, then

(5) $\quad d(X_1,X_2) \leq |\theta_1 - \theta_2|.$

Proofs of (4) and (5) are given in [2]. We then have:

Proposition 1. (Hodges and LeCam [5]) Let $B_1,\ldots,B_n$ *be* n *independent events and let* $N = \sum_{i=1}^{n} I_{B_i}$ *be the number which occurs. Let*

$$\theta = \sum_{i=1}^{n} P(B_i) \quad \textit{and} \quad \varepsilon = \sum_{i=1}^{n} P^2(B_i).$$

Suppose $\varepsilon < \frac{1}{2}$. *Then there exists a constant* C *such that*

$$d(N,Y(\theta)) \leq C\varepsilon.$$

If the $\{X_j(b_n)\}$ were indeed independent, then by Proposition 1 and (3),

$$d\Big(\sum_{j=1}^{n} I_m(X_j(b_n)),Y\Big(\sum_{j=1}^{n} a_j\Big)\Big) = o(1).$$

Noting that

$$d\Big(Y\Big(\sum_{j=1}^{n} a_j\Big),Y\Big(\frac{e^{-x}}{(m-1)!}\Big)\Big) = o(1)$$

completes the proof of Theorem 1.

The difficulty with the previous argument, of course, is that the

$\{X_j(b_n)\}$ are not independent. To get around this problem we use the following property of the Poisson process. It is convenient to state the result in greater generality than needed for Theorem 1. Suppose we have a finite or countable number of urns, and at each event of a Poisson process with unit parameter we independently pick an urn according to some specified distribution $\{p(j)\}$ and place a ball in it. Let

$\tilde{X}_j(t)$ = number of balls in urn j at time t.

We then have:

Proposition 2. For each t, *the* $\{\tilde{X}_j(t)\}$ *are independent random variables with* $\tilde{X}_j(t)$ *distributed as Poisson with parameter* $p(j)t$.

A proof of Proposition 2 can be found in [7].

For Theorem 1 there are n urns with $p(j) \equiv \frac{1}{n}$. Since the $\{\tilde{X}_j(b_n)\}$ are independent, we can apply Proposition 1 to conclude that

$$(6) \qquad d\Big(\sum_{j=1}^{n} I_m(\tilde{X}_j(b_n)), Y\Big(\sum_{j=1}^{n} \tilde{a}_j\Big)\Big) = O\Big(\frac{1}{n}\Big),$$

where

$$(7) \qquad \tilde{a}_j = P(\tilde{X}_j(b_n) \leq m-1) = \sum_{k=0}^{m-1} q_k\Big(\frac{b_n}{n}\Big) = \frac{e^{-x}}{n(m-1)!}\,(1 + o(1)),$$

$$1 \leq j \leq n.$$

Also, by (5),

$$(8) \qquad d\Big(Y\Big(\sum_{k=1}^{n} \tilde{a}_k\Big), Y\Big(\frac{e^{-x}}{(m-1)!}\Big)\Big) \leq \Big|\sum_{k=1}^{n} \tilde{a}_k - \frac{e^{-x}}{(m-1)!}\Big| = O\Big[\frac{\text{loglog } n}{\log n}\Big]$$

It follows from (6) and (8) that

$$d\Big(\sum_{j=1}^{n} I_m(\tilde{X}_j(b_n)), Y(e^{-x}/(m-1)!)\Big) = o(1)$$

It remains only to show that

$$(9) \qquad d\Big(\sum_{j=1}^{n} I_m(\tilde{X}_j(b_n)), \sum_{j=1}^{n} I_m(X_j(b_n))\Big) = o(1).$$

This is accomplished using the Central Limit Theorem, and the details are carried out in Section 2.

It turns out that the technique used to prove Theorem 1 has wide applicability. This is because of the generality of Propositions 1 and 2. A typical result is stated and proven in Section 3.

The motivation for this work comes from a paper of Karlin [6] where the same idea was used. The proof of Theorem 1 has its origin in Karlin's paper.

2. COMPLETION OF THE PROOF OF THEOREM 1

As shown in Section 1, we need to prove (9). To simplify notation, let

$$H(i) = \sum_{j=1}^{n} I_m(X_j(i)), \qquad i \geq 1.$$

$\tilde{H}(i)$ is defined analogously. We then have the following lemma.

Lemma 2.1. Let D *be any positive constant. Then,*

$$\eta_D = \sup_{|i-b_n|<D\sqrt{b_n}} P(H(i) \neq H(b_n)) = O\Big(\frac{D\sqrt{b_n}}{n}\Big).$$

Proof. Observe for $|i - b_n| < D\sqrt{b_n}$,

$$P(H(i) \neq H(i+1)) \leq \sum_{j=1}^{n} P(\text{urn } j \text{ has } m-1 \text{ balls after } i \text{ tosses and the } (i+1) \text{ toss is into urn } j)$$

$$= n\binom{i}{m-1}\left(\frac{1}{n}\right)^{m-1}\left(1 - \frac{1}{n}\right)^{i-(m-1)} \frac{1}{n}$$

$$\leq \left(\frac{b_n + D\sqrt{b_n}}{n}\right)^{m-1} \exp\left\{\frac{b_n - D\sqrt{b_n} - (m-1)}{n}\right\}$$

$$\leq \frac{C_1}{n} ,$$

where C_1 is some constant independent of i. Thus

$$\eta_D \leq \sum_{|i-b_n|<D\sqrt{b_n}} P(H(i) \neq H(b_n)) \leq 2C_1 \frac{D\sqrt{b_n}}{n}$$

Q.E.D.

We now prove (9). Recalling the definition of the $\{\tilde{X}_j(b_n)\}$, we have for any subset A,

$$|P(\tilde{H}(b_n) \in A) - P(H(b_n) \in A)| \leq \sum_{i=0}^{\infty} |P(H(i) \in A) - P(H(b_n) \in A)| q_i(b_n) .$$

Let D be any positive constant. It then follows by the Berry-Esséen Theorem [3, p. 201] that

$$(10) \qquad \sum_{|i-b_n| \geq D\sqrt{b_n}} q_i(b_n) \leq \frac{C_2}{\sqrt{b_n}} + \int_{|z|>D} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz.$$

Also, by Lemma 2.1,

$$(11) \qquad \sum_{|i-b_n|<D\sqrt{b_n}} |P(H(i) \in A) - P(H(b_n) \in A)|\, q_i(b_n) \leq \eta_D = O(\frac{D\sqrt{b_n}}{n}).$$

(9) is now a consequence of (10) and (11) and so Theorem 1 is complete.

Remark. Since there is an estimate of the error in Proposition 1, we can easily make an estimate of the error in Theorem 1. Indeed, it follows from Proposition 1, (6), (8), (10), and (11) that

$$d(\sum_{j=1}^{n} I_m(X_j(b_n)), Y(\frac{e^{-x}}{(m-1)!})) = O(\frac{\text{loglog } n}{\log n})$$

If $m = 1$, the bound is sharper. Indeed, one has that it is

$$O(\frac{D\sqrt{b_n}}{n}) + \int_{|z|>D} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz).$$

3. ANOTHER APPLICATION

In this section we allow a countable number of urns and examine the behavior of

$L_r(n)$ = number of urns containing at least r balls after n tosses $(r \geq 1)$.

The set-up is the following. Let K be a positive integer and suppose that we perform a sequence of experiments, where, for the ℓ-th experiment, we toss $\sum_{k=1}^{K} n_{k,\ell}$ balls independently into urns according to some distribution $\{p_\ell(j)\}$. To simplify notation we suppress the ℓ and just write n_k for $n_{k,\ell}$ and $p(j)$ for $p_\ell(j)$. It will always be assumed that

(12) $\lim_{\ell\to\infty} [\min_{1\le k\le K} n_k] = \infty$ and $\sup_{j\ge 1} p(j) = o(1)$.

Define:

$$s_k = \sum_{i=1}^{k} n_i, \qquad 1 \le k \le K;$$

$$s_0 = 0;$$

$$\theta_k(r) = \sum_{j=1}^{\infty} \frac{(s_k p(j))^r}{r!}, \qquad 1 \le k \le K, \quad r \ge 1;$$

$$\theta_0(r) = 0.$$

$\bigotimes_{k=1}^{K} Y(\eta_k)$ denotes the random vector whose components are independent Poissons with parameters (η_i). For any two random vectors $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ whose components assume nonnegative integer values, let $d(\underset{\sim}{X}_1, \underset{\sim}{X}_2)$ denote the natural analog of the d metric, i.e.,

$$d(\underset{\sim}{X}_1, \underset{\sim}{X}_2) = \sup_A \left| P(\underset{\sim}{X}_1 \in A) - P(\underset{\sim}{X}_2 \in A) \right|,$$

where A ranges over all subsets of K-tuples whose components are nonnegative integers. Note that (4) is still valid. Finally, let

$$I_r(y) = \begin{cases} 1 & \text{if } y \ge r \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that

$$L_r(n) = \sum_{j=1}^{\infty} I_r(X_j(n)).$$

We now state our result.

Theorem 2. Assume (12) and

$$(13) \qquad s_k \sup_{j \geq 1} p(j) = o(1).$$

Also assume that for some integer $r \geq 2$,

$$\lim_{\ell \to \infty} \theta_k(r) = \theta_k < \infty, \qquad 0 \leq k \leq K.$$

Let $\underset{\sim}{L}_r = (L_r(s_1), L_r(s_2), \ldots, L_r(s_K))$. *Then*

$$d(\underset{\sim}{L}_r, \underset{\sim}{W}) = o(1),$$

where

$$\{W_k - W_{k-1}\}_{k=1}^{K} \sim \bigotimes_{k=1}^{K} Y(\theta_k - \theta_{k-1}).$$

Proof of Theorem 2. In the spirit of the proof of Theorem 1, we consider K independent Poisson processes, each with unit parameter, and at each event of these processes an urn is independently picked according to $\{p(j)\}$ and a ball is placed in it. Let

$$\tilde{X}_{jk}(t) = \text{number of balls in box } j \text{ by time } t \text{ for the } k\text{-th process;}$$
$$1 \leq k \leq K, \; j \geq 1$$

and define

$$\tilde{L}_r(s_k) = \sum_{j=1}^{\infty} I_r\Big(\sum_{i=1}^{k} \tilde{X}_{ji}(n_i) \Big), \qquad 1 \leq k \leq K.$$

Let

$$\tilde{L}_{\sim r} = (\tilde{L}_r(s_1),\ldots,\tilde{L}_r(s_K)).$$

We first study the behavior of $\tilde{L}_{\sim r}$. In order to do this we need to generalize Proposition 1. Let

$$V_{jk} = I_r\Big(\sum_{i=1}^{k} \tilde{X}_{ji}(n_i)\Big), \qquad 1 \le k \le K,$$

and

$$V_{\sim j} = (V_{j1},\ldots,V_{jK}), \qquad j \ge 1.$$

Note that the $\{V_{\sim j}\}$ are independent random vectors and that

$$\tilde{L}_{\sim r} = \sum_{j=1}^{\infty} V_{\sim j}.$$

Following the ideas of Hodges and LeCam [5], we need to find a convenient representation for the $\{V_{\sim j}\}$. Indeed, assume that we have constructed on a common probability space independent random vectors $\{W_{\sim j}\}$ such that

$$\{W_{jk} - W_{jk-1}\} \sim \bigotimes_{k=1}^{K} Y(\eta_{jk}), \qquad j \ge 1.$$

where the $\{\eta_{jk}\}$ are constants to be determined. Define

$$Z_{jk} = \prod_{i=1}^{k-1} I_{\{W_{ji}=0\}} \prod_{\ell=k}^{K} I_{\{W_{j\ell}\ge 1\}}, \qquad 1 \le k \le K,$$

and set

$$\underset{\sim}{Z}_j = (Z_{j1},\ldots,Z_{jK}).$$

We now want to choose the $\{\eta_{jk}\}$ so that $\underset{\sim}{Z}_j$ and $\underset{\sim}{V}_j$ have the same distribution. Let

$$b_{jk} = \sum_{i \geq r} q_i(s_k p(j)), \quad 1 \leq k \leq K$$

$$b_{j0} \equiv 0.$$

In view of (12) and (13) it is not difficult to show that

$$\sup_{\substack{j \geq 1 \\ 1 \leq k \leq K}} b_{jk} = o(1) \tag{14}$$

and

$$\sum_{j=1}^{\infty} b_{jk} \to \theta_k, \qquad 1 \leq k \leq K. \tag{15}$$

$\underset{\sim}{V}_j$ can only take on the values $\{\underset{\sim}{e}_\ell\}_{\ell=0}^{K}$, where typically $\underset{\sim}{e}_\ell$ has its first ℓ components equal to 0 and the remaining ones equal to 1. Furthermore,

$$P(\underset{\sim}{V}_j = \underset{\sim}{e}_\ell) = b_{j\ell+1} - b_{j\ell}, \qquad 0 \leq \ell \leq K-1. \tag{16}$$

Also, it is not difficult to check that

$$P(\underset{\sim}{Z}_j = \underset{\sim}{e}_\ell) = e^{-\sum_{i=1}^{\ell} \eta_{ji}} (1 - e^{-\eta_{j\ell+1}}), \qquad 0 \leq \ell \leq K-1. \tag{17}$$

Simple algebra shows that for the r.h.s. of (16) and (17) to be equal,

it must be true that

$$(18) \qquad \eta_{ji} = -\log \frac{1 - b_{ji}}{1 - b_{ji-1}}, \qquad 1 \leq i \leq K.$$

Let $\underset{\sim}{W}' = \sum_{j=1}^{\infty} \underset{\sim}{W}_j$. We then have

$$\begin{aligned} d(\tilde{\underset{\sim}{L}}_r, \underset{\sim}{W}') &= d\Big(\sum_{j=1}^{\infty} \underset{\sim}{V}_j, \sum_{j=1}^{\infty} \underset{\sim}{W}_j\Big) \\ &= d\Big(\sum_{j=1}^{\infty} \underset{\sim}{Z}_j, \sum_{j=1}^{\infty} \underset{\sim}{W}_j\Big) \\ &\leq \sum_{j=1}^{\infty} P(\text{some } W_{jk} \geq 2;\ 1 \leq k \leq K) \\ &\leq \sum_{j=1}^{\infty} \sum_{k=1}^{K} P(W_{jk} \geq 2) \\ &\leq \sum_{j=1}^{\infty} \sum_{k=1}^{K} (-\log (1 - b_{jk}))^2 = o(1). \end{aligned}$$

The last inequality follows from (14), (15), and the fact that $W_{jk} \sim Y(-\log (1 - b_{jk}))$. Also, note that because of (14) and (15),

$$\begin{aligned} d(\underset{\sim}{W}, \underset{\sim}{W}') &\leq P(\underset{\sim}{W} \neq \underset{\sim}{W}') \\ &\leq \sum_{k=1}^{K} P(W_k \neq W_k') \\ &\leq \sum_{k=1}^{K} \Big|\theta_k + \sum_{j=1}^{\infty} \log (1 - b_{jk})\Big| \\ &= o(1). \end{aligned}$$

Thus we have

$$d(\tilde{\underset{\sim}{L}}_r, W) = o(1).$$

It remains only to prove that

$$d(\tilde{\underset{\sim}{L}}_r, \underset{\sim}{L}_r) = o(1). \tag{19}$$

We first prove the analog of Lemma 2.1.

Lemma 3.1. For every positive constant D,

$$\eta_D = \max_{1 \le k \le K} \sup_{|i - s_k| \le D\sqrt{s_k}} P(L_r(i) \neq L_r(s_k)) = o(1).$$

Proof. Since $L_r(i)$ is an increasing function of i,

$$P(L_r(i) \neq L_r(s_k)) = P(|L_r(i) - L_r(s_k)| \ge \tfrac{1}{2})$$

$$\le 2|E(L_r(i)) - E(L_r(s_k))|.$$

A straightforward calculation shows

$$E(L_r(i)) = R_1(i) + R_2(i), \tag{20}$$

where

$$R_1(i) = \sum_{j=1}^{\infty} \binom{i}{r} p^r(j)(1 - p(j))^{i-r},$$

$$R_2(i) = \sum_{j=1}^{\infty} \sum_{t=r+1}^{i} \binom{i}{t} p^t(j)(1 - p(j))^{i-t}.$$

Using (12) and (13), it is not hard to prove that for any D > 0,

$$(21)\qquad \max_{1\le k\le K} \sup_{|i-s_k|\le D\sqrt{s_k}} |R_1(i) - \theta_k| = o(1)$$

and

$$(22)\qquad \max_{1\le k\le K} \sup_{|i-s_k|\le D\sqrt{s_k}} |R_2(i)| = o(1).$$

This proves the lemma. Q.E.D.

One can now complete the proof of Theorem 2 by arguing in essentially the same way as in Section 2. The details are left to the reader.

Remarks.

(1) Theorem 2 is a generalization of a result of Hafner [4]. He showed that

$$\{L_r(s_k) - L_r(s_{k-1})\}_{k=1}^{K} \sim \bigotimes_{k=1}^{K} Y(\theta_k - \theta_{k-1}).$$

(2) Just as for Theorem 1, it is possible to give an estimate of the error.

(3) The case r = 1 deserves special attention. The reason for this is that $E(L_1(s_k)) \to \infty$, $1 \le k \le K$. If r = 2, we do have that

$$(23)\qquad d(n_1 - L_1(n_1), Y(\theta_1)) = o(1).$$

To prove (23) we argue as follows. Define

$$H_r(n) = \text{number of urns with exactly r balls after n throws;}$$
$$= L_r(n) - L_{r+1}(n);\ r \ge 1.$$

Next, observe that

$$n_1 = \sum_{r=1}^{n_1} rH_r(n_1) \quad \text{and} \quad L_1(n_1) = \sum_{r=1}^{n_1} H_r(n_1).$$

Thus

$$\begin{aligned} n_1 - L_1(n_1) &= H_2(n_1) + \sum_{r=3}^{n_1} (r-1)H_r(n_1) \\ &= L_2(n_1) + \sum_{r=3}^{n_1} (r-2)H_r(n_1). \end{aligned}$$

Therefore,

$$\begin{aligned} d(n_1 - L_1(n_1), L_2(n_1)) &\le P\left(\sum_{r=3}^{n_1} (r-2)H_r(n_1) \ge 1\right) \\ &\le E\left(\sum_{r=3}^{n_1} (r-2)H_r(n_1)\right) = o(1). \end{aligned}$$

The last equality is an easy calculation left to the reader. (23) now follows from Theorem 2 since

$$d(L_2(n_1), Y(\theta_1)) = o(1).$$

BIBLIOGRAPHY

[1] P. Erdos and A. Renyi, On a classical problem in probability theory. Magyar Tud. Akad. Kutato Int. Kozl 6 (1961), pp. 215-219.

[2] D. Freedman, The Poisson approximation for dependent events. Ann. of Prob. 2 (1974), pp. 256-269.

[3] B. V. Gnedenko and A. N. Kolmogorov, Limit Distributions for Sums of Independent Random Variables (1954), Addison-Wesley, Reading, Mass.

[4] R. Hafner, Convergence of point complexes in Z^n to Poisson processes. Theory Prob. Appl. 1 (1973), pp. 131-148.

[5] J. L. Hodges and L. LeCam, The Poisson approximation to the binomial distribution. Ann. Math. Stat. 31 (1960), pp. 737-740.

[6] S. Karlin, Central limit theorems for certain infinite urn schemes. J. of Math. and Mech. 4 (1967), pp. 373-401.

[7] S. Karlin and J. McGregor, The number of mutant forms maintained in a population. Proc. 5th Berkeley Symp., Vol. IV (1967), pp. 415-438.