J. Appl. Prob. **48,** 189–199 (2011)
Printed in England

GENERALIZED COUPON COLLECTION: THE SUPERLINEAR CASE

R. T. SMYTHE,* *Oregon State University*

Abstract

We consider a generalized form of the coupon collection problem in which a random number, S, of balls is drawn at each stage from an urn initially containing n white balls (coupons). Each white ball drawn is colored red and returned to the urn; red balls drawn are simply returned to the urn. The question considered is then: how many white balls (uncollected coupons) remain in the urn after the k_n draws? Our analysis is asymptotic as $n \to \infty$. We concentrate on the case when k_n draws are made, where $k_n/n \to \infty$ (the superlinear case), although we sketch known results for other ranges of k_n. A Gaussian limit is obtained via a martingale representation for the lower superlinear range, and a Poisson limit is derived for the upper boundary of this range via the Chen–Stein approximation.

Keywords: Urn model; martingale; occupancy problem; coupon collection; central limit theorem; Poisson limit

2010 Mathematics Subject Classification: Primary 60F05; 60G42
Secondary 05A05; 60C05

1. Introduction to the coupon collection problem

In its simplest form, the coupon collection problem has a long history, beginning at least with De Moivre and Laplace. Consider the problem of placing k_n balls in n cells, independently and at random. Two of the basic questions are then: what is the distribution of the number of empty cells and how large must k_n be (on average) in order that there be no empty cells? Stadje (1990) provided references to the early history of the problem, and the book by Kolchin *et al.* (1978) gives details of much of the prior work in the field.

The connection with collecting coupons is perhaps better seen if we model this problem as an urn, initially containing n white balls (thought of as coupons). A single ball is picked at random from the urn; it is colored red and replaced in the urn. Then another ball is drawn; if it is white, it is painted red and returned to the urn; otherwise, the red ball drawn is simply returned to the urn. A third ball is drawn, and so forth. After k_n draws, the number of white balls remaining in the urn (uncollected coupons) corresponds to the number of empty cells in the occupancy problem described above.

The basic problem has been generalized in a myriad of ways, resulting in an enormous literature concerned with 'coupon collection problems', often described as occupancy problems. The 'birthday problem' and the 'Dixie cup problem' (the distribution of the number of draws required to draw some ball or, respectively, each ball at least $j > 1$ times) are among the best-known variants of the problem. In another extension of the problem, s balls are drawn at a time (for some integer $s \geq 1$), the white balls in the sample painted red, and all balls returned

Received 5 August 2009; revision received 2 December 2010.
* Postal address: Department of Statistics, Oregon State University, Corvallis, OR 97331-4606, USA.
Email address: smythe@science.oregonstate.edu

to the urn. In a well-known paper Pólya (1930) gave a formula for the average waiting time until all balls are colored red (all coupons collected), but this result is difficult to use for large values of n.

2. Coupon collection with a random number of draws

This paper is concerned with a further generalization, in which the number of balls drawn each time is a random variable S, taking values in $\{1, 2, \ldots, n\}$. The S balls are taken as a sample without replacement. The consecutive draws of size S are independent and identically distributed. At each draw, each white ball among the S drawn is painted red, and the entire sample returned to the urn. The same questions asked above can be asked for this problem; we consider the case where both n and k_n grow without bound, and ask what is the (asymptotic) distribution of the number of red balls; for the superlinear case, our result appears to be new.

Sellke (1995), Ivchenko (1998), and Adler and Ross (2001) studied the waiting time until all coupons are collected; the survey of Kobza *et al.* (2007) is a useful reference for the random sample size problem. (In this work, we assume that all balls in the urn have the same probability of being drawn; for the case $S \equiv 1$, unequal probabilities have been considered in Chistyakov (1964), Rosén (1969), and Holst (1971), among others.)

We classify the number k_n into five cases; in this (although not in our nomenclature) we follow Kolchin *et al.* (1978).

Case 1. Lower sublinear range: $k_n \to \infty$ and $k_n = o(\sqrt{n})$.

Case 2. Upper sublinear range: $k_n = o(n)$ and $\sqrt{n} = o(k_n)$.

Case 3. Linear range: $k_n \approx \alpha_n n$, where $\alpha_n > 0$ is bounded above and below.

Case 4. Lower superlinear range: $k_n = o(n \log(n))$.

Case 5. Upper superlinear range: $n \log(n) = o(k_n)$.

Note that there are two 'boundary cases' absent from this classification: $k_n = \Theta(\sqrt{n})$ and $k_n = \Theta(n \log(n))$. We will see that both of these correspond to 'phase changes' in the asymptotic behavior of the number of red balls (equivalently, the number of uncollected coupons).

In a recent paper, Mahmoud (2010) studied the asymptotic distribution, for random S, of the number of red balls in cases 1–3; we give a quick summary of results for these cases in Section 3, and concentrate in Section 4 on the superlinear cases. In Section 5 we deal with the boundary case in the superlinear range, and Section 6 contains some concluding remarks.

Mahmoud's approach makes use of a martingale central limit theorem (see Hall and Heyde (1980)). The present work also makes heavy use of a martingale central limit theorem, although our martingale is constructed differently from Mahmoud's and makes calculations relatively simple.

3. The sublinear and linear cases

3.1. Lower sublinear case

Let R_j and W_j respectively denote the number of red and white balls in the urn after j samples of (random) size S have been drawn. Since $R_j = n - W_j$, the mean and variance of either R_j or W_j follow immediately from those of the other. In the *lower sublinear* range, letting $S_1, S_2, \ldots, S_{k_n}$ denote the k_n independent and identically distributed realizations of

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

draws from the urn, Mahmoud (2010) showed that

$$R_j = \sum_{i=1}^{j} S_i + o_{\mathrm{P}}(1) \quad \text{for } 0 \le j \le k_n.$$

From this, it immediately follows that, if μ and σ^2 respectively denote the mean and variance of S,

$$\frac{R_{k_n} - \mu k_n}{\sqrt{k_n}} \to N(0, \sigma^2) \quad \text{in distribution.}$$

Note that, if $S \equiv s$, we do not have asymptotic normality of R_{k_n} here, but instead the result that

$$\mathrm{P}(R_{k_n} = sk_n) \to 1 \quad \text{as } n \to \infty.$$

At the 'boundary value' $k_n = \sqrt{n}$, when S is deterministic, the limit distribution is Poisson: if $S \equiv s$, and if $k_n^2 s/2n \to \lambda$, then

$$sk_n - R_{k_n} \to \text{Poisson}(\lambda) \quad \text{in distribution.}$$

Kolchin *et al.* (1978) attributed this result to Békéssy (1963) when $s = 1$; Mikhaĭlov (1977) stated the result for general s.

3.2. Upper sublinear case

The asymptotic normality result in this case is slightly more subtle. We will need an exact result for $\mathrm{E}(R_{k_n})$, and we also record, for later use, $\mathrm{var}(R_{k_n})$.

Lemma 3.1. (Mahmoud (2010).) *We have*

$$\mathrm{E}(R_{k_n}) = n\left(1 - \left(1 - \frac{\mu}{n}\right)^{k_n}\right),$$

$$\mathrm{var}(R_{k_n}) = n\left[(n-1)\left(\frac{(n-\mu)(n-\mu-1)+\sigma^2}{n(n-1)}\right)^{k_n} + \left(\frac{n-\mu}{n}\right)^{k_n}\right] - \left(\frac{n-\mu}{n}\right)^{2k_n} n^2.$$

When S is genuinely random ($\sigma^2 > 0$), the following result, also due to Mahmoud, holds in the upper sublinear range:

$$\frac{R_{k_n} - n(1-(1-\mu/n)^{k_n})}{\sqrt{k_n}} \to N(0, \sigma^2) \quad \text{in distribution.}$$

This result also holds at the boundary $k_n = c\sqrt{n}$, so the limiting behavior at the boundary is different for random S than in the deterministic case.

When S is not random, a smaller normalizing factor suffices in the denominator for asymptotic normality in the upper sublinear range: for fixed sample size s,

$$\frac{R_{k_n} - n(1-(1-s/n)^{k_n})}{k_n/\sqrt{n}} \to N\left(0, \frac{1}{2}s^2\right) \quad \text{in distribution.}$$

For the case $s = 1$, this result is due to Rényi (1962); for general s, Kolchin *et al.* (1978, p. 215) appears to have given the first proof.

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

3.3. Linear case

The proof of asymptoptic normality was first given in this case, for $S \equiv 1$, by Weiss (1958). For the case of random S, let

$$v_n \equiv n\mathrm{e}^{-2\mu\alpha_n}(\mathrm{e}^{\mu\alpha_n} + \alpha_n(\sigma^2 - \mu) - 1).$$

Mahmoud (2010) showed that

$$\frac{R_{k_n} - n(1 - \mathrm{e}^{-\mu\alpha_n})}{\sqrt{v}_n} \to N(0, 1) \quad \text{in distribution.}$$

In the linear region, unlike the upper sublinear region, a nondegenerate normal limit follows from the above result by setting $S \equiv s$.

4. The superlinear case

4.1. Preliminaries

We will need a few technical results in preparation for the main theorem.

Lemma 4.1. *In the superlinear range,*

$$\frac{\mathrm{var}(W_{k_n})}{n\mathrm{e}^{-\mu k_n/n}} \to 1 \quad \textit{as } n \to \infty.$$

Proof. Using the result for $\mathrm{var}(W_{k_n})$ in Lemma 3.1, the ratio of the two quantities can be expressed as

$$1 + n\left(\left(1 - \frac{\mu}{n-1} + \frac{\sigma^2}{n(n-1)(1-\mu/n)}\right)^{k_n} - \left(1 - \frac{\mu}{n}\right)^{k_n}\right) - O(\mathrm{e}^{-k_n/(n-1)}),$$

and the second and third terms go to 0 in the superlinear range. (As a consequence of this result, in the superlinear range the variance of S does not enter into the asymptotic distribution of R_{k_n}.)

The next two results apply for k_n in all ranges.

Lemma 4.2. (i) *There exists a* $C > 1$ *such that* $\mathrm{var}(W_{k_n}) \le C\,\mathrm{E}(W_{k_n})$ *for* $n > N(S)$.

(ii) *If* $\sigma^2 < \mu$, C *may be taken to equal* 1 *in (i).*

Proof. See Appendix A.

Lemma 4.3. *Let* S *be random with mean* μ. *If* $\limsup k_n/n \log(n) < 1/\mu$ *then, for* $j \le k_n$,

$$\frac{W_j}{n} = \mathrm{e}^{-j\mu/n} + o_{L_1}(\mathrm{e}^{-j\mu/n}).$$

Proof. Fix $\varepsilon > 0$, and take $n > N(S)$, as in Lemma 4.2. Then

$$\mathrm{P}\left(\left|\frac{W_j}{n} - \mathrm{E}\left(\frac{W_j}{n}\right)\right| > \varepsilon\mathrm{e}^{-j\mu/n}\right) \le \frac{\mathrm{var}(W_j/n)\mathrm{e}^{2j\mu/n}}{\varepsilon^2} \le \frac{C\mathrm{e}^{-j\mu/n}}{n}\frac{\mathrm{e}^{2j\mu/n}}{\varepsilon^2} \le \frac{C}{\varepsilon^2}\frac{\mathrm{e}^{j\mu/n}}{n},$$

and the last term converges to 0 under the conditions given. Hence, we have

$$\mathrm{P}\left(\left|\frac{W_j}{n\,\mathrm{E}(W_j/n)} - 1\right| > \varepsilon\frac{\mathrm{e}^{-j\mu/n}}{\mathrm{E}(W_j/n)}\right) \to 0,$$

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

i.e.

$$\mathrm{P}\left(\left|\frac{W_j}{n\mathrm{e}^{-j\mu/n}} - 1\right| > \varepsilon\right) \to 0.$$

Hence, $\mathrm{E}\,|W_j/n - \mathrm{e}^{-j\mu/n}|\mathrm{e}^{j\mu/n} \to 0$ and $\mathrm{E}\,|W_j/n - \mathrm{e}^{-j\mu/n}| = o(\mathrm{e}^{-j\mu/n})$.

4.2. The basic martingale

We are finally ready to construct the basic martingale. For a given n, let $\mathcal{F}_{in}$ denote the sigma-field generated by the first i draws from the urn that initially contains n white balls. Let X_{in} denote the number of white balls drawn in the ith draw, when the number of balls drawn is random with mean μ and variance σ^2. Let

$$\Delta X_{in} = X_{in} - \mathrm{E}(X_{in} \mid \mathcal{F}_{i-1,n}).$$

Conditional on $S_i = s$, and on $\mathcal{F}_{i-1,n}$, the distribution of X_{in} is hypergeometric, and we have

$$\mathrm{E}(X_{in} \mid \mathcal{F}_{i-1,n},\ S_i = s) = \frac{s}{n}W_{i-1}, \qquad \mathrm{var}(X_{in} \mid \mathcal{F}_{i-1,n},\ S_i = s) = s\frac{n-s}{n-1}\frac{R_{i-1}}{n}\frac{W_{i-1}}{n}.$$

Noting that $R_{k_n} = \sum_{i=1}^{k_n} X_{in}$ and that $\mathrm{E}(X_{in} \mid \mathcal{F}_{i-1,n})$ is a linear function of the $\{X_{in}\}$, we can find constants $\{b_{in}\}$ such that

$$R_{k_n} - \mathrm{E}(R_{k_n}) = \sum_{i=1}^{k_n} b_{in}\Delta X_{in}, \tag{4.1}$$

and the right-hand side will then be a martingale for each value of n. Recalling Lemma 3.1, a bit of algebra then gives $b_{in} = (1-\mu/n)^{k_n - i}$. We also have

$$\begin{aligned}
\mathrm{E}(\Delta X_{in}^2 \mid \mathcal{F}_{i-1,n}) &= \mathrm{var}(\Delta X_{in} \mid \mathcal{F}_{i-1,n}) \\
&= \mathrm{var}(X_{in} \mid \mathcal{F}_{i-1,n}) \\
&= \mathrm{E}(\mathrm{var}(X_{in} \mid \mathcal{F}_{i-1,n}; S_i)) + \mathrm{var}(\mathrm{E}(X_{in} \mid \mathcal{F}_{i-1,n}; S_i)) \\
&= \mathrm{E}\left(S_i\frac{n-S_i}{n-1}\frac{W_{i-1}}{n}\left(1-\frac{W_{i-1}}{n}\right)\right) + \frac{\sigma^2 W_{i-1}^2}{n^2} \\
&= \left(\frac{n}{n-1}\mu - \frac{\sigma^2+\mu^2}{n-1}\right)\frac{W_{i-1}}{n}\left(1-\frac{W_{i-1}}{n}\right) + \frac{\sigma^2 W_{i-1}^2}{n^2}.
\end{aligned} \tag{4.2}$$

The convergence of the martingale in (4.1) (suitably normalized) to a normal limit will be shown via a martingale central limit theorem in the form presented in Hall and Heyde (1980, p. 58). Let $\Delta Y_{in} = b_{in}\Delta X_{in}$. The sufficient conditions for this theorem are that, for some positive increasing sequence λ_n and all $\varepsilon > 0$,

$$U_n := \sum_{i=1}^{k_n} \mathrm{E}\left(\left(\frac{\Delta Y_{in}}{\lambda_n}\right)^2 \mathbf{1}_{\{\Delta|Y_{in}|>\varepsilon\lambda_n\}} \,\middle|\, \mathcal{F}_{i-1,n}\right) \to 0 \quad \text{in probability,}$$

and that a conditional variance condition requiring (in our case) that

$$V_n := \sum_{j=1}^{k_n} \mathrm{E}\left(\left(\frac{\Delta Y_{in}}{\lambda_n}\right)^2 \,\middle|\, \mathcal{F}_{i-1,n}\right) \to c^2 \quad \text{in probability}$$

holds, where c^2 is a positive constant.

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

4.3. The central limit theorem in the superlinear case

Here is the main result on asymptotic normality in the lower superlinear case.

Theorem 4.1. *Let μ denote the mean of S, the random number of draws, and suppose that $k_n/n \to \infty$ and $\limsup k_n/n \log(n) < 1/\mu$. Then*

$$\frac{R_{k_n} - \mathrm{E}(R_{k_n})}{\sqrt{n\mathrm{e}^{-\mu k_n/n}}} \to N(0,1) \quad \textit{in distribution.}$$

Proof. We take $\lambda_n^2 = n\mathrm{e}^{-\mu k_n/n}$. For any i, $W_i/n \le 1$, and, using (4.2),

$$\mathrm{E}(\Delta X_{in}^2 \mid \mathcal{F}_{i-1,n}) \le 2\mu + \sigma^2 + O\left(\frac{1}{n}\right).$$

By hypothesis, there exists a $\delta > 0$ such that, for sufficiently large n, $k_n/n\log(n) < (1-\delta)/\mu$. This implies that

$$n\mathrm{e}^{-\mu k_n/n} \ge n^{\delta}$$

for large n; since $|b_{in}| \le 1$, it follows easily that $|\Delta Y_{in}^2| \le \varepsilon^2\lambda_n^2$ when n is sufficiently large, and the condition $U_n \to 0$ is satisfied.

To examine the quantity V_n, we divide the sum from 1 to k_n into two sums: the first goes from 1 to Mn, where M is an arbitrary positive integer, and the second goes from Mn to k_n. Denote these sums by V_{n1} and V_{n2}, respectively.

First consider V_{n1}. Here we use the bound of Lemma 4.3, with $\rho_n \equiv n/(n-\mu)$, to write

$$V_{n1} \le 2\frac{(1-\mu/n)^{2k_n}}{\lambda_n^2}\sum_{j=1}^{Mn}\rho_n^{2j}(\mu\mathrm{e}^{-\mu j/n} + \sigma^2\mathrm{e}^{-2\mu j/n} + o_{L_1}(\mathrm{e}^{-\mu j/n})).$$

Note first that

$$\sum_{j=1}^{Mn}\mu(\rho_n^2\mathrm{e}^{-\mu/n})^j = n(1+o(1))(\mathrm{e}^{-M\mu}\rho_n^{2Mn} - 1),$$

so that taking $\lambda_n^2 = n\mathrm{e}^{-\mu k_n/n}$ means that this contribution to the sum is $O(\mathrm{e}^{-\mu k_n/n})$, which converges to 0 in the range of k_n considered.

For the second sum, observe that

$$\sum_{j=1}^{Mn}\sigma^2(\rho_n^2\mathrm{e}^{-2\mu/n})^j = (n^2)(1+o(1))\frac{\sigma^2}{\mu^2}(\mathrm{e}^{-2\mu M}\rho_n^{2Mn} - 1) = 2nM\sigma^2 + o(n),$$

so the second part of V_{n1} is also $O(\mathrm{e}^{-\mu k_n/n})$. Putting the two parts together gives $V_{n1} \to 0$.

We turn now to V_{n2}, the sum from Mn to k_n. Using (4.2) and Lemma 4.3, V_{n2} is asymptotically equivalent to

$$\frac{(1-\mu/n)^{2k_n}}{n\mathrm{e}^{-\mu k_n/n}}\sum_{j=Mn}^{k_n}\mu(\rho_n^2\mathrm{e}^{-\mu/n})^j + \sum_{j=Mn}^{k_n}(\sigma^2-\mu)(\rho_n^2\mathrm{e}^{-2\mu/n})^j.$$

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

The first of these two sums is equal to

$$\frac{(1-\mu/n)^{2k_n}}{n\mathrm{e}^{-\mu k_n/n}} n(1+o(1))((\rho_n^2\mathrm{e}^{-\mu/n})^{k_n} - (\rho_n^2\mathrm{e}^{-\mu/n})^{Mn})$$
$$= \mathrm{e}^{-\mu k_n/n}\left(\mathrm{e}^{\mu k_n/n}\left(1+O\left(\frac{k_n}{n^2}\right)\right) - \mathrm{e}^{\mu M}\left(1+O\left(\frac{1}{n}\right)\right)\right) + o(1),$$

which converges to 1 as n tends to ∞. The second of these two sums converges to 0. The contribution to V_{n2} from the second sum is asymptotically equal to

$$\frac{(1-\mu/n)^{2k_n}}{n\mathrm{e}^{-k_n/n}} \frac{n^2(1+o(1))(\sigma^2-\mu)}{\mu^2}((\rho_n^2\mathrm{e}^{-2\mu/n})^{k_n} - (\rho_n^2\mathrm{e}^{-2\mu/n})^{Mn})$$
$$= \frac{n^2(1+o(1))(\sigma^2-\mu)}{n\mu^2\mathrm{e}^{-\mu k_n/n}}\left(\mathrm{e}^{-2\mu k_n/n} - \rho_n^{2Mn}\mathrm{e}^{-2\mu M}\left(1-\frac{\mu}{n}\right)^{2k_n}\right)$$
$$= \frac{n(1+o(1))(\sigma^2-\mu)}{\mu^2\mathrm{e}^{-\mu k_n/n}}\mathrm{e}^{-2\mu k_n/n}\left(1-\left(1-\frac{\mu}{n}\right)^{2k_n}\rho_n^{2Mn}\mathrm{e}^{2\mu(k_n/n-M)}\right)$$
$$= \frac{n(1+o(1))(\sigma^2-\mu)}{\mu^2}\mathrm{e}^{-\mu k_n/n}\left(1-\mathrm{e}^{2\mu k_n/n}\left(1-\frac{\mu}{n}\right)^{2k_n}\left(1+O\left(\frac{1}{n}\right)\right)\right)$$
$$= \frac{n(1+o(1))(\sigma^2-\mu)}{\mu^2}\mathrm{e}^{-\mu k_n/n}\left(O\left(\frac{k_n}{n^2}\right)\right)$$
$$= \frac{\sigma^2-\mu}{\mu^2}\mathrm{e}^{-\mu k_n/n}O\left(\frac{k_n}{n}\right) + o(1),$$

and this goes to 0 as $n \to \infty$. Hence,

$$V_n \to 1 \quad \text{in probability,}$$

completing the proof.

For $S \equiv 1$, the result was established in Zubkov and Mikhaĭlov (1974); the extension to $S \equiv s$ is attributed to Kolchin *et al.* (1978, p. 221).

4.4. Extension to $S = S(n)$

So far we have assumed that the same random variable S is used to determine the number of draws for each value of n. Some extensions of our results hold for the case when S has a possibly different distribution depending on n. Suppose that the draws from the urn of size n are governed by a random variable $S(n)$ with mean μ_n and variance σ_n^2. Under some fairly strong conditions on the growth of μ_n and σ_n^2, we can state the following corollary to Theorem 4.1.

Corollary 4.1. *Suppose that $S(n)$ has mean μ_n and variance σ_n^2. Assume that*

(i) $\limsup \mu_n k_n/n \log(n) < 1$;

(ii) *either $\{\sigma_n^2\}$ is bounded, or $\sigma_n^2 < \mu_n$ for $n > n_0$ and $\sigma_n^2 = o(\mathrm{e}^{\mu_n k_n/n})$.*

Then

$$\frac{R_{k_n} - \mathrm{E}(R_{k_n})}{\sqrt{n\mathrm{e}^{-\mu_n k_n/n}}} \to N(0,1) \quad \textit{in distribution.}$$

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

Proof. The proof follows along the same lines as that of Theorem 4.1. We note that condition (i) implies, since $k_n/n \to \infty$, that $\mu_n = o(\log(n))$. The strong condition on the variances is imposed by our proof of Lemma 4.2, which is key to the basic inequality in Lemma 4.3.

4.5. The upper superlinear case

To complete the picture for k_n superlinear, we state briefly a result in the upper superlinear range. In the upper superlinear range, as in the lower sublinear range, the limit behavior of R_{k_n} is degenerate.

Lemma 4.4. *Suppose that* $\liminf k_n/n\log(n) > 1/\mu$. *Then*

$$W_{k_n} = n - R_{k_n} \to 0 \quad \textit{in probability}.$$

Proof. We have $\mathrm{P}(W_{k_n} \geq 1) \leq \mathrm{E}(W_{k_n}) = n(1-\mu/n)^{k_n}$ and this goes to 0 in the range of k_n specified.

5. The superlinear boundary

If $S \equiv s$ and $k_n = (n/s)\log(n/\lambda)$, a special case of the results of Mikhaĭlov (1977) is that

$$W_{k_n} = n - R_{k_n} \to \mathrm{Poisson}(\lambda) \quad \text{in distribution.}$$

Kolchin *et al.* (1978, p. 30) attributed this result for the case $S \equiv 1$ to von Mises (1939). Here we present, with a simple proof using the Chen–Stein Possion approximation, a corresponding result due to Ivchenko (1998) for the case of random S. We begin with an easy lemma.

Lemma 5.1. *Suppose that* $\mu k_n/n\log(n/\lambda) \to 1$ *for* $\lambda > 0$ *and that* $\phi(n)$ *is a function of* n *that increases without bound. Then, for* $\varepsilon > 0$,

$$\mathrm{P}(W_{k_n} > \varepsilon\phi(n)) \to 0 \quad \textit{in probability}.$$

Proof. We have $\mathrm{P}(W_{k_n} > \varepsilon\phi(n)) \leq \mathrm{E}(W_{k_n})/\varepsilon\phi(n)$. However, under the assumptions, $\mathrm{E}(W_{k_n}) = n\mathrm{e}^{-\mu k_n/n} \approx \lambda$ and $1/\phi(n) \to 0$.

Theorem 5.1. *Suppose that* $\mu k_n/n\log(n/\lambda) \to 1$ *for* $\lambda > 0$. *Then*

$$W_{k_n} = n - R_{k_n} \to \mathrm{Poisson}(\lambda) \quad \textit{in distribution}.$$

Proof. We use a Poission approximation theorem, as presented in Arratia *et al.* (1990). Start with n white balls in the urn, labeled $1, 2, \ldots, n$, and let k_n draws of a random number, S, of balls be made from the urn, where $n = o(k_n)$. Let $Y_i := 1$, $i = 1, 2, \ldots, n$, if ball i is never drawn in the k_n draws. Then

$$p_i := \mathrm{P}(Y_i = 1) = \prod_{j=1}^{k_n}\left(1 - \frac{S_j}{n}\right),$$

where S_j denotes the number of balls drawn in the jth draw. Then, because the $\{Y_i\}$ are exchangeable and $\sum_{i=1}^n p_i = \mathrm{E}(W_{k_n}) = n(1-\mu/n)^{k_n} \to \lambda$, we must have $p_i = (1-\mu/n)^{k_n}$.

In the approach of Arratia *et al.* (1990, p. 405), a 'neighborhood' of each point i is specified and three quantities, b_1, b_2, and b_3, evaluated; the total variation distance between the law of

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

W_{k_n} and a Poisson(λ) distribution is then bounded by $2(b_1 + b_2 + b_3)$. We take the singleton $\{i\}$ as the neighborhood of the point i. Using this notation, we have

$$b_1 = \sum_{i=1}^{n} p_i^2 = n\left(1 - \frac{\mu}{n}\right)^{2k_n} \le n\mathrm{e}^{-2\mu k_n/n} = \frac{\lambda^2}{n} + o\left(\frac{1}{n}\right)$$

and

$$b_2 = \sum_{i=1}^{n} \sum_{j \ne i \varepsilon \{i\}}^{n} \mathrm{E}(\mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=1\}}) = 0.$$

The exchangeability of the Y_i gives $b_3 = n\,\mathrm{E}(|\,\mathrm{E}(Y_i - p_i) \mid Y_j,\ j \ne i|)$. It remains to compute $\mathrm{E}(Y_i \mid Y_j,\ j \ne i)$. This conditional expectation will depend on $\{Y_j,\ j \ne i\}$ only through $\sum_{j \ne i} Y_j$. If this sum equals m, where m must satisfy $m \le n - \max(S_1, \ldots, S_{k_n})$, the conditional expectation will be

$$\prod_{i=1}^{k_n}\left(1 - \frac{S_i}{n-m}\right).$$

Hence,

$$\mathrm{E}\left(Y_i \,\middle|\, \sum_{j \ne i} Y_j = m\right) = \mathrm{e}^{-\mu k_n/(n-m)}\, \mathbf{1}_{\{m \le n - \max\{S_j\}\}}\, \Theta(1)$$

and

$$\mathrm{E}\left(Y_i \,\middle|\, \sum_{j \ne i} Y_j\right) = \exp\left[-\frac{\mu k_n}{n - \sum_{j \ne i} Y_j}\right] \mathrm{P}\left(n - \sum_{j \ne i} Y_j \ge \max\{S_j\}\right)\Theta(1).$$

Then, from Lemma 5.1,

$$\mathrm{E}(Y_i \mid Y_j, j \ne i) - p_i = \left(1 - \frac{\mu}{n}\right)^{k_n} \Theta_p(1)\left(\frac{\exp[-\mu k_n/(n - \sum_{j \ne i} Y_j)]}{(1-\mu/n)^{k_n}} - 1\right).$$

Since $(1 - \mu/n)^{k_n} = (\lambda/n)(1 + o(1))$, to prove that $b_3 \to 0$, it suffices to show that

$$\mathrm{E}\left|\frac{\exp[-\mu k_n/(n - \sum_{j \ne i} Y_j)]}{(1-\mu/n)^{k_n}} - 1\right| \to 0. \tag{5.1}$$

It is easily shown that the fraction inside the expectation in (5.1) is bounded in n. We have $R_{k_n} \le n - \sum_{j \ne i} Y_j \le n$, so to finish the proof that $b_3 \to 0$, we invoke Lemma 5.1 to show that $\mathrm{e}^{\mu k_n/R_{k_n}}/\mathrm{e}^{-\mu k_n/n} \to 1$ in probability.

We now have $b_1 \to 0$, $b_2 = 0$, and $b_3 \to 0$ as $n \to \infty$, so convergence in distribution to Poisson(λ) is established.

6. Concluding remarks

At the risk of belaboring the obvious, we comment on two points regarding our problem. One concerns the symmetry between the sublinear and superlinear cases when S is deterministic, with degenerate limits in the extreme regions, Poisson limits at the two boundary cases, and Gaussian limiting distributions in the upper sublinear and lower superlinear ranges. This symmetry does not persist in the case of random S, where a Gaussian limit holds in the entire sublinear range. The second (related) point is the role of the variance σ^2 of S, the random

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

number of draws. In the sublinear range, the influence of σ^2 is considerable; for the upper sublinear range, the sequence of normalizing constants that give a limit distribution in the case of deterministic s is of a smaller order of magnitude than in the random case. In the linear range, a nonzero variance serves to 'tweak' the normalizing constants, but does not affect their order of magnitude; in the superlinear range, the variance of S does not affect the asymptotic results. As a result, the discontinuity between the lower and upper superlinear regimes persists in the random case, in contrast to the sublinear regime, where the discontinuity disappears when the variance of S becomes positive.

Finally, we observe that the basic martingale constructed in Section 4 gives rather easily, using an approach similar to that of Theorem 4.1, the known Gaussian limits for the sublinear and linear ranges of k_n.

Appendix A. Proof of Lemma 4.2

Using Lemma 3.1, let $A_n = (1 - \mu/n)$ and $B_n = \sigma^2/n(n-1)$. Then

$$\operatorname{var}(W_{k_n}) \le n^2\{(A_n^2 + B_n)^{k_n} - A_n^{2k_n} + nA_n^{k_n}\}.$$

Fix $\varepsilon > 0$. Then the claim is that

$$n^2\{(A_n^2 + B_n)^{k_n} - A_n^{2k_n}\} \le (C-1)nA^{k_n} \quad \text{if } n > N(\varepsilon, S). \tag{A.1}$$

Dividing both sides of (A.1) by $nA_n^{k_n}$, we obtain

$$n\left\{\left(A_n + \frac{B_n}{A_n}\right)^{k_n} - A_n^{k_n}\right\} \le C - 1$$

for sufficiently large n. Note that $B_n/A_n = \sigma_S^2/(n-1)(n-\mu_S) = O(1/n^2)$.

By the mean value theorem,

$$\left(A_n + \frac{B_n}{A_n}\right)^{k_n} - A_n^{k_n} = \frac{B_n}{A_n} k_n \psi_n^{k_n - 1}$$

for some ψ_n satisfying $A_n \le \psi_n \le A_n + B_n/A_n$, so that

$$\psi_n \le 1 - \frac{\mu}{n} + \frac{\sigma^2}{(n-1)(n-\mu)}.$$

We have $\psi_n^n \le C_1 \mathrm{e}^{-\mu}$ for $n > N(S)$, so that

$$n\left\{\left(A_n + \frac{B_n}{A_n}\right)^{k_n} - A_n^{k_n}\right\} \le C_2 \frac{k_n}{n} \mathrm{e}^{-\mu k_n/n}$$

for $n > N(S)$. The quantity $(k_n/n)\mathrm{e}^{-\mu k_n/n}$ is bounded by $\mathrm{e}^{-1}/\mu \le \mathrm{e}^{-1}$, so, for sufficiently large n,

$$n\left\{\left(A_n + \frac{B_n}{A_n}\right)^{k_n} - A_n^{k_n}\right\} \le C_2 \frac{k_n}{n} \mathrm{e}^{-\mu(k_n/n)} < C - 1.$$

To prove part (ii) of the lemma, it suffices to show that

$$\left(\frac{(n-\mu)(n-1-\mu_)+\sigma^2}{n(n-1)}\right)^{k_n} \le \left(\frac{n-\mu}{n}\right)^{2k_n}$$

https://doi.org/10.1239/jap/1300198144 Published online by Cambridge University Press

for large enough n, when $\sigma^2 < \mu$. This holds if

$$\frac{(n-\mu)(n-\mu-1)+\sigma^2}{n(n-1)} \le \left(\frac{n-\mu}{n}\right)^2,$$

or

$$\frac{n-1-\mu}{n-1} + \frac{\sigma^2}{(n-1)(n-\mu)} \le \frac{n-\mu}{n},$$

or

$$(n-1-\mu) + \frac{\sigma^2}{n-\mu} \le \frac{n-1}{n}(n-\mu),$$

which reduces to

$$\frac{\sigma^2}{\mu} < 1 - \frac{\mu}{n},$$

and this holds if $n > N(S)$.

Acknowledgement

The author thanks Hosam Mahmoud for valuable comments on an earlier draft of this manuscript.

References

Adler, I. and Ross, S. M. (2001). The coupon subset collection problem. *J. Appl. Prob.* **38,** 737–746.

Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen–Stein method. *Statist. Sci.* **5,** 403–434.

Békéssy, A. (1963). On classical occupancy problems. I. *Magyar Tud. Akad. Mat. Kuktató Int. Közl.* **8,** 59–71.

Chistyakov, V. P. (1964). On the calculation of the power of the test of empty boxes. *Theory Prob. Appl.* **9,** 648–653.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications*. Academic Press, New York.

Holst, L. (1971). Limit theorems for some occupancy and sequential occupancy problems. *Ann. Math. Statist.* **42,** 1671–1680.

Ivchenko, G. I. (1998). How many samples does it take to see all of the balls in an urn? *Math. Notes* **64,** 49–54.

Kobza, J. E., Jacobson, S. H. and Vaughan, D. E. (2007). A survey of the coupon collector's problem with random sample sizes. *Methodology Comput. Appl. Prob.* **9,** 573–584.

Kolchin, V. F., Sevast'yanov, B. A. and Chistyakov, V. P. (1978). *Random Allocations.* Winston, Washington, DC.

Mahmoud, H. M. (2010). Gaussian phases in generalized coupon collection. *Adv. Appl. Prob.* **42,** 994–1012.

Mikhaĭlov, V. (1977). A Poisson limit theorem in the scheme of group disposal of particles. *Theory Prob. Appl.* **22,** 152–156.

Pólya, G. (1930). Eine Wahrscheinlichkeitsaufgabe zur Kunderwerbung. *Z. Angew. Math. Mech.* **10,** 96–97.

Rényi, A. (1962). Three new proofs and a generalization of a theorem of Irving Weiss. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **7,** 203–214.

Rosén, B. (1969). Asymptotic normality in a coupon collector's problem. *Z. Wahrscheinlichkeitsth.* **13,** 256–279.

Sellke, T. (1995). How many i.i.d. samples does it take to see all the balls in a box? *Ann. Appl. Prob.* **5,** 294–309.

Stadje, W. (1990). The collector's problem with group drawings. *Adv. Appl. Prob.* **22,** 866–882.

Weiss, I. (1958). Limiting distributions in some occupancy problems. *Ann. Math. Statist.* **29,** 878–884.

Von Mises, R. (1939). Über aufteilungs- und besetzungs-Wahrscheinlichkieten. *Revu de la Faculté des Sciences de l'Université d'Istanbul*, Vol. 4, pp. 145–163.

Zubkov, A. M. and Mikhaĭlov, V. G. (1974). Limit distributions of random variables associated with long duplications in a sequence of independent trials. *Theory Prob. Appl.* **19,** 172–179.