How Many Geometric (p) Samples Does It Take to See All the Balls in a Box?

by

Thomas Sellke and John Overdeck

Technical Report #92-51

Department of Statistics
Purdue University

November 1992
Revised March 1993

*

Abstract

Reach into a box containing m balls and pull out a geometric (p) - sized sample. Then put the balls back into the box and sample again. Let X be the number of samples needed to see all m balls. We derive nonrecursive approximation formulas for the mean and standard deviation of X.

1. Introduction

A box contains m identical white balls. Let $K_1, K_2, \ldots$ be independent geometric (p) random variables, so that $P\{K_i = k\} = (1-p)^{k-1}p = q^{k-1}p$ for $k = 1, 2, \ldots$. We sample repeatedly as follows. Sample $K_1 \wedge m$ balls from the box without replacement, paint them red, and then return them to the box. Then sample $K_2 \wedge m$ balls from the box without replacement, paint them red, and put them back into the box, etc. We wish to determine the mean and variance of X, the number of samples needed to paint all the balls red.

As was pointed out to us by Larry Shepp, it is straightforward to derive exact recursive formulas for the mean and variance of the remaining number of samples needed to paint all the balls red when j of the m balls are still white. These formulas are given in the Appendix.

The focus of this paper, however, will be to derive good nonrecursive approximations for EX and the standard deviation σ_X. Our main results are given in Propositions 3.9 (for EX) and 4.18 (for σ_X). In Section 5, exact values for EX and σ_X computed using Shepp's recursions will be compared with our approximations for several values of m ranging from 10 to 300 for $p = \frac{1}{2}$.

2. The (Z, W) process

Our arguments will relate the original sampling-and-painting process described in Section 1 to the following alternative process. Let $Z_1, Z_2, \ldots$ be independent random variables uniformly distributed on $\{1, 2, \ldots, m\}$. Let $W_1, W_2, \ldots$ be Bernoulli (p) random variables, independent of each other and of the Z_i's. The *iid* sequence $\{(Z_i, W_i)\}_{i=1}^{\infty}$ will be referred to as the (Z, W) process.

The sampling-and-painting process of Section 1 can be constructed from the (Z, W) process. Suppose the m balls in the box are numbered $1, 2, \ldots, m$. The Z_i's can be thought of as the result of drawing balls from the box one at a time, with replacement. Suppose also that after each ball draw we flip a coin which has probability p of coming up heads. We may think of W_i as being the indicator of the event that the i^{th} coin toss was heads.

Sampling k balls without replacement can of course be done by drawing balls one at a time, *with replacement*, and ignoring draws of previously drawn balls until k distinct balls have been obtained. Following this approach, define the first sample to be those balls drawn before the first "counted" head, where ball draws and the associated coin flips are "counted" only if the ball drawn does not repeat a previous draw (of the current sample.) To get the second sample, we start anew according to the same rules after the first sample has been completed. The process of generating the first sample usually ends with the first counted head. However, if the coin flips following the first m counted draws are all tails (corresponding to the event $\{K_1 > m\}$), then the process of generating the first sample

ends after the m^{th} counted draw, (and the first sample contains all m balls.)

3. The expectation of X

Let D_i be the number of single-ball draws (counted and uncounted) needed to generate the i^{th} sample. Note that the D_i's are *iid.* Let $\mathcal{F}_i$ be the σ-field generated by all observations made while generating the first i samples. Note that X is an $\mathcal{F}_i$ stopping time. The total number of draws needed to generate the first X samples is $\sum_1^X D_i$. Let $q = 1 - p$.

Lemma 3.1 $E(\sum_1^X D_i) = (ED_1)EX$

Proof: Wald's identity. □

Lemma 3.2

$$ED_1 = \sum_{i=0}^{m-1} \frac{m}{m-i} q^i.$$

Proof

If $K_1 \wedge m > i$, then the number of draws needed to obtain the $(i+1)^{st}$ distinct ball after i distinct balls have already been obtained is a geometric $(\frac{m-i}{m})$ random variable with mean $\frac{m}{m-i}$. Thus,

$$E(D_1|K_1) = \sum_{i=0}^{m-1} \frac{m}{m-i} I\{K_1 > i\}. \tag{3.3}$$

Now take expectations in (3.3) □

Let τ be the number of single-ball draws needed to obtain every ball at least once.

Lemma 3.4

$$E\tau = \sum_{i=0}^{m-1} \frac{m}{m-i} = m \sum_{k=1}^{m} \frac{1}{k}$$

$$\text{var}(\tau) = \sum_{i=0}^{m-1} \frac{i}{m} (\frac{m}{m-i})^2 = m^2 \sum_{k=1}^{m} \frac{1}{k^2} - m \sum_{k=1}^{m} \frac{1}{k}$$

Proof

This situation is sometimes called the coupon collector problem. As in Lemma 3.2, τ is a sum of geometric $(\frac{m-i}{m})$ random variables, here with $i = 0, 1, \ldots, m-1$. Adding the expectations gives the formula for $E\tau$. These geometric random variables are independent, so adding their variances gives the formula for var (τ). □

Now we need to relate $E(\sum_i^X D_i)$ to $E\tau$. The sum $\sum_1^X D_i$ and τ are of course equal except for the "overshoot" given by the number of draws needed to complete the last

sample after the τ^{th} draw. Let V be the size of the overshoot, so that

$$V = (\sum_{1}^{X} D_i) - \tau.$$

From Lemmas 3.1, 3.2, and 3.4

$$\begin{aligned} EX &= \frac{E\sum_{1}^{X} D_i}{ED_1} \\ &= \frac{E\tau + EV}{ED_1} \\ &= \frac{m\sum_{k=1}^{m}\frac{1}{k} + EV}{m\sum_{i=0}^{m-1}\frac{1}{m-i}q^i} \end{aligned} \tag{3.5}$$

Let J be the number of distinct balls in the last sample after the τ^{th} draw. (The τth draw is of course "counted" since the ball drawn was by definition never drawn before. The τ^{th} draw is the J^{th} counted draw in the X^{th} sample.) The expectation of V is easy to find in terms of the distribution of J.

Lemma 3.6

$$EV = \sum_{j=1}^{m-1} P\{J=j\} \sum_{i=j}^{m-1} \frac{m}{m-i} q^{i-j+1}.$$

Proof

The argument is again like the coupon collector argument used for Lemmas 3.2 and 3.4. Given $J = j$ and $K_X = k$ (with $j \leq k$, necessarily), V is a sum of geometric $(\frac{m-i}{m})$ random variables for $j \leq i < k$. Thus,

$$E(V|J, K_X) = \sum_{i=1}^{m-1} \frac{m}{m-i} I\{J \leq i < K_X\}. \tag{3.7}$$

But given $J = j, K_X - J + 1$ is a geometric (p) random variable, by the memoryless property of geometric distributions. Thus, taking expectations with respect to K_X in (3.7) yields

$$\begin{aligned} E(V|J) &= \sum_{i=1}^{m-1} \frac{m}{m-i} I\{J \leq i\} q^{i-J+1} \\ &= \sum_{j=1}^{m-1} I\{J=j\} \sum_{i=j}^{m-1} \frac{m}{m-i} q^{i-j+1}. \end{aligned}$$

Taking expectations on both sides yields Lemma 3.6. □

Lemma 3.8

$$EV = \frac{q}{p} - \frac{q}{p}\sum_{j=1}^{m-1} P\{J=j\}\frac{q^{m-j}}{p} + \sum_{j=i}^{m-1} P\{J=j\}\sum_{i=j}^{m-1}\frac{i}{m-i}q^{i-j+1}.$$

Proof: Straightforward algebra. □

Remark The number of additional "counted" draws needed to complete the last sample after the τ^{th} draw is, except for truncation, a "number of failures" geometric random variable with success probability p and therefore with mean $\frac{q}{p}$. This is where the first term in Lemma 3.8 comes from. The second term reflects the truncation due to the fact that no sample can contain more than m balls. The last term in Lemma 3.8 is the expected number of "uncounted" (i.e., repeat) draws obtained in the course of completing the X^{th} sample after the τ^{th} draw.

When m is large, uncounted draws should be unusual. Assuming that J is seldom very large, the last term in Lemma 3.7 looks like it should be of order m^{-1} for large m, and the second term looks like it might be of order q^m.

Proposition 3.9 As $m \to \infty$ with p fixed,

$$EX = \frac{\sum_{i=0}^{m-1}\frac{1}{m-i} + \frac{p}{1-q^m}\sum_{i=1}^{m-1}\frac{i}{m-i}q^i}{\sum_{i=0}^{m-1}\frac{1}{m-i}q^i} + 0(m^{-2}).$$

Proof

Let N_X be the number of "virgin" balls in the X^{th} sample which did not appear in any previous sample. For large m, N_X should equal 1 except on a set of probability $0(m^{-1})$. The formula in Proposition 3.9 is based on approximating the distribution of J by the distribution of J conditioned on $N_X = 1$.

Lemma 3.10 For $n = 1, 2, \ldots, m$, and $n \le j \le m$,

$$P\{J=j|N_X=n\} = \frac{\binom{j-1}{n-1}q^{j-1}}{\sum_{i=n}^{m}\binom{i-1}{n-1}q^{i-1}}.$$

Proof of Lemma 3.10

Suppose there are n "virgin" balls which have not been seen yet. As far as J is concerned, conditioning on $N_X = n$ here is the same as conditioning on the next sample containing all of the remaining virgin balls.

One way to obtain the next sample is to draw *all* the balls from the box, one by one and without replacement, while flipping a p-coin after each draw. The sample ends with the first heads. The probability that the last virgin ball is drawn on the j^{th} draw here is proportional to $\binom{j-1}{n-1}$. The probability that the first j balls all get into the sample is q^{j-1}. Lemma 3.10 follows. □

Continuation of Proof of Proposition 3.9

For $n = 1$ in Lemma 3.10,

$$P\{J = j | N_X = 1\} = \frac{q^{j-1}}{\sum_{i=1}^{m} q^{i-1}} = \frac{pq^{j-1}}{1-q^m}, 1 \le j \le m. \tag{3.11}$$

Substituting $pq^{j-1}/(1-q^m)$ for $P\{J = j\}$ in Lemma 3.6 and then substituting the result into (3.5) for EV produces the formula in Proposition 3.9. (after cancellation of m's in the numerator and denominator.) To finish the proof, we need to show that the substitution of $P\{J = j | N_X = 1\}$ for $P\{J = j\}$ in Lemma 3.6 (or, equivalently, in Lemma 3.8) causes a change which is $0(m^{-2})$. (Recall that $D_1 \ge 1$, so that the ED_1 in the denominator of (3.5) is greater than 1.)

By using the moment generating function of τ and the Markov inequality, one can show that

$$P\{\tau \ge m^2\} < 2m^{\frac{1}{2}} e^{-m/2}. \tag{3.12}$$

(See Lemma 3.23 of Sellke (1992).) Since $X \le \tau$,

$$P\{X \ge m^2\} < 2m^{\frac{1}{2}} e^{-m/2}, \tag{3.13}$$

and therefore

$$\begin{aligned} P\{K_X > m/2\} &\le P\{\max_{i \le X} K_i > m/2\} \\ &\le m^2 P\{K_1 > m/2\} + P\{X \ge m^2\} \\ &\le m^2 q^{m/2-1} + 2m^{\frac{1}{2}} e^{-m/2}. \end{aligned} \tag{3.14}$$

It follows from (3.14) and $J \le K_X$ that the second term in Lemma 3.8 is $o(m^{-2})$. It is obvious from (3.11) that this second term is still $o(m^{-2})$ if $P\{J = j\}$ is replaced by $P\{J = j | N_X = 1\}$.

Now consider the situation when the second-to-last virgin ball has just been drawn. What is the probability that the last virgin ball is drawn in the same sample as the second-to-last? As long as the size of the current sample is $\le m/2$, the (conditional, given the past) probability that the next counted draw will be the last virgin ball is $\le 2/m$.

Each counted draw (including the one on which the second-to-last virgin ball was drawn) of course has probability p of ending the current sample. It follows that

$$P\{N_X > 1 \quad \text{and} \quad K_X \le \frac{m}{2}\} < \frac{\frac{2}{m}}{p + \frac{2}{m}} < \frac{2}{pm}. \tag{3.15}$$

Combining (3.14) and (3.15) yields

$$P\{N_X > 1\} < \frac{2}{pm} + m^2 q^{\frac{m}{2}-1} + 2m^{\frac{1}{2}} e^{-m/2}. \tag{3.16}$$

A similar argument shows

$$P\{N_X > 3\} < \frac{48}{p^3 m^3} + m^2 q^{m/2-1} + 2m^{\frac{1}{2}} e^{-m/2}. \tag{3.17}$$

Now write the third term of Lemma 3.8 as

$$\sum_{n=1}^{m} P\{N_X = n\} \sum_{j=1}^{m-1} P\{J = j | N_X = n\} \sum_{i=j}^{m-1} \frac{i}{m-i} \; q^{i-j+1} \tag{3.18}$$

By (3.17), the contributions in (3.18) for $n > 3$ are collectively $0(m^{-2})$. By Lemma 3.10 and (3.16) the contributions for $n = 2$ and $n = 3$ are also $0(m^{-2})$. Finally, (3.11) and (3.16) imply that replacing $P\{N_X = 1\}$ by 1 in (3.18) changes the sum by $0(m^{-2})$. Putting all this together establishes Proposition 3.9. □

Remark Sellke (1992) uses a more complicated Markov-chain coupling argument to derive the approximation

$$P\{J = j\} \approx \frac{q^{j-1} \frac{1}{m-j+1}}{\sum_{i=0}^{m-1} \frac{1}{m-i} q^i},$$

which approximates $P\{J = j\}$ much better than does (3.11) above.
The resulting approximation

$$EX \approx \frac{\sum_{k=1}^{m} \frac{1}{k}}{\sum_{i=0}^{m-1} \frac{1}{m-i} q^i} + \frac{\sum_{r=1}^{m-1} \frac{1}{m-r} q^r \sum_{j=1}^{r} \frac{1}{m-j+1}}{[\sum_{i=0}^{m-1} \frac{1}{m-i} q^i]^2}$$

is shown in Sellke (1992) to have an approximation error which converges to 0 exponentially fast (in m).

4. The variance and standard deviation of X

As was the case with EX in Section 3, we will approximate var (X) by exploiting the relationship with the (Z, W) process. The trick here will be to define a new and different sampling scheme for which samples end *either* when a counted flip produces a heads *or* when a "virgin" ball is drawn. (Again, a "virgin" ball is one which has never been drawn before.) With this new sampling scheme, the number $\tilde{X}$ of samples needed to see all the balls is a sum of independent geometric random variables, so the variance of $\tilde{X}$ is the sum of the variances of the summand geometric random variables. It then remains to estimate var(X)− var $(\tilde{X})$.

Let $\{(Z_n, \tilde{W}_n)\}_{n=1}^{\infty}$ be exactly the same as the $\{(Z_n, W_n)\}_{n=1}^{\infty}$ process, except that $\tilde{W}_n$ is set equal to 1 each time the corresponding Z_n is different from all previous Z_i's, $i < n$. (Otherwise, $\tilde{W}_n = W_n$.) In terms of ball draws and coin flips, the story is just as before, except that we don't see the results of coin flips following draws of virgin balls. We pretend that the unseen flips are all heads. Now define a "repeated sampling process" in

terms of the $(Z, \tilde{W})$ process according to the same rules applied in Section 2 to the (Z, W) process. Again draws and flips are "counted" only when the ball drawn does not repeat a previous ball of the current sample, and $\tilde{W}_n = 1$ for a "counted" flip signals the end of the current sample. When there is danger of ambiguity, these samples will be referred as abbreviated samples, since they are abbreviated by the draws of virgin balls. Let $\tilde{X}$ be the number of abbreviated samples needed to see all the balls.

For $k = 0, 1, \ldots, m-1$, let π_k be the probability that the next abbreviated sample contains a virgin ball when exactly k balls have been seen already, (so that there are $m-k$ virgin balls left.)

Lemma 4.1 For $k = 0, 1, \ldots, m-1$,

$$\pi_k = E\frac{m-k}{m-T_k},$$

where T_k is a binomial (k, q) random variable. (As before, $q = 1-p$.)

Proof

Think of the next abbreviated sample as being obtained as follows. For each of the k nonvirgin balls, we flip the p-coin once, *before* drawing any more balls. Each nonvirgin ball becomes a "heads" ball or a "tails" ball, depending on the result of the coin flip corresponding to that ball. Let T_k be the number of tails among the k flips, so that $T_k \sim$ binomial (k, q). Then we draw balls one at a time, without replacement, until we get either a virgin ball or a nonvirgin "heads" ball, either of which ends the current sample. This sampling protocol produces proper abbreviated samples, since it doesn't matter whether the coin flipping is done while drawing the balls or before drawing the balls.

Since these are $m-k$ virgin balls among the $m-T_k$ balls which will terminate the sample, the conditional probability, given T_k, that a virgin ball terminates the sample is $(m-k)/(m-T_k)$. Taking the expectation gives the unconditional probability that the sample contains a virgin ball. □

Lemma 4.2 For $k = 0, 1, \ldots, m-1$ and any $n = 2, 3, \ldots$

$$\pi_k = \frac{m-k}{m-k_q}\{1 + E(\frac{T_k - kq}{m-kq})^2 + \ldots + E(\frac{T_k - kq}{m-kq})^n\}$$
$$+ E\{(\frac{T_k - kq}{m-kq})^{n+1}\frac{m-k}{m-T_k}\}$$

Proof

Note that by algebra

$$\frac{m-k}{m-T_k} = \frac{m-k}{m-kq} \cdot \frac{1}{1 - \frac{T_k - kq}{m-kq}}.$$

Applying

$$\frac{1}{1-x} = 1 + x + \ldots + x^n + \frac{x^{n+1}}{1-x}$$

to the last factor leads to

$$\frac{m-k}{m-T_k} = \frac{m-k}{m-kq}\{1 + \frac{T_k - kq}{m-kq} + \ldots + (\frac{T_k - kq}{m-kq})^n\}$$

$$+(\frac{T_k - kq}{m-kq})^{n+1}\frac{m-k}{m-T_k}.$$

Now take expectations, noting that $E(T_k - kq) = 0$. □

Lemma 4.3 The relative error committed in approximating π_k and also $1 - \pi_k$ using

$$\pi_k \approx \frac{m-k}{m-k_q}\{1 + \frac{kpq}{(m-k_q)^2} + \frac{kpq(q-p)}{(m-k_q)^3} + \frac{3k^2p^2q^2 + kpq^{(1-\sigma pq)}}{(m-k_q)^4}\}$$

is $0(m^{-3})$, uniformly in $0 \leq k < m$ and in p bounded away from 0.

Proof

Take $n = 7$ in Lemma 4.2, and use the formulas for central moments of binomials given in Kendall and Stuart (1969), pages 121-3. □

Recall that τ is the number of single-ball draws (with replacement) needed to see all the balls. Note that $X - 1$ is the number of "counted" heads flips (based on the (Z, W) process) among the first $\tau - 1$ flips. Likewise, $\tilde{X} - 1$ is the number of "counted" heads flips (based on the $(Z, \tilde{W})$ process) among the first $\tau - 1$ flips. Let Δ_1 equal the number of $\tilde{W}_n$'s, $n < \tau$, which equal 1 when W_n is 0, so that

$$\Delta_1 =: \sum_{n=1}^{\tau-1} \tilde{W}_n - W_n.$$

Let $\tilde{\mathcal{G}}$ be the $\sigma-$ field generated by the entire $(Z, \tilde{W})$ process. Obviously, $\tilde{X}$ is $\tilde{\mathcal{G}}$ measurable, while the distribution of Δ_1, given $\tilde{\mathcal{G}}$, is binomial $(m-1, q)$. Thus, $\tilde{X}$ and Δ_1 are independent, and

$$\begin{aligned} \text{var}(\tilde{X} - \Delta_1) &= \text{var}(\Delta_1) + \text{var}(\tilde{X}) \\ &= (m-1)pq + \sum_{k=0}^{m-1} \frac{1-\pi_k}{\pi_k^2}. \end{aligned} \tag{4.4}$$

Now define

$$\Delta_2 =: (\tilde{X} - \Delta_1) - X,$$

so that

$$X = \tilde{X} - \Delta_1 - \Delta_2. \tag{4.5}$$

When the W_n's are replaced by $\tilde{W}_n$'s, Δ_1 tails flips among the first $\tau - 1$ flips are replaced by heads flips. Since the corresponding ball draws all produced virgin balls, these flips were

all "counted" flips in either scheme. *All* "counted" heads in the original scheme (using the (Z, W) process) stay "counted" heads in the new scheme (with the $(Z, \tilde{W})$ process.) However, there may be a few ($= \Delta_2$) "uncounted" heads in the original (Z, W) scheme among the first $\tau - 1$ flips which become "counted" heads in the new $(Z, \tilde{W})$ scheme. (Recall that a coin flip is "uncounted" when the corresponding ball drawn was drawn earlier in the current sample.) When a "virgin tails" draw in the (Z, W) process becomes a "heads" draw in the $(Z, \tilde{W})$ process, subsequent draws in the current, unabbreviated sample which repeated earlier balls of that sample can become non-repeat (and therefore "counted") draws in the $(Z, \tilde{W})$ process.

Now it remains to get a handle on the effect of Δ_2 on the variance of $X = \tilde{X} - \Delta_1 - \Delta_2$.

Heuristic reasoning suggests that the variance of Δ_2 should be bounded, uniformly in m for p bounded away from 0. Indeed, the probability that a repeated draw occurs in the course of generating a particular unabbreviated sample should be $0(m^{-1})$, with multiple repeats having probability $0(m^{-2})$. Since there are m or fewer (unabbreviated) samples which contain virgin balls, the total number of repeated (= uncounted) draws for these samples should be $0_p(1)$. Note that Δ_2 is $\leq$ the number of repeated draws in the samples containing virgin balls. Finally, it seems plausible that Δ_2 should be essentially uncorrelated with $\tilde{X} - \Delta_1$. Thus, one would guess that

$$\begin{aligned} \text{var } X &= \text{ var}(\tilde{X} - \Delta_1) + \text{ var}(\Delta_2) + o(1) \\ &= (m-1)pq + \sum_{k=0}^{m-1} \frac{1-\pi_k}{\pi_k^2} + \text{ var}(\Delta_2) + o(1), \end{aligned} \tag{4.6}$$

so that

$$(m-1)pq + \sum_{k=0}^{m-1} \frac{1-\pi_k}{\pi_k^2} \tag{4.7}$$

should be an underestimate of var X with uniformly bounded error for all m and for p bounded away from 0. We will show that var (Δ_2) is bounded. We will also give an argument showing that the correlation between Δ_2 and $\tilde{X} - \Delta_1$ goes to 0 as $m \to \infty$, but the rate will not be enough to prove (4.6). However, it will be good enough to prove the following result for the standard deviation σ_X of X.

Proposition 4.8 As $m \to \infty$,

$$\sigma_X = \{(m-1)pq + \sum_{k=0}^{m-1} \frac{1-\pi_k}{\pi_k^2}\}^{\frac{1}{2}} + o(1)$$

uniformly for p bounded away from 0.

Remark The error in this approximation for σ_X remains $o(1)$ when the expression in Lemma 4.3 is used to approximate π_k. A more explicit formula for σ_X resulting from application of Lemma 4.3 is given in Proposition 4.18.

Lemma 4.9 For each $\varepsilon > 0$, there exists a constant C_ε so that

$$\text{var}(\Delta_2) \leq C_\varepsilon$$

for all m whenever $p \geq \varepsilon$.

Proof

Let R be the number of repeat (uncounted) draws in all (unabbreviated) samples containing virgin balls. Note again that $\Delta_2 \leq R$. For $1 \leq i \leq m$, let R_i be the number of repeat draws in the i^{th} sample containing a virgin ball, if there is an i^{th} such sample. Set $R_i = 0$ if there are no such sample. Thus, $R = R_1 + \ldots + R_m$. Let $K_{(i)}$ be the size of the i^{th} sample containing a virgin ball, with $K_{(i)} = 0$ if there is no such sample.

Let $\mathcal{G}_i$ be the σ-field generated by everything that happens up to and including the completion of the sample containing the i^{th} virgin ball. Then, given $\mathcal{G}_{i-1}, K_{(i)}$ is stochastically less than or equal to K^*, where

$$P\{K^* = k\} = \begin{cases} (1-q^m)^{-1} k p^2 q^{k-1} & \text{if } 1 \leq k \leq m-1 \\ (1-q^m)^{-1} m p q^{m-1} & \text{if } k = m \\ 0 & \text{else.} \end{cases}$$

Indeed, the K^* distribution is the exact distribution of $K_{(i)}$ (given $\mathcal{G}_{i-1}$) when exactly one virgin ball remains to be drawn. When more than one virgin ball is left, the distribution of $K_{(i)}$ is easily shown to be stochastically smaller.

The conditional distribution of $R_{(i)}$, given $\mathcal{G}_{i-1}$ and $K_{(i)} = k$, is that of a sum of independent geometric $(\frac{m-\ell}{m})$ "number of failures" random variables with with $0 \leq \ell < k$. In symbols,

$$\mathcal{L}(R_{(i)} | \mathcal{G}_{i-1}, K_{(i)} = k) = \sum_{\ell=o}^{k-1} \text{Geom}_F(\frac{m-\ell}{m}).$$

This is because the number of repeat draws made between the ℓ^{th} and $(\ell+1)^{st}$ distinct balls of the sample is $\text{Geom}_F(\frac{m-\ell}{m})$, independent of what came before. Thus,

$$E(R_{(i)} | \mathcal{G}_{i-1}) \leq (1-q^m)^{-1} \sum_{k=1}^{m} k q^{k-1} \sum_{\ell=o}^{k-1} \frac{\ell}{m-\ell}$$

But there exist $\tilde{q}, q < \tilde{q} < 1$, and B (depending on $\varepsilon > 0$) so that $k^7 q^{k-1} < B\tilde{q}^k$ for all $k \geq 0$ and all $q \leq 1 - \varepsilon$. Thus, we easily get the bound (writing $\tilde{p} = 1 - \tilde{q}$)

$$\begin{aligned} E(R_{(i)} | \mathcal{G}_{i-1}) &\leq B(1-\tilde{q}^m)^{-1} \sum_{k=1}^{m} \tilde{q}^k \sum_{\ell=0}^{k-1} \frac{1}{(m-\ell)k^2} \\ &\leq B\tilde{p}^{-1} \sum_{k=1}^{m} \tilde{q}^k \frac{2}{m} \\ &\leq (B\tilde{p}^{-2}) \frac{2}{m}, \end{aligned} \tag{4.10}$$

where the second inequality follows from $(m-\ell)k \geq (m-k+1)k \geq m/2$ *for* $0 \leq \ell < k \leq m$.

From

$$\text{var}\{\text{Geom}_F(\frac{m-\ell}{m})\} = \frac{\ell m}{(m-\ell)^2},$$

we get

$$E(R_{(i)}^2|\mathcal{G}_i, K_{(i)} = k) = \sum_{\ell=o}^{k-1} \frac{\ell m}{(m-\ell)^2} + \{\sum_{\ell=o}^{k-1} \frac{\ell}{m-\ell}\}^2,$$

so

$$\begin{aligned} E(R_{(i)}^2|\mathcal{G}_{(i)}) &\leq (1-q^m)^{-1} \sum_{k=1}^{m} kq^{k-1} \left[\sum_{\ell=o}^{k-1} \frac{\ell m}{(m-\ell)^2} + \{\sum_{\ell=o}^{k-1} \frac{\ell}{m-\ell}\}^2\right] \\ &\leq B\tilde{p}^{-1} \sum_{k=1}^{m} \tilde{q}^k \left[\sum_{\ell=o}^{k-1} \frac{m}{(m-\ell)^2 k^3} + \{\sum_{\ell=o}^{k-1} \frac{1}{(m-\ell)k^2}\}^2\right] \\ &\leq B\tilde{p}^{-2}(\frac{4}{m} + \frac{4}{m^2}) \leq (8B\tilde{p}^{-2})\frac{1}{m}. \end{aligned} \tag{4.11}$$

Bounds (4.10) and (4.11) imply

$$\begin{aligned} E(R^2) &= E\{(R_1 + \ldots + R_m)^2\} \\ &= \sum_{i=1}^{m} E(R_i^2) + 2\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} E\{R_i E(R_j|\mathcal{G}_{j-1})\} \\ &\leq 8B\tilde{p}^{-2} + 4B^2\tilde{p}^{-4}, \end{aligned}$$

which proves Lemma 4.9. □

Lemma 4.12

$$\textit{covariance } (\tilde{X} - \Delta_1, \Delta_2) = o(m), \textit{ uniformly in } p \textit{ for } p > \varepsilon, \varepsilon > 0.$$

Proof of Lemma 4.12

First note from (4.4) and Lemma 4.3 (or see (4.18) below) that

$$\text{var}(\tilde{X} - \Delta_1) = m^2p^2 \sum_{j=1}^{m} \frac{1}{j^2} + 0(m \log m) \tag{4.13}$$

uniformly in p.

The idea here is that most of the variability in Δ_2 comes from what happens early in the sampling process, while most of the variability in $\tilde{X} - \Delta_1$ comes from what happens later in the sampling process. For each $m \geq 3$, let $i^* = i^*(m)$ be the greatest integer less than $m\{1 - (\log m)^{-1}\}$. Again let $\mathcal{G}_i$ be the σ-field generated by the (Z, W) process up

through the completion of the sample containing the i^{th} distinct (=virgin) ball. Then it is easy to show that

$$\frac{E\{\text{var}(\tilde{X} - \Delta_1|\mathcal{G}_{i^*})\}}{\text{var}(\tilde{X} - \Delta_1)} \to 1$$

as $m \to \infty$. This in turn implies

$$\text{var}\{E(\tilde{X} - \Delta_1|\mathcal{G}_{i^*})\} = o\{\text{var}(\tilde{X} - \Delta_1)\} \tag{4.14}$$

since

$$\text{var}(\tilde{X} - \Delta_1) = \text{var}\{E(\tilde{X} - \Delta_1|\mathcal{G}_{i^*})\} + E\{\text{var}(\tilde{X} - \Delta_1|\mathcal{G}_{i^*})\}.$$

On the other hand,

$$E\{\text{var}(\Delta_2|\mathcal{G}_{i^*})\} = o\{\text{var}(\Delta_2)\}, \tag{4.15}$$

since given $\mathcal{G}_{i*}$, there are only about $m/\log m$ remaining samples which will contain virgin balls.

Now write

$$(\tilde{X}-\Delta_1)-E(\tilde{X}-\Delta_1) = \{E(\tilde{X}-\Delta_1|\mathcal{G}_{i^*})-E(\tilde{X}-\Delta_1)\}+\{(\tilde{X}-\Delta_1)-E(\tilde{X}-\Delta_1|\mathcal{G}_{i^*})\} \tag{4.16}$$

and

$$\Delta_2 - E(\Delta_2) = \{E(\Delta_2|\mathcal{G}_{i^*}) - E(\Delta_2)\} + \{\Delta_2 - E(\Delta_2|\mathcal{G}_{i^*})\}. \tag{4.17}$$

By (4.13), (4.14), (4.15), Lemma 4.9, and the Cauchy-Schwarz inequality, the expectation of the product of (4.16) and (4.17) is $o(m)$, uniformly in $p > \varepsilon > 0$, which proves the lemma. □

Proof of Proposition 4.8

By Lemma 4.9 and Lemma 4.12,

$$\begin{aligned} \text{var } X &= \text{var}(\tilde{X} - \Delta_1 - \Delta_2) \\ &= \text{var}(\tilde{X} - \Delta_1) + o(m). \end{aligned}$$

By (4.13),

$$\sigma_X = \sigma_{\tilde{X}-\Delta_1} + o(1),$$

which together with (4.4) proves Proposition 4.8. □

Proposition 4.18

$$\sigma_X = \{(m^2p^2 - 2mpq)\sum_{j=1}^{m}\frac{1}{j^2} + mp(q-p)\sum_{j=1}^{m}\frac{1}{j}\}^{\frac{1}{2}} + o(1)$$

as $m \to \infty$, uniformly for $p > \varepsilon > 0$.

Proof

Replacing π_k in (4.4) by

$$\pi_k \approx \frac{m-k}{m-kq}\{1+\frac{kpq}{(m-kq)^2}\}$$

and using Lemma 4.3 implies

$$\text{(4.19)} \qquad \text{var}(\tilde{X}-\Delta_1) = (m^2p^2-2mpq)\sum_{j=1}^{m}\frac{1}{j^2} + mp(q-p)\sum_{j=i}^{m}\frac{1}{j} + 0(\log m)$$

after some calculation. The proposition now follows from Lemmas 4.9 and 4.12 as before in Proposition 4.8. □

Remark It seems likely that the covariance of $X-\Delta_1$ and Δ_2 is uniformly bounded or perhaps even $o(1)$ as $m \to \infty$ for $p > \varepsilon > 0$. Thus, from (4.19) and Lemma 4.9 we conjecture that

$$\text{(4.20)} \qquad \text{var}(X) = (m^2p^2-2mpq)\sum_{j=1}^{m}\frac{1}{j^2} + mp(q-p)\sum_{j=1}^{m}\frac{1}{j} + 0(\log m)$$

uniformly for $p > \varepsilon > 0$.

5. Numerical results for $p = \frac{1}{2}$

Here is how the exact values for EX (with $p = \frac{1}{2}$) compare with the approximation in Proposition 3.9 for several values of m. (The exact values here and in the next table were calculated from Larry Shepp's recursive formulas.)

Table 1

m	E(X)	E(X)-(3.9)
10	13.3812	0.05997
20	34.5149	0.00649
50	110.5647	0.000524
100	257.2316	0.000113
150	417.0115	0.000048
200	585.3392	0.000026
250	760.0133	0.000017
300	939.7404	0.000012

When $p = \frac{1}{2}$, the approximation for σ_X in Proposition 4.18 equals

$$\text{(5.1)} \qquad \frac{\pi}{\sqrt{24}}(m-1) - \frac{\sqrt{6}}{4\pi}$$

to within $0(m^{-1})$.

The following table shows how the exact values for σ_X (for $p = \frac{1}{2}$) compare with (5.1).

Table 2

m	σ_X	σ_X-(5.1)
10	5.8086	0.2320
20	12.1008	0.1115
50	31.2789	0.0514
100	63.3212	0.0299
150	95.3768	0.0217
200	127.4361	0.0173
250	159.4970	0.0145
300	191.5588	0.0125

APPENDIX Shepp's recursive formulas.

Let e_j^m be defined as the expected number of additional samples needed to paint all the balls red when there are m balls in the urn, of which j remain white. Let N° be the size of the next sample and Y° be the number of white balls in the next sample. Conditioned on $N^\circ = n, Y^\circ$ is hypergeometric with

$$P(Y^\circ = k | N^\circ = n) = \frac{\binom{j}{k}\binom{m-j}{n-k}}{\binom{m}{n}}$$

for $0 \vee (n-(m-j)) \leq k \leq j \wedge n$. We observe $e_\circ^m = 0$ and

$$\begin{aligned} e_j^m &= 1 + \sum_{n=1}^{m} \sum_{k=0\vee[n+j-m]}^{j\wedge n} P(N^\circ = n) P(Y^\circ = k | N^\circ = n) e_{j-k}^m \\ &= 1 + \sum_{n=1}^{m-1} \sum_{k=0\vee[n+j-m]}^{j\wedge n} 2^{-n} \frac{\binom{j}{k}\binom{m-j}{n-k}}{\binom{m}{n}} e_{j-k}^m. \end{aligned}$$

Therefore, gathering e_j^m terms and dividing,

$$e_j^m = \frac{1 + \sum_{n=1}^{m-1} \sum_{k=1\vee[n+j-m]}^{j\wedge n} 2^{-n} \frac{\binom{j}{k}\binom{m-j}{n-k}}{\binom{m}{n}} e_{j-k}^m}{1 - \sum_{n=1}^{m-j} 2^{-n} \frac{\binom{m-j}{n}}{\binom{m}{n}}}.$$

Likewise, if X_j^m is the random number of additional samples needed to paint all m balls red if there are j remaining white balls, we can get a recursive formula for $f_j^m \triangleq E(X_j^m)^2$

in terms of f_j^m, $i = 0, 1, \ldots, j-1$ and e_j^m. Again, $f_\circ^m = 0$. Also, for $j \geq 1$,

$$\begin{aligned} f_j^m &= E[(1 + X_{j-Y^\circ})^2] \\ &= 1 + 2E(X_{j-Y^\circ}) + E(X_{j-Y^\circ}^2) \\ &= 2e_j^m - 1 + E(X_{j-Y^\circ}^2) \\ &= 2e_j^m - 1 + \sum_{n=1}^{m} \sum_{k=0 \vee [n+j-m]}^{j \wedge n} P(N^\circ = n) P(Y^\circ = k | N^\circ = n) f_{j-k}^m \\ &= 2e_j^m - 1 + \sum_{n=1}^{m-1} \sum_{k=0 \vee [n+j-m]}^{j \wedge n} 2^{-n} \frac{\binom{j}{k}\binom{m-j}{n-k}}{\binom{m}{n}} f_{j-k}^m. \end{aligned}$$

Thus,

$$f_j^m = \frac{2e_j^m - 1 + \sum_{n=1}^{m-1} 2^{-n} \sum_{k=1 \vee (n+j-m)}^{j \wedge n} \frac{\binom{j}{k}\binom{m-j}{n-k}}{\binom{m}{n}} f_{j-k}^m}{1 - \sum_{n=1}^{m-j} 2^{-n} \frac{\binom{m-j}{n}}{\binom{m}{n}}}$$

Clearly, $EX = e_m^m$; $\sigma_X = \sqrt{f_m^m - (e_m^m)^2}$. Using these equations, we can obtain tables for EX and σ_X.

ACKNOWLEDGEMENTS

The authors would like to thank M. Jacobsen for his helpful comments, and M.A.Martin for his computer assistance. We also appreciate the careful work of the referee.

Reference

Kendall, M. G., and Stuart, A. (1969), *The Advanced Theory of Statistics*, volume 1, 3rd ed., Hafner, New York.

Sellke, T.M. (1992), How many *iid* samples does it take to see all the balls in a box? Purdue University Technical Report 92-47.