The Classical Occupancy Distribution: Computation and Approximation

Ben O'Neill

To cite this article: Ben O'Neill (2020): The Classical Occupancy Distribution: Computation and Approximation, The American Statistician, DOI: [10.1080/00031305.2019.1699445](https://doi.org/10.1080/00031305.2019.1699445)

To link to this article: <https://doi.org/10.1080/00031305.2019.1699445>

Accepted author version posted online: 19
Dec 2019.
Published online: 06 Jan 2020.

Submit your article to this journal [↗](#)

Article views: 73

View related articles [↗](#)

View Crossmark data [↗](#)

The Classical Occupancy Distribution: Computation and Approximation

Ben O'Neill

Research School of Population Health, Australian National University, Canberra, Australia

ABSTRACT

We examine the discrete distributional form that arises from the “classical occupancy problem,” which looks at the behavior of the number of occupied bins when we allocate a given number of balls uniformly at random to a given number of bins. We review the mass function and moments of the classical occupancy distribution and derive exact and asymptotic results for the mean, variance, skewness and kurtosis. We develop an algorithm to compute a cubic array of log-probabilities from the classical occupancy distribution. This algorithm allows the computation of large blocks of values while avoiding underflow problems in computation. Using this algorithm, we compute the classical occupancy distribution for a large block of values of balls and bins, and we measure the accuracy of its asymptotic approximation using the normal distribution. We analyze the accuracy of the normal approximation with respect to the variance, skewness and kurtosis of the distribution. Based on this analysis, we give some practical guidance on the feasibility of computing large blocks of values from the occupancy distribution, and when approximation is required.

ARTICLE HISTORY

Received August 2019
Accepted November 2019

KEYWORDS

Approximation; Birthday problem; Classical occupancy distribution; Computation; Coupon collector problem; Distribution and moments

1. Introduction

The “classical occupancy problem” (also called the “classical shot problem”) is an important problem in urn models in probability. The problem is straight-forward: find the marginal and conditional distributions of the “occupancy number”—the number of bins occupied when we allocate a given number of balls at random to a given number of bins. This problem is closely related to the “birthday problem,” which looks at the probability that the occupancy is less than the total number of balls (i.e., that at least two balls share a bin), and the “coupon-collector problem,” which looks at the number of randomly allocated balls needed to occupy all the bins. The occupancy problem is fully specified by the specified number of bins and balls, so it leads to a discrete distributional family for the occupancy number, indexed by these two parameters. Specification of the marginal and conditional distributions of the occupancy number solves the occupancy problem for all possible numbers of bins and balls, and also gives related solutions to the birthday problem and coupon collector’s problem.

The classical occupancy problem can be traced back to de Moivre’s treatise *De Mensura Sortis* in 1712, so it has an antique genesis in the history of probability.¹ The problem is examined in Feller (1950) to yield a known distributional form for the

number of empty bins, which gives a trivial correspondence to the distribution of the number of occupied bins. Detailed analysis of the distribution, including asymptotic approximations, is given in Johnson and Kotz (1977) and Kolchin, Sevast’yanov, and Chistyakov (1978), and Williamson et al. (2009) gives a recent summary of the problem for educational purposes, including noting its applications to the technique of bootstrapping.² Other areas of practical application include estimating counts of unseen species in animal samples (Good and Toulmin 1956; Bunge and Fitzpatrick 1993), the number of unused words in sample of vocabulary (McNeil 1973; Efron and Thisted 1978), or the number of times that lotteries yield multiple prizes to a single winner (Hanley 1992).³

Notwithstanding a wide variety of statistical applications, and its core role in several important probability problems, the distributional form arising from the classical occupancy problem has not received widespread attention. In particular, this family of distributions is not included among the discrete distributions taught in standard statistical courses, and generally does not appear in statistical textbooks that include other common discrete distributions. Unlike other common families of discrete probability distributions, at the time of writing, it does not have its own Wikipedia page and does not appear in

CONTACT Ben O'Neill ben.oneill@hotmail.com, ben.oneill@anu.edu.au Research School of Population Health, Australian National University, Canberra, ACT 0200, Australia.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.

¹First published in English as De Moivre (1718), Problems XXIX–XXX, pp. 73–77. These problems are slightly different to the classical occupancy problem as it is presently conceived. Nevertheless, both Hald, de Moivre, and McClintock (1984, p. 232) and Holst (1986, p. 16) give De Moivre credit for the genesis of the occupancy problem on the basis of these two problems. [Note that there are different editions of De Moivre’s work which label the problems with different numbers. Correspondence of the problems is shown in Table 1 of Hald, de Moivre, and McClintock (1984, p. 234).]

²Other analysis of the distribution can be found in Arfwedson (1951), Weiss (1958), Harris (1968), Park (1972), Samuel-Cahn (1974), Chao (1984), and Holst (1986).

³This example is particularly interesting; it involves a strange result that occurred in the Quebec “Super Loto” on July 25, 1982. The lottery allocated 500 prize cars by random sampling with replacement to 2.4 million unclaimed lottery tickets. One participant was lucky enough to win two cars from his single ticket! Hanley notes that this outcome is less improbable than might be imagined.

lists of discrete distributions on that platform.⁴ The result of this obscurity is that the distribution remains largely unknown, even amongst applied statisticians. Even among articles in the statistical literature that look at the properties of the distribution, there is variation in examining the distribution of the number of *occupied* bins or the number of *unoccupied* bins (which is essentially the same problem since it merely reverses the values in the probability mass function). The resulting distribution is given a wide variety of names and is sometimes left unnamed.⁵

In the classical occupancy problem, each of the balls is allocated to one of a finite number of bins via a discrete uniform distribution. This problem has been extended by considering the general case where there is a probability vector taken over a finite or infinite number of bins (see, e.g., Gneden, Hansen, and Pitman 2007). In the present article, we look only at the classical case, but we will expound the foundational aspects of this case in a way that can be easily generalized.

2. The Classical Occupancy Distribution

To facilitate easy generalization to other cases, we frame our analysis of the classical occupancy problem within the broader case where sampling of the balls can occur according to any distribution over a finite number of bins. (In the classical problem the sampling is done using a discrete uniform distribution.)

We consider the case where we have $m \in \mathbb{N}$ bins and we allocate $n \in \mathbb{N}$ balls to these bins with probability vector $\mathbf{p} \equiv (p_1, \dots, p_m)$.⁶ We study the *occupancy number* $1 \leq K \leq \min(n, m)$, which is the number of bins occupied by the balls after random allocation. We can represent this mathematically in terms of an underlying series $X_1, X_2, X_3, \dots \sim \text{iid categorical}(\mathbf{p})$ that gives random allocation of an infinite sequence of balls. For any number of balls n we define the bin counts over $k = 1, \dots, m$ and occupancy number, respectively, by

$$Y_k \equiv Y_k(n, \mathbf{p}) \equiv \sum_{i=1}^n \mathbb{I}(X_i = k)$$

$$K \equiv K(n, \mathbf{p}) \equiv \sum_{k=1}^m \mathbb{I}(Y_k(n, \mathbf{p}) > 0).$$

⁴List of distributions at https://en.wikipedia.org/wiki/List_of_probability_distributions. At the time of writing there is no page for the occupancy distribution, but there is a page for the coupon collector's problem. [Search undertaken on January 13, 2019.]

⁵The name "classical occupancy distribution" was used in Johnson and Kotz (1977, p. 110). In an earlier book (Johnson and Kotz 1969, p. 251) they noted that it is sometimes called "Arfwedson's distribution" after Arfwedson (1951). Williamson et al. (2009) proposed to call the distribution the "Stirling2" distribution, due to the presence of the Stirling numbers of the second kind. The present author is of the view that James Stirling has already received sufficient credit for the Stirling numbers, and since he made no specific contribution to the analysis of the occupancy problem, it is preferable to use the more impersonal terminology used by Johnson and Kotz. This name is already established in some of the literature, and has the advantage of linking the distribution directly to the problem from which it is derived. It also avoids adding another case of Stigler's "law of eponymy": that no scientific discovery is named after its original discoverer (see Stigler 1980).

⁶The probability vector $\mathbf{p}$ determines the number of bins m through its length, and so there is no need to specify the latter as a parameter when the former is already included. For this reason our later distribution notation will not include explicit dependence on m , but this is implicit, through the fact that there is dependence on $\mathbf{p}$.

In the classical occupancy problem, the balls are allocated to the bins using the discrete uniform distribution, so the probability vector $\mathbf{p}$ is $\mathbf{u} \equiv (\frac{1}{m}, \dots, \frac{1}{m})$. In this case we denote the occupancy number as $K(n, m)$ instead of $K(n, \mathbf{p})$.

Before looking further at the occupancy distribution, we briefly note the connection of the occupancy problem to other famous probability problems. In the "birthday problem" we are interested in the event that $K(n, \mathbf{p}) < n$, which means that at least two of the balls share a bin. In the "coupon-collector problem" we are interested in the number of balls needed to get full occupancy, which is $T \equiv T(\mathbf{p}) \equiv \min\{n \in \mathbb{N} | K(n, \mathbf{p}) = m\}$. The latter is closely related to the occupancy problem and its solution is closely related to the occupancy distribution.

In the case of uniform sampling (i.e., the classical occupancy problem), the probability mass function for the occupancy value $K \equiv K(n, m)$ is well-known and can be obtained by recourse to a simple combinatorial argument. The event $K = k$ means that there are k bins occupied by the balls after n balls have been allocated randomly to m bins. We split the possible outcomes in the event into a choice of bins, and then an arrangement of balls into those bins. The resulting number of ways each step can be done is as follows:

Choose an ordered set of k bins from m bins $(m)_k$,

Choose a partition of n balls into k nonempty bins $S(n, k)$.

(The values $(m)_k = \prod_{i=0}^{k-1} (m - i)$ are the falling factorials, which count permutations, and the values $S(n, k)$ are the Stirling numbers of the second kind, which count the number of ways to partition n distinct objects into k nonempty sets.) Under uniform allocation of balls there are m^n equiprobable outcomes for the n balls, so if we apply the multiplication rule of counting to this procedure, we obtain the occupancy distribution $K \sim \text{Occ}(n, m)$ defined below.

Definition (The classical occupancy distribution). This family of discrete distributions is defined by the probability mass function:

$$\text{Occ}(k|n, m) = \frac{(m)_k \cdot S(n, k)}{m^n}$$

for all integers $1 \leq k \leq \min(n, m)$,

where $m \in \mathbb{N}$ is the *space parameter* (bins) and $n \in \mathbb{N}$ is the *size parameter* (balls).

Remark. Rather than using explicit reference to the Stirling numbers of the second kind, this distribution is often written in terms of the k th forward difference Δ^k of the monomial function y^n , which can be written as⁷

$$(\Delta^k y^n)(x) = \sum_{i=0}^k \binom{k}{i} (-1)^i (x + k - i)^n.$$

⁷Notation for this function is awkward, since the forward difference operator operates on the monomial function y^n to yield a new *function*, and the latter is then evaluated at a particular argument value x . We have chosen to use different variable symbols for these two parts to give a clear distinction between the monomial function and the argument used for evaluation of the k th forward difference Δ^k of the monomial function. In our analysis there will be occasions where we will alter the monomial function in this expression, and occasions where we evaluate the result at a particular

The Stirling numbers of the second kind can be written explicitly as $S(n, k) = (\Delta^k y^n)(0)/k!$ and we also have $(\Delta^k y^n)(x) = 0$ for $k > n$ (Ruiz 1996). Thus, the classical occupancy distribution can be written (without any further range restriction) as

$$\begin{aligned} \text{Occ}(k|n, m) &= \frac{1}{m^n} \binom{m}{k} (\Delta^k y^n)(0) \\ &= \frac{1}{m^n} \binom{m}{k} \sum_{i=0}^k \binom{k}{i} (-1)^i (k-i)^n. \end{aligned}$$

We note that this representation is the most common explicit form used for the Stirling numbers of the second kind, and is a common formula for their evaluation.

3. Moments and Approximation of the Classical Occupancy Distribution

The probability mass function for the classical occupancy distribution is tricky to work with, since it contains the Stirling numbers of the second kind. It is nonetheless possible to obtain expressions for the expected value of any polynomial in K (and thus all the moments of the distribution). To facilitate analysis, we define terms $E_r \equiv (1 - r/m)^n$ for all $r = 0, 1, 2, \dots$ and we note that we have the asymptotic equivalence $E_r \sim e^{-rn/m}$ as $n \rightarrow \infty$, which allows us to easily combine products of these terms with different indexes. The simplest way to derive the moments is via a simple formula for the expected value of the following falling factorial (proofs of theorems are in the Appendix).

Theorem 1A (Expected values of falling factorials). For integers $0 \leq r < m$ we have

$$\mathbb{E}((m-K)_r) = (m)_r E_r \quad E_r \equiv (1 - r/m)^n.$$

The falling factorials are polynomials of degree r in K , so we can use their expected values to find the moments of K according to Theorem 1B below. There are many other articles in the literature that give the mean and variance of the distribution (see, e.g., Johnson and Kotz 1969, p. 256; Park 1972, p. 1700; Johnson and Kotz 1977, sec. 6.25; Charalambides 2005, pp. 170–172; Williamson et al. 2009, p. 358; Williamson 2012, p. 1503), but we are unaware of any other article that gives formulas for its skewness and kurtosis (shown in our Theorem 2B).

Theorem 1B (Raw moments). For all powers $p \in \mathbb{N}$ we have raw moments:

$$\mathbb{E}(K^p) = \sum_{r=0}^p (-1)^r a_{p,r}(m) (m)_r E_r,$$

where each $a_{p,r}$ is a polynomial of degree $p - r$ with coefficients that are integer values (see Appendix for further details on these values).

argument value x . The notation we use allows us to clearly distinguish those two things. Our notation contrasts with the notation used in Johnson and Kotz (1977) where they write our $(\Delta^k y^n)(0)$ in a shorthand as $\Delta^k 0^n$ (the k th difference of the monomial evaluated at zero), and use similar shorthand when they make changes in the monomial function. We have chosen to be more explicit to prevent any confusion.

Theorem 2A (Mean and variance). The mean and variance of the distribution are

$$\begin{aligned} \mu_{n,m} &\equiv \mathbb{E}(K) = m(1 - E_1), \\ \sigma_{n,m}^2 &\equiv \mathbb{V}(K) = m[(m-1)E_2 + E_1 - mE_1^2]. \end{aligned}$$

As $n \rightarrow \infty$ we have $E_r \sim e^{-rn/m}$ which gives asymptotic mean and variance:

$$\begin{aligned} \mu_{n,m} &\sim \mu_{n,m}^* \equiv m(1 - e^{-n/m}), \\ \sigma_{n,m}^2 &\sim \sigma_{n,m}^{*2} \equiv me^{-n/m}(1 - e^{-n/m}). \end{aligned}$$

Theorem 2B (Skewness and kurtosis). The skewness and kurtosis of the distribution are⁸

$$\begin{aligned} \gamma_{n,m} &\equiv \text{Skew}(K) \\ &= -\frac{E_1 - 3E_2 + 2E_3 + m \left(\frac{3(E_2 - E_1^2)}{+2m(E_1^3 - E_1E_2)} + (m-3)(E_3 - E_1E_2) \right)}{m^{1/2} [(m-1)E_2 + E_1 - mE_1^2]^{3/2}}, \\ \kappa_{n,m} &\equiv \text{Kurt}(K) \\ &= \frac{\left(E_1 - 4mE_1^2 + 6m^2E_1^3 - 3m^3E_1^4 \right. \\ &\quad \left. + 7(m-1)E_2 + 6(m-1)(m-2)E_3 \right. \\ &\quad \left. + (m-1)(m-2)(m-3)E_4 \right. \\ &\quad \left. - 12m(m-1)E_1E_2 + 6m^2(m-1)E_1^2E_2 \right. \\ &\quad \left. - 4m(m-1)(m-2)E_1E_3 \right)}{m[(m-1)E_2 + E_1 - mE_1^2]^2}. \end{aligned}$$

As $n \rightarrow \infty$ we have $E_r \sim e^{-rn/m}$ which gives asymptotic skewness and kurtosis:

$$\begin{aligned} \gamma_{n,m} &\sim \gamma_{n,m}^* \equiv -\frac{1}{\sqrt{m}} \cdot \frac{1 - 2e^{-n/m}}{\sqrt{e^{-n/m}(1 - e^{-n/m})}}, \\ \kappa_{n,m} &\sim \kappa_{n,m}^* \equiv 3 + \frac{1}{m} \cdot \frac{1 - 6e^{-n/m}(1 - e^{-n/m})}{e^{-n/m}(1 - e^{-n/m})}. \end{aligned}$$

If $n \rightarrow \infty$ and $m \rightarrow \infty$ in a way that yields a fixed limit $n/m \rightarrow \lambda \in (0, \infty)$ then we have $\gamma_{n,m} \rightarrow 0$ and $\kappa_{n,m} \rightarrow 3$ (so the distribution is asymptotically unskewed and mesokurtic).

There is a broad literature on asymptotic results pertaining to the occupancy distribution, but in this article, we will be interested in its approximation by the normal distribution. (There are also several results showing that the distribution of the number of empty cells converges to a Poisson distribution in certain limiting cases.) The simplest and most powerful theorem of this kind is a uniform convergence result in Hwang and Janson (2008, Theorems 2.1 and 2.4) which we reproduce here in a slightly modified form. This theorem shows that the occupancy distribution can be approximated well by the normal distribution so long as $\sigma_{n,m}^2 \rightarrow \infty$.

⁸Our skewness and kurtosis are $\text{Skew}(K) \equiv \mathbb{E}(Z^3)$ and $\text{Kurt}(K) \equiv \mathbb{E}(Z^4)$ where $Z \equiv (K - \mu_{n,m})/\sigma_{n,m}$ (i.e., we refer to the third and fourth standardized moments).

Theorem 3A (Bounds on approximation). This theorem is adapted from Hwang and Janson (2008). There exists a constant $0 < c < \infty$ such that, for all $n > 1$ and $m > 1$ we have

$$\sup_k |\text{Occ}(k|n, m) - N(k|\mu_{n,m}, \sigma_{n,m}^2)| \leq \frac{c}{\sigma_{n,m}^2},$$

$$\sup_k |\text{Occ}(k|n, m) - N(k|\mu_{n,m}^*, \sigma_{n,m}^{*2})| \leq \frac{c}{\sigma_{n,m}^{*2}}.$$

Theorem 3B (Uniform convergence). If $\sigma_{n,m}^2 \rightarrow \infty$ (or $\sigma_{n,m}^{*2} \rightarrow \infty$, respectively) we have the respective limits:

$$\sup_k |\text{Occ}(k|n, m) - N(k|\mu_{n,m}, \sigma_{n,m}^2)| \rightarrow 0,$$

$$\sup_k |\text{Occ}(k|n, m) - N(k|\mu_{n,m}^*, \sigma_{n,m}^{*2})| \rightarrow 0.$$

Theorem 3B establishes uniform convergence between the classical occupancy distribution and the normal distribution when $\sigma_{n,m}^2 \rightarrow \infty$. (Our result here is only for the *classical* occupancy distribution, but we note that Hwang and Janson (2008) establish this uniform convergence result for the more general case where the balls can be allocated by nonuniform sampling.) Since the normal distribution is continuous with support over the real numbers, asymptotic approximation of the classical occupancy distribution is usually done by evaluating the normal density at discrete points, and then scaling to ensure that the probability vector sums exactly to one. If one prefers to use an alternative asymptotic distribution, it is easy to use the triangle inequality to show uniform convergence to any other distribution that has uniform convergence to the normal.

Remark. It is useful to consider parameter sequences where we have $\max(n, m) \rightarrow \infty$ and $n/m \rightarrow \lambda \in [0, \infty]$ (i.e., where at least one parameter value approaches infinity, and their ratio approaches a fixed limit). Within this framework, the condition $0 < \lambda < \infty$ is a necessary and sufficient condition for $\sigma_{n,m}^2 \rightarrow \infty$. Contrarily, if $\lambda = 0$ then we have $\mu_{n,m} \rightarrow n$ and $\sigma_{n,m}^2 \rightarrow 0$, and if $\lambda = \infty$ we have $\mu_{n,m} \rightarrow m$ and $\sigma_{n,m}^2 \rightarrow 0$. In these two cases the occupancy distribution approaches a point-mass distribution. In the first case, when the number of bins is infinite, each ball is allocated to a different bin with probability one, so the occupancy will be equal to the number of balls. In the second case, an infinite number of balls allocated over a finite number of bins will occupy all bins with probability one.

4. Computing the Classical Occupancy Distribution

Although the probability mass function for the occupancy distribution has an apparently simple form, as soon as the values of n or m become even moderately large it is difficult to compute the value of the mass function directly from the formula for the mass function. There are two difficulties here. The first difficulty is that the explicit form of the Stirling numbers is both computationally intensive, and involves canceling of large terms. The second problem is that, even once the Stirling numbers are computed, the mass function is a ratio of large integers, which leads to arithmetic overflow problems, and the values in the tail of the distribution become so small that they lead to arithmetic underflow problems.

To avoid underflow problems in computation, we will represent the occupancy distribution by its log-probabilities (i.e., the logarithms of the values in the mass function). To calculate these values we need to compute the relevant Stirling numbers, or some simple variation of these. We will do this by defining the *reduced occupancy distribution* as an instance of the classical occupancy distribution where we have an identical number of balls and bins, which reduces it to the one-parameter distribution:

$$\text{Occ}(k|n) \equiv \text{Occ}(k|n, n) = \frac{(n)_k \cdot S(n, k)}{n^n}$$

for all integers $1 \leq k \leq n$.

We can write the mass values in the classical occupancy distribution in terms of the mass values in the reduced occupancy distribution as follows:

$$\begin{aligned} \text{Occ}(k|n, m) &= \text{Occ}(k|n) \cdot \frac{n^n}{m^n} \cdot \frac{(m)_k}{(n)_k} \\ &= \text{Occ}(k|n) \cdot \left(\frac{n}{m}\right)^n \cdot \prod_{i=0}^{k-1} \left(\frac{n-i}{m-i}\right). \end{aligned}$$

This gives the corresponding log-probability equation:

$$\begin{aligned} \log \text{Occ}(k|n, m) &= \log \text{Occ}(k|n) + n [\log(n) - \log(m)] \\ &\quad - \sum_{i=1}^{k-1} [\log(n-i) - \log(m-i)]. \end{aligned}$$

This simple log-probability correspondence between the classical occupancy distribution and the reduced occupancy distribution allows us to create an algorithm for the former values that hinges on values of the latter. Since the reduced occupancy distribution has a single argument value and a single parameter, we can store its log-probabilities up to some given maximum parameter value in a matrix (as opposed to a cubic array) and this will make it feasible to store a large number of log-probabilities. We can use these to rapidly compute the log-probabilities in the classical occupancy distribution.

In accordance with this line of reasoning, our algorithm for the classical occupancy distribution will start with of an algorithm to precompute a large matrix of log-probability values from the reduced occupancy distribution. For brevity, we now denote the probability mass values from the two distributions as $p_{k,n,m} \equiv \text{Occ}(k|n, m)$ and $p_{k,n} \equiv \text{Occ}(k|n)$. For a large value $N \in \mathbb{N}$, we will compute the matrix of reduced occupancy log-probabilities:

$$L_N = \begin{bmatrix} \log p_{1,1} & \log p_{1,2} & \log p_{1,3} & \cdots & \log p_{1,N} \\ -\infty & \log p_{2,2} & \log p_{2,3} & \cdots & \log p_{2,N} \\ -\infty & -\infty & \log p_{3,3} & \cdots & \log p_{3,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\infty & -\infty & -\infty & \cdots & \log p_{N,N} \end{bmatrix}.$$

In **Algorithm 1**, this matrix of values is computed recursively using the characterization in **Theorem 4**. We compute the first row of the matrix using the first equation, then compute the main diagonal using the second equation, and then compute each remaining upper diagonal—going outward from the main diagonal—using the third equation.

Algorithm 1 Reduced occupancy matrix**Input:** A positive integer N **Output:** An $N \times N$ matrix of reduced occupancy log-probabilities

```

#Create output matrix
ROCC <- Matrix with dimensions  $N \times N$ , with initial values all set to  $-\infty$ 

#Compute the first row and main diagonal
for each  $n = 1, \dots, N$  {
  ROCC[1,  $n$ ] <-  $-(n-1) \log(n)$  }
for each  $n = 2, \dots, N$  {
  ROCC[ $n, n$ ] <-  $(n-1) \log(1-1/n) + \text{ROCC}[n-1, n-1]$  }

#Compute the other diagonals
for each  $i = 2, \dots, N-1$  {
  for each  $k = 2, \dots, N-i+1$  {
     $n \leftarrow k+i-1$ 
    LOGTERM1 <-  $\log(k) - \log(n-k) + \text{ROCC}[k, n-1]$ 
    LOGTERM2 <-  $\text{ROCC}[k-1, n-1]$ 
    ROCC[ $k, n$ ] <-  $(n-1) \log(1-1/n) + \text{logsumexp}(\text{LOGTERM1}, \text{LOGTERM2})$  } }
ROCC <- ROCC -- logsumexp(ROCC)

return ROCC
#ROCC[ $k, n$ ] =  $\log(\text{Occ}(k|n))$ 

```

Theorem 4 (Recursive characterization). The reduced occupancy distribution satisfies the recursive equations⁹:

$$\begin{aligned}
 \log p_{1,n} &= -(n-1) \log(n), \\
 \log p_{n,n} &= (n-1) \log(1-1/n) + \log p_{n-1,n-1}, \\
 \log p_{k,n} &= (n-1) \log(1-1/n) \\
 &\quad + \text{logsumexp} \left(\begin{array}{c} \log(k) - \log(n-k) + \log p_{k,n-1}, \\ \log p_{k-1,n-1} \end{array} \right).
 \end{aligned}$$

The classical occupancy distribution can be written in terms of these values as

$$\begin{aligned}
 \log p_{k,n,m} &= \log p_{k,n} + n [\log(n) - \log(m)] \\
 &\quad - \sum_{i=0}^{k-1} [\log(n-i) - \log(m-i)].
 \end{aligned}$$

Below we construct algorithms to compute the classical occupancy distribution. [Algorithm 1](#) computes the matrix L_N using the recursive equations in [Theorem 4](#) and [Algorithm 2](#) uses the last equation to convert these values to log-probabilities of the classical occupancy distribution. The log-probabilities generated by these algorithms are not exact (due to rounding of the $\log$ and logsumexp functions) but they are extremely close to the exact values represented using arbitrary precision arithmetic. These algorithms avoid problems with arithmetic overflow and underflow by using log-probabilities and using operations directly on these logarithms.

By precomputing the matrix L_N with [Algorithm 1](#), we allow rapid computation of the classical occupancy distribution in [Algorithm 2](#). The drawback is that the precomputed object takes up storage space on the computer. It is feasible to generate this matrix for quite large values of N (e.g., with $N = 25,000$ the matrix can be computed and stored on an ordinary computer, though the file is quite large).¹⁰ The algorithm we have developed is limited to computing the classical occupancy distribution for $n \leq N$, so it is impractical when the size parameter is so large as to make computation of the reduced occupancy distribution prohibitive.

Our goal in this article is to develop a general algorithm to compute the occupancy distribution over any parameter values. However, for analysts who only need to compute the distribution for specific parameter values, it is not necessary to store the entire matrix L_N , and all that is needed for computation is the column of this matrix corresponding to the parameter n in the analysis. It is relatively straightforward to modify [Algorithm 1](#) to compute a single column of the matrix L_N without having to store the entire matrix. Each column of the matrix is defined recursively by the previous column, so the simplest way to do this is to compute one column of the matrix at a time using the recursive equations, and discard each previous column when a new one has been computed. In this case the storage limitation is alleviated substantially, since a single column of the matrix is much smaller than the entire matrix. It would be feasible to store the required column up to extremely large values of N (squaring the feasible size of the matrix suggests that something in the

⁹This is computed as $\text{logsumexp}(l_1, l_2) = \max(l_1, l_2) + \log(1 + \exp(-|l_1 - l_2|))$ where the function $\log(1 + \exp(-x))$ can be computed with high accuracy using its Maclaurin series. The function logsumexp is directly available in most statistical computing programs.

¹⁰With $N = 25,000$ this matrix of log-probabilities takes up 2.19 GB as an rmd file in R. Computation on a standard personal computer at the time of writing takes less than an hour.

Algorithm 2 Classical occupancy distribution

Input: Positive integers n and m and logical value **log**
Output: Vector of occupancy log-probabilities

```

#Create output vector
MAX <- min(n,m)
COCC <- Vector with MAX elements, with initial values all set to -∞

#Compute the classical occupancy distribution (or approximation)
#ROCC is from Algorithm 1 and COCC_APPROX is from Algorithm 3
if n ≤ N {
  TERM1 <- log(n+1-1:MAX) --log(m+1-1:MAX)
  TERM2 <- cumsum(TERM1)
  for each k = 1,...,MAX { COCC[k] <- ROCC[k,n] + n*[log(n) --log(m)] -- TERM2[k] } }
if n > N {
  for each k = 1,...,MAX { COCC[k] <- COCC_APPROX[k,n,m] } }
COCC <- COCC -- logsumexp(COCC)

if (log = TRUE) return COCC
else return exp(COCC)
#COCC[k,n,m] = log(Occ(k|n,m))

```

order of $N = 6.25 \times 10^8$ would be feasible for a single column.) Computing the relevant column of L_N for large N could still take substantial computation time, but it would lighten the storage limitation of our method. This could be used to compute the distribution for a specific set of parameter values, but it is not useful if one wishes to compute the distribution over a large range of parameter values.

5. The Normal Approximation and Its Accuracy

Our algorithm computes the classical occupancy distribution for $n \leq N$ where the upper limit is determined by storage considerations. In the case where $n > N$ we can appeal to asymptotic results, using the normal approximation or higher order approximations. We examine these approximations in greater detail in this section.

We have already seen from [Theorems 3A–B](#) that the normal density provides a good asymptotic approximation to the classical occupancy distribution, so long as the variance of the occupancy distribution is large. We know from this theorem that there is uniform convergence between the true probabilities and the approximation as the variance of the occupancy distribution tends to infinity. The normal approximation can be improved to adjust for higher order moments using various asymptotic expansions, but we will not find this necessary, since we can already compute the distribution for all but high values of the size parameter.

In this section we examine the accuracy of the normal approximation to the classical occupancy distribution. We will quantify the accuracy of the approximation for each finite combination of n and m in a large finite range, and observe how the accuracy improves as the parameters grow over this range. Our analysis extends simulations done in [Williamson \(2012\)](#), which undertook comparisons using selected values of the

parameters. In that article, accuracy of the normal and Poisson approximations were tested against the exact mass values from the classical occupancy distribution with small parameters, but against a simulation of the distribution for large parameters. By contrast, in our analysis, all comparisons are made using log-probabilities generated by our occupancy algorithm, which is able to compute the mass function for much higher values of the size parameter.

The normal approximation to the occupancy distribution is a continuous approximation to a discrete distribution, so there are various ways that it can be constructed. In our analysis we will use the density points from the continuous approximating distribution, but we scale these to ensure that we obtain a vector of probabilities that sums exactly to one. The approximation we use is the *scaled normal density approximation* defined as¹¹

$$\widehat{\text{Occ}}(k|n, m) = \frac{N(k|\mu_{n,m}, \sigma_{n,m}^2)}{\sum_k N(k|\mu_{n,m}, \sigma_{n,m}^2)}.$$

Below we construct an algorithm to approximate the classical occupancy distribution for the case where the size parameter is too large to feasibly compute using our recursive calculation. [Algorithm 3](#) computes the mean and variance of the classical occupancy distribution directly from their formulas and then computes the scaled normal approximation.

It is possible to adjust this algorithm to approximate using a higher order asymptotic expansion, such as the Gram–Charlier

¹¹ In the special case where $\min(n, m) = 1$ the occupancy distribution is a point-mass distribution at $k = 1$ and we have $\mu_{n,m} = 1$ and $\sigma_{n,m}^2 = 0$. In this special case, we take the density $N(k|\mu_{n,m}, \sigma_{n,m}^2)$ to be the Dirac function at $k = \mu_{n,m} = 1$ which means that the approximation is a point mass at $k = 1$ (which is the exact distribution). In this case both the occupancy distribution and its approximation are point-mass distributions at the same value, so there is no approximation error. This will be programmed as a special case in our algorithm.

Algorithm 3 Classical occupancy distribution (approx)

Input: Positive integers n and m and logical value **log**
Output: Vectors of approximate occupancy log-probabilities

```

#Create output vector
MAX          <- min(n,m)
COCC_APPROX <- Vector with MAX elements, with initial values all set to -∞

#Calculate the moments
E1           <- (1-1/m)^n
E2           <- (1-2/m)^n
COCC_MEAN    <- m*(1 - E1)
COCC_VAR     <- m*[(m-1)*E2 + E1 - m*E1^2]

#Compute the approximation vector
if MAX = 1 {
  COCC_APPROX[1] <- 0 }
if MAX > 1 {
  for each k = 1,...,MAX {
    COCC_APPROX[k] <- log(N(k| mean = COCC_MEAN, var = COCC_VAR)) }
COCC_APPROX <- COCC_APPROX -- logsumexp(COCC_APPROX)

if (log = TRUE) return COCC_APPROX
  else return exp(COCC_APPROX)
#COCC_APPROX[k,n,m] ≈ log(Occ(k|n,m))

```

expansion or the Edgeworth expansion (e.g., adjusting for skewness and kurtosis). This involves programing more of the moments into the algorithm, and then making the relevant adjustments to the normal approximation. [Theorem 2B](#) shows that, under broad limit conditions, the asymptotic skewness $\gamma_{n,m}^* = \mathcal{O}(m^{-1/2})$ and asymptotic kurtosis $\kappa_{n,m}^* = 3 + \mathcal{O}(m^{-1})$, which means that the classical occupancy distribution is asymptotically unskewed and mesokurtic. (Of course, this is a necessary condition for the convergence to the normal distribution, so it is weaker than the convergence results in [Theorems 3A–B](#).) The author has undertaken analysis of higher order approximations using the skewness and kurtosis formulas, but found that these adjustments are unnecessary. Specifically, we find that good performance is obtained by storing the reduced occupancy distribution for a high value of N and relying on the normal approximation for values outside this range.

In [Figure 1](#), we compare the log-probabilities from the classical occupancy distribution to the log-probabilities from the scaled normal density approximation for $n = m = 1000$. The solid line shows the log-probabilities of the classical occupancy distribution computed using [Algorithm 2](#), and the dashed line shows the scaled normal density approximation. We have also computed the log-probabilities in the distribution using multiple-precision floating-point arithmetic with 2^8 bits (using the `Rmpfr` package in R). Our algorithm for the classical occupancy distribution gives values extremely close to the values calculated from multiple-precision floating-point arithmetic; inaccuracy in this computation is ignorable, relative to the much smaller approximation error from using the normal approximation.

As can be seen from the figure, the normal approximation gives log-probabilities that are close to the true values in the center of the distribution, but they are overestimated in the tails. (The occupancy distribution with these parameters is slightly platykurtic, so it has thinner tails than the normal distribution.) When converted to the raw probability scale the approximation is extremely good, owing to the fact that the tails have very low probabilities, and so inaccuracy in the tails is relatively less important. (Though it may be important when trying to calculate higher order moments, etc.) The selected parameter values in this figure are quite low relative to the upper limit when we will switch to using the normal approximation. When we increase the values of the parameters the approximation becomes more accurate (we will establish this soon).

The purpose of [Figure 1](#) is to establish that our recursive computation method is essentially as accurate as exact computation using multiple-precision floating-point arithmetic; any deviation from the exact computation is so tiny that it is visually imperceptible in a graphical depiction of the log-probabilities, and it is small relative to the inaccuracy in the normal approximation. Thus, for practical purposes, we are concerned with the accuracy of the normal approximation relative to our recursive algorithm, but we are not concerned with the accuracy of our recursive algorithm relative to exact computation using multiple-precision floating-point arithmetic.

To examine the accuracy of the normal approximation, we will look at the classical occupancy distribution over all parameter combinations in the range $1 \leq n, m \leq 5000$. It is feasible to store the precomputed matrix of log-probabilities for the reduced occupancy distribution over this parameter range, and

Figure 1. Values of $\log \text{Occ}(k|n, m)$ for $n = 1000$ and $m = 1000$. The solid line is the exact value and the dashed line is the normal approximation.

Figure 2. Values of $\text{LRMSE}(n, m)$ measuring the exact classical occupancy distribution against the normal approximation, shown over the parameter range $1 \leq n \leq 1000$ and $1 \leq m \leq 1000$. The reader will observe that there are ridges with lower accuracy (higher LRMSE) for certain parameter ratios.

the storage requirement is manageable (it takes up 0.088GB as an rmd file). This allows us to compute the classical occupancy distribution using our recursive algorithm over the same parameter range, and compare this to the normal approximation. We will measure the accuracy of the normal approximation with the log-root-mean-square-error $\text{LRMSE}(n, m) \equiv \log \text{RMSE}(n, m)$ where the root-mean-square-error is:

$$\text{RMSE}(n, m) \equiv \sqrt{\frac{\sum_{k=1}^{\min(n, m)} [\text{Occ}(k|n, m) - \widehat{\text{Occ}}(k|n, m)]^2}{\min(n, m)}}.$$

Figure 2 shows a heat-map of our accuracy measure over the stated parameter range. From [Theorems 3A–B](#) we know that the

normal approximation converges uniformly to the true classical occupancy distribution along any path leading to an infinite limit for the variance of the distribution. This includes any path in which the parameter ratio n/m converges to a fixed value. Moreover, from [Figure 2](#) we can see that the only problematic regions for the normal approximation occur for some ridges of parameter combinations where the LRMSE does not appear to be decreasing rapidly. Our uniform convergence theorem means that the only ridges that do not lead to perfect accuracy are those where the variance does not converge to infinity.

To get a clearer understanding of the accuracy of the approximation, it is useful to reduce the two-dimensional heat-map in [Figure 2](#) to assessment along a single dimension. To do this we will examine the root-mean-squared-error as the value $s = \max(n, m)$ gets bigger. For each $s \in \mathbb{N}$ we define the “mean,” “maximum,” and “diagonal” RMSE values, respectively, by

$$\begin{aligned} \text{Mean RMSE}(s) &\equiv \text{RMSE}(s) = \frac{1}{|\mathcal{J}(s)|} \\ &\times \sum_{(n, m) \in \mathcal{J}(s)} \text{RMSE}(n, m), \end{aligned}$$

$$\text{Maximum RMSE}(s) \equiv \overline{\text{RMSE}}(s) = \max_{(n, m) \in \mathcal{J}(s)} \text{RMSE}(n, m),$$

$$\text{Diagonal RMSE}(s) \equiv \widetilde{\text{RMSE}}(s) = \text{RMSE}(s, s).$$

where $\mathcal{J}(s) \equiv \{(n, m) \in \mathbb{N}^2 \mid \max(n, m) = s\}$. These functions allow us to see the accuracy of the normal approximation by reducing the parameter space to a single dimension, measuring accuracy against the maximum parameter value. [Figure 3](#) shows these two functions, after discarding some burn-in values for low s . We can see from the figure that both the mean and maximum RMSE decrease as s increases. It is comforting to note that the maximum RMSE measures the accuracy of the *least accurate* approximation for a given value s , and even this accuracy measure is decreasing over the parameter range

Figure 3. Plot of the mean and maximum (and diagonal) RMSE over the parameter range $1 \leq \max(n, m) \leq 5000$. The reader will observe that.

we examine. (Note that this function shows the RMSE in the heat-map over the ridge of parameter values with the worst accuracy.)

We can see from the plot that the RMSE of the normal approximation diminishes rapidly along the main diagonal, the mean RMSE diminishes a bit more slowly, and the maximum RMSE reduces much more slowly. Using precomputed values up to $s = \max(n, m) = 5000$ in Algorithm 1 gives a maximum RMSE of 2.524% and mean RMSE of 0.046%. In other words, if we take the normal approximation to the classical occupancy distribution for a parameter combination with this maximum value, then on average we will have a distributional error of about one-twentieth of a percent, and we will have a maximum distributional error of about two-and-a-half percent.

Computational limits make it difficult to compute the full heat-map of LRMSE values over a parameter range that is substantially larger than what we have shown here, and thus, it is also computationally difficult to extend the functions RMSE and $\overline{\text{RMSE}}$ substantially. Nevertheless, imposing appropriate caveats on the dangers of extrapolation, log-linear regression analysis of these functions (see Appendix) shows that they conform very closely to the following functions:

$$\begin{aligned} \text{RMSE}(s) &\approx \exp(-2.12982 \cdot s^{1/15.58907}) & R^2 &= 0.9996, \\ \overline{\text{RMSE}}(s) &\approx \exp(-3.24423 \cdot s^{1/9.837861}) & R^2 &= 0.9984. \end{aligned}$$

Diagnostic plots for this regression show evidence of nonlinearity and heteroscedasticity, but this nonlinearity is a roughly damped oscillation with diminishing variance, suggesting that the regression line might actually become more accurate for larger values of s . If one is willing to extrapolate the present regression lines to values of s beyond the range of the inputs, then these functions suggests that a maximum parameter value of $s = \max(n, m) = 25,000$ would give a maximum RMSE of 1.694% and a mean RMSE of 0.011%. (This maximum

parameter value is close to the limits of what we can store as precomputed values for Algorithm 1 in R.)

There are many parameter paths that can be taken as $\max(n, m) \rightarrow \infty$, and most (but not all) of these paths satisfy the variance limit $\sigma_{n,m}^2 \rightarrow \infty$. In particular, we have previously noted that if the parameter ratio n/m approaches a fixed limit $0 < \lambda < \infty$ then the variance approaches infinity. The uniform convergence result in Theorem 3B applies in this case, which implies that $\text{RMSE} \rightarrow 0$. Since this variance limit applies for “almost all” parameter paths, we would therefore expect that mean RMSE must approaches zero in the limit, but maximum RMSE will not. Nevertheless, it is comforting to see that even the latter decreases (albeit quite slowly) over the parameter range we are examining.

6. Concluding Remarks

In this article, we have reviewed some of the basic properties of the classical occupancy problem and the classical occupancy distribution that arises from this problem. We have derived the moments of this distribution, including exact formulas for its skewness and kurtosis. We have also developed an algorithm that efficiently generates log-probabilities/probabilities from the classical occupancy distribution for any parameter values. For cases where the size parameter is “small” ($n \leq N$) our algorithm uses a recursive calculation method that gives values that are close to the exact values calculated from multiple-precision floating-point arithmetic. For cases where the size parameter is “large” ($n > N$) our algorithm uses the normal approximation, giving values close to the true values.

Our analysis of the latter approximation shows that we can obtain high accuracy when the size parameter is large. This means that our algorithm can guarantee high accuracy by setting N to be sufficiently large to ensure that the approximation

applies only for large size values. So long as storage space is not an issue, we can feasibly set $N = 25,000$ and compute and store an $N \times N$ matrix of log-probabilities of the reduced occupancy distribution, which ensures that our algorithm has high accuracy for “small” and “large” size parameters. If storage space is an issue then we can set $N = 5000$ to obtain reasonable accuracy for “large” size values.

The algorithm developed in this article can be used to create functions dealing with the classical occupancy distribution. We have given a function to compute log-probabilities/probabilities from the mass function of the distribution, but this can easily be extended to create functions to compute the cumulative distribution or quantile function, or to generate random variables from the distribution. Our algorithm makes it feasible to create a suite of functions of this kind to mimic the functionality that is available in statistical packages for standard distributions.

Appendix: Proof of Theorems

Proof of Theorem 1A. Expanding the falling factorial we have

$$(m - k)_r = \underbrace{(m - k)(m - k - 1) \cdots (m - k - r + 1)}_{r \text{ terms}}.$$

If $k > m - r$ then one of these terms is zero, and so the falling factorial is zero; otherwise the falling factorial is positive. This means that we can write:

$$\frac{(m - k)_r}{(m - k)!} = \frac{\mathbb{I}(k \leq m - r)}{(m - k - r)!}.$$

Thus, we have

$$\begin{aligned} \mathbb{E}((m - K)_r) &= \sum_{k=1}^{\min(n, m)} (m - k)_r \cdot \text{Occ}(k|n, m) \\ &= \sum_{k=1}^{\min(n, m)} (m - k)_r \cdot \frac{m!S(n, k)}{m^n (m - k)!} \\ &= \sum_{k=1}^{\min(n, m)} \frac{(m - k)_r}{(m - k)!} \cdot \frac{m!S(n, k)}{m^n} \\ &= \sum_{k=1}^{\min(n, m)} \mathbb{I}(k \leq m - r) \cdot \frac{m!S(n, k)}{m^n (m - k - r)!} \\ &= \sum_{k=1}^{\min(n, m-r)} \frac{m!S(n, k)}{m^n (m - r - k)!} \\ &= \frac{m!}{(m - r)!} \left(\frac{m - r}{m}\right)^n \sum_{k=1}^{\min(n, m-r)} \text{Occ}(k|n, m - r) \\ &\quad \times \frac{(m - r)!S(n, k)}{(m - r)^n (m - r - k)!} \\ &= \frac{m!}{(m - r)!} \left(\frac{m - r}{m}\right)^n \sum_{k=1}^{\min(n, m-r)} \text{Occ}(k|n, m - r) \\ &= \frac{m!}{(m - r)!} \left(\frac{m - r}{m}\right)^n \\ &= (m)_r \left(1 - \frac{r}{m}\right)^n = (m)_r E_r, \end{aligned}$$

which was to be shown. $\square$

Lemma 1A. For all powers $p \in \mathbb{N}$ and all integer values $1 \leq k \leq m$ we have

$$\begin{aligned} k^p &= \sum_{r=0}^p (-1)^r a_{p,r}(m) \cdot (m - k)_r \\ a_{p,r}(m) &= \sum_{q=r}^p (-1)^{q-r} \binom{p}{q} S(q, r) m^{p-q}. \end{aligned}$$

Proof of Lemma 1A. Start with the polynomial $(x - y)^p$ and expand it as follows:

$$\begin{aligned} (x - y)^p &= \sum_{q=0}^p \binom{p}{q} (-1)^q x^{p-q} y^q \\ &= \sum_{q=0}^p \binom{p}{q} (-1)^q x^{p-q} \sum_{r=0}^q S(q, r) (y)_r \\ &= \sum_{q=0}^p \sum_{r=0}^q (-1)^q \binom{p}{q} S(q, r) x^{p-q} (y)_r \\ &= \sum_{r=0}^p (-1)^r \sum_{q=r}^p (-1)^{q-r} \binom{p}{q} S(q, r) x^{p-q} (y)_r. \end{aligned}$$

(The first line is the binomial expansion and the second line uses the formula for expressing powers as sums of the Stirling numbers of the second kind.) Now we substitute $x = m$ and $y = m - k$ to get the required result. $\square$

Proof of Theorem 1B. These result follow using Lemma 1A and Theorem 1A via linearity of the expected value operator. $\square$

Lemma 1B. The first four raw moments of the occupancy distribution are

$$\begin{aligned} \mathbb{E}(K) &= m[1 - E_1], \\ \mathbb{E}(K^2) &= m[m - (2m - 1)E_1 + (m - 1)E_2], \\ \mathbb{E}(K^3) &= m \begin{bmatrix} m^2 - (3m^2 - 3m + 1)E_1 \\ + 3(m - 1)(m - 1)E_2 \\ - (m - 1)(m - 2)E_3 \end{bmatrix}, \\ \mathbb{E}(K^4) &= m \begin{bmatrix} m^3 - (4m^3 - 6m^2 + 4m - 1)E_1 \\ + (6m^2 - 12m + 7)(m - 1)E_2 \\ - (4m - 6)(m - 1)(m - 2)E_3 \\ + (m - 1)(m - 2)(m - 3)E_4 \end{bmatrix}. \end{aligned}$$

Proof of Lemma 1B. To obtain the first four moments we use the formula in Theorem 1B, which requires us to calculate the polynomials $a_{p,r}$ for $0 \leq r \leq p \leq 4$. From the general result it can easily be established that $a_{p,0} = m^p$ and $a_{p,p} = 1$. With a bit of algebra we obtain the other required polynomials, which are:

$$\begin{aligned} a_{2,1} &= 2m - 1, \\ a_{3,1} &= 3m^2 - 3m + 1 & a_{3,2} &= 3m - 3, \\ a_{4,1} &= 4m^3 - 6m^2 + 4m - 1 & a_{4,2} &= 6m^2 - 12m + 7 & a_{4,3} &= 4m - 6. \end{aligned}$$

Substitution into Theorem 1B and a bit of algebra yields the results. $\square$

Proof of Theorem 2A. Follows trivially from the raw moments in Lemma 1B. $\square$

Proof of Theorem 2B. With a substantial amount of algebra using the raw moments in Lemma 1B, the third and fourth central moments can

be shown to be

$$\begin{aligned} \mathbb{E}\left((K - \mu_{n,m})^3\right) &= -m \left[\begin{pmatrix} E_1 + 3(m-1)E_2 \\ + (m-1)(m-2)E_3 \end{pmatrix} - \begin{pmatrix} 3mE_1^2 - 2m^2E_1^3 \\ + 3m(m-1)E_1E_2 \end{pmatrix} \right], \\ \mathbb{E}\left((K - \mu_{n,m})^4\right) &= m \left[\begin{pmatrix} E_1 - 4mE_1^2 + 6m^2E_1^3 - 3m^3E_1^4 \\ + 7(m-1)E_2 + 6(m-1)(m-2)E_3 \\ + (m-1)(m-2)(m-3)E_4 \\ - 12m(m-1)E_1E_2 + 6m^2(m-1)E_1^2E_2 \\ - 4m(m-1)(m-2)E_1E_3 \end{pmatrix} \right]. \end{aligned}$$

With a bit more algebra, the first of these can be rewritten as

$$\begin{aligned} \mathbb{E}\left((K - \mu_{n,m})^3\right) &= -m \left[E_1 - 3E_2 + 2E_3 + m \begin{pmatrix} 3(E_2 - E_1^2) \\ + 2m(E_1^3 - E_1E_2) \\ + (m-3)(E_3 - E_1E_2) \end{pmatrix} \right]. \end{aligned}$$

These formulas lead to the stated skewness and kurtosis values in the corollary, and the asymptotic expressions readily follow with a bit more algebra. $\square$

Remark. It is important to note that there is a substantial amount of algebra involved in the derivation of the expressions for the skewness and kurtosis, and this gives rise to a large danger of algebraic error. To ensure that our algebraic forms are correct, we computed the output of these formulas over a range of values of n and m and compared these with computations of the skewness and kurtosis obtained by the alternative method of summing over the generated values of the occupancy distribution. Our analysis showed that these values match over the range of values we used, to within tolerances that are explainable by rounding error. We are therefore confident that these expressions are correct.

Proof of Theorem 3. This is a complicated proof; see Theorems 2.1 and 2.4 of Hwang and Janson (2008) for full details and some discussion of methods. The theorem stated here is slightly less general than those results insofar as we take the supremum only over the integers in the support of the occupancy distribution, whereas Hwang and Janson give a result with the supremum taken over all the real numbers (with the argument value of the mass function of the occupancy distribution obtained by appropriate rounding). $\square$

Proof of Theorem 4. For all $n \in \mathbb{N}$ we have $S(n, 1) = 1$ so we have the base values:

$$p_{1,n} = \frac{(n)_1 \cdot S(n, 1)}{n^n} = \frac{n \cdot 1}{n^n} = \frac{n}{n^n} = \frac{1}{n^{n-1}}.$$

Now, since $S(n, n) = 1$ for all $n \in \mathbb{N}$ we can also establish that:

$$p_{n,n} = \frac{(n)_n \cdot S(n, n)}{n^n} = \frac{n! \cdot 1}{n^n} = \frac{n!}{n^n}.$$

Hence, we have the recursive equation:

$$\begin{aligned} p_{n,n} &= \frac{n!}{n^n} = \frac{(n-1)!}{n^{n-1}} = \left(1 - \frac{1}{n}\right)^{n-1} \\ &\cdot \frac{(n-1)!}{(n-1)^{n-1}} = \left(1 - \frac{1}{n}\right)^{n-1} \cdot p_{n-1,n-1}. \end{aligned}$$

Finally, applying the recursive equation $S(n, k) = k \cdot S(n-1, k) + S(n-1, k-1)$ for the Stirling numbers of the second kind, we obtain:

$$p_{k,n} = \text{Occ}(k|n) = \frac{(n)_k \cdot S(n, k)}{n^n} = \frac{(n-1)_{k-1} \cdot S(n, k)}{n^{n-1}}$$

$$\begin{aligned} &= \left(1 - \frac{1}{n}\right)^{n-1} \cdot \frac{(n-1)_{k-1} \cdot S(n, k)}{(n-1)^{n-1}} \\ &= \left(1 - \frac{1}{n}\right)^{n-1} \cdot \frac{(n-1)_{k-1}}{(n-1)^{n-1}} \cdot [kS(n-1, k) + S(n-1, k-1)] \\ &= \left(1 - \frac{1}{n}\right)^{n-1} \cdot \left[\frac{k}{n-k} \cdot \frac{(n-1)_k \cdot S(n-1, k)}{(n-1)^{n-1}} \right. \\ &\quad \left. + \frac{(n-1)_{k-1} \cdot S(n-1, k-1)}{(n-1)^{n-1}} \right] \\ &= \left(1 - \frac{1}{n}\right)^{n-1} \cdot \left[\frac{k}{n-k} \cdot p_{k,n-1} + p_{k-1,n-1} \right] \end{aligned}$$

Taking logarithms of both sides of these three equations establishes the recursive equations in the theorem. Finally, as noted in the main body of the article, we have

$$p_{k,n,m} = p_{k,n} \cdot \frac{n^n}{m^n} \cdot \frac{(m)_k}{(n)_k} = p_{k,n} \cdot \left(\frac{n}{m}\right)^n \cdot \prod_{i=0}^{k-1} \left(\frac{n-i}{m-i}\right).$$

Taking logarithms of both sides establishes the last equation in the theorem. $\square$

Log-linear regression for mean and maximum LRMSE: To fit the functions shown in Figure 3, we used data for the mean and maximum LRMSE (discarding the first 100 values as burn-in) and fit these to the following log-linear regression models:

$$\begin{aligned} \log|\text{LRMSE}(s)| &= \beta_0 + \beta_1 \log s + \sigma \cdot \varepsilon(s) \quad \varepsilon(s) \sim \text{IID } N(0, 1), \\ \log|\overline{\text{LRMSE}}(s)| &= \beta_0 + \beta_1 \log s + \sigma \cdot \varepsilon(s) \quad \varepsilon(s) \sim \text{IID } N(0, 1). \end{aligned}$$

The data-frame DATA consisted all cases with $100 < \min(n, m) \leq 5000$ with the variables as defined in the body of the article. The R code for the model, and the resulting output (omitting the model call and p-values for brevity) are below.

```
> MODEL1 <- lm(log(-Max_LRMSE) ~ log(s),
+ data = DATA);
>
> summary(MODEL1);
```

Residuals:			
Min	1Q	Median	3Q
-0.0015724	-0.0004535	0.0001440	0.0003290
Max			
0.0140668			
Coefficients:			
	Estimate	Std. Error	t value
(Intercept)	0.75603727	0.00013619	5551
log(s)	0.06414752	0.00001782	3599

```
Residual standard error: 0.001029 on 4898 degrees
of freedom

Multiple R-squared: 0.9996, Adjusted
R-squared: 0.9996
```

```
F-statistic: 1.296e+07 on 1 and
4898 DF, p-value: < 0.0000000000000000022

> MODEL2 <- lm(log(-Mean_LRMSE) ~ log(s),
+ data = DATA);
>
> summary(MODEL2);
```

Residuals:			
Min	1Q	Median	3Q
-0.0194787	-0.0021772	0.0005202	0.0028199
Max			
0.0038111			

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	1.17687813	0.00044686	2634
log(s)	0.10164811	0.00005847	1738

Residual standard error: 0.003376 on 4898 degrees of freedom

Multiple R-squared: 0.9984, Adjusted R-squared: 0.9984

F-statistic: 3.022e+06 on 1 and 4898 DF, p-value: < 0.00000000000000022

The stated results in the body of the article are the regression functions fitted with these models. Diagnostic plots for these models do not accord to standard regression assumptions; there are patterns that are the result of additional deterministic relationships not in the model.

Acknowledgments

The author would like to thank two anonymous journal referees and the associate editor for suggestions that improved the article from a previous draft.

References

- Arfwedson, G. (1951), "A Probability Distribution Connected With Stirling's Second Class Numbers," *Scandinavian Actuarial Journal*, 34, 121–132. [1,2]
- Bunge, J., and Fitzpatrick, M. (1993), "Estimating the Number of Species: A Review," *Journal of the American Statistical Association*, 88, 364–373. [1]
- Chao, A. (1984), "Nonparametric Estimation of the Number of Classes in a Population," *Scandinavian Journal of Statistics*, 11, 265–270. [1]
- Charalambides, C. A. (2005), *Combinatorial Methods in Discrete Probability*, Hoboken, NJ: Wiley. [3]
- de Moivre, I. (1718), *The Doctrine of Chances*, London: Printed by W. Pearfor [Publicly Available on Google Books]. [1]
- Efron, B., and Thisted, R. (1978), "Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?," *Biometrika*, 63, 435–447. [1]
- Feller, W. (1950), *An Introduction to Probability Theory and Its Applications*, New York: Wiley. [1]
- Gnedin, A., Hansen, B., and Pitman, J. (2007), "Notes on the Occupancy Problem With Infinitely Many Boxes: General Asymptotics and Power Laws," *Probability Surveys*, 4, 146–171. [2]
- Good, I. J., and Toulmin, G. H. (1956), "The Number of New Species and the Increase in Population Coverage, When a Sample Is Increased," *Biometrika*, 43, 45–63. [1]
- Hald, A., de Moivre, A., and McClintock, B. (1984), "A. de Moivre: 'De Mensura Sortis' or 'On the Measurement of Chance,'" *International Statistical Review/Revue Internationale de Statistique*, 52, 229–262. [1]
- Hanley, J. A. (1992), "Jumping to Coincidences: Defying Odds in the Realm of the Preposterous," *The American Statistician*, 46, 197–202. [1]
- Harris, B. (1968), "Statistical Inference in the Classical Occupancy Problem Unbiased Estimation of the Number of Classes," *Journal of the American Statistical Association*, 63, 837–847. [1]
- Holst, L. (1986), "On Birthday, Collectors' Occupancy and Other Classical Urn Problems," *International Statistical Review/Revue Internationale de Statistique*, 54, 15–27. [1]
- Hwang, H. K., and Janson, S. (2008), "Local Limit Theorems for Finite and Infinite Urn Models," *The Annals of Probability*, 36, 992–1022. [3,4,11]
- Johnson, N. L., and Kotz, S. (1969), *Discrete Distributions*, New York: Wiley, pp. 251–253. [2,3]
- (1977), *Urn Models and Their Applications*, New York: Wiley, pp. 107–175, 318–370. [1,2,3]
- Kolchin, V. F., Sevast'yanov, B. A., and Chistyakov, V. P. (1978), *Random Allocations*, New York: Wiley. [1]
- McNeil, D. R. (1973), "Estimating an Author's Vocabulary," *Journal of the American Statistical Association*, 68, 92–96. [1]
- Ruiz, S. (1996), "An Algebraic Identity Leading to Wilson's Theorem," *The Mathematical Gazette*, 80, 579–582. [3]
- Park, C. J. (1972), "A Note on the Classical Occupancy Problem," *The Annals of Mathematical Statistics*, 43, 1698–1701. [1,3]
- Samuel-Cahn, E. (1974), "Asymptotic Distributions for Occupancy and Waiting Time Problems With Positive Probability of Falling Through the Cells," *The Annals of Probability*, 2, 515–521. [1]
- Stigler, S. M. (1980), "Stigler's Law of Eponymy," *Transactions of the New York Academy of Sciences*, 39, 147–158. [2]
- Weiss, I. (1958), "Limiting Distributions in Some Occupancy Problems," *The Annals of Mathematical Statistics*, 29, 878–884. [1]
- Williamson, P. P. (2012), "Usefulness of Asymptotic Distributions in the Classical Occupancy Problem," *Communications in Statistics—Simulation and Computation*, 41, 1501–1517. [3,6]
- Williamson, P. P., Mays, D. P., Abay Asmerom, G., and Yang, Y. (2009), "Revisiting the Classical Occupancy Problem," *The American Statistician*, 63, 356–360. [1,2,3]