Properties of a model of sequential random allocation

Laura Stewart

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics & Statistics
College of Science and Engineering
University of Glasgow

September 2022

Abstract

Probabilistic models of allocating shots to boxes according to a certain probability distribution have commonly been used for processes involving agglomeration. Such processes are of interest in many areas of research such as ecology, physiology, chemistry and genetics. Time could be incorporated into the shots-and-boxes model by considering multiple layers of boxes through which the shots move, where the layers represent the passing of time. Such a scheme with multiple layers, each with a certain number of occupied boxes is naturally associated with a random tree. It lends itself to genetic applications where the number of ancestral lineages of a sample changes through the generations. This multiple-layer scheme also allows us to explore the difference in the number of occupied boxes between layers, which gives a measure of how quickly merges are happening. In particular, results for the multiple-layer scheme corresponding to those known for a single-layer scheme, where, under certain conditions, the limiting distribution of the number of occupied boxes is either Poisson or normal, are derived.

To provide motivation and demonstrate which methods work well, a detailed study of a small, finite example is provided. A common approach for establishing a limiting distribution for a random variable of interest is to first show that it can be written as a sum of independent Bernoulli random variables as this then allows us to apply standard central limit theorems. Additionally, it allows us to, for example, provide an upper bound on the distance to a Poisson distribution. One way of showing that a random variable can be written as a sum of independent Bernoulli random variables is to show that its probability generating function (p.g.f.) has all real roots. Various methods are presented and considered for proving the p.g.f. of the number of occupied boxes in any given layer of the scheme has all real roots. By considering small finite examples some of these methods could be ruled out for general N.

Finally, the scheme for general N boxes and n shots is considered, where again a uniform allocation of shots is used. It is shown that, under certain conditions, the distribution of the number of occupied boxes tends towards either a normal or Poisson limit. Equivalent results are also demonstrated for the distribution of the difference in the number of occupied boxes between consecutive layers.

Contents

List of Figures

Acknowledgements

Firstly, I would like to thank my supervisors, Dr Alexey Lindo and Dr Vincent Macaulay for their time, patience and encouragement. Our meetings provided some much needed normality during the last few years!
I am very grateful to the EPSRC for their generous funding via a studentship.
I would like to thank my examiners, Dr Mayetri Gupta and Dr Denis Denisov, for taking the time to read my thesis and for their suggested improvements.
Also, thanks to Dr. Theodore Papamarkou for his help with applying for a PhD and securing an EPSRC scholarship.
Thank you to the School of Mathematics and Statistics and my fellow PhD students for making this time more enjoyable.
To my new colleagues at UWS, thank you for your patience and understanding as I finished writing this thesis.
Last, but definitely not least, I would like to thank my parents for their support and Effie and David for making Glasgow feel more like home.

Declaration

I, Laura Stewart, declare that this thesis titled *Properties of a model of sequential random allocation* and the work presented in it are my own. I confirm that:

- This work was done wholly while a candidate for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Chapter 1

Introduction

1.1 Motivation

Consider allocating n shots independently into N boxes according to a certain probability distribution. We can alter the behaviour of this process by changing, for example, n, N or the probability distribution. Traditionally, when using this scheme the main focus has been on the number of empty boxes under certain conditions ([1]-[5]). This model can be used in many settings such as computational linguistics [6], ecology [7] and biology [8]. For example, in ecological problems boxes may be viewed as representing species and then the number of occupied boxes can be interpreted as the ecological diversity [7]. As well as applications where the objective is to determine the occupancy count, there are other problems where the resulting probabilities of landing in specific boxes are the subject of interest. Consider the following:

1. Disclosure risk limitation [9]: In general, disclosure risk limitation involves editing data prior to it becoming publicly available in order to prevent the possibility of releasing personally identifiable information. Here the shots would represent people or companies whose details (not including information that would easily identify them) are stored. The boxes are then the unique combinations of values of variables that would allow for them to be identified.

2. Database query optimisation [10]: The main aim of database query optimisation is to find the most efficient way of executing a given query. Here the shots would represent entries in the columns of a database. The boxes would then be all of the distinct values appearing in the column.

3. Literature [6]: Here the shots would represent the words in everything that has been written by a particular author. The boxes are then all of the distinct words the author has used. The goal here is to try and identify or verify the author of (often historical) work which was previously not attributed to anyone.

We are mainly interested in applications involving agglomeration. Processes involving agglomeration occur in a broad range of fields, such as ancestry (tracing back a sample of individuals to a common ancestor [11]), physiology (blood clotting [12]) and statistics (hierarchical clustering [13]). In particular, coalescence theory has been used in the field of population genetics for a long time with it being used more recently to analyse DNA sequence data - see [11] for details. Note that it is common in the literature to see boxes referred to as urns ([14], [15]) or cells ([3], [16]). Similarly, shots are often referred to as balls ([8], [16]) or particles [3].

One of the earliest results relating to this scheme was established by Weiss in 1958 [1]. He proved that, supposing n and N tend to infinity in such a way that n and N are proportional to one another, the number of empty boxes has a normal limiting distribution. David and Barton then extended this result in 1962 to show that the normal limit also holds for the number of boxes containing exactly m shots [17]. In the same year, Rényi was able to show that the condition proposed by Weiss (that the number of shots and boxes are proportional) can be relaxed and he established that a sufficient condition is that the variance of the number of empty boxes tends to infinity [5]. Further, in 1963, Békéssy then extended Rényi's result by showing that it holds more generally for the number of boxes containing exactly m shots [4].

In 1978, Kolchin et al. [2] explored this scheme of shots and boxes in detail for both uniform and non-uniform allocations of shots. That is to say, under a uniform allocation of shots the probability of landing in any given box is $1/N$ whereas in the non-uniform case the probabilities for each box are no longer equal. In particular, for the uniform scenario they were able to establish five distinct domains where the limiting behaviour of the distribution of the number of empty boxes is either Poisson or normal. For the non-uniform case, they showed three regions where again the limiting distribution is either Poisson or normal under certain conditions on the probabilities assigned to the boxes. Chapter four of this thesis reviews these regions and gives examples.

One paper that proves very useful for the work in this thesis in terms of the method of proof used was written by Vatutin and Mikhailov in 1982 [3]. Their proof is discussed in detail in Chapter four but the main idea is that if you want to prove distributional results for a certain random variable it is very helpful if it can be shown that this random variable is equivalent to a sum of independent Bernoulli random variables (not necessarily with the same probability of success). This can be done by first showing that its probability generating function (p.g.f.) has all real roots. One way to achieve this is by showing it can be obtained by applying a sequence of real-root-preserving transformations to a polynomial which is known to have all real roots. From this point normal and Poisson limiting results follow [3].

As well as establishing limit theorems, later papers were also interested in proving results for the speed of convergence to these limiting distributions. The number of occupied boxes (when scaled and centred) is known to be asymptotically normal as the number of shots tends to infinity in the uniform case with n proportional to N, and a Berry-Esseen bound for the discrepancy

from the normal, tending to zero at the optimum rate, was ascertained in 1981 by Englund [18] (see also [3]). In 1982, Quine and Robinson then proved a similar result for the more general scheme with a non-uniform allocation of shots to boxes [19].

Most papers focus on either the number of empty or occupied boxes but a more recent paper by Penrose in 2009 [20] showed results for the number of boxes containing a single shot and illustrated why this is of interest. Consider the well-known birthday paradox where the number of boxes containing a single shot represents the number of individuals in the group who have a different birthday to the other group members. We mention some other applications where the number of isolated shots could be of particular interest. If the shots represent individuals and the boxes represent their classification in a database according to certain characteristics (see [8]), then the number of isolated shots represents the number of individuals which can be identified uniquely from their classification. If the shots represent particles or biological individuals and the boxes represent their spatial locations, each shot occupying one of N locations chosen at random, and if two or more shots sharing the same location annihilate one another, then the number of isolated shots is the number of surviving particles [20]. If the shots represent the users of a set of communication channels at a given instant (these could be physical channels or wavebands for wireless communications), and each user randomly selects one of N available channels, and two or more attempted users of the same channel interfere with each other and are unsuccessful, then the number of isolated shots is the number of successful users [20]. When a uniform allocation of shots is used, a normal limiting result was obtained for the number of boxes containing one shot and it was shown that the distance to the normal distribution tends to zero at the optimum rate when the number of shots and boxes are proportional to one another. This relates to an earlier result established by Englund in 1981 [18]. He showed that for the distribution of the number of occupied boxes the distance to the normal distribution tends to zero at the optimum rate when n and N are proportional to one another. A central limit theorem for the number of boxes containing a single shot was also proven in the non-uniform case under certain conditions on the probability distribution used. As well as being different in terms of the random variable being considered, the methods of proof used in this paper were also different to what had come before as they were based on a method previously used by Goldstein and Penrose for a problem in stochastic geometry [21]. Most other papers up to this point had either used characteristic functions [2] or a Poissonization technique to deal with dependence between boxes ([8], [22]).

In the papers mentioned above the number of boxes was always finite although allowed to become arbitrarily large. In contrast, some authors instead considered having an infinite number of boxes from the outset with a non-uniform allocation of shots. The motivation behind this was that such models can be used to approximate sampling from a large finite population. One of the first studies of this infinite scheme was conducted by Karlin in 1967 [23]. It should also be noted that others such as Bahadur [24] and Darling [25] were also studying this infinite scheme

around the same time as Karlin. To deal with the additional complexity created by having an infinite number of boxes, the conditions were more restrictive [23]. However, in 1989, Dutko [26] showed that Karlin's results actually hold under the simple assumption that the variance of the number of occupied boxes tends to infinity. Note this mirrors the way in which such results were obtained for the finite scheme where first limiting results were proven under certain conditions but it was later shown that a sufficient condition was that the variance of the random variable being considered tends to infinity. More recently, in 2007, Gnedin et al. [8] provided an overview of the known results relating to the infinite scheme. Additionally, in 2008, Hwang and Janson were able to produce local limit theorems for the number of empty boxes in both the finite and infinite schemes with the single condition that the variance of the number of empty boxes tends to infinity [22].

This thesis will focus on a multiple-layer scheme with a finite number of boxes but it is interesting to understand how things change with an infinite number of boxes and the methods used to establish results in this case. This multiple-layer scheme is set up as follows. We start by allocating n shots independently to N boxes as is done in the standard shots and boxes scheme used in the above papers. We shall specify that a uniform probability distribution is used so that the probability of a particular shot landing in any box is given by $1/N$. Now, we want to add additional layers and this is where the scheme starts to differ from the established literature. In every layer the number of boxes will be the same so that N is fixed throughout the layers. Also, in every layer, shots are thrown independently according to the equiprobable distribution. From layer two onwards, the number of shots thrown into that layer is given by the number of occupied boxes in the previous layer. As in the single layer shots-and-boxes scheme, we have that if two or more shots lands in the same box in any given layer then they merge together. Let $S(k)$ be a random variable denoting the number of shots thrown into layer k and let $R^*(k,N)$ be the number of occupied boxes in layer k of the scheme with N boxes. Then the number of shots thrown into each layer is dictated as follows:

1. The first layer of the scheme involves n shots being scattered into N boxes giving $R^*(1,N)$.
2. For layer two $S(2) = R^*(1,N)$ shots are allocated in N boxes.
3. This process is continued throughout the layers where $S(k) = R^*(k-1,N)$. That is, the number of shots thrown into layer k is equal to the number of occupied boxes in the previous layer.
4. So, the probability distribution for $S(k)$ is given by the probability distribution for $R^*(k-1,N)$.

1.2 Connection to iterations of random functions

Although the multiple-layer scheme has not been studied in terms of shots and boxes it has been studied in the context of iterations of random functions ([27]-[29]). This thesis does however provide novel results for this scheme and these will be discussed in detail in the relevant chapters. The application of a single random function corresponds to the allocation of shots in a single layer. Additionally, layer k of this multiple-layer scheme corresponds to the composition of k random functions. One of the questions we might be interested in answering using this multiple-layer scheme of shots and boxes is the following:

Suppose we observe the sequences of the maternally inherited mitochondrial DNA for a finite sample of females, how far back in time do we have to go in order to find their Most Recent Common Ancestor (MRCA)?

Here, the shots would be the finite sample we start with, the boxes would be the population of females and the number of occupied boxes in a given layer corresponds to the number of ancestral females of the sample in a certain generation. Suppose we start with four individuals (labelled A, B, C and D, Figure 1.1) and in the not too distant past individuals C and D share a common maternal ancestor. Then, some time before this B also shares a common ancestor with C and D. Finally, after a certain length of time - the time until a most recent common ancestor - all four individuals have a common ancestor. Hence, the time until a most recent common ancestor (the "coalescence" time) corresponds to the first time when there is just one occupied box in the shots and boxes scheme. Now, in terms of iterations of random functions this would correspond to the minimum number of functions we need to take the composition of until the size of the output is one.

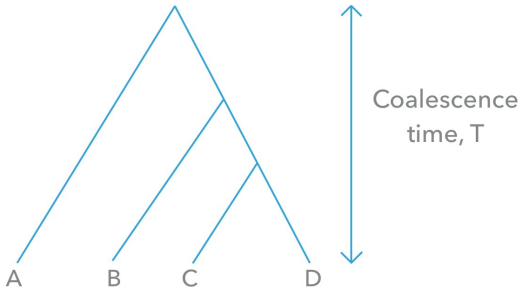

Figure 1.1: Example of tracing back a sample of four individuals to a MRCA.

In 2015, Zubkov and Serov provided lower and upper bounds for the mean size of the image after k compositions [27] (i.e., the number of occupied boxes after k layers in terms of shots and boxes). They also established an upper bound for the variance of the size of this image after k compositions. In 2016, Serov then explored the scenario where you have two schemes running at the same time where they both consist of k compositions of random functions [28]. Within

each scheme the mappings are independent but the two schemes are dependent. He established a variety of bounds on the associated probabilities and expected values. In 2018, Zubkov and Serov continued their work from 2015 and established recurrence relations and inequalities for a random variable that is equivalent to the number of occupied boxes in layer k of the multiple-layer shots and boxes scheme presented in this thesis [29]. This contrasts with my approach as I aim to prove distributional results for the number of occupied boxes in any given layer of the multiple-layer shots and boxes scheme.

1.3 Connection to the Wright-Fisher model

The Wright-Fisher (WF) model [11] specifies that, in a population of N gene copies, the number of descendants of each gene copy is $\text{Pois}(1)$, but in addition, the number of descendants must sum to N (i.e., population size is assumed to be constant). It follows, using Bayes' theorem that the vector of the number of descendants of each gene copy is $\text{Mult}(N; 1/N, \ldots, 1/N)$. This assumes a labelling of the ancestors, say $1, 2, \ldots, N$. Let X_i be the random number of descendants of ancestor i. Note that we are just counting the number of descendants and are not interested in who the descendants are. Let us now label the gene copies in the descendant generation $1, 2, \ldots, N$, and randomly choose which of these are the descendants that were counted by the X_is. It is a random partition of the first N integers given the sizes of the sets in the partition.

We can represent this as a random bipartite graph between the sets of labelled ancestor nodes and descendant nodes where an edge joins an ancestor to a descendant. Every descendant has order one and the order of each ancestor i is X_i. This fully labelled WF graph can be generated from the description just outlined.

However, a graph with the same probability can be generated in an equivalent way which we will now outline. Suppose you go through the descendant nodes one at a time and connect them to ancestors (with replacement) uniformly and independently. This then leads to the same Multinomial distribution for the vector of the order of all of the ancestral nodes (i.e., the number of their descendants).

Now we can connect this to the multiple-layer shots and boxes scheme. The labelled ancestors and descendants are boxes, in layers two and one. The descendants boxes are all filled. Their shots are thrown one at a time uniformly and independently at the ancestral boxes. Connecting the origin of each shot to its endpoint creates a graph equivalent to (having the same probability distribution as) the WF graph.

The equivalence through iterated layers follows too. At an arbitrary generation suppose only n of the gene copies have descendants in the final generation. We can find their ancestors in the same way. That is, by randomly connecting to n ancestors uniformly with replacement from the previous generation (just as we chose N ancestors above). This is exactly what we do with shots in boxes. The n gene copies with descendants in the final generation map into the n filled boxes

at that layer, which are used to fill (uniformly and independently) the next layer up.

1.4 Aims and structure of this thesis

The aims of this thesis are:

1. to build on the known results for a single-layer allocation of shots to boxes;
2. to illustrate the multiple-layer scheme of shots and boxes and produce explicit formulae for the expectation, variance and probability generating function for the number of occupied boxes in any given layer for a small example;
3. to extend this to the multiple-layer scheme for the more general scenario with an arbitrary fixed number of boxes;
4. to establish limit theorems for the distribution of the number of occupied boxes in a given layer of a multiple-layer shots and boxes model;
5. to indicate where this work fits within the literature and how it could be further developed.

In more detail, Chapter two refreshes some standard techniques for showing that a polynomial has real roots. Chapter three looks at the multiple-layer shots-and-boxes model for three boxes as well as considering equivalent formulations of the scheme. I establish explicit, non-recursive formulae for the expectation, variance and p.g.f. for the number of occupied boxes in any given layer of the scheme. Chapter four reviews established results for the single-layer process in more detail to provide necessary context before considering a multiple-layer scheme with more boxes. In particular, we are interested in the various limit theorems that have been established and the conditions under which they hold. Chapter five returns to the multiple-layer setting where we first look at small extensions of the three box case before looking at general N. It becomes apparent that even adding one or two boxes adds complexity and many of the methods of proof that worked for three boxes are no longer applicable. This chapter contains the main novel results. This is where I prove that the p.g.f for the number of occupied boxes in any given layer of scheme has all real roots and hence establish both normal and Poisson limiting results. Chapter six mirrors the work undertaken for the number of occupied boxes in this multiple-layer scheme but we instead look at the difference in the number of occupied boxes in consecutive layers. As the literature overwhelmingly treats a single-layer process, this quantity does not naturally arise. Here, I prove that the p.g.f. for the difference in the number of occupied boxes between consecutive layers has all real roots and once again, normal and Poisson limiting results are established. Chapter seven summarises and discusses my work as well as considering future possible extensions.

Chapter 2

Technical background

One of the main aims of this thesis is to establish limit theorems for the distribution of the number of occupied boxes in a given layer of a multiple-layer shots and boxes model. In order to prove such results I will utilise a method of proof that was used by Vatutin and Mikhailov in 1982 [3]. In simple terms their approach was to show that the p.g.f. of interest has all real roots by demonstrating it can be obtained by applying a sequence of real-root preserving transformations to a polynomial which is known to have all real roots. Once we know the p.g.f. has all real roots we have established that the random variable in question is equivalent to a sum of independent Bernoulli random variables (not necessarily with the same probability of success) and then standard limit theorem results follow. In this chapter, various methods of proving a given polynomial has all real roots will be considered and compared with each other.

2.1 Methods of proving a polynomial has all real roots

Here I want to explore various methods of proving a polynomial has all real roots. By considering a small number of boxes in the upcoming chapters, the limitations of some of these approaches will become apparent. The polynomial of specific interest will ultimately be the probability generating function for the number of occupied boxes or empty boxes.

2.1.1 Discriminant

Let Δ_d be the discriminant for a polynomial of degree d. For the quadratic polynomial $ax^2 + bx + c$ we can determine the nature of the roots using $\Delta_2 = b^2 - 4ac$. If $\Delta_2 > 0$ then there are two distinct real roots. For $\Delta_2 = 0$, those roots becomes identical. For the cubic polynomial $ax^3 + bx^2 + cx + d$ we consider $\Delta_3 = 18abcd - 4b^3d + b^2c^2 - 4ac^3 - 27a^2d^2$. If $\Delta_3 > 0$ then the cubic has three distinct real roots. For the quartic polynomial $ax^4 + bx^3 + cx^2 + dx + e$ its roots

are either all real or all non-real if

$$\begin{aligned}\Delta_4 = {} & 256a^3e^3 - 192a^2bde^2 - 128a^2c^2e^2 + 144a^2cd^2e - 27a^2d^4 + 144ab^2ce^2 - 6ab^2d^2e \\ & - 80abc^2de + 18abcd^3 + 16ac^4e - 4ac^3d^2 - 27b^4e^2 + 18b^3cde - 4b^3d^3 - 4b^2c^3e \\ & + b^2c^2d^2 > 0.\end{aligned}$$

2.1.2 Kurtz's theorem

Kurtz established a nice way of checking a polynomial has all real roots provided that the polynomial in question has positive coefficients like a p.g.f. [30]. Consider a polynomial of degree N of the form

$$\gamma_N x^N + \gamma_{N-1}x^{N-1} + \cdots + \gamma_0,$$

where $\gamma_i > 0$ for all $i = 0, 1, \ldots, N$.

Theorem 2.1.1. (Kurtz, Theorem 1 in [30]) Suppose we have a polynomial of degree N ($N \geq 2$) with positive coefficients. If

$$\gamma_i^2 - 4\gamma_{i-1}\gamma_{i+1} > 0,$$

($i = 1, 2, \ldots, N-1$) then the roots of the polynomial are all real and distinct.

That is, Kurtz's theorem says that if we apply the discriminant for a quadratic to each ordered triplet of coefficients and they are all positive then we have all real roots.

2.1.3 Wronskian approach

Let f and g be polynomials with real coefficients. Then, the Wronskian of f and g is defined by

$$W(f,g) = fg' - gf',$$

where $'$ denotes the first derivative.

Theorem 2.1.2. (Alotaibi, Theorem 5.9 in [31]) If we have a pair of polynomials (f_0, f_1) which both have real coefficients then, if the following conditions hold, f_1 will have all real roots.

1. f_0 must have all real roots.
2. The degree of f_1 must be one more than the degree of f_0.
3. $W(f_1, f_0)$ must be negative everywhere.

2.1.4 Vatutin and Mikhailov's approach

The general idea of Vatutin and Mikhailov's approach was to show that the polynomial under consideration can be written as a sequence of real-root-preserving transformations applied to a polynomial known to have all real roots [3]. By real-root-preserving transformations we mean those that preserve the number of real roots (including multiplicities) not the roots themselves. Their method is discussed in detail in Chapter four.

2.1.5 Comparing these methods

We can summarise the main advantage and disadvantage of each of these methods to prove that a polynomial has all real roots.

Method	Main advantage	Main disadvantage
Discriminant	Simple to check for small N	Cannot be used for polynomials of degree greater than four
Kurtz's theorem	Simple to check for small N	It provides a sufficient (but not necessary) condition for real roots and is not satisfied for N above small values
Wronksian	Avoids limitations of the discriminant and Kurtz's theorem	Difficult to generalise to all N and k
Vatutin and Mikhailov's approach	Worked for all N for one layer	Requires that the polynomial can be written as real root-preserving-transformations applied to a polynomial known to have real roots

2.2 Transformations of polynomials that preserve real roots

Definition 2.2.1. Suppose we have a polynomial p of degree α with all real roots (i.e., with α real roots). We shall define a transformation T of p to be *real-root-preserving* if, after applying T to p, the transformed polynomial also has α real roots.

Let $p(x)$ take the general form

$$p(x) = a_\alpha x^\alpha + a_{\alpha-1}x^{\alpha-1} + \cdots + a_1 x + a_0.$$

Define the transformation T_s as follows:

$$T_s[p](x) = p(x+s), \text{ where } s \in \mathbb{R}. \tag{2.1}$$

Note that we are just interested in whether the number of real roots has been preserved and not the values of the roots themselves. As $T_s[p]$ represents a shift of the argument of p by s, if p has all real roots, then so too will $T_s[p]$. Since p has α real roots, we can write it as

$$p(x) = c(x+r_1)(x+r_2)\cdots(x+r_\alpha),$$

where $r_i \in \mathbb{R}$ and $c \in \mathbb{R}$. Then,

$$p(x+s) = c(x+s+r_1)(x+s+r_2)\cdots(x+s+r_\alpha),$$

so $T_s[p]$ will have real roots at $-r_i - s$.
We also define the transformation T_d by

$$T_d[p](x) = x\frac{\mathrm{d}}{\mathrm{d}x}p(x). \tag{2.2}$$

If the polynomial p has α real roots then the derivative $\mathrm{d}p/\mathrm{d}x$ will have $\alpha - 1$ real roots by Rolle's theorem (Chapter 1, [32]). Multiplying by x then gives us an additional root at zero for a total of α real roots.
Let T_f be the transformation which flips the order of the coefficients of the polynomial. That is,

$$T_f[p](x) = x^\alpha p(x^{-1}). \tag{2.3}$$

The transformation T_f preserves the number of real roots when applied to a polynomial with all real non-zero roots. To see why this is true consider the following. Suppose that $r \neq 0$ is a root of

$$p(x) = \sum_{i=0}^{\alpha} a_i x^i, \text{ so } \sum_{i=0}^{\alpha} a_i r^i = 0 \text{ and } \sum_{i=0}^{\alpha} a_i r^{i-\alpha} = 0.$$

Then, changing the summation index by letting $j = \alpha - i$ gives

$$\sum_{j=0}^{\alpha} a_{\alpha-j} r^{-j} = 0.$$

Hence, if $r \neq 0$ is a root of the original polynomial, then $1/r$ is a root of the polynomial after T_f has been applied. Additionally, we know that if r is real then so too is $1/r$, so that if all the roots of the original polynomial are real then so are the roots of the polynomial after applying the transformation T_f. Another key property of the transformation T_f is that T_f is self-inverse when applied to a polynomial with no roots at zero. Suppose $p(x)$ is a polynomial of degree α

which has no roots at zero. Then, $p(x)$ will have the following general form (where $c_0 \neq 0$):

$$p(x) = c_\alpha x^\alpha + c_{\alpha-1}x^{\alpha-1} + c_{\alpha-2}x^{\alpha-2} + \ldots + c_1 x + c_0.$$

Applying T_f once gives

$$T_f[p](x) = x^\alpha \left\{ c_\alpha x^{-\alpha} + c_{\alpha-1}x^{1-\alpha} + c_{\alpha-2}x^{2-\alpha} + \ldots + c_1 x^{-1} + c_0 \right\}.$$

Then, applying T_f again returns the original polynomial,

$$\begin{aligned} T_f \circ T_f[p](x) &= x^\alpha \left\{ c_\alpha + c_{\alpha-1}x^{-1} + c_{\alpha-2}x^{-2} + \ldots + c_1 x^{1-\alpha} + c_0 x^{-\alpha} \right\} \\ &= c_\alpha x^\alpha + c_{\alpha-1}x^{\alpha-1} + c_{\alpha-2}x^{\alpha-2} + \ldots + c_1 x + c_0 \\ &= p(x). \end{aligned}$$

Chapter 3

Multiple-layer shots and boxes scheme with three boxes

3.1 Chapter outline

This chapter begins by setting up the multiple-layer scheme for three boxes and introducing the relevant notation. An example illustrating this sequential allocation process over layers is given in Figure 3.1. I then consider alternative formulations of essentially the same model which is used in different contexts. Before looking at properties of the p.g.f. for the number of occupied boxes in any given layer of the scheme, I first derive non-recursive formulae for the probabilities of having a certain number of occupied boxes. When trying to obtain limiting results for a distribution one approach is to first show that the random variable of interest can be written as a sum of independent Bernoullis. The aim is to show this for general N but first I prove it for this small finite case with three boxes. Finally, a summary of all the results obtained for this multiple-layer scheme with three boxes is provided.

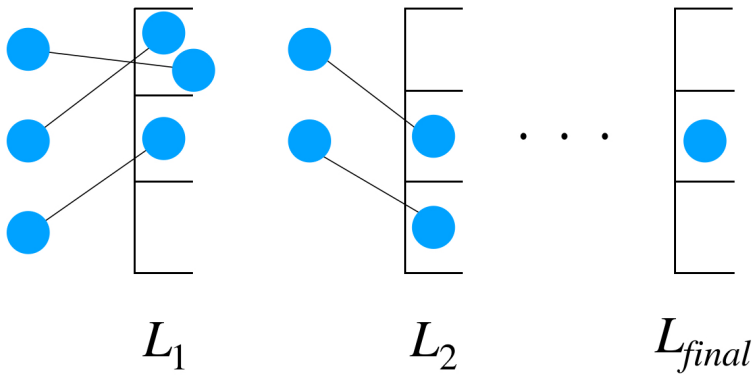

Figure 3.1: Example of the sequential allocation process for three boxes (where L_{final} is the first layer where there is just one occupied box).

3.2 Setting up the scheme

The shots and boxes scheme with multiple layers will be studied. As layers are added or the number of boxes is increased, it becomes more challenging to obtain exact results so we want to start by looking at a simple example. Shots will be allocated independently and uniformly. We start the scheme by throwing three shots into three boxes. If two or more shots land in the same box then they merge together so that there will be either one, two or three occupied boxes. We will call this the first layer of the scheme and the probability distribution for the number of occupied boxes can be easily obtained. Let $R^*(k,N)$ denote the number of occupied boxes in layer k of the scheme with N boxes. Then,

$$\mathbb{P}\Big(R^*(1,3)=1\Big)=1\times\frac{1}{3}\times\frac{1}{3}=\frac{1}{9}.$$

We throw the first shot and it lands in any of the three boxes. The second shot lands in the same box as the first with probability $\frac{1}{3}$. Finally, the third shot lands in this box with probability $\frac{1}{3}$. To have three occupied boxes, each shot that is thrown needs to land in a different box, which happens with probability

$$\mathbb{P}\Big(R^*(1,3)=3\Big)=1\times\frac{2}{3}\times\frac{1}{3}=\frac{2}{9}.$$

That is, we throw the first shot into any box and the second then must land in one of the two other boxes. The third shot then needs to land in the one remaining empty box. Finally,

$$\mathbb{P}\Big(R^*(1,3)=2\Big)=1-\mathbb{P}\Big(R^*(1,3)=1\Big)-\mathbb{P}\Big(R^*(1,3)=3\Big)=1-\frac{1}{9}-\frac{2}{9}=\frac{2}{3}.$$

Alternatively, to have two occupied boxes, we need two of the shots to merge so we need to count the number of ways of selecting these two: $\binom{3}{2}$. So,

$$\mathbb{P}\Big(R^*(1,3)=2\Big)=\binom{3}{2}\times 1\times\frac{1}{3}\times\frac{2}{3}=\frac{2}{3}.$$

These probabilities can be used to compute the expectation of $R^*(1,3)$ in the usual way:

$$\mathbb{E}\Big[R^*(1,3)\Big]=1\mathbb{P}\Big(R^*(1,3)=1\Big)+2\mathbb{P}\Big(R^*(1,3)=2\Big)+3\mathbb{P}\Big(R^*(1,3)=3\Big)=\frac{19}{9}.$$

So far we are still in the well-studied regime where we allocate a fixed number of shots to a fixed number of boxes [2]. Now, we want to add additional layers and this is where our model starts to differ from the established literature. In every layer the number of boxes will be the same, so that here $N=3$ throughout the layers. Also, in every layer, shots are thrown independently according to an equiprobable distribution. From layer two onwards, the number of shots thrown into that layer is given by the number of occupied boxes in the previous layer. As in the single layer shots and boxes scheme we have that if two or more shots lands in the same box in any

given layer then they merge together. Let $S(k)$ be a random variable denoting the number of shots thrown into layer k and let $R^*(k,3)$ be the number of occupied boxes in layer k of the scheme with 3 boxes. Then the number of shots thrown into each layer is dictated as follows:

1. Run the first layer of the scheme and record $R^*(1,3)$.
2. For layer two we allocate $S(2) = R^*(1,3)$ shots.
3. Continue this process throughout the layers where $S(k) = R^*(k-1,3)$. That is, the number of shots thrown into layer k is equal to the number of occupied boxes in the previous layer.
4. So, the probability distribution for $S(k)$ is given by the probability distribution for $R^*(k-1,3)$.

3.2.1 Layer two

To get the probabilities for $R^*(2,3)$ we need to account for the fact that the number of shots is now random. Using the Law of Total Probability,

$$\begin{aligned}
\mathbb{P}\Big(R^*(2,3)=1\Big) &= \sum_{i=1}^{3} \mathbb{P}\Big(R^*(2,3)=1 | S(2)=i\Big)\mathbb{P}\Big(S(2)=i\Big) \\
&= \sum_{i=1}^{3} \mathbb{P}\Big(R^*(2,3)=1 | R^*(1,3)=i\Big)\mathbb{P}\Big(R^*(1,3)=i\Big) \\
&= 1\times\frac{1}{9}+\frac{1}{3}\times\frac{2}{3}+\frac{1}{9}\times\frac{2}{9}=\frac{29}{81}.
\end{aligned}$$

Recall that if two or more shots land in the same box then they merge together so that the number of shots is non-increasing as we move through the layers. Hence, to have two occupied boxes in the second layer we must have thrown at least two shots.

$$\begin{aligned}
\mathbb{P}\Big(R^*(2,3)=2\Big) &= \sum_{i=2}^{3} \mathbb{P}\Big(R^*(2,3)=2 | S(2)=i\Big)\mathbb{P}\Big(S(2)=i\Big) \\
&= \sum_{i=2}^{3} \mathbb{P}\Big(R^*(2,3)=2 | R^*(1,3)=i\Big)\mathbb{P}\Big(R^*(1,3)=i\Big) \\
&= \frac{2}{3}\times\frac{2}{3}+\frac{2}{3}\times\frac{2}{9}=\frac{16}{27}.
\end{aligned}$$

Finally, the only way we can have three occupied boxes in the second layer is if we had three shots to throw.

$$\begin{aligned}\mathbb{P}\Big(R^*(2,3)=3\Big) &= \mathbb{P}\Big(R^*(2,3)=3|S(2)=3\Big)\mathbb{P}\Big(S(2)=3\Big)\\ &= \mathbb{P}\Big(R^*(2,3)=3|R^*(1,3)=3\Big)\mathbb{P}\Big(R^*(1,3)=3\Big)\\ &= \frac{2}{9}\times\frac{2}{9}=\frac{4}{81}.\end{aligned}$$

3.2.2 Layer k

For layer k in general we then have

$$\begin{aligned}\mathbb{P}\Big(R^*(k,3)=1\Big) &= \sum_{i=1}^{3}\mathbb{P}\Big(R^*(k,3)=1|S(k)=i\Big)\mathbb{P}\Big(S(k)=i\Big)\\ &= \sum_{i=1}^{3}\mathbb{P}\Big(R^*(k,3)=1|R^*(k-1,3)=i\Big)\mathbb{P}\Big(R^*(k-1,3)=i\Big)\\ &= \mathbb{P}\Big(R^*(k-1,3)=1\Big)+\frac{1}{3}\mathbb{P}\Big(R^*(k-1,3)=2\Big)+\frac{1}{9}\mathbb{P}\Big(R^*(k-1,3)=3\Big).\\ \mathbb{P}\Big(R^*(k,3)=2\Big) &= \sum_{i=2}^{3}\mathbb{P}\Big(R^*(k,3)=2|S(k)=i\Big)\mathbb{P}\Big(S(k)=i\Big)\\ &= \sum_{i=2}^{3}\mathbb{P}\Big(R^*(k,3)=2|R^*(k-1,3)=i\Big)\mathbb{P}\Big(R^*(k-1,3)=i\Big)\\ &= \frac{2}{3}\mathbb{P}\Big(R^*(k-1,3)=2\Big)+\frac{2}{3}\mathbb{P}\Big(R^*(k-1,3)=3\Big).\\ \mathbb{P}\Big(R^*(k,3)=3\Big) &= \mathbb{P}\Big(R^*(k,3)=3|S(k)=3\Big)\mathbb{P}\Big(S(k)=3\Big)\\ &= \mathbb{P}\Big(R^*(k,3)=3|R^*(k-1,3)=3\Big)\mathbb{P}\Big(R^*(k-1,3)=3\Big)\\ &= \frac{2}{9}\mathbb{P}\Big(R^*(k-1,3)=3\Big).\end{aligned}$$

These are recursive formulae for the occupancy probabilities for each layer which it would be useful to solve. In particular, this will be helpful when trying to prove results that hold for all layers with three boxes. Before doing this we can also look at alternative formulations of our process in terms of Markov chains, iterations of random functions [33] and the Wright-Fisher model [34].

3.2.3 Markov chain formulation

Recall that we are interested in the number of occupied boxes but not which particular boxes are occupied. Additionally, the number shots allocated in layer k is given by the number of occupied boxes in the previous layer. This means that when we consider layer k of the scheme we only

need to refer to layer $k-1$ and can disregard what came before it. As a result of this we can also think of our scheme in terms of Markov chains. Let R^* be the number of occupied boxes. We define the transition probabilities as follows,

$$p_{i,j}^{(k)}(N) = \mathbb{P}(\text{going from } R^* = i \text{ to } R^* = j \text{ in } k \text{ layers in } N \text{ boxes}).$$

The one-layer transition matrix for three boxes is given by T and we can also visualise this as in Figure 3.2.

$$T = \begin{bmatrix} p_{33}^{(1)}(3) & p_{32}^{(1)}(3) & p_{31}^{(1)}(3) \\ p_{23}^{(1)}(3) & p_{22}^{(1)}(3) & p_{21}^{(1)}(3) \\ p_{13}^{(1)}(3) & p_{12}^{(1)}(3) & p_{11}^{(1)}(3) \end{bmatrix} = \begin{bmatrix} 2/9 & 2/3 & 1/9 \\ 0 & 2/3 & 1/3 \\ 0 & 0 & 1 \end{bmatrix}.$$

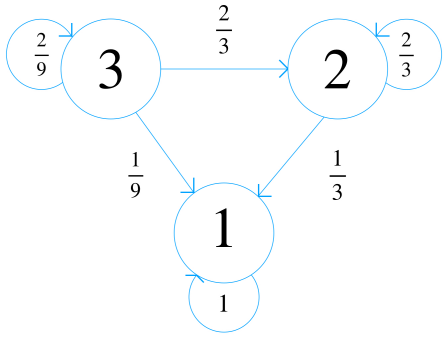

Figure 3.2: Markov chain representation for the three boxes scheme.

Since we start by throwing exactly three shots then the initial distribution is

$$\Pi_0 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}.$$

So, to find the probabilities of having one, two or three occupied boxes after k layers we need

$$\Pi_k = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2/9 & 2/3 & 1/9 \\ 0 & 2/3 & 1/3 \\ 0 & 0 & 1 \end{bmatrix}^k.$$

We can simplify the k-layer transition matrix as follows

$$T^k = \begin{bmatrix} (2/9)^k & a_k & b_k \\ 0 & (2/3)^k & 1-(2/3)^k \\ 0 & 0 & 1 \end{bmatrix},$$

where

$$a_k = a_{k-1}\left\{\left(\frac{2}{9}\right)^{k-1} + \left(\frac{2}{3}\right)^{k-1}\right\},$$

and

$$b_k = b_{k-1}\left[\left(\frac{2}{9}\right)^{k-1} + a_{k-1}\left\{1 - \left(\frac{2}{3}\right)^{k}\right\} + 1\right].$$

Note, that as $a_k \to 0$ and $b_k \to 1$ as $n \to \infty$, so

$$T^{\infty} = \lim_{k\to\infty} T^k = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence, the limiting distribution is

$$\Pi_{\infty} = \Pi_0 T^{\infty} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}.$$

That is, in the limit over layers there will be just one occupied box as expected, since shots are merging.

3.2.4 Iterations of random functions

Let $x_s^{(0)}$ denote the box that shot s is in at the beginning. Suppose we specify that we start with one shot in each box so we can take $x_1^{(0)} = 1$, $x_2^{(0)} = 2$ and $x_3^{(0)} = 3$. Now, we can apply a function f_1 to each of $x_1^{(0)}$, $x_2^{(0)}$ and $x_3^{(0)}$, where $f_1(x_i^{(0)}) = j$ with probability $1/3$ (for $j = 1,2,3$). For step two onwards let f_k be a particular realisation of the random function which maps from i ($i = 1,2,3$) to j with probability $1/3$ for $j = 1,2,3$. Then, after k steps we will have applied a composition of realisations of this random function to the starting points $x_1^{(0)} = 1$, $x_2^{(0)} = 2$ and $x_3^{(0)} = 3$. Note that once two points have been mapped to the same value they will then continue to be mapped to the same point for the remaining steps. For example, suppose f_1 and f_2 have the realisations

$$f_1(1) = 1,\ f_1(2) = 2,\ f_1(3) = 1,$$

$$f_2(1) = 3,\ f_2(2) = 1,\ f_2(3) = 2.$$

Then, after two steps

$$f_2 \circ f_1(x_1^{(0)}) = 3,\ f_2 \circ f_1(x_2^{(0)}) = 1,\ f_2 \circ f_1(x_3^{(0)}) = 3.$$

Now, let $x_s^{(k)}$ denote the box that the shot that started in box s is in after k steps. If we count the number of unique elements in the set

$$\left\{x_1^{(k)}, x_2^{(k)}, x_3^{(k)}\right\},$$

then, this is equivalent to counting the number of occupied boxes in layer k of the shots and boxes scheme with three boxes.

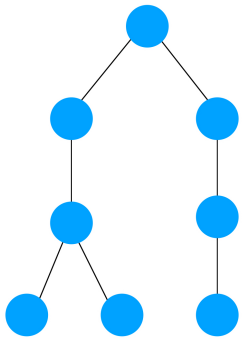

Figure 3.3: Example of a tree structure generated using either model.

Example 3.2.1. Consider the tree shown in Figure 3.3 where the bottom layer corresponds to layer one and as we move upwards the layer index increases. The number of nodes in each layer gives us the number of occupied boxes. Recall we are not interested in the location of the shots, we just want to count the number of occupied boxes. We can calculate the probability of obtaining this tree under our shots and boxes scheme as

$$p_{3,2}^{(1)}(3)p_{2,2}^{(1)}(3)p_{2,1}^{(1)}(3) = \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} = \frac{4}{27}.$$

Now, we can obtain the probability of observing this tree using iterations of random functions. We need our first function to map two of the inputs to the same output (note it does not matter which two as we are just interested in the number of unique elements in the range space of our output). The first input maps to any output with probability one. The second maps to the same value as the first with probability $1/3$. The third input maps to a different output to the first two with probability $2/3$. There are $\binom{3}{2}$ ways of choosing the two inputs that map to the same output so putting all of this together gives the probability for the first step,

$$\binom{3}{2} \times \frac{2}{3} \times \frac{1}{3} = \frac{2}{3}.$$

Now, the two inputs that were mapped to the same output in the first step will automatically go to the same output regardless of what function is used in step two. Then, the remaining input is mapped to a different output with probability $2/3$. Finally, in step three we need that this

remaining input now maps to the same output as the other two which happens with probability $1/3$. Putting all of this together gives the probability of this tree using iterations of random functions,

$$\frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} = \frac{4}{27}.$$

So, the probability of observing this particular tree structure is the same regardless of whether it is generated using the shots and boxes scheme or iterations of random functions.

3.2.5 Wright-Fisher model

Suppose the population size is set to three. Recall the earlier discussion of the connection between the multiple-layer shots and boxes scheme and the Wright-Fisher model (1.3). This allows us to think of the WF model as follows here. In the present, each individual picks an ancestor in the previous generation with probability $1/3$. We repeat this process of descendants picking ancestors uniformly at random from the previous generation. We can then trace the ancestry of our starting sample of three individuals to obtain a tree. Note that the first point where our tree has a single branch is the point at which we reach the MRCA. Counting the number of branches at a particular step in this model then corresponds to counting the number of occupied boxes at a particular layer of the shots and boxes scheme. These two models are equivalent in the sense that the probability of obtaining a certain tree structure will be the same for both.

Example 3.2.2. Suppose we consider the same tree structure as we did for iterations of random functions. In the first layer the first individual picks any ancestor with probability one. The second picks the same ancestor as the first with probability $1/3$. The third picks a different ancestor to the first two with probability $2/3$. We also need to multiply this by the number of ways of choosing the two individuals that pick the same ancestor in the previous generation which is $\binom{3}{2}$. So, for the first layer the probability is,

$$\binom{3}{2} \times \frac{1}{3} \times \frac{2}{3} = \frac{2}{3}.$$

Now, we have two individuals. The first picks any ancestor in the previous generation with probability one. The second picks a different ancestor with probability $2/3$. For the final step we need that the second individual picks the same ancestor as the first which happens with probability $1/3$. Putting this together, the probability of observing this tree structure under the Wright-Fisher model is

$$\frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} = \frac{4}{27}.$$

Note that this is the same probability that we obtained for both the shots and boxes scheme and iterations of random functions.

3.2.6 Finding non-recursive formulae for the occupancy probabilities

Since the only way there can be three occupied boxes in layer k is if we have three shots to throw from the previous layer, it is easy to see that

$$\mathbb{P}\Big(R^*(k,3)=3\Big) = p^{(k)}_{3,3}(3) = \Big(p^{(1)}_{33}\Big)^k = \left(\frac{2}{9}\right)^k.$$

The probability of having two occupied boxes in layer 2 is

$$\mathbb{P}\Big(R^*(2,3)=2\Big) = p^{(1)}_{33}(3)p^{(1)}_{32}(3) + p^{(1)}_{32}(3)p^{(1)}_{22}(3) = \frac{2}{9}\times\frac{2}{3}+\frac{2}{3}\times\frac{2}{3}.$$

For layer three,

$$\begin{aligned}\mathbb{P}\Big(R^*(3,3)=2\Big) &= p^{(1)}_{33}(3)p^{(1)}_{33}(3)p^{(1)}_{32}(3) + p^{(1)}_{33}(3)p^{(1)}_{32}(3)p^{(1)}_{22}(3) + p^{(1)}_{32}(3)p^{(1)}_{22}(3)p^{(1)}_{22}(3)\\ &= \left(\frac{2}{9}\right)^2\times\frac{2}{3}+\frac{2}{9}\times\left(\frac{2}{3}\right)^2+\left(\frac{2}{3}\right)^3.\end{aligned}$$

In general we have,

$$\mathbb{P}\Big(R^*(k,3)=2\Big) = \left(\frac{2}{3}\right)^k + \left(\frac{2}{3}\right)^{k-1}\left(\frac{2}{9}\right) + \left(\frac{2}{3}\right)^{k-2}\left(\frac{2}{9}\right)^2 + \cdots + \left(\frac{2}{3}\right)\left(\frac{2}{9}\right)^{k-1},$$

each term corresponding to the drop from three to two shots in a distinct layer. Using,

$$x^{j-1}+x^{j-2}y+\cdots+xy^{j-2}+y^{j-1} = \frac{x^j-y^j}{x-y} \qquad (x\neq y),$$

we get

$$\mathbb{P}\Big(R^*(k,3)=2\Big) = \frac{\left(\frac{2}{3}\right)^{k+1}-\left(\frac{2}{9}\right)^{k+1}}{\frac{2}{3}-\frac{2}{9}} - \left(\frac{2}{9}\right)^k.$$

Finally, $\mathbb{P}\Big(R^*(k,3)=1\Big)$ follows from the condition that, for any given k,

$$\sum_{i=1}^{3}\mathbb{P}\Big(R^*(k,3)=i\Big) = 1.$$

Now, we have non-recursive formulas for the occupancy probabilities in every layer ($k \geq 1$) of the scheme with three boxes:

$$\mathbb{P}(R^*(k,3)=3) = p_{3,3}^{(k)}(3) = \left(\frac{2}{9}\right)^k, \tag{3.1}$$

$$\mathbb{P}(R^*(k,3)=2) = p_{3,2}^{(k)}(3) = \frac{3}{2}\left\{\left(\frac{2}{3}\right)^k - \left(\frac{2}{9}\right)^k\right\}, \tag{3.2}$$

$$\mathbb{P}(R^*(k,3)=1) = p_{3,1}^{(k)}(3) = 1 - \frac{1}{2}\left\{3\left(\frac{2}{3}\right)^k - \left(\frac{2}{9}\right)^k\right\}. \tag{3.3}$$

As expected, the probability of having only one occupied box increases with k whilst the probability of two or three decreases as we move through the layers (Figure 3.4).

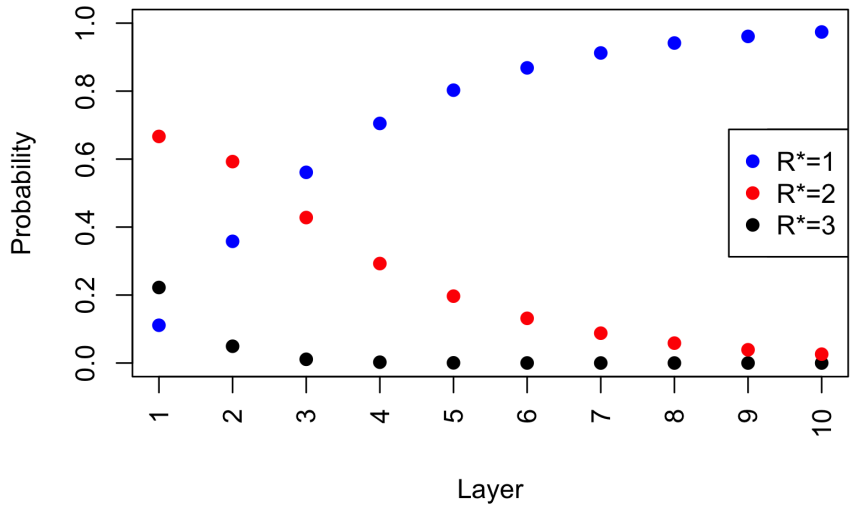

Figure 3.4: Probabilities of occupancy numbers for the first 10 layers when $N = 3$.

Using (3.1)-(3.3) we can obtain a non-recursive formula for the expected number of occupied boxes in layer k:

$$\mathbb{E}\left[R^*(k,3)\right] = 1 + \frac{3}{2}\left(\frac{2}{3}\right)^k + \frac{1}{2}\left(\frac{2}{9}\right)^k.$$

Since

$$\mathbb{E}\left(\{R^*(k,3)\}^2\right) = 1 + \frac{9}{2}\left(\frac{2}{3}\right)^k + \frac{7}{2}\left(\frac{2}{9}\right)^k,$$

then

$$\mathrm{Var}\left(R^*(k,3)\right) = \frac{3}{2}\left(\frac{2}{3}\right)^k + \frac{5}{2}\left(\frac{2}{9}\right)^k - \frac{9}{4}\left(\frac{4}{9}\right)^k - \frac{3}{2}\left(\frac{4}{27}\right)^k - \frac{1}{4}\left(\frac{4}{81}\right)^k.$$

We can consider how both the expectation and variance evolve as we move through layers (Figure 3.5). These non-recursive formulae for the probabilities and the first two moments are not in the existing literature as previous authors were focussing on throwing a fixed number of shots into a single layer, although Kolchin et al. considered a scheme where a random number of shots was allocated to a single layer and produced a formula for the expected number of occupied boxes given the distribution for the number of shots [2].

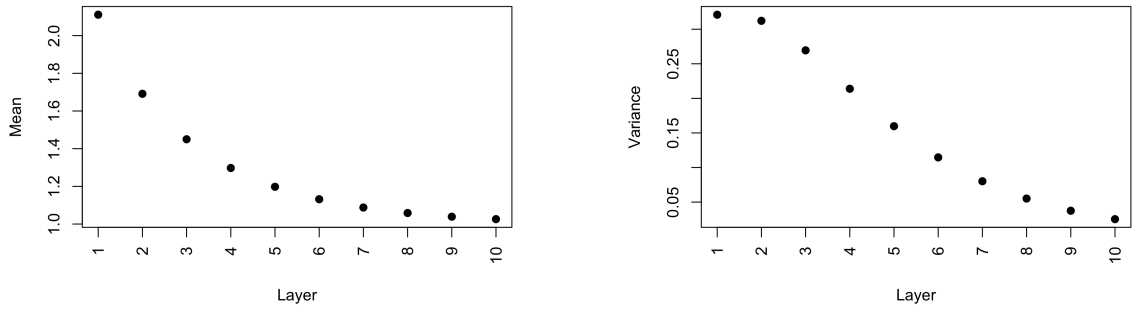

Figure 3.5: Expectation (left) and variance (right) of $R^*(k,3)$ for the first 10 layers.

3.2.7 Parameter reduction

In the first layer we have only three possible transitions to consider:

$$3 \to 1,\ 3 \to 2 \text{ and } 3 \to 3.$$

Then, when we move to the second layer and beyond we also need to consider

$$2 \to 1 \text{ and } 2 \to 2,$$

since, for example.

$$p_{32}^{(2)} = p_{33}^{(1)} p_{32}^{(1)} + p_{32}^{(1)} p_{22}^{(1)}.$$

Recall that the first layer is effectively characterised by two parameters since the three occupancy probabilities must sum to one. The probabilities for two shots in three boxes will have the same constraint so that they are only adding one additional parameter. However, this is where selecting an equiprobable distribution for allocating the shots proves useful. The transition probabilities for three shots can be written in terms of those for two shots. For example, to have one occupied box after throwing three shots we take the probability of having one occupied box after two shots and multiply it by the probability of this third shot landing in the same box as the previous two. That is,

$$p_{31}^{(1)}(3) = p_{21}^{(1)}(3) \times \frac{1}{3}.$$

So

$$p_{21}^{(1)}(3) = 3p_{31}^{(1)}(3) \text{ and } p_{22}^{(1)}(3) = 1 - p_{21}^{(1)}(3) = 1 - 3p_{31}^{(1)}(3).$$

So, the entire probability tree for three boxes for any number of layers can be parameterised in terms of just two probabilities.

3.3 Real roots

3.3.1 Motivation

One approach for proving distributional results of a random variable is to first show that its probability generating function (p.g.f.) has all real roots [3]. As a result of this, the random variable in question can be written as a sum of independent Bernoulli random variables. Then, standard limit theorems can be applied to ascertain, for example, normal or Poisson limiting distributions. Vatutin and Mikhailov [3] proved that for a single layer with a fixed number of boxes and shots the p.g.f. for the number of occupied boxes has all real roots. I want to be able to prove equivalent results for the multi-layer scheme where the additional complexity now is that from the second layer onwards the number of shots thrown is random. Before considering the scenario for general N it will be instructive to provide explicit proofs for the scheme with three boxes.

Discriminant approach

For three boxes there will be either one, two or three occupied boxes so that the p.g.f for the number of occupied boxes in any layer of the process will be of the form

$$a_1x^3 + b_1x^2 + c_1x = x(a_1x^2 + b_1x + c_1).$$

Suppose we impose the condition that this p.g.f. must have real roots. From the above, we certainly have one real root at $x = 0$. If we have all real roots then the discriminant $b_1^2 - 4a_1c_1 \geq 0$. Applying the Law of Total Probability we get that the p.g.f for the number of occupied boxes in the next layer will be

$$c_1x + b_1x\left(\frac{2}{3}x + \frac{1}{3}\right) + a_1x\left(\frac{2}{9}x^2 + \frac{2}{3}x + \frac{1}{9}\right)$$

$$= x\left\{x^2\left(\frac{2}{9}a_1\right) + x\left(\frac{2}{3}b_1 + \frac{2}{3}a_1\right) + \left(c_1 + \frac{1}{3}b_1 + \frac{1}{9}a_1\right)\right\}.$$

Again, there is one real root at $x = 0$. For the remaining two roots, consider the discriminant of the quadratic factor

$$\begin{aligned}\Delta_2 &= \left(\frac{2}{3}b_1 + \frac{2}{3}a_1\right)^2 - 4\left(\frac{2}{9}a_1\right)\left(c_1 + \frac{1}{3}b_1 + \frac{1}{9}a_1\right) \\ &= \frac{28}{81}a_1^2 + \frac{4}{9}b_1^2 + \frac{16}{27}a_1b_1 - \frac{8}{9}a_1c_1.\end{aligned}$$

From the assumption that the p.g.f. for the number of occupied boxes in the previous layer has all real roots, we have

$$b_1^2 \geq 4a_1c_1 \implies \frac{4}{9}b_1^2 \geq \frac{16}{9}a_1c_1 \geq \frac{8}{9}a_1c_1.$$

Therefore,

$$\frac{28}{81}a_1^2 + \frac{4}{9}b_1^2 + \frac{16}{27}a_1b_1 - \frac{8}{9}a_1c_1 \geq \frac{28}{81}a_1^2 + \frac{16}{27}a_1b_1 \geq 0.$$

Hence, the p.g.f. for the number of occupied boxes will have all real roots if the p.g.f. from the previous layer had all real roots. Now, recall for three boxes the p.g.f. for layer one is

$$\frac{2}{9}x^3 + \frac{2}{3}x^2 + \frac{1}{9}x = x\left(\frac{2}{9}x^2 + \frac{2}{3}x + \frac{1}{9}\right).$$

We have one real root at $x = 0$ and

$$\left(\frac{2}{3}\right)^2 - 4\left(\frac{2}{9} \times \frac{1}{9}\right) > 0,$$

so that all the roots are real. Then, by the above proof the roots for the p.g.f. for $R^*(2,3)$ will all be real which in turn means the p.g.f. for $R^*(3,3)$ has all real roots and so on. Therefore, by induction, I have established for three boxes that the p.g.f. for $R^*(k,3)$ will have all real roots for all k. This method works nicely for three boxes but we do not have useful expressions for the discriminant of a polynomial with degree greater than four [35].

Now, we want to explore different methods of showing that the p.g.f. has all real roots for all layers with the aim of extending our proof to schemes with a larger number of boxes. Note the condition of the p.g.f. for the number of thrown shots having all real roots is sufficient but not necessary for the p.g.f. of the number of occupied boxes to have all real roots. We have already shown it is sufficient above but to show that it is not necessary we can consider an example. Note that the coefficients used below were obtained by considering a range of possible values in search of a counter-example.

Example 3.3.1. Suppose the p.g.f. for the number of shots thrown into layer k was

$$x\left(0.2929293x^2 + 0.4713805x + 0.2356902\right).$$

Then, we have one real root at $x = 0$ but

$$0.4713805^2 - 4 \times 0.2929293 \times 0.2356902 < 0,$$

so there are two complex roots. Now, when we obtain the p.g.f. for $R^*(k,3)$ we get

$$0.2356902x + 0.4713805x\left(\frac{2}{3}x + \frac{1}{3}\right) + 0.2929293x\left(\frac{2}{9}x^2 + \frac{2}{3}x + \frac{1}{9}\right)$$

$$\approx x\left(0.065x^2 + 0.510x + 0.425\right).$$

Again, we have one real root at $x = 0$. Now, the discriminant is positive, however. Hence, we have an example where the p.g.f. for $R^*(k,3)$ has all real roots but the p.g.f. for the number of shots thrown had non-real roots.

3.3.2 Kurtz's theorem

We want to find a method of showing that our p.g.f. has all real roots without using the discriminant as we need to apply this to general N. Kurtz established a way of checking a polynomial has all real roots provided that the polynomial in question has positive coefficients [30]. Consider a general polynomial of degree N with the following form

$$\gamma_N x^N + \gamma_{N-1} x^{N-1} + \cdots + \gamma_0,$$

where $\gamma_i > 0$ for all $i = 0, 1, \ldots, N$.

Theorem 3.3.1. (Kurtz, Theorem 1 in [30]) Suppose we have a polynomial of degree N ($N \geq 2$) with positive coefficients. If

$$\gamma_i^2 - 4\gamma_{i-1}\gamma_{i+1} > 0,$$

(where $i = 1, 2, \cdots, N-1$) then the roots of the polynomial are all real and distinct.

Kurtz's theorem says that if we apply the discriminant formula for a quadratic to each ordered triplet of coefficients and they are all positive then we have all real roots. The result applies to polynomials of degree at least two which have all positive coefficients, both of which are properties satisfied by our p.g.f.s. Since we have already seen that the discriminant is positive for all layers for three boxes then we know that Kurtz's theorem applies. However, as we start increasing N, we find that the conditions of this theorem are no longer satisfied. As these are sufficient but not necessary conditions, this does not mean that the p.g.f. does not have real roots in these examples, but we can not use this theorem to prove we do.

Example 3.3.2. Suppose we start with ten boxes and shots and want to check Kurtz's condition holds for the first layer. We obtain the following p.g.f. for the first layer by considering all of the

probabilities individually (the coefficients come from the probability of zero merges, one merge, etc when we start with throwing ten shots into ten boxes),

$$\begin{aligned} g_{R^*(1,10)}(x) = {} & 0.00036288x^{10} + 0.0163296x^9 + 0.13608x^8 + 0.3556224x^7 \\ & + 0.34514424x^6 + 0.1285956x^5 + 0.01718892x^4 + 0.00067176x^3 \\ & + 0.000004599x^2 + 10^{-9}x. \end{aligned}$$

We can now check Kurtz's condition for each triplet of coefficients. It turns out that the condition does not hold here since, for example, when $i = 5$,

$$0.01718892^2 - 4 \times 0.1285956 \times 0.00067176 < 0.$$

Therefore, although Kurtz's method is an effective way of showing the p.g.f. has all real roots for small N, it fails even for ten boxes.

3.3.3 Wronskian approach

Yet another approach that can be used to prove a polynomial has all real roots is using a quantity called the Wronskian. Let f and g be polynomials with real coefficients. Then, the Wronskian of f and g is defined by

$$W(f,g) = fg' - gf',$$

where $'$ denotes the first derivative.

Theorem 3.3.2. (Alotaibi, Theorem 5.9 in [31]) If we have a pair of polynomials (f_0, f_1) which both have real coefficients then, if the following conditions hold, f_1 will have all real roots [31].

1. f_0 must have all real roots.
2. The degree of f_1 must be one more than the degree of f_0.
3. $W(f_1, f_0)$ must be negative everywhere.

Now, before being able to use this theorem we need to decide how to define f_0 and f_1 for each choice of N and k. For now let us fix $N = 3$ and let $f_0^{(k)}$, $f_1^{(k)}$ be the f_0 and f_1 we take for the kth layer. Let $R^*(k,N)$ be the number of occupied boxes in layer k of the scheme for N boxes. Then, the p.g.f. for $R^*(k,3)$ is given by

$$g_{R^*(k,3)}(x) = \sum_{i=1}^{3} \mathbb{P}\Big(R^*(k,3) = i\Big)x^i = x\sum_{i=1}^{3} \mathbb{P}\Big(R^*(k,3) = i\Big)x^{i-1}.$$

Take

$$f_0^{(k)}(x) = \frac{1}{x} g_{R^*(k,2)}(x) = \sum_{i=1}^{2} \mathbb{P}\Big(R^*(k,2) = i\Big)x^{i-1},$$

and

$$f_1^{(k)}(x) = \frac{1}{x} g_{R^*(k,3)}(x) = \sum_{i=1}^{3} \mathbb{P}\Big(R^*(k,3) = i\Big) x^{i-1}. \tag{3.4}$$

Recall we let $p_{ij}^{(k)}(N)$ denote the probability of going from i shots to j over k layers in the scheme with N boxes. For layer k,

$$f_0^{(k)}(x) = p_{22}^{(k)}(2)x + p_{21}^{(k)}(2) = \left(\frac{1}{2}\right)^k x + \left[1 - \left(\frac{1}{2}\right)^k\right].$$

This is because if we have two shots in a given layer then we must have had two shots in all the previous layers since the number of shots is non-increasing as we move through layers. Now, from (3.4),

$$f_1^{(k)}(x) = p_{33}^{(k)}(3)x^2 + p_{32}^{(k)}(3)x + p_{31}^{(k)}(3).$$

The aim of applying this theorem is to show that the p.g.f. for layer k with three boxes has all real roots. This p.g.f. is $xf_1^{(k)}(x)$. There is clearly one real root at $x = 0$ but to show all the roots are real we need to show that f_1 has real roots. To do this with the Wronskian approach, the three conditions of Theorem 3.3.2 must be satisfied.

Lemma 3.3.1. $g_{R^*(1,3)}(x)$ has all real roots.

Proof. For the first layer with three boxes the Wronskian is

$$W\Big(f_1^{(1)}, f_0^{(1)}\Big)(x) = -\frac{1}{9}x^2 - \frac{2}{9}x - \frac{5}{18}.$$

One way of showing this is always negative is to show it has a negative maximum value. To find the maximum of this we take the first derivative and set it equal to zero to get the turning point at $x = -1$ (and the second derivative is negative there so we do indeed have a maximum). So

$$W\Big(f_1^{(1)}, f_0^{(1)}\Big)(x) \leq W\Big(f_1^{(1)}, f_0^{(1)}\Big)(-1) = -\frac{1}{6} < 0.$$

Hence for the first layer the p.g.f. for the number of occupied boxes has all real roots. □

Lemma 3.3.2. $g_{R^*(k,3)}(x)$ has all real roots ($k \geq 1$).

Proof. For any layer k, the Wronskian takes the form

$$\begin{aligned} W\Big(f_1^{(k)}, f_0^{(k)}\Big)(x) = x^2 \left\{-\left(\frac{1}{9}\right)^k\right\} + x\left\{\left(\frac{1}{9}\right)^k (2) + \left(\frac{2}{9}\right)^k (-2)\right\} \\ + \left\{-\frac{3}{2}\left(\frac{2}{3}\right)^k + \left(\frac{1}{2}\right)^k + \frac{3}{2}\left(\frac{2}{9}\right)^k - \left(\frac{1}{9}\right)^k\right\}, \end{aligned}$$

which has a turning point at $x = 1-2^k$ (where the second derivative is negative). Now, plugging this back in will give us the maximum

$$W\left(f_1^{(k)}, f_0^{(k)}\right)\left(1-2^k\right) = -\left(\frac{1}{9}\right)^k\left(1-2^k\right)^2 - \left\{2\left(\frac{2}{9}\right)^k - 2\left(\frac{1}{9}\right)^k\right\}\left(1-2^k\right) + \left\{-\frac{3}{2}\left(\frac{2}{3}\right)^k + \left(\frac{1}{2}\right)^k + \frac{3}{2}\left(\frac{2}{9}\right)^k - \left(\frac{1}{9}\right)^k\right\}.$$

After expanding and simplifying we compare the magnitude of the positive and negative terms

$$+ve : 2^{k+1}\left(\frac{2}{9}\right)^k + \left(\frac{1}{9}\right)^k + \left(\frac{1}{2}\right)^k + \frac{3}{2}\left(\frac{2}{9}\right)^k,$$

$$|-ve| : 2^{2k}\left(\frac{1}{9}\right)^k + 2\left(\frac{2}{9}\right)^k + \frac{3}{2}\left(\frac{2}{3}\right)^k + \left(\frac{1}{9}\right)^k.$$

We have

$$\frac{2^{k+1}\left(\frac{2}{9}\right)^k + \left(\frac{1}{2}\right)^k}{\frac{3}{2}\left(\frac{2}{3}\right)^k} = \frac{4}{3}\left(\frac{2}{3}\right)^k + \frac{2}{3}\left(\frac{3}{4}\right)^k.$$

Now,

$$\frac{4}{3}\left(\frac{2}{3}\right)^k < \frac{1}{2} \text{ for } k \geq 3,$$

$$\frac{2}{3}\left(\frac{3}{4}\right)^k < \frac{1}{2} \text{ for } k \geq 2.$$

Hence, for $k \geq 3$,

$$\frac{2^{k+1}\left(\frac{2}{9}\right)^k + \left(\frac{1}{2}\right)^k}{\frac{3}{2}\left(\frac{2}{3}\right)^k} < 1.$$

Also, we have,

$$\frac{\frac{3}{2}\left(\frac{2}{9}\right)^k}{2\left(\frac{2}{9}\right)^k} = \frac{3}{4} < 1.$$

Therefore we have shown that for $k \geq 3$ the maximum value the Wronskian can take is negative. For $k = 1$ we have already explicitly done these calculations so all that remains is to show this holds for the second layer. In layer two, the maximum occurs at $x = -3$ where

$$W\left(f_1^{(2)}, f_0^{(2)}\right)(-3) = -\frac{79}{324}.$$

□

3.4 Bernoulli sum

Recall that the reason to show that the p.g.f. for the number of occupied boxes has all real roots was that this then means it can be written as a sum of independent Bernoulli random variables. Then, we can apply standard limit theorems to, for example, establish a normal or Poisson limiting distribution. We could instead show directly that the number of occupied boxes in a given layer can be written as a sum of independent Bernoulli random variables. As usual when we have an idea for the general scheme we first test it out on a small number of boxes and ensure two things:

1. the result we are hypothesising does indeed hold for the first layer with three boxes, and,
2. we can prove the result holds for all layers for three boxes.

In the first layer with three boxes we have

$$\mathbb{P}\Big(R^*(1,3)=1\Big) = p^{(1)}_{3,1}(3) = \frac{1}{9},$$

$$\mathbb{P}\Big(R^*(1,3)=2\Big) = p^{(1)}_{3,2}(3) = \frac{2}{3},$$

$$\mathbb{P}\Big(R^*(1,3)=3\Big) = p^{(1)}_{3,3}(3) = \frac{2}{9}.$$

We want to represent $R^*(1,3)$ as a sum of three independent Bernoulli random variables. Let

$$R^*(1,3) = X^{(1)}_1 + X^{(1)}_2 + X^{(1)}_3,$$

where

$$X^{(1)}_1 \sim \text{Be}(1),\ X_2 \sim \text{Be}\left(\theta^{(1)}_2\right),\ X_3 \sim \text{Be}\left(\theta^{(1)}_3\right).$$

Note that one of the Bernoulli random variables always takes the value one since there is always at least one occupied box. Setting the probabilities for each value of $R^*(1,3)$ equal to the corresponding Bernoulli probabilities gives

$$\begin{aligned}
&\mathbb{P}(R^*(1,3)=1) = p^{(1)}_{3,1}(3) = \mathbb{P}(X^{(1)}_2=0)\mathbb{P}(X^{(1)}_3=0),\\
&\mathbb{P}(R^*(1,3)=2) = p^{(1)}_{3,2}(3) = \mathbb{P}(X^{(1)}_2=1)\mathbb{P}(X^{(1)}_3=0) + \mathbb{P}(X^{(1)}_2=0)\mathbb{P}(X^{(1)}_3=1),\\
&\mathbb{P}(R^*(1,3)=3) = p^{(1)}_{3,3}(3) = \mathbb{P}(X^{(1)}_2=1)\mathbb{P}(X^{(1)}_3=1).
\end{aligned}$$

So, for layer one,

$$p_{3,1}^{(1)}(3) = \frac{1}{9} = \left(1-\theta_2^{(1)}\right)\left(1-\theta_3^{(1)}\right),$$
$$p_{3,2}^{(1)}(3) = \frac{2}{3} = \theta_2^{(1)}\left(1-\theta_3^{(1)}\right) + \theta_3^{(1)}\left(1-\theta_2^{(1)}\right),$$
$$p_{3,3}^{(1)}(3) = \frac{2}{9} = \theta_2^{(1)}\theta_3^{(1)}.$$

Solving for $\theta_2^{(1)}$ and $\theta_3^{(1)}$ gives

$$\theta_2^{(1)} = \frac{5+\sqrt{7}}{9} \text{ and } \theta_3^{(1)} = \frac{5-\sqrt{7}}{9}.$$

Therefore,

$$R^*(1,3) = X_1^{(1)} + X_2^{(1)} + X_3^{(1)},$$

with

$$X_1^{(1)} \sim \text{Be}(1),\ X_2^{(1)} \sim \text{Be}\left(\frac{5+\sqrt{7}}{9}\right),\ X_3^{(1)} \sim \text{Be}\left(\frac{5-\sqrt{7}}{9}\right),$$

(where the order of the θs is not important). We now consider the second layer of the process to give an idea of how the θs change through the layers. For the second layer we have to consider the transition probabilities over two layers:

$$\mathbb{P}(R^*(2,3)=1) = p_{3,1}^{(2)}(3) = p_{3,1}^{(1)}(3)p_{1,1}^{(1)}(3) + p_{3,2}^{(1)}(3)p_{2,1}^{(1)}(3) + p_{3,3}^{(1)}(3)p_{3,1}^{(1)}(3) = \frac{29}{81},$$
$$\mathbb{P}(R^*(2,3)=2) = p_{3,2}^{(2)}(3) = p_{3,2}^{(1)}(3)p_{2,2}^{(1)}(3) + p_{3,3}^{(1)}(3)p_{3,2}^{(1)}(3) = \frac{16}{27},$$
$$\mathbb{P}(R^*(2,3)=3) = p_{3,3}^{(2)}(3) = p_{3,3}^{(1)}(3)p_{3,3}^{(1)}(3) = \frac{4}{81}.$$

We write

$$R^*(2,3) = X_1^{(2)} + X_2^{(2)} + X_3^{(2)},$$

where,

$$X_1^{(2)} \sim \text{Be}(1),\ X_2^{(2)} \sim \text{Be}\left(\theta_2^{(2)}\right),\ X_3^{(2)} \sim \text{Be}\left(\theta_3^{(2)}\right).$$

Setting the probabilities for the values of $R^*(2,3)$ equal to the corresponding Bernoulli probabilities and solving gives

$$\theta_2^{(2)} = \frac{28+2\sqrt{115}}{81} \text{ and } \theta_3^{(2)} = \frac{28-2\sqrt{115}}{81}.$$

So, the distribution of $R^*(2,3)$ can be written as a sum of independent Bernoulli random variables with parameters

$$\theta_1^{(2)} = 1,\ \theta_2^{(2)} = \frac{28+2\sqrt{115}}{81},\ \theta_3^{(2)} = \frac{28-2\sqrt{115}}{81}.$$

Repeating this process for the third layer gives

$$\theta_1^{(3)} = 1,\ \theta_2^{(3)} = \frac{164+2\sqrt{5266}}{729},\ \theta_3^{(3)} = \frac{164-2\sqrt{5266}}{729}.$$

Note that there is always one parameter equal to one so that we expect the remaining parameters tend to zero since the number of shots is non-increasing as we move through the layers. This is consistent with the pattern in the first three layers:

$$\text{Layer 1}:\ \theta_1^{(1)} = 1,\ \theta_2^{(1)} \approx 0.85,\ \theta_3^{(1)} \approx 0.26,$$

$$\text{Layer 2}:\ \theta_1^{(2)} = 1,\ \theta_2^{(2)} \approx 0.61,\ \theta_3^{(2)} \approx 0.08,$$

$$\text{Layer 3}:\ \theta_1^{(3)} = 1,\ \theta_2^{(3)} \approx 0.42,\ \theta_3^{(3)} \approx 0.026.$$

But it is not clear what form $\theta_2^{(k)}$ and $\theta_3^{(k)}$ would take in general. We can, however derive a non-recursive formula for the parameters in layer k. Note that

$$\mathbb{E}\left[R^*(k,3)\right] = 1+\theta_2^{(k)}+\theta_3^{(k)} \implies \theta_3^{(k)} = \mathbb{E}\left[R^*(k,3)\right] = 1-\theta_2^{(k)}.$$

Recall,

$$\mathbb{P}\Big(R^*(k,3)=3\Big) = p_{3,3}^{(k)}(3) = \left(\frac{2}{9}\right)^k = \theta_2^{(k)}\theta_3^{(k)}.$$

So, from the above,

$$\theta_2^{(k)}\left(\mathbb{E}\left[R^*(k,3)\right]-1-\theta_2^{(k)}\right) = \left(\frac{2}{9}\right)^k$$

$$\implies \theta_2^{(k)}\mathbb{E}\left[R^*(k,3)\right]-\theta_2^{(k)}-\left\{\theta_2^{(k)}\right\}^2 = \left(\frac{2}{9}\right)^k$$

$$\implies \left\{\theta_2^{(k)}\right\}^2-\theta_2^{(k)}\left\{\mathbb{E}\left[R^*(k,3)\right]-1\right\}+\left(\frac{2}{9}\right)^k = 0.$$

This means that, for layer k, the two θs that will allow us to write $R^*(k,3)$ as a sum of independent Bernoullis are given by

$$\frac{1}{2}\left\{\mathbb{E}\left[R^*(k,3)\right]-1\pm\sqrt{\left\{1-\mathbb{E}\left[R^*(k,3)\right]\right\}^2-4\left(\frac{2}{9}\right)^k}\right\}.$$

Plugging the expression for the expected value of $R^*(k,3)$ into the above gives a non-recursive formula for the two θs in every layer for three boxes:

$$\frac{1}{2}\left[\left(\frac{2}{9}\right)^k+\frac{9}{4}\left\{\left(\frac{2}{3}\right)^{k+1}-\left(\frac{2}{9}\right)^{k+1}\right\}\pm\sqrt{\left\{-\left(\frac{2}{9}\right)^k-\frac{9}{4}\left\{\left(\frac{2}{3}\right)^{k+1}-\left(\frac{2}{9}\right)^{k+1}\right\}\right\}^2-4\left(\frac{2}{9}\right)^k}\right].$$

How important is the structure of the p.g.f.s for $R^*(k,N)$

It can be easier to prove a more general result and here we can try and understand how strict we need to be in the conditions imposed on the p.g.f.s. In particular, instead of focussing on the specific form of the p.g.f.s for $R^*(k,N)$ for three boxes, we could instead start with a general quadratic and try to understand under what conditions these will still have real roots as we move through the layers. We can start off with minimal conditions and try to find a counter-example. If we can find such an example then we need to add more conditions but if we cannot then we can try and prove that these conditions are sufficient. Start with a quadratic with all real roots ($b^2-4ac\geq 0$) whose coefficients are probabilities. We then multiply the coefficients by a sequence of positive, monotonically increasing numbers. The resulting quadratic has coefficients which are probabilities and they sum to one.

Example 3.4.1. Suppose we start with the following quadratic,

$$\frac{1}{5}x^2+\frac{3}{5}x+\frac{1}{5}.$$

Then, the discriminant is

$$\Delta_2=\left(\frac{3}{5}\right)^3-4\left(\frac{1}{5}\right)\left(\frac{1}{5}\right)=\frac{1}{5}>0.$$

Now, we take this first coefficient multiplied by $1/5$, the second multiplied by $2/5$ and the third by $18/5$. This gives us a new quadratic

$$\frac{1}{25}x^2+\frac{6}{25}x+\frac{18}{25}.$$

We can check the discriminant,

$$\Delta_2=\left(\frac{6}{25}\right)^2-4\left(\frac{1}{25}\right)\left(\frac{18}{25}\right)=-\frac{36}{625}<0.$$

These initial conditions were not sufficient and we need to consider something else. Recall there are simple formulae for $\mathbb{P}(N\to N)$ and $\mathbb{P}(N\to 1)$ for the first layer of our scheme which

would correspond to the starting quadratic here. For layer one,

$$\mathbb{P}(N \to N) = \frac{(N-1)!}{N^{N-1}},\ \mathbb{P}(N \to 1) = \frac{1}{N^{N-1}}.$$

Hence,

$$\mathbb{P}(N \to N) = (N-1)!\mathbb{P}(N \to 1).$$

For $N = 3$,

$$\mathbb{P}(3 \to 3) = 2\mathbb{P}(3 \to 1).$$

That is, in addition to our original conditions we could start with a quadratic in which the first coefficient is twice the last one.

Example 3.4.2. Suppose we start with the following quadratic,

$$\frac{2}{6}x^2 + \frac{1}{2}x + \frac{1}{6}.$$

We can check that this does indeed give us real roots,

$$\Delta_2 = \left(\frac{1}{2}\right)^2 - 4\left(\frac{2}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36} > 0.$$

Now, we take the first term multiplied by $3/5$, the second by $4/5$ and the third by $12/5$. Then, the resulting quadratic is,

$$\frac{1}{5}x^2 + \frac{2}{5}x + \frac{2}{5}.$$

We can check the discriminant,

$$\Delta_2 = \left(\frac{2}{5}\right)^2 - 4\left(\frac{1}{5}\right)\left(\frac{2}{5}\right) = -\frac{4}{25} < 0.$$

Again, we can see that we need to impose more conditions. We can add another condition that the first coefficient for the jth transformation is the first coefficient from the quadratic we start with raised to the power $j+1$.

Example 3.4.3. Suppose we start with the following quadratic,

$$\frac{2}{6}x^2 + \frac{1}{2}x + \frac{1}{6}.$$

Then,

$$\Delta_2 = \left(\frac{1}{2}\right)^2 - 4\left(\frac{2}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36} > 0.$$

We can multiply the first coefficient by $2/6$, the second by $3/6$ and the third by $23/6$. We then

get the new quadratic,

$$\frac{1}{9}x^2+\frac{1}{4}x+\frac{23}{36}.$$

We can check if we have real roots,

$$\Delta_2=\left(\frac{1}{4}\right)^2-4\left(\frac{1}{9}\right)\left(\frac{23}{36}\right)=-\frac{287}{1296}<0.$$

Again, we need stronger conditions. We can add the condition that for the first transformation the second coefficient is given by the second coefficient from the quadratic we started with multiplied by the sum of the first two coefficients fro the same quadratic. Then, for transformation j, we have that the second coefficient is equal to the second coefficient from the quadratic we started with multiplied by the sum of the first two coefficients from transformation $j-1$.

Example 3.4.4. Suppose we start with the following quadratic,

$$\frac{2}{6}x^2+\frac{1}{2}x+\frac{1}{6}.$$

We can check that we have real roots,

$$\Delta_2=\left(\frac{1}{2}\right)^2-4\left(\frac{2}{6}\right)\left(\frac{1}{6}\right)=\frac{1}{36}>0.$$

We can multiply the first term by $2/6$, the second by $5/6$ and the third by $17/6$. The resulting quadratic is,

$$\frac{1}{9}x^2+\frac{5}{12}x+\frac{17}{36}.$$

We can check the discriminant,

$$\Delta_2=\left(\frac{5}{12}\right)^2-4\left(\frac{1}{9}\right)\left(\frac{17}{36}\right)=-\frac{47}{1296}<0.$$

We still need more conditions. Suppose we add the condition that the coefficient of the x term in the quadratic we start with is greater than 0.52. We can now prove by induction that after any integer number of transformations under all of the above conditions we preserve the property of having real roots. To be clear our (somewhat contrived) conditions are the following.

1. Start with a quadratic with real roots with coefficients that are probabilities so they are between zero and one and sum to one.

2. Multiply the coefficients by a sequence of positive, monotonically increasing numbers.

3. The resulting quadratic has coefficients which are probabilities so they are between zero and one and sum to one.

4. In the quadratic we start with we require that the first coefficient is twice the last one.

5. The first coefficient for the jth transformation if the first coefficient from the quadratic we start with raised to the power $j+1$.

6. For the first transformation the second coefficient is given by the second coefficient from the quadratic we started with multiplied by the sum of the first coefficients from that same quadratic. Then, for transformation j, we have that the second coefficient is equal to the second coefficient from the quadratic we started with multiplied by the sum of the first two coefficients from transformation $j-1$.

7. The coefficient of the x term in the quadratic we start with is greater than 0.52.

Proof. We start with a quadratic of the form,

$$2ax^2+bx+a,$$

where,

$$\Delta_2 = b^2-4(2a)(a) = b^2-8a^2 > (0.52)^2-8(0.16)^2 = 0.0656.$$

Note that when $b=0.52$ we must have $a=0.16$ since we require that $3a+b=1$. After the first transformation we get,

$$4a^2x^2+b(b+2a)x+(1-4a^2-b^2-2ab).$$

The discriminant of this is,

$$\Delta_2 = \left(b^2+2ab\right)^2-4\left(4a^2\right)\left(1-4a^2-b^2-2ab\right)$$

$$= b^4+4ab^3+4a^2b^2-16a^2+64a^4+16a^2b^2+32a^3b.$$

Since we know that when $b=0.52$ we have $a=0.16$ we can plug this in to get a lower bound for the discriminant:

$$\Delta_2 > \frac{801}{390625} = 0.00205056 > 0.$$

Assumption step: We will assume that the roots are real for the jth transformation. That is, after transformation j we get a quadratic of the form,

$$c_1x^2+c_2x+c_3,$$

where $c_2^2-4c_1c_3>0$ and c_1,c_2,c_3 are all probabilities (between zero and one and sum to one). Since we know the first coefficient follows a nice pattern we can rewrite this as:

$$(2a)^{j+1}x^2+c_2x+c_3.$$

Now, after performing transformation $j+1$ we get,

$$(2a)^{j+2}x^2 + b\left\{(2a)^{j+1} + c_2\right\}x + \left\{1 - (2a)^{j+2} - (2a)^{j+1} - c_2\right\}.$$

Now, we need to check the discriminant of this transformed polynomial,

$$\Delta_2 = (2a)^{2(j+1)} + 2(2a)^{j+1}c_2 + c_2^2 - 4(2a)^{j+2}\left\{1 - (2a)^{j+2} - (2a)^{j+1} - c_2\right\}$$

$$= (2a)^{2(j+1)} + 2(2a)^{j+1}c_2 + c_2^2 - 4(2a)^{j+2} + 4(2a)^{2(j+2)} + 4(2a)^{2j+3} + 4c_2(2a)^{j+2}.$$

By assumption we have,

$$c_2^2 - 4(2a)^{j+1}c_3 > 0.$$

We can rewrite this as,

$$c_2^2 - 4(2a)^{j+1}\left\{1 - c_2 - (2a)^{j+1}\right\} > 0$$

$$\implies c_2^2 - 4(2a)^{j+1} + 4c_2(2a)^{j+1} + 4(2a)^{2(j+1)} > 0.$$

As a result of this,

$$2ac_2^2 - 4(2a)^{j+2} + 4c_2(2a)^{j+2} + 4(2a)^{2j+3} > 0.$$

Now, we know $a < 0.16$ (since $b > 0.52$ and we have that $3a+b=1$) so we have that $c_2^2 > 2ac_2^2$. Using this we can obtain a lower bound for Δ,

$$\Delta_2 > (2a)^{2(j+1)} + 2(2a)^{j+1}c_2 + 2ac_2^2 - 4(2a)^{j+1} + 4(2a)^{2(j+2)} + 4(2a)^{2j+3} + 4c_2(2a)^{j+2}$$

$$= \left\{2ac_2^2 - 4(2a)^{j+2} + 4c_2(2a)^{j+2} + 4(2a)^{2j+3}\right\} + (2a)^{2(j+1)} + 2(2a)^{j+1}c_2 + 4(2a)^{2(j+2)} > 0.$$

From our assumptions we know that the bracketed term is greater than zero and then we are adding a sequence of terms which are all greater than zero (since $a > 0$, $c_2 > 0$) hence the result is obtained. Therefore, by induction we get all real roots for transformation j where $j \in \mathbb{N}$ under the described conditions. □

Therefore, even in the case where we are considering a quadratic, the conditions required to ensure real roots are not straightforward.

3.5 Summary of results for three boxes

Here I recap the main new results from this chapter.

1. The p.g.f. for the number of occupied boxes in any given layer k of the scheme with three boxes is given by,

$$g_{R^*(k,3)}(x) = \left(\frac{2}{9}\right)^k x^2 + \frac{3}{2}\left\{\left(\frac{2}{3}\right)^k - \left(\frac{2}{9}\right)^k\right\}x + \left(1 - \frac{1}{2}\left\{2\left(\frac{2}{3}\right)^k - \left(\frac{2}{9}\right)^k\right\}\right).$$

2. The expected number of occupied boxes in any given layer k of the scheme with three boxes is given by

$$\mathbb{E}[R^*(k,3)] = 1 + \frac{3}{2}\left(\frac{2}{3}\right)^k + \frac{1}{2}\left(\frac{2}{9}\right)^k.$$

3. The variance of the number of occupied boxes in any given layer k of the scheme with three boxes is given by

$$\mathrm{Var}[R^*(k,3)] = \frac{3}{2}\left(\frac{2}{3}\right)^k + \frac{5}{2}\left(\frac{2}{9}\right)^k - \frac{9}{4}\left(\frac{4}{9}\right)^k - \frac{3}{2}\left(\frac{4}{27}\right)^k - \frac{1}{4}\left(\frac{4}{81}\right)^k.$$

4. The p.g.f. of the occupancy number has all real roots for all layers and so the number of occupied boxes in any layer can be expressed as a sum of independent Bernoulli random variables. We have shown this using several different methods:

   i Discriminant approach,

   ii Kurtz's theorem,

   iii Wronskian approach,

Chapter 4

One layer of the shots and boxes scheme for general N

Before trying to work with multiple layers of the scheme with N boxes we first need to fully understand how one layer of this scheme behaves and it has been well-studied both in the finite [2] and infinite cases ([8], [23]). Note that the single layer scheme is simpler since in this case both the number of shots and boxes can be fixed in advance. In the multiple layer scheme we do fix the number of shots and boxes for the first layer but from the second layer onwards the number of shots will be random. This means that even if we start with a simple equiprobable allocation of shots in the first layer we still have to account for the fact that the number of shots will have a probability distribution of its own after the initial allocation.

4.1 Allocation using a uniform probability distribution

We start with the simplest version of the scheme before adding complexity. Suppose we select a uniform distribution so that the chance of a shot landing in any given box is the same and equal to $1/N$. We want to know what happens under this scheme as we let both n and N tend to infinity. Kolchin et al. (Chapter 1, [2]) studied this scheme in detail and were able to identify five domains under which the limiting distribution of the number of empty boxes is either Poisson or normal. In this section we will discuss the conditions required for the number of shots and boxes to be in each of these domains and I will provide examples which satisfy them.

Notation for the finite uniform scheme

Before going any further we shall outline the notation used in this section. Let n be the number of shots and N be the number of boxes. The ratio of shots to boxes will be denoted by α, where $\alpha := n/N$. The number of empty boxes when n shots have been thrown into N boxes is given by $R_0(n,N)$. We shall refer to this simply as R_0 throughout. Similarly, $R_s(n,N)$ is the number

of boxes containing exactly s shots after n shots have been thrown into N boxes. We will refer to this as R_s throughout. Also, the number of occupied boxes after n shots have been allocated to N boxes is given by $R^*(n,N)$ and we will refer to this as R^* for simplicity. The probability any single shot lands in box i is denoted by p_i. Finally, we shall use $\xrightarrow{\text{d}}$ to denote convergence in distribution.

4.1.1 Results for the number of empty boxes

Types of behaviour

Using the ratio of shots to boxes and conditions on the variance of the number of empty boxes we can distinguish between five domains where the behaviour of the distribution of the number of empty boxes differs (Chapter 1, [2]). In particular, it can be shown that the limiting distribution of the number of empty boxes is either Poisson or normal in each of these domains. We shall discuss each of these domains and their corresponding behaviour in detail.

Left-Hand 0-Domain

By definition, if the ratio of the number of shots to boxes tends to zero and the variance of the number of empty boxes tends to a finite positive constant as we increase n and N, then we are in the Left-Hand 0-Domain (I.1, [2]). That is, we are in the Left-Hand 0-Domain if, as $n, N \to \infty$,

$$\alpha \to 0 \qquad \text{and} \qquad \text{Var}(R_0) \to \lambda_0.$$

In the Left-Hand 0-Domain we have a small number of shots relative to the number of boxes meaning that we expect that most shots will land in their own box (since the probability of two or more shots landing in the same box will be small). Additionally, the variance of the number of empty boxes tends to a finite positive constant. Hence, we may expect to observe a Poisson limiting distribution here. When proving the limiting result for this domain, the difference between the number of empty boxes we actually observe and the number of empty boxes we would have if every shot went to its own box was considered. It can be shown that the limiting value of the expectation and variance of this difference is the same as that for R_0 itself. An example of parameter choices which satisfy the necessary conditions of this domain is provided.

Example 4.1.1. Let n be a function of N. Then, an example in the Left-Hand 0-Domain is given by

$$n = c\left\lfloor \sqrt{N} \right\rfloor,$$

where c is a finite positive integer and $\lfloor \sqrt{N} \rfloor$ denotes applying the floor function to the square root of N. Recall that α is the ratio of the number of shots to boxes. To check that we satisfy the conditions for the Left-Hand 0-Domain we need to first ensure that α tends to zero as n and N

increase:

$$\alpha = \frac{n}{N} = \frac{c\left\lfloor\sqrt{N}\right\rfloor}{N} \to 0 \quad (\text{as } n, N \to \infty).$$

Kolchin et al. established an asymptotic formula for the variance of the number of empty boxes (I.1, [2]) and we can check this quantity satisfies the conditions of the Left-Hand 0-Domain:

$$\frac{n^2}{2N} = \frac{\left(c\left\lfloor\sqrt{N}\right\rfloor\right)^2}{2N} \leq \frac{c^2 N}{2N} = \frac{c^2}{2} < \infty.$$

In 1963, Békéssy proved that under the conditions of the Left-Hand 0-Domain, a Poisson limiting distribution is obtained [4]. An alternative proof was later provided in 1982 by Vatutin and Mikhailov [3]. They were able to show this result also holds for a more general scheme where we allocate shots in groups. Here we are considering that we just allocate shots individually so that our groups are all of size one.

Theorem 4.1.1. [4] In the Left-Hand 0-Domain the local limiting distribution of $R_0 - \{N-n\}$ is Poisson. As $n, N \to \infty$,

$$\mathbb{P}(R_0 - \{N-n\} = i) \to \frac{\lambda_0^i e^{-\lambda_0}}{i!},$$

where $\lambda_0 = \lim \mathbb{E}(R_0)$.

Left-Hand Intermediate 0-Domain

By definition, if the ratio of the number of shots to boxes tends to zero and the variance of the number of empty boxes tends to infinity as n and N increase then we are in the Left-Hand Intermediate 0-Domain (I.1, [2]). That is, we are in the Left-Hand Intermediate 0-Domain if, as $n, N \to \infty$,

$$\alpha \to 0 \qquad \text{and} \qquad \text{Var}(R_0) \to \infty.$$

In the Left-Hand Intermediate 0-Domain it is still specified that the number of shots is small compared with the number of boxes. However, in this domain, both the expected value and the variance of the number of empty boxes grow with n and N. This suggests that a Poisson limiting distribution may not be suitable here. To get an idea of what this behaviour looks like in terms of the number of shots being a function of the number of boxes we can consider an example.

Example 4.1.2. Let

$$n = \left\lfloor \sqrt{N}\log(N) \right\rfloor.$$

Note that

$$\alpha = \frac{n}{N} = \frac{\left\lfloor\sqrt{N}\log(N)\right\rfloor}{N} \to 0 \quad (\text{as } n, N \to \infty).$$

We also need to check that the variance condition is satisfied. Using the asymptotic formula for the variance of the number of empty boxes, we can check that this quantity tends to infinity as

N increases:

$$\frac{n^2}{2N} = \frac{\left\lfloor \sqrt{N}\log(N) \right\rfloor^2}{2N} \approx \frac{N\{\log(N)\}^2}{2N} = \frac{\{\log(N)\}^2}{2} \to \infty \text{ as } N \to \infty.$$

The limiting result for the Left-Hand Intermediate 0-Domain was originally proved by Rényi in 1962 [5] as an extension to the result proved for the Central Domain by Weiss [1].

Theorem 4.1.2. [5] In the Left-Hand Intermediate 0-Domain the limiting distribution of the standardised version of R_0 is the standard Normal distribution. That is, as $n, N \to \infty$,

$$\frac{R_0 - \mathbb{E}(R_0)}{\sqrt{\text{Var}(R_0)}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Central Domain

By definition, if the ratio of shots to boxes is bounded above and below by finite positive constants then we are in the Central Domain (I.1, [2]). That, is we are in the Central Domain if, as $n, N \to \infty$,

$$0 < a \leq \alpha \leq b < \infty,$$

where a and b are finite positive constants. In the Central Domain we have that the number of shots and boxes are proportional to one another (as $n, N \to \infty$). It also holds that the expected value and variance of the number of empty boxes tends to infinity as n and N increase. Hence, we may expect to observe a normal limiting distribution here. We can look at an example in this domain by considering a function that fluctuates between finite positive constants.

Example 4.1.3. Suppose

$$n = N\lfloor 2 + \sin(N) \rfloor.$$

The ratio of shots to boxes is bounded above and below by finite positive constants:

$$\alpha = \frac{n}{N} = \frac{N\lfloor 2 + \sin(N) \rfloor}{N} = \lfloor 2 + \sin(N) \rfloor.$$

The limiting distribution in the Central Domain was originally proven in 1958 by Weiss [1] who used the method of moments. Later, in 1967, Kolchin et al. provided an alternative proof using characteristic functions (I.3, [2]).

Theorem 4.1.3. [1] In the Central Domain the limiting distribution of the standardised version of R_0 is the standard Normal distribution. That is, as $n, N \to \infty$,

$$\frac{R_0 - \mathbb{E}(R_0)}{\sqrt{\text{Var}(R_0)}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Right-Hand Intermediate 0-Domain

By definition, if both the ratio of the number of shots to boxes and the expected number of empty boxes tend to infinity as n and N increase then we are in the Right-Hand Intermediate 0-Domain (I.1, [2]). That is, we are in the Right-Hand Intermediate 0-Domain if, as $n, N \to \infty$,

$$\alpha \to \infty \qquad \text{and} \qquad \mathbb{E}(R_0) \to \infty.$$

In the Right-Hand Intermediate 0-Domain there is a large number of shots relative to the number of boxes. Further, the expected value and variance of the number of empty boxes tends to infinity with n and N. Hence, as in the Left-Hand Intermediate 0-Domain and the Central Domain, we may expect a limiting distribution which is normal here.

Example 4.1.4. Consider

$$n = N \lfloor \text{loglog}(N) \rfloor .$$

Then

$$\alpha = \frac{n}{N} = \frac{N \lfloor \text{loglog}(N) \rfloor}{N} = \lfloor \text{loglog}(N) \rfloor \to \infty \ \ (\text{as } n, N \to \infty).$$

We also need to check that the condition on the expectation is satisfied,

$$\mathbb{E}(R_0) = N \left(1 - \frac{1}{N}\right)^{N \lfloor \text{loglog}(N) \rfloor} \to \infty \ \ (\text{as } n, N \to \infty).$$

The limiting result for the Right-Hand Intermediate 0-Domain was originally proven by Rényi in 1962 [5] as an extension to the result proven for the Central Domain by Weiss [1].

Theorem 4.1.4. [5] In the Right-Hand Intermediate 0-Domain the limiting distribution of the standardised version of R_0 is the standard Normal distribution. That is, as $n, N \to \infty$,

$$\frac{R_0 - \mathbb{E}(R_0)}{\sqrt{\text{Var}(R_0)}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Right-Hand 0-Domain

By definition, if the ratio of shots to boxes tends to infinity and the expected number of empty boxes tends to a finite positive constant as n, N increase then we are in the Right-Hand 0-Domain (I.1, [2]). That is, we are in the Right-Hand 0-Domain if, as $n, N \to \infty$,

$$\alpha \to \infty \qquad \text{and} \qquad \mathbb{E}(R_0) \to \lambda_0 < \infty.$$

Finally, in the Right-Hand 0-Domain the number of shots is large relative to the number of boxes. However, the expected value and variance of the number of empty boxes tend to the same finite positive constant. This suggests that a Poisson limiting distribution may be appropriate here.

Example 4.1.5. Consider

$$n = N^2.$$

Note that

$$\alpha = \frac{n}{N} = \frac{N^2}{N} = N \to \infty \text{ as } N \to \infty.$$

The limiting distribution for the Right-Hand 0-Domain was originally proven by Mises in 1939 [36].

Theorem 4.1.5. [36] In the Right-Hand 0-Domain the local limiting distribution of R_0 is Poisson. As $n, N \to \infty$,

$$\mathbb{P}(R_0 = i) \to \frac{\lambda_0^i e^{-\lambda_0}}{i!},$$

where $\lambda_0 = \lim \text{Var}(R_0) = \lim \mathbb{E}(R_0)$.

To summarise (Figure 4.1), we observe a Poisson limiting distribution in both the Left-Hand 0-Domain and the Right-Hand 0-Domain. In the Central Domain and the two Intermediate 0-Domains we saw a limiting distribution which is normal.

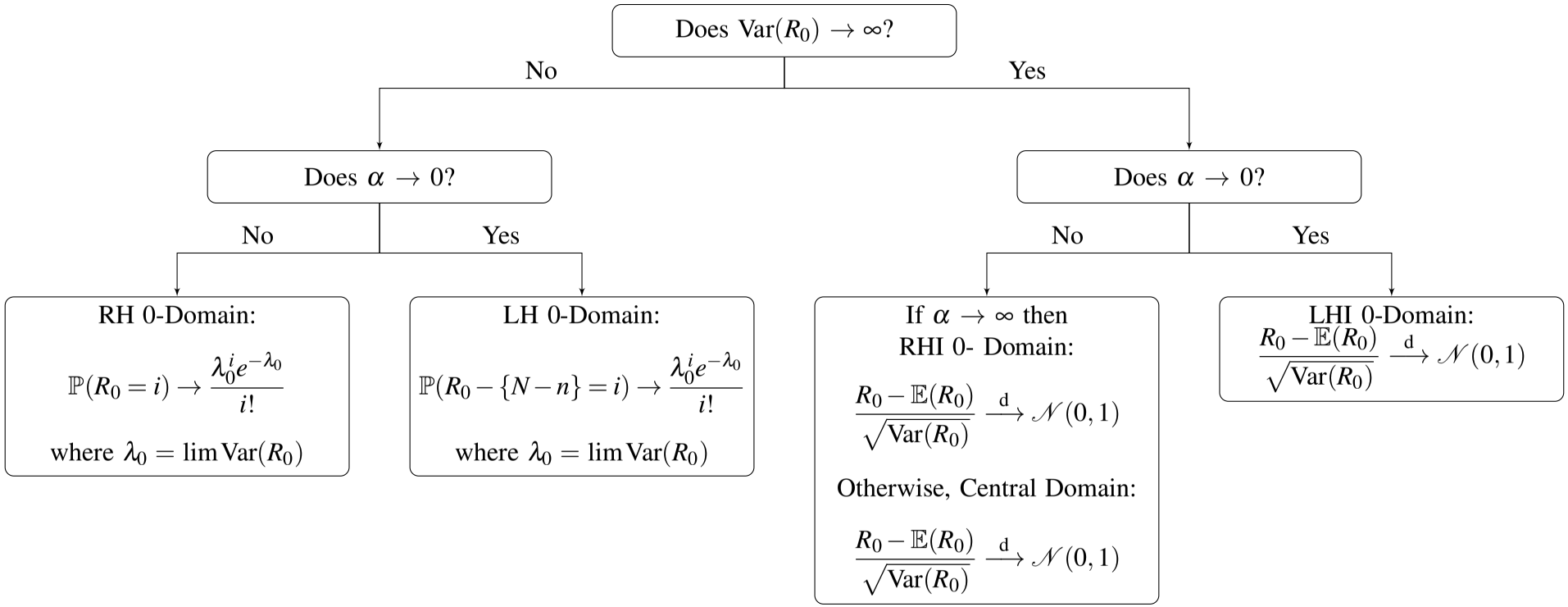

Figure 4.1: Summary of results for the number of empty boxes.

4.1.2 Results for the number of boxes containing two or more shots

Types of behaviour

Note that in what follows the distance is calculated as the sum of the absolute differences between the probabilities in the two distributions being compared. Suppose we are now interested in counting the number of boxes that contain exactly s shots ($s \geq 2$). We can again distinguish between five domains where the scheme behaves differently by imposing certain conditions (II.1, [2]). As before, we will have a strict Left-Hand s-Domain and a strict Right-Hand s-Domain with a Poisson limiting distribution and a Central Domain whose limiting distribution is normal. However, there now are two Intermediate s-Domains with limiting distributions that can be either Poisson or normal. Whether Poisson or normal works better depends on whether we are closer to the strict Left-Hand (Right-Hand) s-Domain or the Central Domain. That is, it has been shown that as we move from the Left-Hand s-Domain to the Central Domain, the distance from the Poisson distribution increases whilst the distance to the normal distribution decreases. Then, as we move from the Central Domain to the Right-Hand s-Domain, the distance from the normal distribution increases whilst the distance to the Poisson distribution decreases (II.5, [2]).

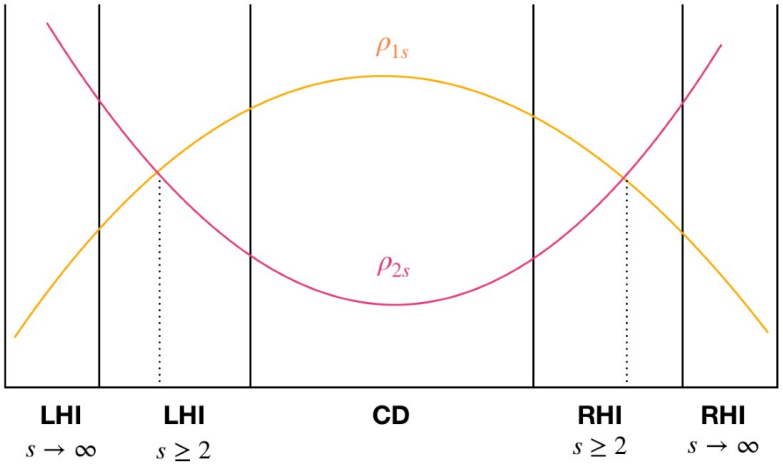

Figure 4.2: Distance to Poisson(ρ_{1s}) and normal distribution(ρ_{2s}) in each domain.

Left-Hand s-Domain

By definition, if the ratio of the number of shots to boxes tends to zero and the expected value of the number of boxes with s shots tends to a finite positive constant as n and N increase then we are in the Left-Hand s-Domain (II.1, [2]). That is, we are in the Left-Hand s-Domain if, as $n, N \to \infty$,

$$\alpha \to 0 \qquad \text{and} \qquad \mathbb{E}(R_s) \to \lambda_s.$$

In the Left-Hand s-Domain the number of shots is small relative to the number of boxes. Also, the expected value and variance of the number of boxes containing s shots tend to the same finite positive constant. Hence, we expect a Poisson limiting distribution may work here. The proof for the limiting distribution of R_s (the number of boxes containing exactly s shots) in the Left-Hand s-Domain was originally given by Kolchin in 1966 [37].

Theorem 4.1.6. [37] In the Left-Hand s-Domain the local limiting distribution of R_s ($s \geq 2$) is Poisson. As $n, N \to \infty$,

$$\mathbb{P}(R_s = i) \to \frac{\lambda_s e^{-\lambda_s}}{i!},$$

where $\lambda_s = \lim \mathrm{Var}(R_s) = \lim \mathbb{E}(R_s)$.

Left-Hand Intermediate s-Domain

By definition, if the ratio of shots to boxes tends to zero and the expected value of the number of boxes containing s shots tends to infinity as n and N increase then we are in the Left-Hand Intermediate s-Domain (II.1, [2]). That is, we are in the Left-Hand Intermediate s-Domain if, as $n, N \to \infty$,

$$\alpha \to 0 \qquad \text{and} \qquad \mathbb{E}(R_s) \to \infty.$$

In the Left-Hand Intermediate s-Domain the number of shots is small relative to the number of boxes. However, now the expected value and variance of the number of boxes containing s shots tends to infinity as n and N increase. Unlike the cases for $s = 0$ and $s = 1$, when we are looking at the number of boxes with s ($s \geq 2$) shots, the limiting distribution in the intermediate domains can be either Poisson or normal. Whether it is closer to Poisson or normal depends on whether we are closer to the strict Left-Hand s-Domain or the Central Domain. The local limiting theorems for the Left-Hand Intermediate s-Domain were proven by Kolchin in 1966 using the saddle-point method [37].

Theorem 4.1.7. [37] Let $s \geq 2$ be fixed and $\mathbb{E}(R_s) \to \infty$ so that we are closer to the Central Domain. Then as n,$N \to \infty$,

$$\mathbb{P}(R_s = k) \sim \frac{1}{\sqrt{2\pi \mathrm{Var}(R_s)}} \exp\left\{ -\frac{1}{2} \frac{[k - \mathbb{E}(R_s)]^2}{\mathrm{Var}(R_s)} \right\}.$$

Theorem 4.1.8. [37] Let $s \geq 2$ be fixed and $\alpha \to 0$ so that we are closer to the strict Left-Hand s-Domain. Then, as $n, N \to \infty$,

$$\mathbb{P}(R_s = k) \to \frac{\lambda_s^k}{k!} e^{-\lambda_s},$$

where $\lambda_s = \lim \mathbb{E}(R_s)$.

Central Domain

By definition, if the ratio of shots to boxes is bounded above and below by finite positive constants then we are in the Central Domain (II.1, [2]). That is, we are in the Central Domain if, as $n\,N \to \infty$,

$$0 < a \leq \alpha \leq b < \infty,$$

where a and b are finite positive constants. In the Central Domain the number of shots and boxes are proportional to one another (as $n, N \to \infty$). Also, in this domain, the expected value and variance of the number of boxes with s shots tend to infinity as n and N grow. Therefore, we expect to have a normal limiting distribution here. The asymptotic normality of R_s in the Central Domain was proven by Békéssy [4]. The method of moments used to prove the asymptotic normality of R_s was discussed by Harris and Park [38].

Theorem 4.1.9. [4] In the Central Domain the limiting distribution of the standardised version of R_s is the standard Normal distribution:

$$\frac{R_s - \mathbb{E}(R_s)}{\sqrt{\mathrm{Var}(R_s)}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

Right-Hand Intermediate s-Domain

By definition, if the ratio of shots to boxes and the expected number of boxes with s shots tend to infinity with n and N then we are in the Right-Hand Intermediate s-Domain (II.1, [2]). That is, we are in the Right-Hand Intermediate s-Domain if, as $n, N \to \infty$,

$$\alpha \to \infty \qquad \text{and} \qquad \mathbb{E}(R_s) \to \infty.$$

In the Right-Hand Intermediate s-Domain there is a large number of shots relative to the number of boxes and both the expected value and the variance of the number of boxes containing exactly s shots tend to infinity as n and N increase. As for the Left-Hand Intermediate s-Domain, the limiting distribution here can be either Poisson or normal. However, here this depends on whether we are closer to the strict Right-Hand s-Domain or the Central Domain. The local limiting theorems for the Right-Hand Intermediate s-Domain were proven by Kolchin in 1966 using the saddle-point method [37].

Theorem 4.1.10. [37] Let $s \geq 2$ be fixed and $\mathbb{E}(R_s) \to \infty$ so that we are closer to the Central Domain. Then, as $n, N \to \infty$,

$$\mathbb{P}(R_s = k) \sim \frac{1}{\sqrt{2\pi \mathrm{Var}(R_s)}} \exp\left\{ -\frac{1}{2} \frac{[k - \mathbb{E}(R_s)]^2}{\mathrm{Var}(R_s)} \right\}.$$

Theorem 4.1.11. [37] Let $s \geq 2$ be fixed and $\alpha \to \infty$ so that we are closer to the strict Right-

Hand s-Domain. Then, as $n, N \to \infty$,

$$\mathbb{P}(R_s = k) \to \frac{\lambda_s^k}{k!} e^{-\lambda_s},$$

where $\lambda_s = \lim \mathbb{E}(R_s)$.

Right-Hand s-Domain

By definition, if the ratio of shots to boxes tends to infinity and the expected number of boxes containing s shots tends to a finite positive constant as n and N increase then we are in the Right-Hand s-Domain (II.1, [2]). That is, we are in the Right-Hand s-Domain if, as $n, N \to \infty$,

$$\alpha \to \infty \qquad \text{and} \qquad \mathbb{E}(R_s) \to \lambda_s < \infty.$$

In the Right-Hand s-Domain the number of shots is large relative to the number of boxes. However, unlike the Right-Hand Intermediate s-Domain, the expected value and variance of R_s tend to the same finite positive constant. Thus, we expect to observe a Poisson limiting distribution here. Erdos and Rényi showed that the limiting distribution of R_s in the Right-Hand s-Domain is Poisson in 1961 [39].

Theorem 4.1.12. [39] In the Right-Hand s-Domain the local limiting distribution of R_s (where $s \geq 2$) is Poisson:

$$\mathbb{P}(R_s = i) \to \frac{\lambda_s^i e^{-\lambda_s}}{i!},$$

where $\lambda_s = \lim \mathrm{Var}(R_s) = \lim \mathbb{E}(R_s)$.

The results for the distribution of the number of boxes containing exactly s ($s \geq 2$) shots are summarised in Figure 4.3.

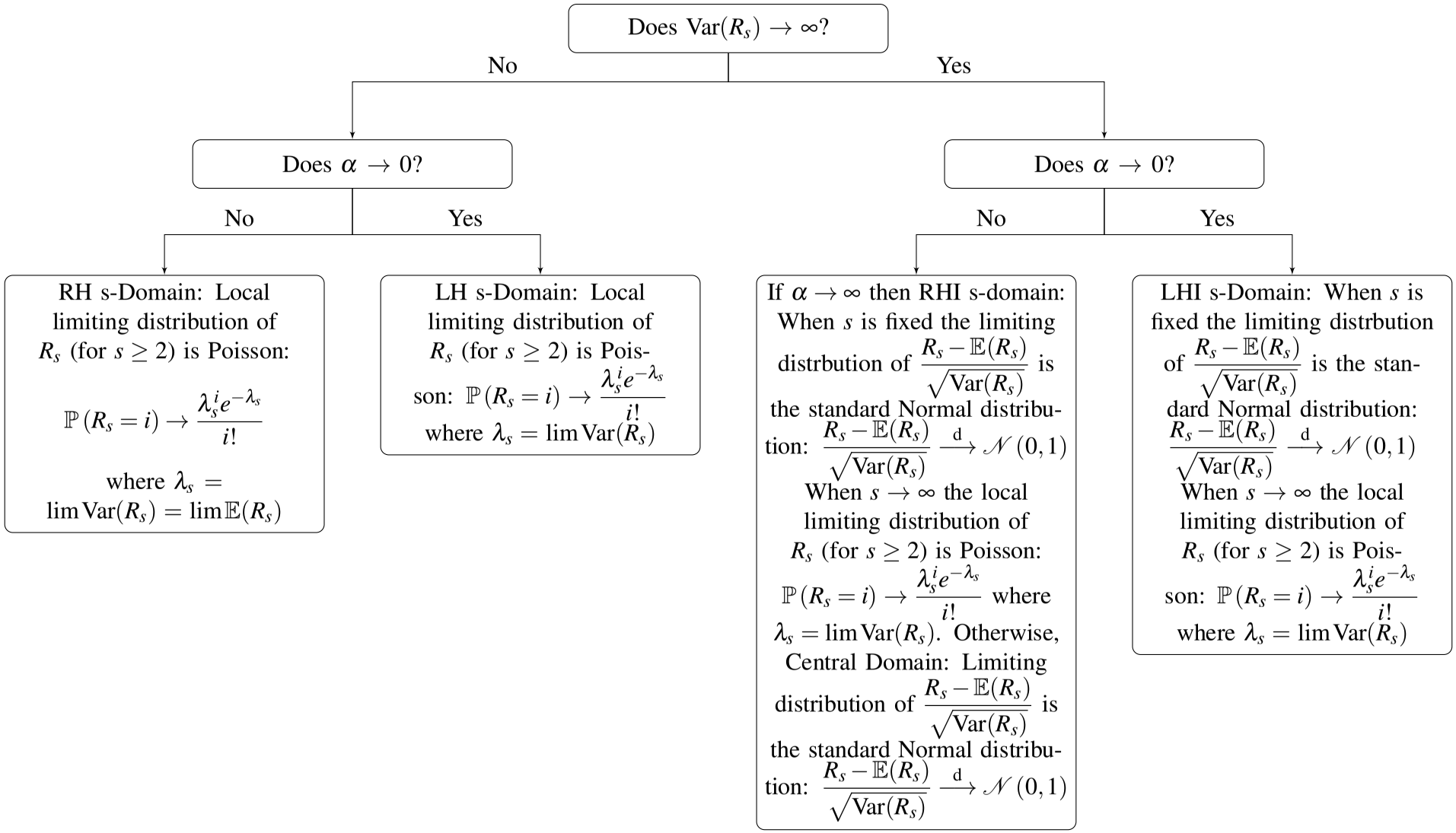

Figure 4.3: Summary of results for R_s where $s \geq 2$.

4.1.3 Relationship between the number of boxes with zero, one or two shots

In the Left-Hand s-Domain, the probability there are any boxes containing more than s shots can be bounded above by the expected number of boxes containing s shots [2]. However, in this domain we have that the expected number of boxes containing more than s shots tends to zero (II.3, [2]). Hence, if we consider, for example, the Left-Hand 2-Domain then the probability of having any boxes containing more than two shots tends to zero. This means that all of the boxes must either contain zero, one or two shots. That is,

$$R_0 + R_1 + R_2 = N \qquad \text{and} \qquad R_1 = n - 2R_2.$$

Then, for,

$$0 < \frac{n^2}{2N} \to \lambda < \infty,$$

the asymptotic properties of R_2, $R_0 - (N-n)$ and $(n-R_1)/2$ are the same. The Right-Hand s-Domain and the Central Domain are defined for all $s \geq 0$ but the Left-Hand s-Domain is only defined for $s \geq 2$. Since we have this relationship described above we can identify the Left-Hand 0-Domain and Left-Hand 1-Domain from the Left-Hand 2-Domain. This is why we considered $R_0 - (N-n)$ instead of just R_0 in the Left-Hand 0-Domain and why we had $(n-R_1)/2$ in the Left-Hand 1-Domain instead of R_1.

4.1.4 Results for the number of boxes containing a single shot

Motivation

Kolchin et al. studied the limiting behaviour of the number of boxes containing a single shot because it exhibited some unusual behaviour, especially as the ratio of shots to boxes tended to zero (II.4, [2]). Later, in 2009, Penrose studied this setting because of its many applications to research questions [20]. Penrose provided a Berry-Esseen bound for the distance between the distribution of R_1 and the normal distribution when allocation is performed using the uniform distribution [20]. Additionally, in the more complex scenario where a non-uniform distribution is used for allocating shots Penrose was still able to bound this distance [20].

Types of behaviour

The number of boxes containing one shot behaves differently from the number of boxes containing at least two shots. We can again use conditions on the ratio of shots to boxes and the variance to define five domains where the behaviour differs (II.4, [2]). Then, the limiting distribution for each domain will be either Poisson or normal.

Left-Hand 1-Domain

By definition, if the ratio of shots to boxes tends to zero and the variance of half the difference between the number of shots and the number of boxes containing one shot tends to a finite positive constant then we are in the Left-Hand 1-Domain (II.4, [2]). That is, we are in the Left-Hand 1-Domain if, as $n, N \to \infty$,

$$\alpha \to 0 \qquad \text{and} \qquad \text{Var}\left(\frac{n-R_1}{2}\right) \to \lambda_1 < \infty.$$

In the Left-Hand 1-Domain we consider half the difference between the number of shots and the number of boxes containing one shot. The reason for considering this is because of the relationship between R_0, R_1 and R_2 in the Left-Hand 2-Domain which we discussed above (II.1, [2]). The number of shots is small relative to the number of boxes and we have that the variance and expected value of the quantity we are considering tend to the same finite positive constant. Hence, we expect to observe a Poisson limiting distribution here.

Theorem 4.1.13. (II.4, [2]) In the Left-Hand 1-Domain the local limiting distribution of $\frac{n-R_1}{2}$ is Poisson. As $n, N \to \infty$,

$$\mathbb{P}\left(\frac{n-R_1}{2} = i\right) \to \frac{\lambda_1^i e^{-\lambda_1}}{i!},$$

where $\lambda_1 = \lim \mathbb{E}\left(\frac{n-R_1}{2}\right)$.

Left-Hand Intermediate 1-Domain

By definition, if the ratio of shots to boxes tends to zero and the variance of the number of boxes with a single shot tends to infinity as n and N increase then we are in the Left-Hand Intermediate 1-Domain (II.4, [2]). That is, we are in the Left-Hand Intermediate 1-Domain if, as $n, N \to \infty$,

$$\alpha \to 0 \qquad \text{and} \qquad \text{Var}(R_1) \to \infty.$$

In the Left-Hand Intermediate 1-Domain the number of shots is small relative to the number of boxes. However, in contrast with the Left-Hand 1-Domain, the variance and expected number of boxes containing one shot tend to infinity as n and N increase. Thus, we may expect to observe a normal limiting distribution here rather than a Poisson one.

Theorem 4.1.14. (II.4, [2]) In the Left-Hand Intermediate 1-Domain the local limiting distribution of R_1 is normal. As $n, N \to \infty$,

$$\mathbb{P}(R_1 = k) \to \frac{1}{\sqrt{2\pi \text{Var}(R_1)}} \exp\left\{-\frac{1}{2}\frac{[k-\mathbb{E}(R_1)]^2}{\text{Var}(R_1)}\right\}.$$

Central Domain

By definition, if the ratio of shots to boxes is bounded above and below by finite positive constants then we are in the Central Domain (II.4, [2]). That is, we are in the Central Domain if, as $n, N \to \infty$,

$$0 < a \leq \alpha \leq b < \infty,$$

where a and b are finite positive constants. In the Central Domain the number of shots and boxes are proportional to one another (as $n, N \to \infty$) and the expected value and variance of the number of boxes containing one shot grow with n and N. Hence, we may expect to have a normal limiting distribution here.

Theorem 4.1.15. (II.4, [2]) In the Central Domain the limiting distribution of the standardised version of R_1 is the standard normal distribution. That is, as $n, N \to \infty$,

$$\frac{R_1 - \mathbb{E}(R_1)}{\sqrt{\text{Var}(R_1)}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Right-Hand Intermediate 1-Domain

By definition, if both the ratio of shots to boxes and the expected number of boxes containing a single shot tend to infinity then we are in the Right-Hand Intermediate 1-Domain (II.4, [2]). That is, we are in the Right-Hand Intermediate 1-Domain if, as $n, N \to \infty$,

$$\alpha \to \infty \qquad \text{and} \qquad \mathbb{E}(R_1) \to \infty.$$

In the Right-Hand Intermediate 1-Domain there is a large number of shots relative to the number of boxes, and the expected value and variance of the number of boxes containing a single shot tend to infinity. Therefore, as in the Left-Hand Intermediate 1-Domain and Central Domain, we anticipate having a limiting distribution which is normal.

Theorem 4.1.16. (II.4, [2]) In the Right-Hand Intermediate 1-Domain the local limiting distribution of R_1 is normal. As $n, N \to \infty$,

$$\mathbb{P}(R_1 = k) \to \frac{1}{\sqrt{2\pi\text{Var}(R_1)}} \exp\left\{ -\frac{1}{2} \frac{[k - \mathbb{E}(R_1)]^2}{\text{Var}(R_1)} \right\}.$$

Right-Hand 1-Domain

By definition, if the ratio of shots to boxes tends to infinity and the expected number of boxes containing one shot tends to a finite positive constant then we are in the Right-Hand 1-Domain (II.4, [2]). That is, we are in the Right-Hand 1-Domain if, as $n, N \to \infty$,

$$\alpha \to \infty \qquad \text{and} \qquad \mathbb{E}(R_1) \to \lambda_1.$$

In the Right-Hand 1-Domain there is a large number of shots compared to the number of boxes. Unlike in the Right-Hand Intermediate 1-Domain the expected value and variance of the number of boxes with one shot tend to the same finite positive constant. Hence, we expect the limiting distribution to be Poisson here.

Theorem 4.1.17. (II.4, [2]) In the Right-Hand 1-Domain the local limiting distribution of R_1 is Poisson. As $n, N \to \infty$,

$$\mathbb{P}(R_1 = i) \to \frac{\lambda_1^i}{i!} e^{-\lambda_1}.$$

To summarise, we expect a Poisson limiting distribution in the Left-Hand 1-Domain and the Right-Hand 1-Domain. In the Central Domain and the two Intermediate 1-Domains the limiting distribution is normal.

4.2 Allocation using a multinomial probability distribution

Setting

We will allocate n shots to N boxes one by one according to a multinomial probability distribution in such a way that the probabilities are all positive and without loss of generality ordered in a non-increasing manner. Note that both n and N are finite here. In the uniform case the parameters that the occupancy numbers depended on were just n and N since every box had the same probability of $1/N$. Now, when we consider this multinomial scheme then the probability distribution also becomes a parameter on which R_0, R_1 and R_s depend. To define the domains where the behaviour is different we previously used the ratio of the number of shots to boxes. However, we must now have equivalent statements involving the probabilities to distinguish, as before, these domains.

4.2.1 Results for the number of empty boxes

Types of behaviour

Previously, when we used a uniform distribution we could completely characterise the behaviour in five distinct domains. When using a multinomial distribution things become more complex and as we shall discuss later in this section, we shall see that there may potentially be more than just five domains. Before looking into this we can first detail the conditions under which we are in the equivalent domains to the uniform scenario (III.1, [2]).

Left-Hand 0-Domain

By definition, if the product of the number of shots and the maximum probability tends to zero and the expected number of empty boxes tends to a finite positive constant as n and N increase

then we are in the Left-Hand 0-Domain. That is, we are in the Left-Hand 0-Domain if, as n, $N \to \infty$,

$$n \times \max p_k \to 0 \qquad \text{and} \qquad \mathbb{E}(R_0) \to \lambda_0 < \infty.$$

In the Left-Hand 0-Domain the product of the number of shots and the maximum probability tends to zero as n and N increase (III.1, [2]). This means that the product of n with any of the probabilities goes to zero as we add more shots and boxes. Further,

$$\max p_k \geq \frac{1}{N} \implies n \times \max p_k \geq \frac{n}{N},$$

$$\text{so,} \quad n \times \max p_k \to 0 \implies \alpha = \frac{n}{N} \to 0.$$

This means that again there is a small number of shots relative to the number of boxes and due to the condition on the product of n and the probabilities we know that we have lots of very small probabilities, much like the uniform setting. Thus, as before, we expect that most shots will go to their own box. In this domain, the variance and expected number of empty boxes tend to the same finite positive constant so we may expect a Poisson limiting distribution here. The Poisson limiting result for the Left-Hand 0-Domain in this multinomial setting was proven by Kolchin et al. [2] using earlier results established by Sevast'yanov in 1972 [43].

Theorem 4.2.1. (III.3, [2]) In the Left-hand 0-Domain the local limiting distribution of $R_0 - \{N-n\}$ is Poisson. As $n, N \to \infty$,

$$\mathbb{P}(R_0 - \{N-n\} = i) \to \frac{\lambda_0^i e^{-\lambda_0}}{i!},$$

where $\lambda_0 = \lim \mathbb{E}(R_0)$.

Left-Hand Intermediate 0-Domain

By definition, if the product of the number of shots and the maximum probability tends to zero and the expected number of empty boxes tends to infinity as n and N increase then we are in the Left-Hand Intermediate 0-Domain. That is, we are in the Left-hand Intermediate 0-Domain if, as $n, N \to \infty$,

$$n \times \max p_k \to 0 \qquad \text{and} \qquad \mathbb{E}(R_0) \to \infty.$$

In the Left-Hand Intermediate 0-Domain the same condition on the product of n and the maximum probability holds but in addition the expected value and variance of the number of empty boxes grows with n and N (III.1, [2]). There are a small number of shots relative to the number of boxes and lots of small probabilities. However, due to the behaviour of the expectation and variance, we anticipate a normal limiting distribution instead of a Poisson one here. The asymptotic normality of R_0 in the Central Domain was proven by Chistyakov in 1964 [40]. Holst then

extended this result to the Left-Hand Intermediate 0-Domain in [41].

Theorem 4.2.2. [41] In the Left-Hand Intermediate 0-Domain the limiting distribution of the standardised version of R_0 is the standard Normal distribution. That is, as $n, N \to \infty$,

$$\frac{R_0 - \mathbb{E}(R_0)}{\sqrt{\text{Var}(R_0)}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Central Domain

By definition, if the ratio of shots to boxes is bounded above and below by finite positive constants and the product of N and any of the probabilities is bounded above by a finite positive constant then we are in the Central Domain. That is, we are in the Central Domain if, as n, $N \to \infty$,

$$0 < a \leq \alpha \leq b < \infty \qquad \text{and} \qquad Np_k \leq c,$$

where a, b and c are finite positive constants. In the Central Domain the number of shots and boxes are still proportional to one another (as $n, N \to \infty$) but in addition the product of N and each probability must be bounded above by a finite positive constant (III.1, [2]). This means that there will be lots of small probabilities as in the uniform setting. Given the above we would expect to obtain a limiting distribution which is normal here. The asymptotic normality of R_0 in the Central Domain was originally proven by Chistyakov in 1964 [40]. A similar theorem was formulated by Kitabatake in 1958 under stronger restrictions on the probabilities [42].

Theorem 4.2.3. [40] In the Central Domain the limiting distribution of the standardised version of R_0 is the standard normal distribution. As $n, N \to \infty$,

$$\frac{R_0 - \mathbb{E}(R_0)}{\sqrt{\text{Var}(R_0)}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Right-Hand Intermediate 0-Domain

By definition, if both the product of the number of shots and the minimum probability and the expected number of empty boxes tend to infinity with n and N then we are in the Right-Hand Intermediate 0-Domain (III.1, [2]). That is, we are in the Right-Hand Intermediate 0-Domain if, as $n, N \to \infty$,

$$n \times \min p_k \to \infty \qquad \text{and} \qquad \mathbb{E}(R_0) \to \infty.$$

In the Right-Hand Intermediate 0-Domain the product of n and the smallest probability tends to infinity with n and N. Further,

$$\min p_k \leq \frac{1}{N} \implies n \times \min p_k \leq \frac{n}{N},$$

$$\text{so} \quad n \times \min p_k \to \infty \implies \frac{n}{N} \to \infty.$$

This means that there are a large number of shots relative to the number of boxes. In this domain the variance and expected number of empty boxes tend to infinity as n and N increase so that, as in the Central Domain, we expect a normal limiting distribution here. In 2008 it was proven by Hwang and Janson that if the variance of the number of empty boxes goes to infinity with n and N then the resulting limiting distribution is normal [22].

Theorem 4.2.4. [22] In the Right-Hand Intermediate 0-Domain the limiting distribution of the standardised version of R_0 is the standard Normal distribution. As $n, N \to \infty$,

$$\frac{R_0 - \mathbb{E}(R_0)}{\sqrt{\text{Var}(R_0)}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Right-Hand 0-Domain

By definition, if the product of the number of shots and the minimum probability tends to infinity and the expected number of empty boxes tends to a finite positive constant as we increase n and N then we are in the Right-Hand 0-Domain (III.1, [2]). That is, we are in the Right-Hand 0-Domain if, as $n, N \to \infty$,

$$n \times \min p_k \to \infty \qquad \text{and} \qquad \mathbb{E}(R_0) \to \lambda_0 < \infty.$$

In the Right-Hand 0-Domain the product of n and the smallest probability tends to infinity with n and N but now the expected value and variance of the number of empty boxes tend to the same finite positive constant. Hence, we may anticipate observing a Poisson limiting distribution here. The Poisson limiting result for the Right-Hand 0-Domain was proven by Kolchin et al. [2] using earlier results established by Sevast'yanov in 1972 [43].

Theorem 4.2.5. (III.3, [2]) In the Right-Hand 0-Domain the local limiting distribution of R_0 is Poisson. As $n, N \to \infty$,

$$\mathbb{P}(R_0 = i) \to \frac{\lambda_0^i e^{-\lambda_0}}{i!},$$

where $\lambda_0 = \lim \mathbb{E}(R_0)$.

To summarise, as in the uniform case, we expect a Poisson limiting distribution in both the Left-Hand and Right-Hand 0-Domains. Then, in the Central Domain and the two Intermediate 0-Domains the limiting distribution is Normal.

4.2.2 Results for the number of boxes containing two or more shots

Left-Hand *s*-Domain

By definition, if the product of the number of shots and the maximum probability tends to zero and the expected number of shots containing s shots tends to a finite positive constant as n and

N increase then we are in the Left-Hand s-Domain (III.1, [2]). That is, we are in the Left-Hand s-Domain if, as $n, N \to \infty$,

$$n \times \max p_k \to 0 \qquad \text{as} \qquad \mathbb{E}(R_s) \to \lambda_s.$$

In the Left-Hand s-Domain there is a small number of shots relative to the number of boxes and lots of small probabilities. The variance and expected value of the number of boxes containing s shots tend to the same finite positive constant. Thus, we expect to have a Poisson limiting distribution here.

Left-Hand Intermediate s-Domain

By definition, if the product of the number of shots and the maximum probability tends to zero and the expected number of boxes with s shots tends to infinity with n and N then we are in the Left-hand Intermediate s-Domain (III.1, [2]). That is, we are in the Left-Hand Intermediate s-Domain if, as $n, N \to \infty$,

$$n \times \max p_k \to 0 \qquad \text{and} \qquad \mathbb{E}(R_s) \to \infty.$$

In the Left-Hand Intermediate s-Domain there is still a small number of shots compared to the number of boxes but now the expected number of boxes containing s shots tends to infinity with n and N. So, we anticipate having a limiting distribution which is normal rather than Poisson here.

Central Domain

By definition, if the ratio of shots to boxes is bounded above and below by finite positive constants and the product of the number of boxes with any of the probabilities is bounded above by a finite positive constant then we are in the Central Domain (III.1, [2]). That is, we are in the Central Domain if, as $n, N \to \infty$,

$$0 < a \leq \alpha \leq b < \infty \qquad \text{and} \qquad Np_k \leq c.$$

In the Central Domain the number of shots and boxes are proportional to one another (as n, $N \to \infty$) and there are lots of small probabilities. It also holds that both the expected value and variance of the number of boxes with s shots grow with n and N, so a normal limiting distribution should work well here.

Right-Hand Intermediate s-Domain

By definition, if the product of the number of shots and the minimum probability and the expected number of boxes containing s shots tends to infinity as n and N increase then we are in the Right-Hand Intermediate s-Domain (III.3, [2]). That is, we are in the Right-Hand Intermediate s-Domain if, as $n, N \to \infty$,

$$n \times \min p_k \to \infty \qquad \text{and} \qquad \mathbb{E}(R_s) \to \infty.$$

In the Right-Hand Intermediate s-Domain there is a large number of shots compared to the number of boxes. The expected number and variance of the number of boxes with s shots tend to infinity as n and N increase. Hence, as in the Central Domain we can anticipate observing a Normal limiting distribution.

Right Hand s-Domain

By definition, if the product of the number of shots and the minimum probability tends to infinity and the expected number of boxes with s shots tends to a finite positive constant then we are in the Right-Hand s-Domain (III.1, [2]). That is, we are in the Right-Hand s-Domain if, as n, $N \to \infty$,

$$n \times \min p_k \to \infty \qquad \text{and} \qquad \mathbb{E}(R_s) \to \lambda_s < \infty.$$

In the Right-Hand s-Domain the number of shots is large relative to the number of boxes. However, now the variance and expected value of the number of boxes containing s shots tend to the same finite positive constant. Therefore, we expect to observe a Poisson limiting distribution here.

4.2.3 Results for the number of boxes containing a single shot

Left-Hand 1-Domain

By definition, if the product of n and the maximum probability tends to zero and the expected value of the number of boxes containing a single shot tends to a finite positive constant then we are in the Left-Hand 1-Domain (III.1, [2]). That is, we are in the Left-Hand 1-Domain if, as n, $N \to \infty$,

$$n \times \max p_k \to 0 \qquad \text{and} \qquad \mathbb{E}\left(\frac{n - R_1}{2}\right) \to \lambda_1 < \infty.$$

In the Left-Hand 1-Domain again there is a small number of shots relative to the number of boxes and lots of small probabilities. Both the variance and expected value of the quantity we are considering tend to a finite positive constant so we may expect to have a Poisson limiting distribution.

Theorem 4.2.6. (III.3, [2]) In the Left-Hand 1-Domain the local limiting distribution of $\frac{1}{2}(n - R_1)$ is Poisson. As $n, N \to \infty$,

$$\mathbb{P}\left(\frac{n-R_1}{2} = i\right) \to \frac{\lambda_1^i e^{-\lambda_1}}{i!},$$

where $\lambda_1 = \lim \mathbb{E}\left(\frac{n-R_1}{2}\right)$.

Left-Hand Intermediate 1-Domain

By definition, if the product of the number of shots and the maximum probability tends to zero and the expected number of boxes occupied by a single shot tends to infinity as n and N increase then we are in the Left-Hand Intermediate 1-Domain (III.1, [2]). That is, we are in the Left-Hand Intermediate 1-Domain if, as $n, N \to \infty$,

$$n \times \max p_k \to 0 \qquad \text{and} \qquad \mathbb{E}(R_1) \to \infty.$$

In the Left-Hand Intermediate 1-Domain the number of shots is still small compared to the number of boxes and there are lots of small probabilities. However, now the variance and expected value of the number of boxes occupied by one shot tend to infinity as n and N grow. Thus, it seems like a normal limiting distribution may be more likely than a Poisson one here.

Central Domain

By definition, if the ratio of the number of shots to boxes is bounded above and below by finite positive constants and the product of the number of boxes and any of the probabilities is bounded above by a finite positive constant then we are in the Central Domain (III.1, [2]). That is, we are in the Central Domain if, as $n, N \to \infty$,

$$0 < a \leq \alpha \leq b < \infty \qquad \text{and} \qquad N p_k \leq c,$$

where a, b and c are finite positive constants. In the Central Domain there are proportional numbers of shots and boxes (as $n, N \to \infty$) and there are lots of small probabilities. In this domain both the expected value and variance of the number of singly occupied boxes grows with n and N. Hence, as in the Left-Hand Intermediate 1-Domain we expect to observe a normal limiting distribution.

Right-Hand Intermediate 1-Domain

By definition, if both the product of the number of shots and the minimum probability and the expected number of boxes containing one shot tend to infinity as n and N increase then we

are in the Right-Hand Intermediate 1-Domain. That is, we are in the Right-Hand Intermediate 1-Domain if, as $n, N \to \infty$,

$$n \times \min p_k \to \infty \qquad \text{and} \qquad \mathbb{E}(R_1) \to \infty.$$

In the Right-Hand Intermediate 1-Domain there is a large number of shots relative to the number of boxes. Both the variance and expected number of boxes which are occupied by a single shot tend to infinity with n and N. Thus, we again anticipate a limiting distribution which is normal.

Right-Hand 1-Domain

By definition, if the product of the number of shots and the minimum probability tends to infinity and the expected number of boxes with a single shot tends to a finite positive constant then we are in the Right-Hand 1-Domain (III.1, [2]). That is, we are in the Right-Hand 1-Domain if, as $n, N \to \infty$,

$$n \times \min p_k \to \infty \qquad \text{and} \qquad \mathbb{E}(R_1) \to \lambda_1 < \infty.$$

In the Right-Hand 1-Domain there is a large number of shots compared with the number of boxes. However, now the expected number and variance of boxes containing one shot tend to the same finite positive constant. Therefore, we expect to have a Poisson limiting distribution here.

Theorem 4.2.7. (III.3, [2]) In the Right-Hand 1-Domain the local limiting distribution of R_1 is Poisson. As $n, N \to \infty$,

$$\mathbb{P}(R_1 = i) \to \frac{\lambda_1^i e^{-\lambda_1}}{i!},$$

where $\lambda_1 = \lim \mathbb{E}(R_1)$.

To summarise, as in the uniform case, we expect to see a Poisson limiting distribution in both the Left-Hand and Right-Hand 1-Domains. In the Central Domain and the two Intermediate 1-Domains the limiting distribution is normal.

Limiting distributions that are neither Poisson or normal

In 1967, Chistyakov proved that under certain conditions in the multinomial setting a limiting distribution which is neither Poisson or normal is obtained [40]. The limiting distribution he derived for the number of empty boxes was the distribution of a sum of a Poisson and Bernoulli random variable.

4.2.4 Review of Vatutin and Mikhailov's (1982) proof

Vatutin and Mikhailov were able to prove for a single layer of the process that the p.g.f. for the number of empty boxes has all real roots [3]. This then led to them establishing various normal

and Poisson limit results. Let $R_0(n,N)$ be a random variable counting the number of empty boxes when n shots are allocated independently to N boxes according to the uniform distribution. We shall refer to this quantity simply as R_0 for the remainder of this section. In order to establish limit theorems for the distribution of R_0 we want to show that R_0 can be written as a sum of N independent (but not identically distributed) Bernoulli random variables. We can do this by showing that the p.g.f. for R_0 has all real roots for all $N,n \in \mathbb{N}$. The main idea of Vatutin and Mikhailov's proof is to show that the p.g.f. can be written as a number of real-root-preserving transformations applied to a polynomial which is known to have all real roots. We know that $(z+1)^N$ has N real roots at -1. Real-root-preserving transformations can be applied to $(z+1)^N$ in order to recover the the p.g.f. for R_0, which we shall denote by g_{R_0}. Note that we include multiplicities when counting the number of real roots. Recall that $\circ$ denotes composition of functions and

$$T_d[p](x) = x\frac{d}{dx}p(x).$$

Also, note that

$$\underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{n \text{ times}}[z^N] = N^n z^N. \tag{4.1}$$

Combining this with the Binomial theorem we have that

$$\begin{aligned}\underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{n \text{ times}}[(z+1)^N] &= \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{n \text{ times}}\left[\sum_{i=0}^{N}\binom{N}{i}1^i z^{N-i}\right] \\ &= \sum_{i=0}^{N}\binom{N}{i}(N-i)^n z^{N-i}. \end{aligned} \tag{4.2}$$

So,

$$\underbrace{T_d \circ T_d \circ \cdots T_d}_{n \text{ times}}(z+1)^N = \sum_{i=0}^{N}\binom{N}{i}(N-i)^n z^{N-i} = z^N v(z^{-1}),$$

where

$$v(z) = \sum_{i=0}^{N}\binom{N}{i}z^i(N-i)^n = \sum_{i=0}^{N-1}\binom{N}{i}z^i(N-i)^n. \tag{4.3}$$

But, note that

$$T_f[z^N v(z^{-1})] = v(z),$$

recalling that T_f is the transformation which flips the order of the coefficients (2.3). We started with $(z+1)^N$ which has all real roots and then applied a sequence of real-root-preserving transformations. First, T_d was applied n times and then finally T_f was applied to produce v. Hence, v has all real roots. We will now show that the p.g.f. for the number of empty boxes can be

obtained by applying real-root-preserving transformations to v. Firstly, let,

$$\eta_i = \begin{cases} 0, & \text{if box } i \text{ is occupied;} \\ 1, & \text{if box } i \text{ is empty.} \end{cases}$$

Then, since $R_0 = \sum_{i=1}^{N} \eta_i$,

$$\mathbb{E}(R_0) = \sum_{i=1}^{N} \mathbb{E}(\eta_i),$$

where

$$\mathbb{E}(\eta_i) = \mathbb{P}(\eta_i = 1) = \left(1 - \frac{1}{N}\right)^n \quad \text{(zero shots land in box } i\text{)}.$$

Hence,

$$\mathbb{E}(R_0) = \sum_{i=1}^{N} \mathbb{E}(\eta_i) = \sum_{i=1}^{N} \left(1 - \frac{1}{N}\right)^n = N\left(1 - \frac{1}{N}\right)^n.$$

We shall use the following notation for a falling factorial:

$$x^{[a]} := x(x-1)(x-2)\cdots(x-a+1),\ x^{[0]} = 1.$$

Note $R_0^{[1]} = R_0$. Recall we defined η_i to be one when box i is empty and zero otherwise. We can show by induction over a that:

$$R_0^{[a]} = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_a=1}^{N} \eta_{i_1}\eta_{i_2}\cdots\eta_{i_a},$$

where the box indices $i_1, i_2, \ldots, i_a$ must be different. For the base case we have already seen that

$$R_0 = \sum_{i=1}^{N} \eta_i.$$

We assume the statement is true for a and now need to prove for $a+1$. Since

$$R_0^{[a+1]} = R_0(R_0-1)(R_0-2)\cdots(R_0-(a+1)+1) = R_0(R_0-1)(R_0-2)\cdots(R_0-a)$$

we have

$$R_0^{[a+1]} = R_0^{[a]}(R_0 - a),$$

where $R_0 - a = 1$ if $R_0 = a+1$ and $R_0 - a = 0$ if $R_0 = a$ so that we can write

$$R_0^{[a+1]} = \sum_{i_1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_{a+1}}^{N} \eta_{i_1}\eta_{i_2}\ldots\eta_{i_a}\eta_{i_{a+1}},$$

where the indices $i_1, i_2, \ldots i_a, i_{a+1}$ are all different. Note that $\eta_{i_1}\eta_{i_2}\cdots\eta_{i_a}$ is only equal to one when all boxes $i_1, i_2, \ldots, i_a$ are empty so

$$\begin{aligned}\mathbb{E}[\eta_{i_1}\eta_{i_2}\cdots\eta_{i_a}] &= \mathbb{P}(\eta_{i_1} = \eta_{i_2} = \cdots = \eta_{i_a} = 1) \\ &= \mathbb{P}(n \text{ shots independently go to boxes other than those with indices } i_1, i_2, \ldots, i_a) \\ &= \left(1 - \frac{a}{N}\right)^n.\end{aligned}$$

Putting this together gives

$$\mathbb{E}\left[R_0^{[a]}\right] = \sum_{i_1}^{N}\sum_{i_2}^{N}\cdots\sum_{i_a}^{N}\left(1 - \frac{a}{N}\right)^n$$

Now we need to count the number of arrangements of indices we can have. For i_1 there are N choices, for i_2 there are $N-1$ and so on. For i_a we will have $N-a+1$ choices so in total we have,

$$N(N-1)\cdots(N-a+1) = N^{[a]} \text{ terms.}$$

Hence,

$$\mathbb{E}\left[R_0^{[a]}\right] = N^{[a]}\left(1 - \frac{a}{N}\right). \tag{4.4}$$

Also (from a Taylor expansion at $z = 1$),

$$g_{R_0}(z) = \mathbb{E}\left(z^{R_0}\right) = \sum_{i=0}^{N-1} g_{R_0}^{(i)}(1)\frac{(z-1)^i}{i!}, \tag{4.5}$$

where $g_{R_0}^{(i)}$ denotes the ith derivative of g_{R_0}. However, since

$$g_{R_0}^{(i)}(z) = \mathbb{E}(R_0^{[i]} z^{R_0 - i}),$$

$$g_{R_0}^{(i)}(1) = \mathbb{E}\left(R_0^{[i]}\right). \tag{4.6}$$

Therefore,

$$\begin{aligned}g_{R_0}(z) &= \sum_{i=0}^{N-1}\mathbb{E}\left(R_0^{[i]}\right)\frac{(z-1)^i}{i!} && \text{[from (4.5) and (4.6)]} \\ &= \sum_{i=0}^{N-1} N^{[i]}\left(1 - \frac{i}{N}\right)^n\frac{(z-1)^i}{i!} && \text{[from (4.4)]} \\ &= \frac{1}{N^n}\sum_{i=0}^{N-1}\binom{N}{i}(z-1)^i(N-i)^n \\ &= \frac{1}{N^n}v(z-1),\end{aligned}$$

Finally,

$$g_{R_0}(z) = \frac{1}{N^n} v(z-1) = \frac{1}{N^n} T_{-1}[v](z),$$

has all real roots since both multiplying by a constant and the transformation T_{-1} preserves the number of real roots (2.1).

Chapter 5

The multiple-layer shots-and-boxes scheme for N boxes

5.1 Chapter outline

We want to obtain results for the multiple-layer scheme for N boxes in general but first we want to understand what happens when we add just one additional box to the scheme with three boxes. Moving from three boxes to four already makes clear how much more complicated things get when trying to prove the p.g.f. for the number of occupied boxes has all real roots. To demonstrate this additional complexity I begin by replicating the results previously obtained for three boxes. Then, with this deeper understanding obtained from considering small examples, I treat the scheme for general N. This is where the most significant novel results of the thesis are established. The proof that the p.g.f. for the number of occupied boxes has all real roots and the consequential limiting results are given.

5.2 Small extensions to the multiple-layer scheme with three boxes

Start by considering the scheme with four boxes. The calculations for the first layer are provided but the steps taken to obtain the explicit formulas for the occupancy probabilities in any given layer are omitted as the same method as was used for three boxes was employed. To still have four shots remaining after allocating shots to the initial layer each shot must have gone to its own box, which happens with probability

$$\mathbb{P}\Big(R^*(1,4)=4\Big)=p^{(1)}_{4,4}(4)=1\times\frac{3}{4}\times\frac{2}{4}\times\frac{1}{4}=\frac{3}{32}.$$

If there are three shots remaining then two shots must have merged and we need to count the number of ways we can select the two that merge:

$$\mathbb{P}\Big(R^*(1,4)=3\Big)=p^{(1)}_{4,3}(4)=\binom{4}{2}\times 1\times\frac{1}{4}\times\frac{3}{4}\times\frac{2}{4}=\frac{9}{16}.$$

If there are two occupied boxes this could have happened if either three shots merged which happens with probability

$$\mathbb{P}\Big(3\text{ shots merged}\Big)=\binom{4}{3}\times 1\times\frac{1}{4}\times\frac{1}{4}\times\frac{3}{4}=\frac{3}{16},$$

or if two boxes contain two shots:

$$\mathbb{P}\Big(2\text{ boxes with 2 shots}\Big)=\binom{4}{2}\times 1\times\frac{1}{4}\times\frac{3}{4}\times\frac{1}{4}\times\frac{1}{2}=\frac{9}{64}.$$

Note that the final factor of $1/2$ comes from the fact that the ordering of the double merges is not important. Putting these together gives

$$\mathbb{P}\Big(R^*(1,4)=2\Big)=p^{(1)}_{4,2}(4)=\frac{3}{16}+\frac{9}{64}=\frac{21}{64}.$$

Finally, the only remaining outcome left to consider is that all shots land in the same box which happens with probability

$$\mathbb{P}\Big(R^*(1,4)=1\Big)=p^{(1)}_{4,1}(4)=1\times\left(\frac{1}{4}\right)^3=\frac{1}{64}.$$

We want to obtain non-recursive formulae for the occupancy probabilities in any given layer k of the scheme with four boxes. Using the same probabilistic arguments that we used for three boxes gives

$$p^{(k)}_{4,4}(4)=\left(\frac{3}{32}\right)^k, \tag{5.1}$$

$$p^{(k)}_{4,3}(4)=2\left\{\left(\frac{3}{8}\right)^k-\left(\frac{3}{32}\right)^k\right\}, \tag{5.2}$$

$$p^{(k)}_{4,2}(4)=\frac{1}{14}\left\{25\left(\frac{3}{4}\right)^k-42\left(\frac{3}{8}\right)^k+17\left(\frac{3}{32}\right)^k\right\}, \tag{5.3}$$

$$p^{(k)}_{4,1}(4)=1-\frac{1}{14}\left\{25\left(\frac{3}{4}\right)^k-14\left(\frac{3}{8}\right)^k+3\left(\frac{3}{32}\right)^k\right\}. \tag{5.4}$$

To visualise what happens to these probabilities as we move through layers we can consider their values for the first ten layers (Figure 5.1). Notice here that, for example, the probability

of having two occupied boxes first increases and then decreases. This contrasts with the three boxes example where we had that each of the occupancy probabilities were either monotonically increasing or decreasing.

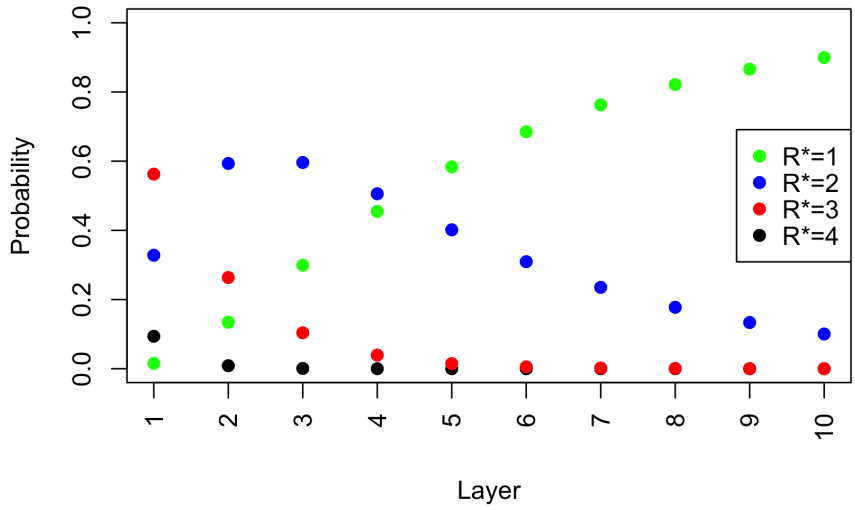

Figure 5.1: How the coefficients of the p.g.f. for four boxes change as we move through layers.

This then gives an expression for the p.g.f. for the number of occupied boxes in any layer of the scheme with four boxes:

$$\begin{aligned} g_{R^*(k,4)}(x) = &\left(\frac{3}{32}\right)^k x^4 + \left(2\left\{\left(\frac{3}{8}\right)^k - \left(\frac{3}{32}\right)^k\right\}\right)x^3 \\ &+ \left(\frac{1}{14}\left\{25\left(\frac{3}{4}\right)^k - 42\left(\frac{3}{8}\right)^k + 17\left(\frac{3}{32}\right)^k\right\}\right)x^2 \\ &+ \left(1 - \frac{1}{14}\left\{25\left(\frac{3}{4}\right)^k - 14\left(\frac{3}{8}\right)^k + 3\left(\frac{3}{32}\right)^k\right\}\right)x. \end{aligned}$$

Using (5.1)-(5.4) we can obtain a non-recursive formula for the expected number of occupied boxes in layer k:

$$\mathbb{E}[R^*(k,4)] = 1 + \frac{25}{14}\left(\frac{3}{4}\right)^k + \left(\frac{3}{8}\right)^k + \frac{3}{14}\left(\frac{3}{32}\right)^k.$$

Since

$$\mathbb{E}\left(\{R^*(k,4)\}^2\right) = 1 + \frac{75}{14}\left(\frac{3}{4}\right)^k + 7\left(\frac{3}{8}\right)^k + \frac{37}{14}\left(\frac{3}{32}\right)^k,$$

then

$$\mathrm{Var}\Big(R^*(k,4)\Big) = \frac{25}{14}\left(\frac{3}{4}\right)^k + 5\left(\frac{3}{8}\right)^k + \frac{31}{14}\left(\frac{3}{32}\right)^k - \frac{625}{196}\left(\frac{9}{16}\right)^k - \frac{25}{7}\left(\frac{9}{32}\right)^k - \frac{75}{98}\left(\frac{9}{128}\right)^k - \left(\frac{9}{64}\right)^k - \frac{3}{7}\left(\frac{9}{256}\right)^k - \frac{9}{196}\left(\frac{9}{1024}\right)^k.$$

Real roots

Discriminant approach

Recall for four boxes the p.g.f. for the number of occupied boxes in layer k is given by

$$x\left(p_{44}^{(k)}(4)x^3 + p_{43}^{(k)}(4)x^2 + p_{42}^{(k)}(4)x + p_{41}^{(k)}(4)\right),$$

where we have non-recursive formulae for $p_{44}^{(k)}(4)$, $p_{43}^{(k)}(4)$, $p_{42}^{(k)}(4)$ and $p_{41}^{(1)}(4)$. Using these we can get a formula for the discriminant of the cubic in layer k,

$$\begin{aligned}\Delta = {} & \frac{625}{49}\left(\frac{81}{1024}\right)^k - 32\left(\frac{27}{512}\right)^k + \frac{450}{7}\left(\frac{27}{1024}\right)^k - \frac{5825}{686}\left(\frac{81}{2048}\right)^k - 12\left(\frac{27}{2048}\right)^k \\ & . - \frac{1054}{49}\left(\frac{81}{4096}\right)^k - 27\left(\frac{9}{1024}\right)^k + \frac{225}{7}\left(\frac{27}{4096}\right)^k - \frac{200}{49}\left(\frac{81}{8192}\right)^k + \frac{12}{7}\left(\frac{27}{8192}\right)^k \\ & - \frac{4523}{1372}\left(\frac{81}{16384}\right)^k - \frac{125}{49}\left(\frac{81}{32768}\right)^k - \frac{1}{7}\left(\frac{27}{32768}\right)^k + \frac{16}{49}\left(\frac{81}{65536}\right)^k - \frac{75}{343}\left(\frac{81}{131072}\right)^k \\ & + \frac{3}{49}\left(\frac{81}{262144}\right)^k + \frac{9}{1372}\left(\frac{81}{1048576}\right)^k.\end{aligned}$$

Lemma 5.2.1. The p.g.f. for the number of occupied boxes in each layer of the scheme with four boxes has all real roots.

Proof. We can show that for $k \geq 1$,

$$\frac{450}{7}\left(\frac{27}{1024}\right)^k > \frac{4523}{1372}\left(\frac{81}{16384}\right)^k + 12\left(\frac{27}{2048}\right)^k + \frac{1054}{49}\left(\frac{81}{4096}\right)^k + 27\left(\frac{9}{1024}\right)^k + \frac{200}{49}\left(\frac{81}{8192}\right)^k + \frac{125}{49}\left(\frac{81}{32768}\right)^k + \frac{1}{7}\left(\frac{27}{32768}\right)^k + \frac{75}{343}\left(\frac{81}{131072}\right)^k.$$

We will proceed using proof by induction. For the base case we have,

$$k = 1: \ \mathrm{LHS} = 1.695033482 > 0.8840953282 = \mathrm{RHS}.$$

Now, we will assume that the result holds for layer k, that is for layer k we assume LHS>RHS.

Finally, using this assumption we need to show that the result holds for layer $k+1$. Firstly,

$$\text{LHS for } k+1 = \frac{27}{1024} \times \text{ LHS for } k.$$

Additionally, we have,

$$\begin{aligned}\text{RHS for layer } k+1 &< \frac{81}{4096} \times \text{ RHS for layer } k \\ &< \frac{81}{4096} \times \text{ LHS for layer } k \\ &< \frac{27}{1024} \times \text{ LHS for layer } k = \text{ LHS for layer } k+1.\end{aligned}$$

Now, in our expression for the discriminant we still have two negative terms that we need to account for. We can show that for $k \geq 3$,

$$\frac{625}{49}\left(\frac{81}{1024}\right)^k > 32\left(\frac{27}{512}\right)^k + \frac{5825}{686}\left(\frac{81}{2048}\right)^k.$$

Again, we shall use proof by induction here. For the base case we have,

$$k = 3: \text{ LHS} = 0.006313048474 > 0.005218128712.$$

We will now assume the statement holds for layer k, that is for layer k we assume LHS>RHS. Now, given this assumption we need to show the result holds for layer $k+1$. We update the LHS as follows,

$$\text{LHS for layer } k+1 = \frac{81}{1024} \times \text{ LHS for layer } k.$$

Then, for the RHS we have,

$$\begin{aligned}\text{RHS for layer } k+1 &< \frac{27}{512} \times \text{RHS for layer } k \\ &< \frac{27}{512} \times \text{ LHS for layer } k \\ &< \frac{81}{1024} \times \text{ LHS for layer } k = \text{LHS for layer } k+1.\end{aligned}$$

All the remaining terms in the discriminant are positive so combining this with the above two proofs tells us that for layer three onwards we have all real roots for four boxes. □

Similar, progressively more tedious, calculations (not shown) can be repeated for the multiple-layer scheme for five boxes. For example, the p.g.f. for any given layer of such a scheme is given

below and the probabilities for the first ten layers are shown in Figure 5.2.

$$
\begin{aligned}
g_{R^*(k,5)}(x) = &\left(\frac{24}{625}\right)^k x^5 + \left(\frac{5}{2}\left\{\left(\frac{24}{125}\right)^k - \left(\frac{24}{625}\right)^k\right\}\right) x^4 \\
&+ \left(\frac{1}{23}\left\{65\left(\frac{12}{25}\right)^k - 115\left(\frac{24}{125}\right)^k + 50\left(\frac{24}{625}\right)^k\right\}\right) x^3 \\
&+ \left(\frac{75}{38}\left(\frac{4}{5}\right)^k - \frac{195}{46}\left(\frac{12}{25}\right)^k + \frac{115}{38}\left(\frac{24}{125}\right)^k - \frac{35}{46}\left(\frac{24}{625}\right)^k\right) x^2 \\
&+ \left(1 - \frac{75}{38}\left(\frac{4}{5}\right)^k + \frac{65}{46}\left(\frac{12}{25}\right)^k - \frac{10}{19}\left(\frac{24}{125}\right)^k + \frac{2}{23}\left(\frac{24}{625}\right)^k\right) x.
\end{aligned}
$$

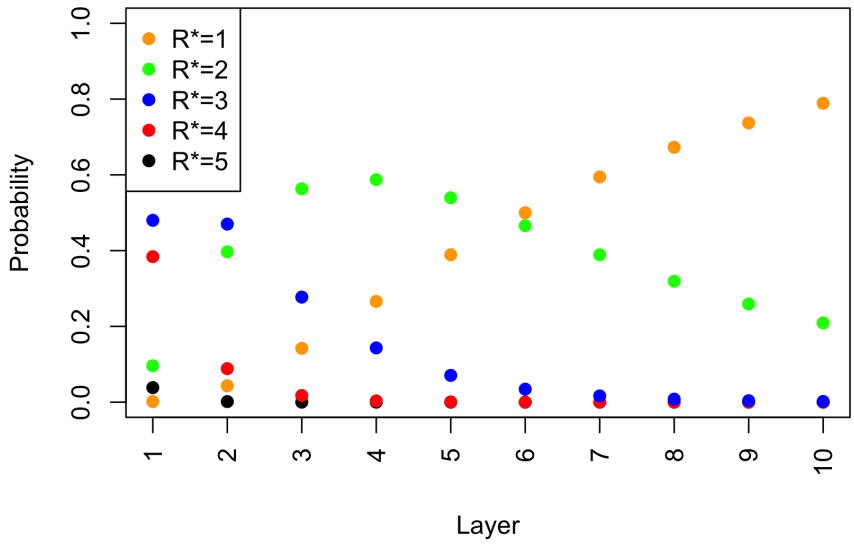

Figure 5.2: How the coefficients of the p.g.f. for five boxes change as we move through layers..

Summary of results for four boxes

The results I have shown for four boxes are summarised below.

1. The p.g.f. for the number of occupied boxes in any given layer k of the scheme with four boxes is given by

$$\begin{aligned} g_{R^*(k,4)}(x) = &\left(\frac{3}{32}\right)^k x^4 + \left(2\left\{\left(\frac{3}{8}\right)^k - \left(\frac{3}{32}\right)^k\right\}\right)x^3 \\ &+ \left(\frac{1}{14}\left\{25\left(\frac{3}{4}\right)^k - 42\left(\frac{3}{8}\right)^k + 17\left(\frac{3}{32}\right)^k\right\}\right)x^2 \\ &+ \left(1 - \frac{1}{14}\left\{25\left(\frac{3}{4}\right)^k - 14\left(\frac{3}{8}\right)^k + 3\left(\frac{3}{32}\right)^k\right\}\right)x. \end{aligned}$$

2. The expected number of occupied boxes in any given layer k of the scheme with four boxes is given by,

$$\begin{aligned} \mathbb{E}\left[R^*(k,4)\right] &= p_{4,1}^{(k)}(4) + 2p_{4,2}^{(k)}(4) + 3p_{4,3}^{(k)}(4) + 4p_{4,4}^{(k)}(4) \\ &= 1 + \frac{25}{14}\left(\frac{3}{4}\right)^k + \left(\frac{3}{8}\right)^k + \frac{3}{14}\left(\frac{3}{32}\right)^k. \end{aligned}$$

3. The variance of the number of occupied boxes in any given layer k of the scheme with four boxes is given by,

$$\begin{aligned} \mathrm{Var}\left[R^*(k,4)\right] = &\frac{25}{14}\left(\frac{3}{4}\right)^k + 5\left(\frac{3}{8}\right)^k + \frac{31}{14}\left(\frac{3}{32}\right)^k - \frac{625}{196}\left(\frac{9}{16}\right)^k - \frac{25}{7}\left(\frac{9}{32}\right)^k \\ &- \frac{75}{98}\left(\frac{9}{128}\right)^k - \left(\frac{9}{64}\right)^k - \frac{3}{7}\left(\frac{9}{256}\right)^k - \frac{9}{196}\left(\frac{9}{1024}\right)^k. \end{aligned}$$

4. The scheme with four boxes has all real roots for all layers and so the number of occupied boxes in any layer can be expressed as a sum of independent Bernoulli random variables. We have shown this using a discriminant approach.

5.3 General formulae for the coefficients of the p.g.f. for the number of occupied boxes

Let $\mathbb{P}\left[R^*(k,N) = N - j\right] = p_{N,N-j}^{(k)}(N)$ denote the probability that the number of occupied boxes has decreased by j over k layers of the scheme with N boxes. For general N we know that when

$j = 0$,

$$\mathbb{P}\Big[R^*(k,N) = N\Big] = p^{(k)}_{N,N}(N) = \Big(p^{(1)}_{N,N}(N)\Big)^k.$$

Probability the occupancy count decreases by one over k layers

$$p^{(k)}_{3,2}(3) = \sum_{i=0}^{k-2} \Big(p^{(1)}_{3,3}(3)\Big)^i p^{(1)}_{3,2}(3) \Big(p^{(1)}_{2,2}(3)\Big)^{k-i-1} + \Big(p^{(1)}_{3,3}(3)\Big)^{k-1} p^{(1)}_{3,2}(3),$$

$$p^{(k)}_{4,3}(4) = \sum_{i=0}^{k-2} \Big(p^{(1)}_{4,4}(4)\Big)^i p^{(1)}_{4,3}(4) \Big(p^{(1)}_{3,3}(4)\Big)^{k-i-1} + \Big(p^{(1)}_{4,4}(4)\Big)^{k-1} p^{(1)}_{4,3}(4),$$

and for general N

$$p^{(k)}_{N,N-1}(N) = \sum_{i=0}^{k-2} \Big(p^{(1)}_{N,N}(N)\Big)^i p^{(1)}_{N,N-1}(N) \Big(p^{(1)}_{N-1,N-1}\Big)^{k-i-1} + \Big(p^{(1)}_{N,N}(N)\Big)^{k-1} p^{(1)}_{N,N-1}(N).$$

Probability that the occupancy count decreases by two over k layers

$$\begin{aligned} p^{(k)}_{4,2}(4) = &\sum_{i=0}^{k-2} \Big(p^{(1)}_{4,4}(4)\Big)^i p^{(1)}_{4,2}(4) \Big(p^{(1)}_{2,2}(4)\Big)^{k-i-1} + \sum_{i=0}^{k-2} \Big(p^{(1)}_{4,4}(4)\Big)^i p^{(1)}_{4,3}(4) p^{(k-i-1)}_{3,2}(4) \\ &+ \Big(p^{(1)}_{4,4}(4)\Big)^{k-1} p^{(1)}_{4,2}(4), \end{aligned}$$

$$\begin{aligned} p^{(k)}_{5,3}(5) = &\sum_{i=0}^{k-2} \Big(p^{(1)}_{5,5}(5)\Big)^i p^{(1)}_{5,4}(5) p^{(k-i-1)}_{4,3}(5) + \sum_{i=0}^{k-2} \Big(p^{(1)}_{5,5}(5)\Big)^i p^{(1)}_{5,3}(5) \Big(p^{(1)}_{3,3}(5)\Big)^{k-i-1} \\ &+ \Big(p^{(1)}_{5,5}(5)\Big)^{k-1} p^{(1)}_{5,3}(5), \end{aligned}$$

and for general N

$$\begin{aligned} p^{(k)}_{N,N-2}(N) = &\sum_{i=0}^{k-2} \Big(p^{(1)}_{N,N}(N)\Big)^i p^{(1)}_{N,N-1}(N) p^{(k-i-1)}_{N-1,N-2}(N) \\ &+ \sum_{i=0}^{k-2} \Big(p^{(1)}_{N,N}(N)\Big)^i p^{(1)}_{N,N-2}(N) \Big(p^{(1)}_{N-2,N-2}(N)\Big)^{k-i-1} + \Big(p^{(1)}_{N,N}(N)\Big)^{k-1} p^{(1)}_{N,N-2}(N). \end{aligned}$$

The general idea is that we consider the possible drops in the number of occupied boxes we can have that would result in us having a certain number of boxes occupied after k layers. We need to sum over both the different possible drops and over the different locations where these drops can happen. For $j \geq 1$ we have (where m is the index for the drop and i is the index for the

location of that drop),

$$\mathbb{P}(R^*(k,N)=N-j)=p^{(k)}_{N,N-j}(N)=\sum_{m=1}^{j}\sum_{i=0}^{k-2}\left(p^{(1)}_{N,N}(N)\right)^i p^{(1)}_{N,N-m}(N)p^{(k-i-1)}_{N-m,N-j}(N) + \left(p^{(1)}_{N,N}(N)\right)^{k-1} p^{(1)}_{N,N-j}(N). \tag{5.5}$$

So, the general form of the p.g.f. for the number of occupied boxes in layer k in the scheme for N boxes is given by,

$$G_{R^*(k,N)}(x)=\left(p^{(1)}_{N,N}(N)\right)^k x^N+\sum_{j=1}^{N-1}x^{N-j}p^{(k)}_{N,N-j}(N).$$

5.3.1 Proving the p.g.f. for the number of occupied boxes has all real roots for all layers

We want to use the same method as Vatutin and Mikhailov (4.2.4, [3]) which means we need to be able to write the p.g.f. for any given layer of the scheme for N boxes as a sequence of real-root-preserving transformations applied to a polynomial which is known to have all real roots. Recall, in layer one we throw n shots into N boxes and from layer two onwards the number of shots thrown into a given layer is equal to the number of occupied boxes in the previous layer. Let the probability of going from the initial n shots to i occupied boxes in k layers be denoted by $p^{(k)}_{n,i}(N)$. Since we start with n shots in layer one, $p^{(k)}_{n,i}(N)$ is just the probability of having i occupied boxes (out of N) in layer k of the scheme. Let $R^*(k,N,n)$ be the number of occupied boxes in layer k of the scheme with N boxes and n shots initially. We shall denote this by $R^*(k)$ when n and N are fixed. Now, let

$$g_{R^*(k)}(z)=\sum_{j=1}^{N}\mathbb{P}\Big[R^*(k)=j\Big]z^j=\sum_{j=1}^{N}p^{(k)}_{n,j}(N)z^j \tag{5.6}$$

be the (unconditional) p.g.f. for the number of occupied boxes in layer k of the scheme with N boxes and n shots initially. Similarly, let $R_0(k,N,n)$ denote the number of empty boxes in layer k of the scheme with N boxes and n shots initially, shortened to $R_0(k)$. Note that $R_0(k)+R^*(k)=N$. Also, let

$$g_{R_0(k)}(z)=\sum_{j=0}^{N-1}\mathbb{P}\Big[R_0(k)=j\Big]z^j=\sum_{j=0}^{N-1}p^{(k)}_{n,N-j}(N)z^j,$$

be the p.g.f. for the number of empty boxes in layer k. Let $S(k)$ be the number of shots thrown into layer k, where $S(1)=n$. We define the conditional p.g.f.s for the number of occupied and

empty boxes as follows:

$$g_{R^*(k)|S(k)=i}(z) = \sum_{j=1}^{N} \mathbb{P}\Big[R^*(k) = j | S(k) = i\Big] z^j = \sum_{j=1}^{N} p_{i,j}^{(1)}(N) z^j, \tag{5.7}$$

$$g_{R_0(k)|S(k)=i}(z) = \sum_{j=0}^{N-1} \mathbb{P}\Big[R_0(k) = j | S(k) = i\Big] z^j = \sum_{j=0}^{N-1} p_{i,N-j}^{(1)}(N) z^j.$$

We can obtain the conditional p.g.f. for the number of occupied boxes from the conditional p.g.f. for the number of empty boxes. Since

$$\begin{aligned} g_{R_0(k)|S(k)=i}(z) = \mathbb{P}\Big[R_0(k) = i | S(k) = i\Big] z^i + \mathbb{P}\Big[R_0(k) = i-1 | S(k) = i\Big] z^{i-1} + \ldots \\ + \mathbb{P}\Big[R_0(k) = 1 | S(k) = i\Big] z + \mathbb{P}\Big[R_0(k) = 0\Big]. \end{aligned}$$

Then, using (2.3) and that the degree of the p.g.f of R_0 is $N-1$,

$$g_{R^*(k)|S(k)=i}(z) = z^N g_{R_0|S(k)=i}(z^{-1}) = z T_f[g_{R_0|S(k)=i}](z).$$

The range space for R_0 is

$$\{0, 1, 2, \ldots, N-1\}.$$

Hence the p.g.f. for R_0 will have the general form as shown below:

$$p_{N-1} z^{N-1} + p_{N-2} z^{N-2} + \ldots + p_1 z + p_0, \text{ where } p_0 > 0.$$

This polynomial has no roots at zero so that the number of real roots will be preserved when T_f is applied (2.3). Hence $g_{R^*(k)|S(k)=i}$ will have $N-1$ real roots from applying T_f to $g_{R_0|S(k)=i}$ along with one additional real root at zero. We shall denote this transformation by T_g in what follows. That is,

$$T_g[p](z) = z T_f[p](z). \tag{5.8}$$

The following lemma is also required for my proof.

Lemma 5.3.1. (Corollary 5.6.7, [32]) Let $f(z) = \sum_{i=0}^{N} a_i z^i$ be a polynomial of degree $N \geq 1$. Let ϕ be a polynomial with only real roots and no roots in the open interval $(0, N)$. Define $h(z) := \sum_{l=0}^{N} a_l \phi(l) z^l$. Then the following statement holds: if all of the roots of f are real, or real and non-negative, or real and non-positive, respectively, then so are those of h.

Theorem 5.3.1. $g_{R^*(k)}$ has all real roots for all k, N and $n \in \mathbb{N}$.

Proof. We shall proceed using proof by induction.
Base case: Setting $k = 1$ gives us the p.g.f. for the first layer of the process where both the initial number of shots, n, and the number of boxes, N, are fixed. This result has already been proved

by Vatutin and Mikhailov (4.2.4, [3]), and their proof was reviewed in Chapter four..

Assumption step: We shall assume that the p.g.f. for the number of occupied boxes for layer k in the scheme with N boxes and n shots initially has all real roots. That is, we assume (5.6) has all real roots.

Inductive step: We need to show for N boxes that the p.g.f. for the number of occupied boxes in layer $k+1$ has all real roots given that the p.g.f. for layer k has all real roots. For layer $k+1$ we only need to know what happened in the previous layer and can disregard all layers before this. In other words, the multiple-layer scheme is Markov with respect to the number of occupied or empty boxes. We have that

$$\begin{aligned}
g_{R^*(k+1)}(z) &= \sum_{j=1}^{N} \mathbb{P}\Big[R^*(k+1)=j\Big] z^j \\
&= \sum_{j=1}^{N}\sum_{i=1}^{N} \mathbb{P}\Big[R^*(k+1)=j | S(k+1)=i\Big]\mathbb{P}\Big[S(k+1)=i\Big] z^j \quad \text{(Law of Total Prob.)} \\
&= \sum_{i=1}^{N} \mathbb{P}\Big[S(k+1)=i\Big] \sum_{j=1}^{N} \mathbb{P}\Big[R^*(k+1)=j | S(k+1)=i\Big] z^j \quad \text{(swap order of sums)} \\
&= \sum_{i=1}^{N} \mathbb{P}\Big[R^*(k)=i\Big] g_{R^*(k+1)|S(k+1)=i}(z) \qquad \text{(from [5.7])} \\
&= \sum_{i=1}^{N} p_{n,i}^{(k)}(N) g_{R^*(1)|n=i}(z), \qquad (5.9)
\end{aligned}$$

where the last equality follows from the property that the conditional p.g.f.s are the same for all layers (the Markov chain is homogeneous). Since we know some nice results for the p.g.f. for the number of empty boxes in the single layer scheme it makes sense to write the p.g.f. for the number of occupied boxes in the first layer in terms of that for empty boxes. If we have j empty boxes then $N-j$ are occupied so we need to apply the transformation T_g (5.8). Using this along with (5.8) gives

$$g_{R^*(k+1)}(z) = \sum_{j=1}^{N} p_{n,j}^{(k)}(N) T_g[g_{R_0(1)|n=j}](z), \qquad (5.10)$$

where T_g will preserve the number of real roots here (5.8). Let

$$v_{n,N}(z) = \sum_{m=0}^{N-1} \binom{N}{m} z^m (N-m)^n. \qquad (5.11)$$

We have already seen (4.2.4) that we can reformulate the p.g.f. for the number of empty boxes in layer one for the scheme with N boxes and a fixed number of shots n as

$$g_{R_0(k)|n=j}(z) = \frac{v_{j,N}(z-1)}{N^j}. \qquad (5.12)$$

Now, plugging this into (5.10),

$$\begin{aligned} g_{R^*(k+1)}(z) &= \sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} T_g[v_{j,N}](z-1) \\ &= \sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} T_g \circ T_{-1}[v_{j,N}](z), \end{aligned} \tag{5.13}$$

where $T_{-1}[p](z) = p(z-1)$ is a real-root-preserving transformation (2.1). Further (see (4.2.4),

$$z^N v_{j,N}(z^{-1}) = \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}}[(z+1)^N]$$

or

$$T_f[v_{j,N}](z) = \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}}[(z+1)^N] = \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}} \circ T_1[z^N], \tag{5.14}$$

where both T_d and T_1 are real-root-preserving transformations (see (2.2) and (2.1)). Recall that T_f is self-inverse when applied to a polynomial with no roots at zero (2.3). So, applying T_f to both sides of (5.14),

$$v_{j,N}(z) = T_f \circ \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}} \circ T_1[z^N].$$

Now, plugging this new representation for $v_{j,N}(z)$ into (5.13) gives

$$\begin{aligned} g_{R^*(k+1)}(z) &= \sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} T_g \circ T_{-1} \circ T_f \circ \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}} \circ T_1[z^N] \\ &= T_g \circ T_{-1} \circ T_f \circ \sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}} \circ T_1[z^N]. \end{aligned}$$

Now, $T_g \circ T_{-1} \circ T_f$ preserves the number of real roots (see (5.8),(2.1) and (2.3)), so we just need to show that

$$\sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}} \circ T_1[z^N] = \sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}}[(z+1)^N]$$

has all real roots. Define

$$T_c[p] = \sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}}[p].$$

Then, since $\underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}}[z^i] = i^j z^i$ (see (4.1)),

$$\begin{aligned} T_c[z^i] &= \sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} \underbrace{T_d \circ T_d \circ \cdots \circ T_d}_{j \text{ times}}[z^i] \\ &= \sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} i^j z^i = z^i \sum_{j=1}^{N} p_{n,j}^{(k)}(N) \left(\frac{i}{N}\right)^j = z^i g_{R^*(k)}\left(\frac{i}{N}\right). \end{aligned}$$

Now, using the Binomial theorem,

$$\begin{aligned} T_c[(z+1)^N] &= T_c\left[\sum_{l=0}^{N} \binom{N}{l} z^l\right] = \sum_{l=0}^{N} \binom{N}{l} T_c[z^l] \\ &= \sum_{l=0}^{N} \binom{N}{l} z^l g_{R^*(k)}\left(\frac{l}{N}\right). \end{aligned} \tag{5.15}$$

In Lemma 5.3.1, let

$$f(z) = \sum_{l=0}^{N} \binom{N}{l} z^l = (z+1)^N,$$

so $a_l = \binom{N}{l}$. Here, we see that f has all its N roots at -1, so all its roots are real and non-positive. Now, let $\phi(z) = g_{R^*(k)}(z/N)$. We need $\phi(z)$ to satisfy the following two conditions:

1. $\phi(z)$ must have all real roots;
2. $\phi(z)$ must have no roots in the interval $(0,N)$.

Recall that in our assumption step, the p.g.f. for the number of occupied boxes in layer k has all real roots so that $g_{R^*(k)}(z)$ has all real roots. Now, mapping z to z/N preserves the number of real roots so that $g_{R^*(k)}(z/N)$ also has all real roots. Additionally, since ϕ is a p.g.f. the coefficients are all non-negative so that the roots are all non-positive and hence lie outside the interval $(0,N)$. By Lemma 5.3.1, all roots of

$$h(z) = \sum_{l=0}^{N} a_l \phi(l) z^l = \sum_{l=0}^{N} \binom{N}{l} g_{R^*(k)}\left(\frac{l}{N}\right) z^l$$

will be real and non-positive. But this is the same as (5.14), so $T_c[(z+1)^N]$ has all real roots which are non-positive. Recall that

$$g_{R^*(k+1)}(z) = T_g \circ T_{-1} \circ T_f \circ T_c[(z+1)^N].$$

We have just shown that $T_c[(z+1)^N]$ has all real roots and then we are applying a composition of transformations which preserve the number of real roots. Therefore, $g_{R^*(k+1)}(z)$ has all real

roots. Hence, by induction, we have shown that $g_{R^*(k)}(z)$ has all real roots for all k, N and $n \in \mathbb{N}$. □

5.3.2 Distributional results for $R^*(k)$ in the multiple-layer scheme with N boxes

Theorem 5.3.2. Again, let $R^*(k)$ be the number of occupied boxes in layer k of the scheme which starts with n shots and has N boxes throughout. Then, there is a constant $c < \infty$ such that

$$\sup_x \left| \mathbb{P}\left\{ \frac{R^*(k) - \mathbb{E}[R^*(k)]}{\sqrt{\mathrm{Var}[R^*(k)]}} \leq x \right\} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy \right| \leq \frac{c}{\sqrt{\mathrm{Var}[R^*(k)]}}. \tag{5.16}$$

Proof. (Theorem 5.3.2): Let $-r_1, -r_2, \ldots, -r_N$ be the roots of $g_{R^*(k)}$. Since $g_{R^*(k)}$ is a p.g.f. the coefficients are non-negative so $r_i \geq 0$. We can therefore write

$$g_{R^*(k)}(z) = \left(\frac{z+r_1}{1+r_1}\right)\left(\frac{z+r_2}{1+r_2}\right)\cdots\left(\frac{z+r_N}{1+r_N}\right).$$

Because of this factorisation, this is the same as the p.g.f. for the sum of N independent Bernoulli random variables each with success probability $\theta_i = 1/(1+r_i)$ $(i = 1, 2, \ldots, N)$. Then, writing

$$R^*(k) = X_1 + X_2 + \cdots + X_N, \text{ where } X_i \sim \mathrm{Bern}\left(\frac{1}{1+r_i}\right),$$

it follows that

$$\mathbb{E}[R^*(k)] = \sum_{i=1}^{N} \mathbb{E}(X_i) = \sum_{i=1}^{N} \frac{1}{1+r_i}, \tag{5.17}$$

$$\mathrm{Var}[R^*(k)] = \sum_{i=1}^{N} \mathbb{E}(X_i)\{1 - \mathbb{E}(X_i)\} = \sum_{i=1}^{N} \frac{r_i}{(1+r_i)^2}. \tag{5.18}$$

Also,

$$\begin{aligned}
\mathbb{E}|X_i - \mathbb{E}(X_i)|^3 &= \sum_{i=1}^{N} \left[|1-\theta_i|^3 \mathbb{P}(X_i = 1) + |0-\theta_i|^3 \mathbb{P}(X_i = 0) \right] \\
&= \sum_{i=1}^{N} \left[(1-\theta_i)^3 \theta_i + \theta_i^3 (1-\theta_i) \right] \qquad (\text{since } 0 \leq \theta_i \leq 1) \\
&= \sum_{i=1}^{N} (1-\theta_i)\left\{ \theta_i (2\theta_i^2 - 2\theta_i + 1) \right\} \leq \sum_{i=1}^{N} (1-\theta_i)\left\{ \theta_i (2\theta_i - 2\theta_i + 1) \right\} \\
&= \sum_{i=1}^{N} (1-\theta_i)\theta_i = \sum_{i=1}^{N} \mathrm{Var}(X_i) = \mathrm{Var}\left[R^*(k) \right].
\end{aligned}$$

That is,

$$\sum_{i=1}^{N} \mathbb{E}\left|X_i - \mathbb{E}(X_i)\right|^3 \leq \mathrm{Var}\left[R^*(k)\right]. \tag{5.19}$$

Since $R^*(k)$ can be written as a sum of independent Bernoulli random variables we can use an Esseen inequality (page 111, [44]) to write:

$$\sup_x \left| \mathbb{P}\left\{ \frac{R^*(k) - \mathbb{E}\left[R^*(k)\right]}{\sqrt{\mathrm{Var}\left[R^*(k)\right]}} \leq x \right\} - \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} e^{-y^2/2} dy \right| \leq \frac{c}{(\mathrm{Var}\left[R^*(k)\right])^{3/2}} \sum_{i=1}^{N} \mathbb{E}\left|X_i - \mathbb{E}(X_i)\right|^3$$

$$\leq \frac{c}{(\mathrm{Var}\left[R^*(k)\right])^{3/2}} \mathrm{Var}\left[R^*(k)\right] \qquad \text{(using (5.19))}$$
$$\leq \frac{c}{\sqrt{\mathrm{Var}\left[R^*(k)\right]}}.$$

□

Corollary 5.3.2.1. Suppose that n and N are varied so that $\mathrm{Var}\left[R^*(k)\right] \to \infty$. Then, $\mathbb{E}\left[R^*(k)\right] \to \infty$ and the distribution of

$$\frac{R^*(k) - \mathbb{E}\left[R^*(k)\right]}{\sqrt{\mathrm{Var}\left[R^*(k)\right]}},$$

tends to the standard normal distribution.

We have shown that $R^*(k)$ can be written as a sum of independent (but not identically distributed) Bernoulli random variables. Let:

$$R^*(k) = X_1 + X_2 + \ldots + X_N, \text{ where } X_i \sim \mathrm{Bern}(\theta_i).$$

Then,

$$\mathbb{E}\left[R^*(k)\right] = \sum_{i=1}^{N} \mathbb{E}(X_i) = \sum_{i=1}^{N} \theta_i, \tag{5.20}$$

$$\mathrm{Var}\left[R^*(k)\right] = \sum_{i=1}^{N} \mathrm{Var}(X_i) = \sum_{i=1}^{N} \theta_i(1-\theta_i) \qquad \text{(since the } X_i\text{s are independent)}. \tag{5.21}$$

So, for each X_i we have, since $0 \leq \theta_i \leq 1$,

$$\mathrm{Var}(X_i) = \theta_i(1-\theta_i) \leq \theta_i \leq \mathbb{E}(X_i).$$

Hence,

$$\mathrm{Var}\left[R^*(k)\right] = \sum_{i=1}^{N} \mathrm{Var}(X_i) \leq \sum_{i=1}^{N} \mathbb{E}(X_i) \leq \mathbb{E}\left[R^*(k)\right], \tag{5.22}$$

so that,

$$\mathrm{Var}\left[R^*(k)\right] \to \infty \implies \mathbb{E}\left[R^*(k)\right] \to \infty.$$

Then, by Theorem 5.3.2 we have,

$$\sup_x \left| \mathbb{P}\left\{ \frac{R^*(k) - \mathbb{E}\left[R^*(k)\right]}{\sqrt{\mathrm{Var}\left[R^*(k)\right]}} \leq x \right\} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy \right| \leq \frac{c}{\sqrt{\mathrm{Var}\left[R^*(k)\right]}}.$$

So, if $\mathrm{Var}\left[R^*(k)\right] \to \infty$ then this upper bound on the distance tends to zero giving the desired result.

Lemma 5.3.2.

$$\rho_\lambda\left(R^*(k)\right) \leq \mathbb{E}\left[R^*(k)\right] - \mathrm{Var}\left[R^*(k)\right], \tag{5.23}$$

where $\lambda = \mathbb{E}\left[R^*(k)\right]$ and $\rho_\lambda(X)$ is the total variation distance between the distribution of X and the Poisson distribution with parameter λ.

For the total variation distance between the distribution of $\sum_{i=1}^{N} X_i$, where $X_i \sim \mathrm{Bern}(\theta_i)$, Le Cam [45] proved an upper bound is

$$d_{TV}\left(\text{distr. of } \sum_i X_i, \mathrm{Pois}\left(\lambda_N = \sum_{i=1}^{N} \mathbb{E}(X_i)\right)\right) \leq \sum_{i=1}^{N} \theta_i^2. \tag{5.24}$$

Using (5.19) and (5.20) we also have

$$\mathbb{E}\left[R^*(k)\right] - \mathrm{Var}\left[R^*(k)\right] = \sum_{i=1}^{N} \left\{\theta_i - \theta_i(1-\theta_i)\right\} = \sum_{i=1}^{N} \theta_i \left\{1 - (1-\theta_i)\right\} = \sum_{i=1}^{N} \theta_i^2. \tag{5.25}$$

Putting this together then gives the desired result.

5.3.3 Simulations

We have looked in detail at the distribution of the number of occupied boxes over layers for small examples and have developed formulae for general N but we can also visualise what is happening for larger values of N via simulations run using RStudio [46]. The $N = 100$, $n = 100$ case was simulated though 600 layers, $100,000$ times and histograms were produced for the number of occupied boxes in various layers. Since we are starting with n and N proportional to one another we expect to be in the Central Domain as described by Kolchin et al. [2]. As expected, given their results, we observe a distribution which looks fairly close to Normal (Figure 5.3). Then, as we move through the layers the distribution seems to move further from a Normal distribution (Figure 5.4). After a certain layer, we observe an almost degenerate distribution which again is to be expected since the more layers the shots have moved through the more likely merges are

to have happened (Figure 5.5). The expected value and variance in each layer shown is provided in Table 5.1.

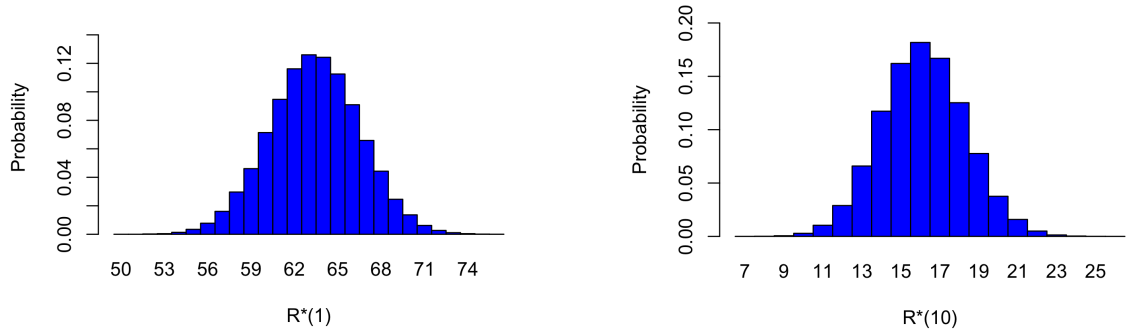

Figure 5.3: Histograms for the number of occupied boxes in layer one (left) and ten (right) across 100,000 simulations.

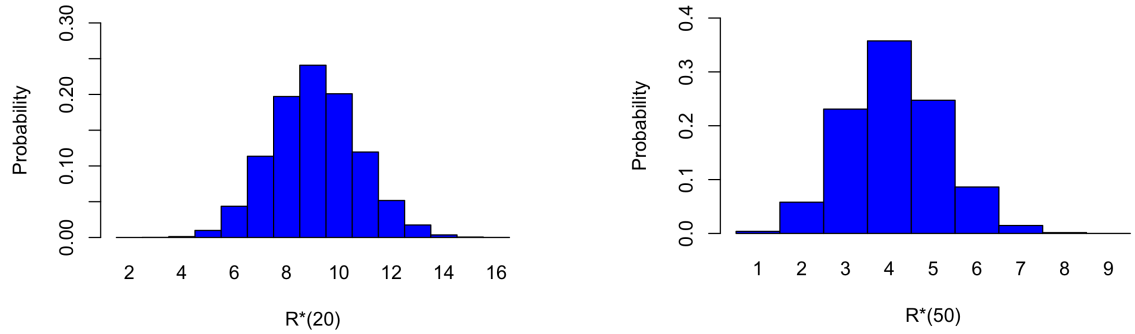

Figure 5.4: Histograms for the number of occupied boxes in layer 20 (left) and 50 (right) across 100,000 simulations.

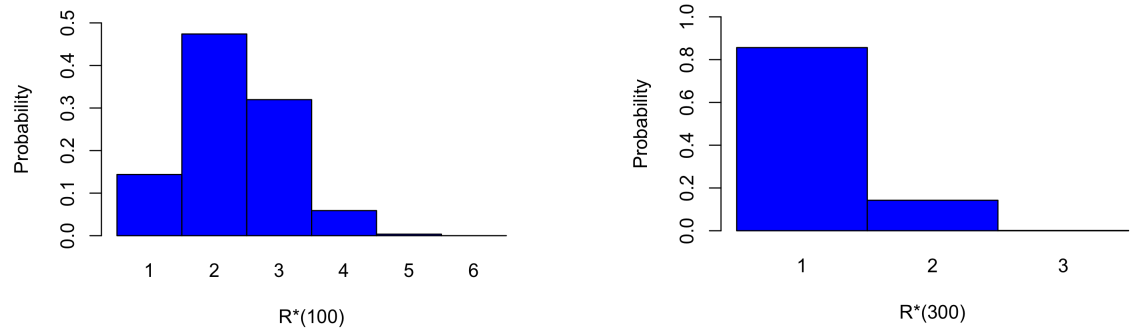

Figure 5.5: Histograms for the number of occupied boxes in layer 100 (left) and 300 (right) across 100,000 simulations.

Layer	$\mathbb{E}[R^*(k)]$	$\mathrm{Var}[R^*(k)]$
1	63.39	9.74
10	16.14	4.71
20	9.09	2.77
50	4.11	1.23
100	2.30	0.64
300	1.15	0.13

Table 5.1: The expected value and variance for $R^*(k)$ across 100,000 runs of the simulations starting with $n = 100$, $N = 100$.

Suppose we again want to consider an example where we start in the Central Domain but with a larger number of shots and boxes. Figures 5.6-5.8 and Table 5.2 demonstrate how the distribution of $R^*(k)$ changes throughout the layers when we start with $n = 1000$, $N = 1000$.

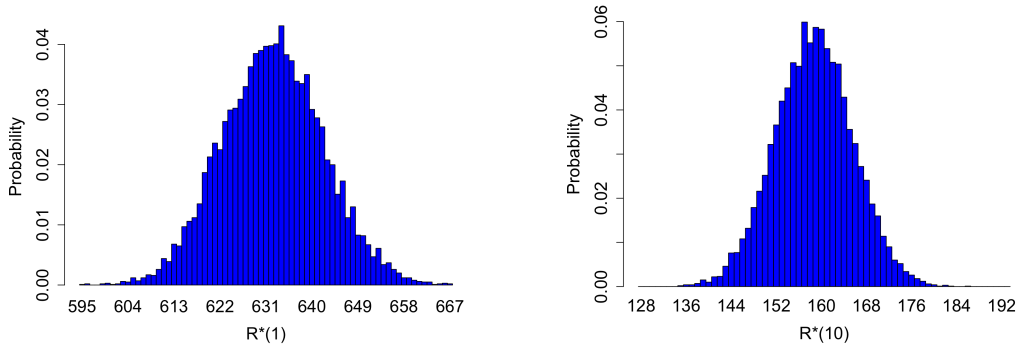

Figure 5.6: Histograms for the number of occupied boxes in layer one (left) and ten (right) across 10,000 simulations.

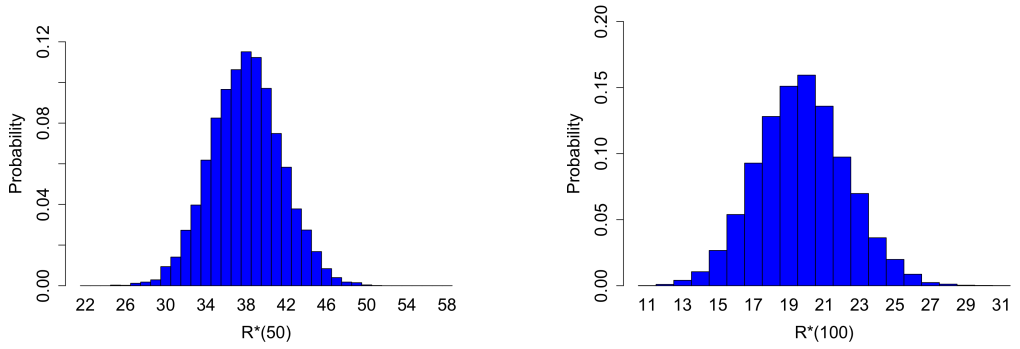

Figure 5.7: Histograms for the number of occupied boxes in layer 50 (left) and 100 (right) across 10,000 simulations.

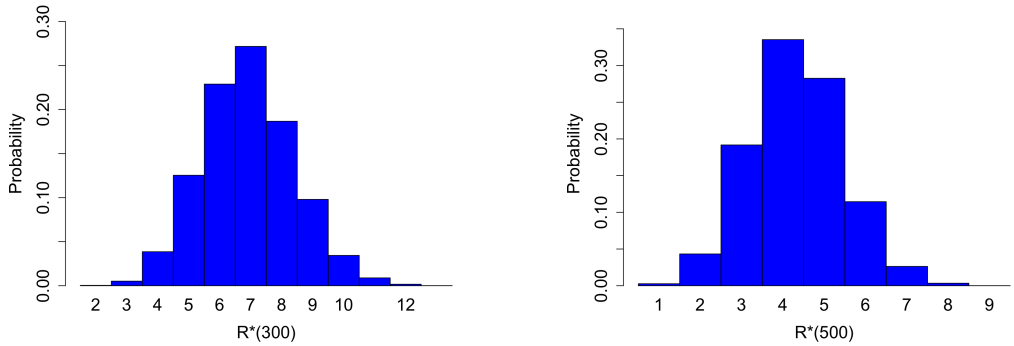

Figure 5.8: Histograms for the number of occupied boxes in layer 300 (left) and 500 (right) across 10,000 simulations.

Layer	$\mathbb{E}[R^*(k)]$	$\text{Var}[R^*(k)]$
1	632.32	97.74
10	158.66	47.47
50	37.99	12.29
100	19.73	6.42
300	6.91	2.24
500	4.32	1.32

Table 5.2: The expected value and variance for $R^* * (k)$ across 10,000 runs of the simulations starting with $n = 1000$, $N = 1000$.

Suppose we instead want to consider an example where we start in the Left-Hand Domain so that the number of shots is much smaller than the number of boxes. Figures 5.9-5.11 and Table 5.3 demonstrate how the distribution of $R^*(k)$ changes throughout the layers when we start with $n = 10$, $N = 100$.

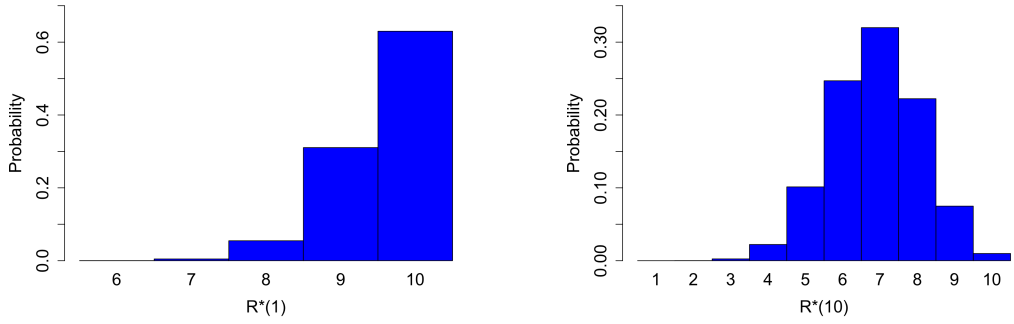

Figure 5.9: Histograms for the number of occupied boxes in layer one (left) and ten (right) across 100,000 simulations.

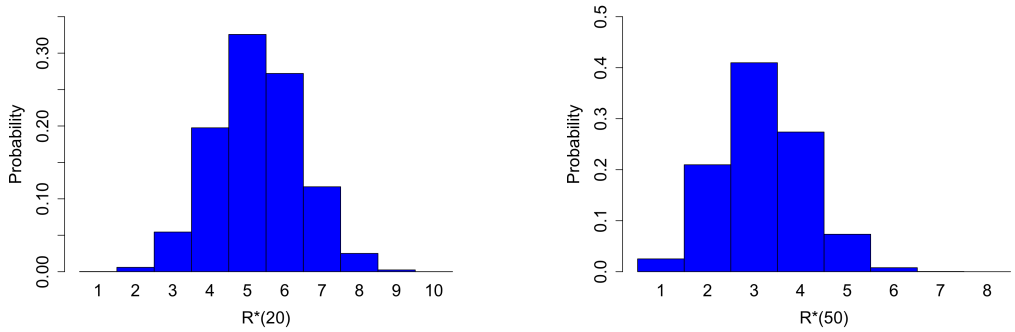

Figure 5.10: Histograms for the number of occupied boxes in layer 20 (left) and 50 (right) across 100,000 simulations.

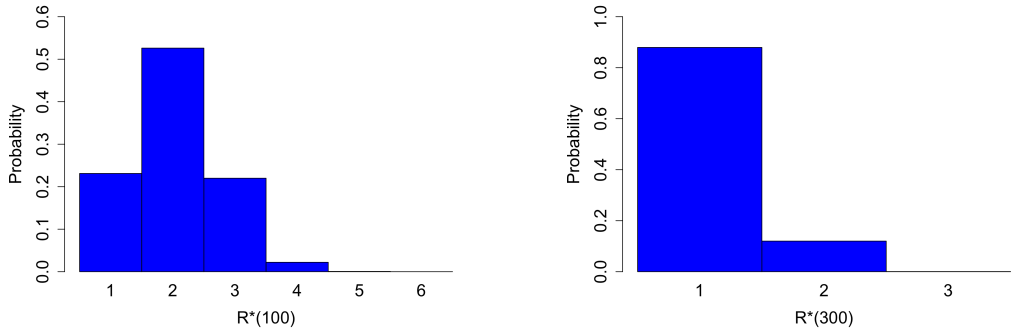

Figure 5.11: Histograms for the number of occupied boxes in layer 100 (left) and 300 (right) across 100,000 simulations.

Layer	$\mathbb{E}[R^*(k)]$	$\text{Var}[R^*(k)]$
1	9.57	0.39
10	6.88	1.49
20	5.27	1.41
50	3.19	0.92
100	2.03	0.54
300	1.12	0.11

Table 5.3: The expected value and variance for $R^*(k)$ across 100,000 runs of the simulations starting with $n = 10$, $N = 100$.

Suppose we again want to consider an example in the Left-Hand Domain but with a larger number of boxes. Figures 5.12-5.14 and Table 5.4 demonstrate how the distribution of $R^*(k)$ changes throughout the layers when we start with $n = 10$, $N = 1000$.

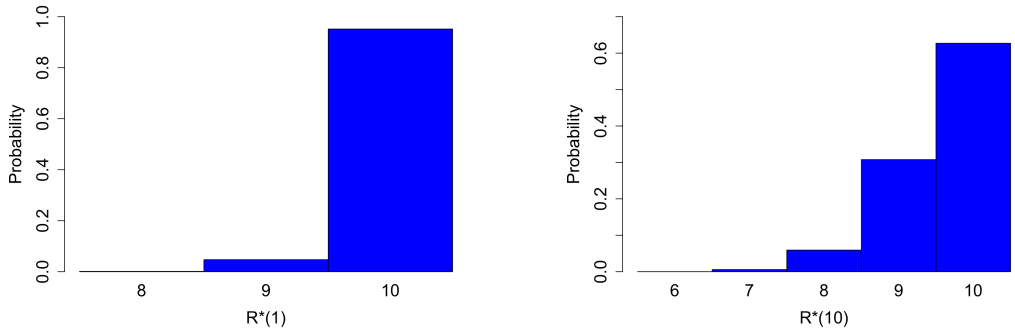

Figure 5.12: Histograms for the number of occupied boxes in layer one (left) and ten (right) across 10,000 simulations.

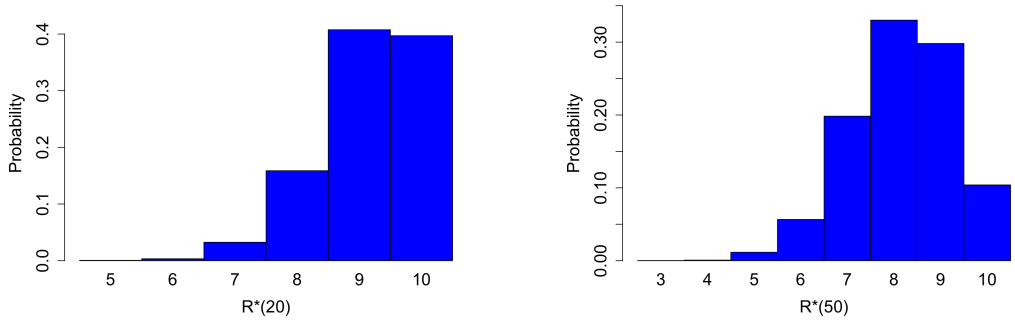

Figure 5.13: Histograms for the number of occupied boxes in layer 20 (left) and 50 (right) across 10,000 simulations.

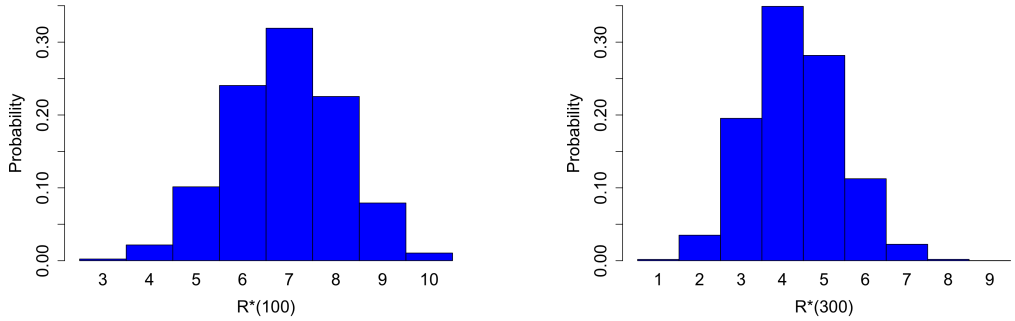

Figure 5.14: Histograms for the number of occupied boxes in layer 100 (left) and 300 (right) across 10,000 simulations.

Layer	$\mathbb{E}[R^*(k)]$	$\mathrm{Var}[R^*(k)]$
1	9.95	0.05
10	9.56	0.40
20	9.16	0.70
50	8.16	1.24
100	6.90	1.50
300	4.31	1.22

Table 5.4: The expected value and variance for $R^*(k)$ across 10,000 runs of the simulations starting with $n = 10$, $N = 1000$.

Another scenario we could consider is when we start with a much larger number of shots than boxes so that we are initially in the Right-Hand Domain. In this case, regardless of how many shots we throw into the first layer the maximum number of occupied boxes is given by N. Then, from layer two onwards, at most N shots will be distributed and the picture is the same as that shown for the Central Domain.

5.3.4 Summary of results for $R^*(k)$ for N boxes

1. The p.g.f. for the number of occupied boxes in any layer k of the scheme with N boxes has all real roots.

2. As a result of the above we show that if the variance of the number of occupied boxes tends to infinity then the distance between the standardised version of $R^*(k)$ and the standard Normal distribution tends to zero. We provide both a classical and local limit theorem.

3. An upper bound for the total variation distance between the distribution of $R^*(k)$ and the Poisson distribution with parameter $\lambda = \mathbb{E}[R^*(k)]$ is

$$\mathbb{E}[R^*(k)] - \mathrm{Var}[R^*(k)].$$

Chapter 6

Properties of the difference in the number of occupied boxes between consecutive layers

6.1 Motivation

In a multiple-layer scheme we can also look at the difference in the number of occupied boxes between consecutive layers. Suppose, for example, we consider the boxes in each layer to be biological species so that the number of occupied boxes is a measure of the species diversity. Then, the difference in the occupancy number between layers may be of interest to address questions such as:

1. Did interventions help to reduce drops in diversity?
2. When did the largest/smallest drops happen?
3. How did certain events affect diversity?

As discussed in Chapter four, Kolchin et al. considered the number of empty boxes, R_0, in a single layer of a scheme where n shots are allocated to N boxes according to the uniform distribution [2]. They defined the Left-Hand 0-Domain to be where, as $n, N \to \infty$,

$$\frac{n}{N} \to 0 \qquad \text{and} \qquad \text{Var}(R_0) \approx \frac{n^2}{2N} \to \lambda < \infty.$$

In this domain they showed that the following asymptotic result holds.

Theorem 6.1.1. [2] As $n, N \to \infty$,

$$\mathbb{P}\{R_0 - (N-n) = i\} \to \frac{\lambda^i e^{-\lambda}}{i!},$$

where $\text{Var}(R_0) \to \lambda$ as $n, N \to \infty$.

Note that $N-n$ would be the number of empty boxes if each shot went to a different box, so that $R_0-(N-n)$ is the difference in the number of empty boxes observed and the number we would have if each shot had its own box. Hence, their result shows that, as n and N tend to infinity in the way specified, the distribution of this difference tends to Poisson, where λ is the expected difference.
In the multiple-layer scheme, let $R_0(k)$ be the number of empty boxes in layer k in the scheme with N boxes. Then, $R_0(1)-(N-n)$ in layer one becomes $R_0(k)-R_0(k-1)$ for layer $k \geq 2$, which can alternatively be written in terms of the number of occupied boxes as $R^*(k-1)-R^*(k)$. We want to be able to prove an equivalent result to Theorem 6.1.1 for this multiple-layer scheme.

6.2 Three boxes

To start, assume $N=3$ and consider the difference between the number of occupied boxes in the first and second layers. Let $R^*_{\text{diff}}(k-1,k)$ be the difference in the number of occupied boxes between layer $k-1$ and k in the scheme with N boxes. Recall that the number of occupied boxes is non-increasing so $R^*(k-1)-R^*(k) \geq 0$. The range space of

$$R^*_{\text{diff}}(k-1,k) \in \{0,1,\ldots N-1\}.$$

For the difference to be two we must have had three occupied boxes in layer one and then one at layer two, so

$$\mathbb{P}\Big(R^*_{\text{diff}}(1,2)=2\Big) = p^{(1)}_{3,3}(3)p^{(1)}_{3,1}(3) = \frac{2}{9}\times\frac{1}{9} = \frac{2}{81}.$$

For the difference to be one, there were either two or three occupied boxes in layer one and then the occupancy count decreased by one after the allocation to layer two, so

$$\mathbb{P}\Big(R^*_{\text{diff}}(1,2)=1\Big) = p^{(1)}_{3,3}(3)p^{(1)}_{3,2}(3)+p^{(1)}_{3,2}(3)p^{(1)}_{2,1}(3) = \frac{2}{9}\times\frac{2}{3}+\frac{2}{3}\times\frac{1}{3} = \frac{10}{27}.$$

Finally, for the difference to be zero, there were either one, two or three occupied boxes in the first layer and then each of these shots must land in its own box in the second layer, so

$$\begin{aligned}\mathbb{P}\Big(R^*_{\text{diff}}(1,2)=0\Big) &= p^{(1)}_{3,3}(3)p^{(1)}_{3,3}(3)+p^{(1)}_{3,2}(3)p^{(1)}_{2,2}(3)+p^{(1)}_{3,1}(3)p^{(1)}_{1,1}(3)\\ &= \frac{2}{9}\times\frac{2}{9}+\frac{2}{3}\times\frac{2}{3}+\frac{1}{9}\times 1 = \frac{49}{81}.\end{aligned}$$

Applying similar arguments for the first few layers gives the following p.g.f.s (the probabilities for the first ten layers are shown in Figure 6.1):

$$\begin{aligned}
g_{R^*_{\text{diff}}(1,2)}(x) &= \frac{2}{81}x^2 + \frac{10}{27}x + \frac{49}{81},\\
g_{R^*_{\text{diff}}(2,3)}(x) &= \frac{4}{729}x^2 + \frac{56}{243}x + \frac{557}{729},\\
g_{R^*_{\text{diff}}(3,4)}(x) &= \frac{8}{6561}x^2 + \frac{328}{2187}x + \frac{5569}{6561},\\
g_{R^*_{\text{diff}}(4,5)}(x) &= \frac{16}{59049}x^2 + \frac{1952}{19683}x + \frac{53177}{59049}.
\end{aligned}$$

In order to construct the p.g.f. for this difference in any given layer of the scheme with three boxes, non-recursive formulae for three probabilities are needed.

i $\mathbb{P}\Big(R^*_{\text{diff}}(k-1,k)=0\Big)$: There could have been either one, two or three shots in the previous layer and then no merges occurred in the final one-layer transition. So,

$$\mathbb{P}\Big(R^*_{\text{diff}}(k-1,k)=0\Big) = p^{(k-1)}_{3,3}(3)p^{(1)}_{3,3}(3) + p^{(k-1)}_{3,2}(3)p^{(1)}_{2,2}(3) + p^{(k-1)}_{3,1}(3)p^{(1)}_{1,1}(3).$$

Using the non-recursive formulae for $p^{(k-1)}_{3,3}(3)$, $p^{(k-1)}_{3,2}(3)$ and $p^{(k-1)}_{3,1}(3)$ (from (3.1)-(3.3)),

$$\begin{aligned}
\mathbb{P}\Big(R^*_{\text{diff}}(k-1,k)=0\Big) &= \frac{2}{9}\left(\frac{2}{9}\right)^{k-1} + \frac{2}{3}\left(\frac{3}{2}\left\{\left(\frac{2}{3}\right)^{k-1} - \left(\frac{2}{9}\right)^{k-1}\right\}\right)\\
&\quad + \left\{1 + \frac{1}{2}\left(\frac{2}{9}\right)^{k-1} - \frac{3}{2}\left(\frac{2}{3}\right)^{k-1}\right\}\\
&= 1 - \frac{5}{4}\left(\frac{2}{9}\right)^{k} - \frac{3}{4}\left(\frac{2}{3}\right)^{k}.
\end{aligned}$$

ii $\mathbb{P}\Big(R^*_{\text{diff}}(k-1,k)=1\Big)$: There could have been two or three shots in the previous layer and then, in the final one-layer transition, the occupancy count has decreased by one, so,

$$\begin{aligned}
\mathbb{P}\Big(R^*_{\text{diff}}(k-1,k)=1\Big) &= p^{(k-1)}_{3,3}(3)p^{(1)}_{3,2}(3) + p^{(k-1)}_{3,2}(3)p^{(1)}_{2,1}(3)\\
&= \frac{3}{4}\left(\frac{2}{9}\right)^{k} + \frac{3}{4}\left(\frac{2}{3}\right)^{k} = \frac{3}{4}\left\{\left(\frac{2}{9}\right)^{k} + \left(\frac{2}{3}\right)^{k}\right\}.
\end{aligned}$$

iii $\mathbb{P}\Big(R^*_{\text{diff}}(k-1,k)=2\Big)$: The only option here is that there were three shots in the previous layer and then, in the final one-layer transition, the occupancy count has decreased by two. Hence,

$$\mathbb{P}\Big(R^*_{\text{diff}}(k-1,k)=2\Big) = p^{(k-1)}_{3,3}(3)p^{(1)}_{3,1}(3) = \frac{1}{2}\left(\frac{2}{9}\right)^{k}.$$

Hence, the p.g.f. for the difference in the number of occupied boxes between consecutive layers of the scheme with three boxes is given by

$$g_{R^*_{\text{diff}}(k-1,k)}(x) = \frac{1}{2}\left(\frac{2}{9}\right)^k x^2 + \frac{3}{4}\left\{\left(\frac{2}{9}\right)^k + \left(\frac{2}{3}\right)^k\right\}x + 1 - \frac{5}{4}\left(\frac{2}{9}\right)^k - \frac{3}{4}\left(\frac{2}{3}\right)^k. \tag{6.1}$$

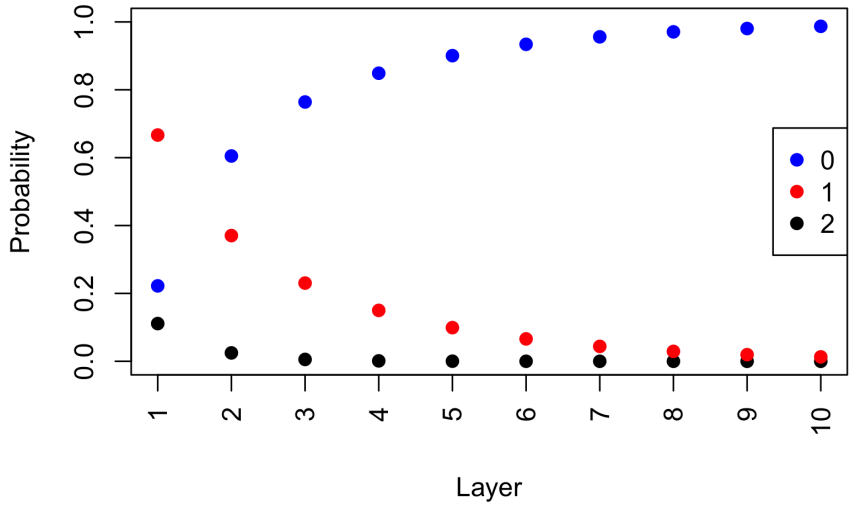

Figure 6.1: Probabilities for the difference in the number of occupied boxes for the first ten layers of the scheme with three boxes.

The expected difference in the number of occupied boxes is then

$$\mathbb{E}\left[R^*_{\text{diff}}(k-1,k)\right] = \frac{3}{4}\left(\frac{2}{3}\right)^k + \frac{7}{4}\left(\frac{2}{9}\right)^k$$

and

$$\text{Var}\left[R^*_{\text{diff}}(k-1,k)\right] = \frac{3}{4}\left(\frac{2}{3}\right)^k - \frac{9}{16}\left(\frac{4}{9}\right)^k + \frac{11}{4}\left(\frac{2}{9}\right)^k - \frac{21}{8}\left(\frac{4}{27}\right)^k - \frac{49}{16}\left(\frac{4}{81}\right)^k$$

(Figure 6.2). The difference

$$\mathbb{E}\left[R^*_{\text{diff}}(k-1,k)\right] - \text{Var}\left[R^*_{\text{diff}}(k-1,k)\right] = \frac{9}{16}\left(\frac{4}{9}\right)^k - \left(\frac{2}{9}\right)^k + \frac{21}{8}\left(\frac{4}{27}\right)^k + \frac{49}{16}\left(\frac{4}{81}\right)^k$$
$$\to 0 \text{ as } k \to \infty.$$

This suggests that a Poisson distribution may be a possible limiting distribution. Compare this with what happens for the number of occupied boxes. In this case, there is always at least one

box occupied and

$$\mathbb{E}\left[R^*(k)\right] - \mathrm{Var}\left[R^*(k)\right] = 1 + 3\left(\frac{2}{3}\right)^k + \left(\frac{2}{9}\right)^k + \frac{9}{4}\left(\frac{4}{9}\right)^k + \frac{3}{2}\left(\frac{4}{27}\right)^k + \frac{1}{4}\left(\frac{4}{81}\right)^k$$
$$\to 1 \text{ as } k \to \infty.$$

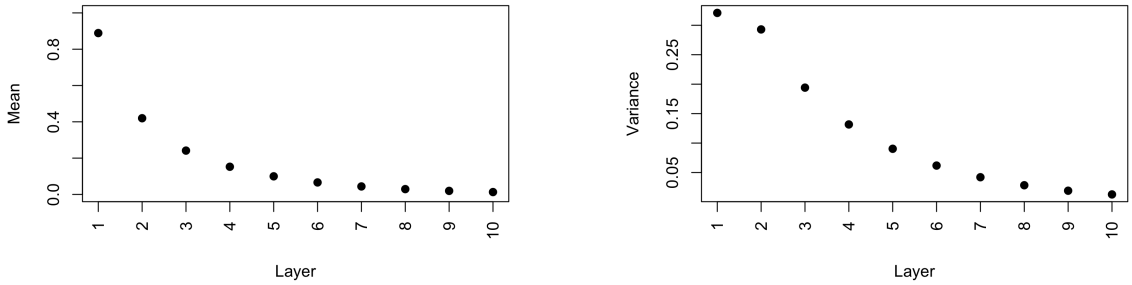

Figure 6.2: Expectation (left) and variance (right) of $R^*_{\text{diff}}(k-1,k)$ for the first ten layers of the scheme with three boxes.

6.2.1 Distance to the Poisson distribution

It is useful to quantify how far away the distribution of the difference in the number of occupied boxes is from a Poisson distribution. For example, consider the difference in the number of occupied boxes between the second and third layer of the scheme, where

$$\mathbb{P}\Big(R^*_{\text{diff}}(2,3) = 0\Big) = \frac{557}{729},\ \mathbb{P}\Big(R^*_{\text{diff}}(2,3) = 1\Big) = \frac{56}{243},\ \mathbb{P}\Big(R^*_{\text{diff}}(2,3) = 2\Big) = \frac{4}{729}.$$

Now, suppose we want to see how far away this is from a Poisson distribution with parameter $\lambda_k = \mathbb{E}\left[R^*_{\text{diff}}(k-1,k)\right]$ for $k = 3$, i.e.,

$$\lambda_3 = \frac{176}{729}.$$

The distance is then 0.0816 (including the contribution from the Poisson tail). Repeating this for layer five gives a distance of 0.0179. The larger the value of k, the smaller the distance and so the closer the distribution is to Poisson. The expected number of layers until we have only one occupied box is approximately twice the number of boxes [11] which is why we chose to look at the above values of k. After all the shots have merged into one that single remaining shot will just continue to be allocated to its own box and the difference in the number of occupied boxes will remain zero from that point onwards.

6.2.2 Discriminant approach

Lemma 6.2.1. The p.g.f. for the difference in occupancy between consecutive layers has all real roots for all layers for three boxes.

Recall from (6.1) that the p.g.f. for the difference in the number of occupied boxes between layer $k-1$ and k with three boxes is

$$g_{R^*_{\text{diff}}(k-1,k)}(x) = \frac{1}{2}\left(\frac{2}{9}\right)^k x^2 + \frac{3}{4}\left\{\left(\frac{2}{9}\right)^k + \left(\frac{2}{3}\right)^k\right\}x + 1 - \frac{5}{4}\left(\frac{2}{9}\right)^k - \frac{3}{4}\left(\frac{2}{3}\right)^k.$$

The discriminant

$$\Delta_2 = \frac{49}{16}\left(\frac{4}{81}\right)^k + \frac{21}{8}\left(\frac{4}{27}\right)^k + \frac{9}{16}\left(\frac{4}{9}\right)^k - 2\left(\frac{2}{9}\right)^k.$$

Since

$$\frac{\frac{9}{16}\left(\frac{4}{9}\right)^k}{2\left(\frac{2}{9}\right)^k} = \frac{9}{32}(2)^k > 1 \text{ for } k \geq 2,$$

$\Delta_2 > 0$ and hence the p.g.f. has all real roots for all layers, when $N = 3$.

6.2.3 Sum of independent Bernoulli random variables

This allows $R^*_{\text{diff}}(1,2)$ for example to be written as a sum of independent Bernoulli random variables:

$$R^*_{\text{diff}}(1,2) = X_1^{(2)} + X_2^{(2)},$$

where,

$$X_1^{(2)} \sim \text{Be}(\theta_1^{(2)}),\ X_2^{(2)} \sim \text{Be}(\theta_2^{(2)}).$$

Setting the probabilities of $R^*(1,2)$ equal to the corresponding Bernoulli probabilities gives

$$\begin{aligned}
\mathbb{P}\left(R^*_{\text{diff}}(1,2) = 0\right) &= \left(1-\theta_1^{(2)}\right)\left(1-\theta_2^{(2)}\right) = \frac{49}{81},\\
\mathbb{P}\left(R^*_{\text{diff}}(1,2) = 1\right) &= \theta_1^{(2)}\left(1-\theta_2^{(2)}\right) + \theta_2^{(2)}\left(1-\theta_1^{(2)}\right) = \frac{10}{27},\\
\mathbb{P}\left(R^*_{\text{diff}}(1,2) = 2\right) &= \theta_1^{(2)}\theta_2^{(2)} = \frac{2}{81},
\end{aligned}$$

giving

$$\theta_2^{(2)} = \frac{17-\sqrt{127}}{81} \text{ and } \theta_1^{(2)} = \frac{17+\sqrt{127}}{81}.$$

Therefore,

$$R^*_{\text{diff}}(1,2) = X_1^{(2)} + X_2^{(2)},$$

with

$$X_1^{(2)} \sim \text{Be}\left(\frac{17+\sqrt{127}}{81}\right),\ X_2^{(2)} \sim \text{Be}\left(\frac{17-\sqrt{127}}{81}\right).$$

Repeating this for the difference in the number of occupied boxes between the third and second layers

$$R^*_{\text{diff}}(2,3) = X_1^{(3)} + X_2^{(3)},$$

with

$$X_1^{(3)} \sim \text{Be}\left(\frac{88}{729}+\frac{1}{2}\sqrt{\frac{19312}{531441}}\right),\ X_2^{(3)} \sim \text{Be}\left(\frac{88}{729}-\frac{1}{2}\sqrt{\frac{19312}{531441}}\right).$$

Comparing the θs for the first three differences:

$$\theta_1^{(2)} \approx 0.349,\ \theta_2^{(2)} \approx 0.0708;$$

$$\theta_1^{(3)} \approx 0.216,\ \theta_2^{(3)} \approx 0.0254;$$

$$\theta_1^{(4)} \approx 0.144,\ \theta_2^{(4)} \approx 0.0085,$$

we see that as the layer index increases, both θ parameters are moving towards zero which is consistent with the expected difference in the number of occupied boxes tending to zero.

6.2.4 Summary of results for three boxes

1. The p.g.f. for the difference in the number of occupied boxes in consecutive layers in the scheme with three boxes is given by

$$g_{R^*_{\text{diff}}(k-1,k)}(x) = \frac{1}{2}\left(\frac{2}{9}\right)^k x^2 + \frac{3}{4}\left\{\left(\frac{2}{9}\right)^k + \left(\frac{2}{3}\right)^k\right\}x + 1 - \frac{5}{4}\left(\frac{2}{9}\right)^k - \frac{3}{4}\left(\frac{2}{3}\right)^k.$$

2. The expected difference in the number of occupied boxes in consecutive layers in the scheme with three boxes is

$$\mathbb{E}\left[R^*_{\text{diff}}(k-1,k)\right] = \frac{3}{4}\left(\frac{2}{3}\right)^k + \frac{7}{4}\left(\frac{2}{9}\right)^k.$$

3. The variance of the difference in the number of occupied boxes in consecutive layers in the scheme with three boxes is

$$\text{Var}\left[R^*_{\text{diff}}(k-1,k)\right] = \frac{3}{4}\left(\frac{2}{3}\right)^k - \frac{9}{16}\left(\frac{4}{9}\right)^k + \frac{11}{4}\left(\frac{2}{9}\right)^k - \frac{21}{8}\left(\frac{4}{27}\right)^k - \frac{49}{16}\left(\frac{4}{81}\right)^k.$$

4. The p.g.f. for the difference in the number of occupied boxes in any layer of the scheme has all real roots so that the difference can be expressed as a sum of independent Bernoulli random variables. We have again shown this using a discriminant approach.

Similar calculations (not shown) were repeated for $N = 4$ and $N = 5$. For example, the probabilities for the differences over the first ten layers are shown in Figure 6.3.

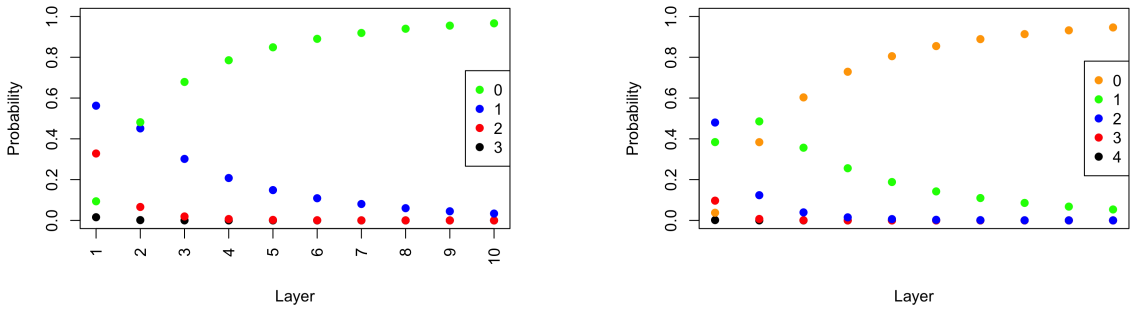

Figure 6.3: Difference probabilities for the first ten layers of the schemes with four (left) and five (right) boxes.

Further, if we continue to increase N and calculate the distance between the distribution of the difference for layer k and a Poisson distribution with its parameter given by the mean

difference in layer k we see it decreases as N increases (Figure 6.4). In order to get the difference probabilities for the larger values of N, the recursive formulae for the probabilities for general N (5.5) were used.

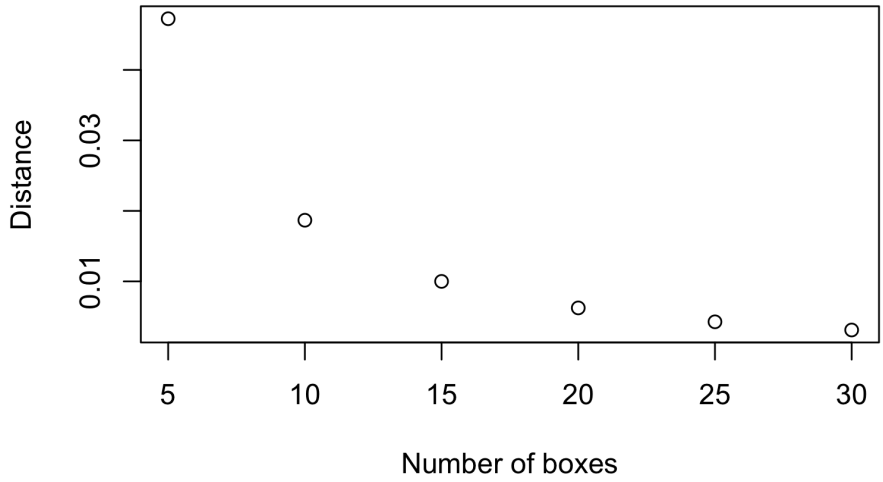

Figure 6.4: Distance between the distribution of $R^*_{\text{diff}}(k-1,k)$ and Poisson distribution for $k=N$.

6.3 General N

A proof demonstrating that the p.g.f. for the number of occupied boxes in layer k has all real roots was given in Chapter five and we now want produce an equivalent result for the p.g.f. for the difference in the number of occupied boxes between consecutive layers of the scheme. Recall that our motivation for proving the p.g.f. has all real roots is that it permits us to represent the difference in the number of occupied boxes as a sum of independent (but not identically distributed) Bernoulli random variables. From this, normal and Poisson limiting results follow and this will be shown in this chapter.

Theorem 6.3.1. The p.g.f. for the difference in the number of occupied boxes between consecutive layers of the scheme has all real roots for all layers.

Proof. Let $g_{R^*_{\text{diff}}}(k,k+1)$ be the p.g.f. for the difference in the number of occupied boxes between layer k and layer $k+1$ (note that the number of occupied boxes is non-increasing so take $R^*_{\text{diff}}(k,k+1) = R^*(k) - R^*(k+1)$). In the previous layer there could have been $1,2,\ldots,N$ occupied boxes. Let $p^{(k)}_{n,i}(N)$ be the probability of having i occupied boxes in layer k when you started by throwing n shots into N boxes. To have a difference of j going from layer k to layer

$k+1$ a one-step transition from i to $i-j$ is required. Then (where $p_{i,r}=0$ for $r\leq 0$),

$$\begin{aligned} g_{R^*_{\text{diff}}}(k,k+1) &= \sum_{i=1}^{N}\sum_{j=0}^{N-1} p_{n,i}^{(k)}(N)p_{i,i-j}^{(1)}(N)z^j \\ &= \sum_{i=1}^{N} p_{n,i}^{(k)}(N)\sum_{j=0}^{N-1} p_{i,i-j}^{(1)}z^j \\ &= \sum_{i=1}^{N} p_{n,i}^{(k)}(N)T_f[g_{R^*(1)|S(k+1)=i}](z) \\ &= \sum_{i=1}^{N} p_{n,i}^{(k)}(N)T_f\circ T_g[g_{R_0(1)|n=i}](z). \end{aligned} \tag{6.2}$$

Let

$$v_{n,N}(z)=\sum_{m=0}^{N-1}\binom{N}{m}z^m(N-m)^n. \tag{6.3}$$

We have already seen (4.2.4) that we can reformulate the p.g.f. for the number of empty boxes in layer one for the scheme with N boxes and a fixed number of shots n as

$$g_{R_0(k)|n=j}(z)=\frac{v_{j,N}(z-1)}{N^j}. \tag{6.4}$$

Plugging this into our expression for $g_{R^*_{\text{diff}}}(k,k+1)$ gives

$$\begin{aligned} g_{R^*_{\text{diff}}}(k,k+1)(z) &= \sum_{j=1}^{N}\frac{p_{n,j}^{(k)}(N)}{N^j}T_f\circ T_g[v_{j,N}](z-1) \\ &= \sum_{j=1}^{N}\frac{p_{n,j}^{(k)}(N)}{N^j}T_f\circ T_g\circ T_{-1}[v_{j,N}](z), \end{aligned} \tag{6.5}$$

where $T_{-1}[p](z)=p(z-1)$ is a real-root-preserving transformation (2.1). Now, applying T_f to both sides of (5.14) gives a new representation for $v_{j,N}(z)$ which can be plugged into (6.5) to give

$$\begin{aligned} g_{R^*_{\text{diff}}}(k,k+1)(z) &= \sum_{j=1}^{N}\frac{p_{n,j}^{(k)}(N)}{N^j}T_f\circ T_g\circ T_{-1}\circ T_f\circ \underbrace{T_d\circ T_d\circ\cdots T_d}_{j\text{ times}}\circ T_1[z^N] \\ &= T_f\circ T_g\circ T_{-1}\circ T_f\circ \sum_{j=1}^{N}\frac{p_{n,j}^{(k)}(N)}{N^j}\underbrace{T_d\circ T_d\circ\cdots T_d}_{j\text{ times}}\circ T_1[z^N]. \end{aligned} \tag{6.6}$$

Now, $T_f\circ T_g\circ T_{-1}\circ T_f$ preserves the number of real roots (since it is a composition of real-root-

preserving transformations) and we showed in Chapter five that

$$\sum_{j=1}^{N} \frac{p_{n,j}^{(k)}(N)}{N^j} \underbrace{T_d \circ T_d \circ \cdots T_d}_{j \text{ times}} \circ T_1[z^N] = T_c[(z+1)^N]$$

has all real roots. So

$$g_{R^*_{\text{diff}}(k,k+1)}(z) = T_f \circ T_g \circ T_{-1} \circ T_f \circ T_c[(z+1)^N].$$

and we are applying a composition of transformations which preserve the number of real roots. Therefore, $g_{R^*_{\text{diff}}(k,k+1)}$ has all real roots. Hence, by induction, we have shown that $g_{R^*_{\text{diff}}(k,k+1)}$ has all real roots for all k, N and $n \in \mathbb{N}$. □

6.3.1 Distributional results for $R^*_{\text{diff}}(k,k+1)$ in the multiple-layer scheme with N boxes

Theorem 6.3.2. Again, let $R^*_{\text{diff}}(k,k+1)$ be the difference in the number of occupied boxes between layer k and $k+1$ of the scheme which starts with n shots and has N boxes throughout. Then, there is a constant $c < \infty$ such that

$$\sup_x \left| \mathbb{P}\left\{ \frac{R^*_{\text{diff}}(k,k+1) - \mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right]}{\sqrt{\text{Var}\left[R^*_{\text{diff}}(k,k+1)\right]}} \leq x \right\} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy \right| \leq \frac{c}{\sqrt{\text{Var}\left[R^*_{\text{diff}}(k,k+1)\right]}}. \tag{6.7}$$

The proof simply follows that for $R^*(k)$ (Theorem 5.3.2).

Corollary 6.3.2.1. Suppose that n and N are varied so that $Var\left[R^*_{\text{diff}}(k,k+1)\right] \to \infty$. Then, $\mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right] \to \infty$ and the distribution of:

$$\frac{R^*_{\text{diff}}(k,k+1) - \mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right]}{\sqrt{\text{Var}\left[R^*_{\text{diff}}(k,k+1)\right]}},$$

tends to the standard normal distribution.

The proof mimics that for $R^*(k)$ (Corollary 5.3.2.1).

Theorem 6.3.3. Let $x(i) = (i - \mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right]) / \sqrt{\text{Var}\left[R^*_{\text{diff}}(k,k+1)\right]}$. Then, there is a constant $c < \infty$ such that:

$$\max_{0 \leq i \leq N} \left| \sqrt{\text{Var}\left[R^*_{\text{diff}}(k,k+1)\right]} \mathbb{P}(R^*_{\text{diff}}(k,k+1) = i) - \frac{1}{\sqrt{2\pi}} e^{-(x(i))^2/2} \right| \leq \frac{c}{\sqrt{\text{Var}\left[R^*_{\text{diff}}(k,k+1)\right]}}. \tag{6.8}$$

Again, this follows in the same as for $R^*(k)$ (Theorem 5.3.3).

Lemma 6.3.1.

$$\rho_\lambda(R^*_{\text{diff}}(k,k+1)) \leq \mathbb{E}[R^*_{\text{diff}}(k,k+1)] - \text{Var}[R^*_{\text{diff}}(k,k+1)], \tag{6.9}$$

where $\lambda = \mathbb{E}(R^*_{\text{diff}}(k,k+1))$ and $\rho_\lambda(X)$ is the total variation distance between the distribution of X and the Poisson distribution with parameter λ.

As above, the proof is the same as that for $R^*(k)$ (Lemma 5.3.2).

6.3.2 Conditions under which this distance to Poisson is small

We have shown that $g_{R^*_{\text{diff}}}(k,k+1)$ has all real roots and can be written as a sum of independent (but not identically distributed) Bernoulli random variables. Then,

$$\rho_\lambda(R^*_{\text{diff}}(k,k+1)) \leq \mathbb{E}[R^*_{\text{diff}}(k,k+1)] - \text{Var}[R^*_{\text{diff}}(k,k+1)].$$

For n and N fixed, as $k \to \infty$,

$$\mathbb{E}[R^*_{\text{diff}}(k,k+1)] \to 0 \quad \text{and} \quad \text{Var}[R^*_{\text{diff}}(k,k+1)] \to 0.$$

Therefore, for n and N fixed, as k increases, the distribution of R^*_{diff} moves closer to a Poisson distribution. That is, as $k \to \infty$,

$$\rho_\lambda[R^*_{\text{diff}}(k,k+1)] \to 0.$$

In contrast, for the number of occupied boxes in a given layer for n and N fixed, as $k \to \infty$,

$$\mathbb{E}[R^*(k)] \to 1 \qquad \text{and} \qquad \text{Var}[R^*(k)] \to 0.$$

In 2015, Zubkov and Serov [27] established a result for iterations of random functions that translates to bounds for the expected number of occupied boxes in a given layer k where we start with n shots and have N boxes in every layer. They showed that

$$n - \binom{n}{2}\frac{k}{N} \leq \mathbb{E}[R^*(k)] < n - \binom{n}{2}\frac{k}{N} + \frac{n^3k^2}{4N^2}. \tag{6.10}$$

Note, these bounds are often not tight but they can give us examples of regions where the Poisson approximation works well. Using (6.10) we can get bounds for the difference in the number of occupied boxes in consecutive layers of the scheme:

$$\mathbb{E}[R^*_{\text{diff}}(k,k+1)] \leq \frac{n^3k^2}{4N^2} + \frac{n^2}{2N} - \frac{n}{2N},$$

and

$$\mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right] \geq -\frac{n^3(k+1)^2}{4N^2} + \frac{n^2}{2N} - \frac{n}{2N}.$$

Putting this together gives:

$$-\frac{n^3(k+1)^2}{4N^2} + \frac{n^2}{2N} - \frac{n}{2N} \leq \mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right] \leq \frac{n^3k^2}{4N^2} + \frac{n^2}{2N} - \frac{n}{2N}. \tag{6.11}$$

If

$$\frac{n^3k^2}{4N^2} \to 0 \text{ and } \frac{n^2}{2N} \to 0,$$

then,

$$\mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right] \to 0 \text{ as } n, N, k \to \infty.$$

Note that

$$\frac{n^2}{2N} = \frac{n}{2}\left(\frac{n}{N}\right) \text{ so } \frac{n^2}{2N} \to 0 \implies \alpha = \frac{n}{N} \to 0.$$

Example 6.3.1. If $n = \left\lfloor N^{1/8} \right\rfloor$ and $k = \left\lfloor N^{1/8} \right\rfloor$ so

$$\frac{n^3k^2}{4N^2} \approx \frac{N^{5/8}}{4N^2} \to 0, \ \frac{n^2}{2N} \approx \frac{N^{2/8}}{2N} \to 0 \ (\text{as } N \to \infty).$$

Consider, for example,

$$N = 256, \ n = 2, \ k = 2: \ \mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right] \leq \frac{33}{8192} \approx 0.004,$$

$$N = 6561, \ n = 3, \ k = 3, \ \mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right] \leq \frac{325}{708588} \approx 0.00046.$$

6.3.3 Simulations

We can consider the multiple-layer scheme for various values of N and calculate the distance between the distribution of the difference and a Poisson distribution with its parameter given by the mean difference in layer k. In order to get the difference probabilities for the larger values of N, the recursive formulae for the probabilities for general N (5.5) were used.

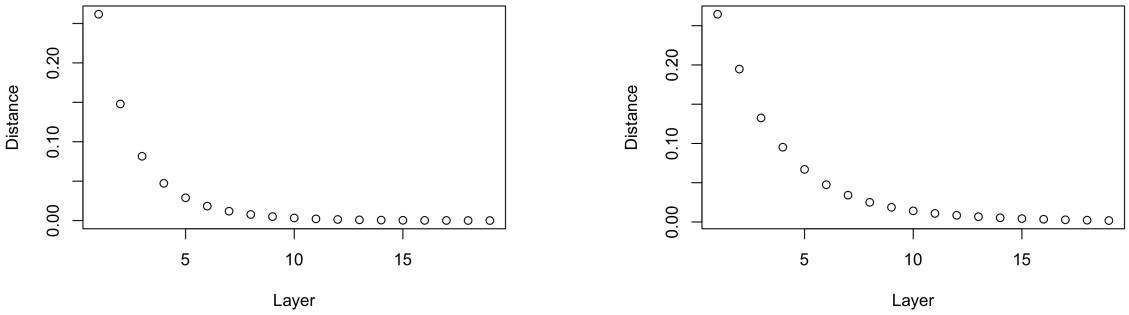

Figure 6.5: Distance to Poisson for the distribution of $R^*_{\text{diff}}(k,k+1)$ for $N=5$ (left) and $N=10$ (right).

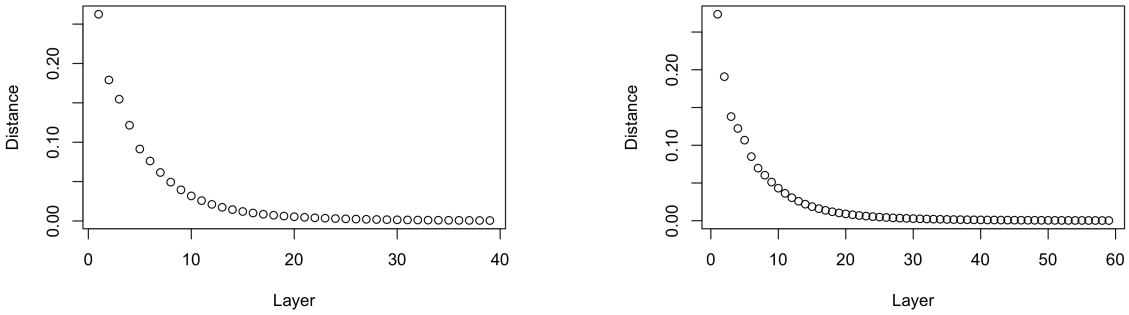

Figure 6.6: Distance to Poisson for the distribution of $R^*_{\text{diff}}(k,k+1)$ for $N=20$ (left) and $N=30$ (right).

6.3.4 Summary of results for $R^*_{\mathbf{diff}}(k,k+1)$ for N boxes

1. The p.g.f. for the difference in the number of occupied boxes between consecutive layers of the scheme with N boxes has all real roots.

2. Let $R^*_{\text{diff}}(k,k+1)$ be the difference in the number of occupied boxes between layer k and $k+1$ of the scheme which starts with n shots and has N boxes throughout. Then, there is a constant $c < \infty$ such that an upper bound on the distance between $R^*_{\text{diff}}(k,k+1)$ (standardised) and the standard normal distribution is

$$\frac{c}{\sqrt{\text{Var}\left[R^*_{\text{diff}}(k,k+1)\right]}} \to 0 \text{ if } \text{Var}\left[R^*_{\text{diff}}(k,k+1)\right] \to \infty.$$

3. An upper bound for the distance from the distribution of $R^*_{\text{diff}}(k,k+1)$ to a Poisson distribution with parameter $\lambda = \mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right]$ is

$$\mathbb{E}\left[R^*_{\text{diff}}(k,k+1)\right] - \text{Var}\left[R^*_{\text{diff}}(k,k+1)\right].$$

Chapter 7

Summary and discussion

The allocation of n shots to N boxes according to either a uniform or more general non-uniform distribution has been extensively studied by various authors (for example, [2], [5], [20]). The work in this thesis considered a multiple-layer extension of the uniform scenario. In the single-layer scheme the number of shots was fixed. I also fix the initial number of shots but from layer two onwards the number of shots being scattered into each layer becomes random (although necessarily non-increasing). This additional complexity is why the simpler uniform allocation was specified from the outset. Before moving to the general scheme we started with small finite examples allowing us to try various methods of proof and to get a better understanding of how the distribution of the number of occupied boxes behaves moving through the layers.

7.1 Conclusions

The initial aim of this research was to build on the work done for a single-layer shots-and-boxes scheme and to expand this to a multiple-layer scheme. One motivation of such a scheme is that it then allows us to generate a model for a tree describing the way objects might sequentially coalesce which opens up applications to, for example, studying how ancestral lineages evolve over time in some population of organisms [47]. More generally, having multiple layers allows time to be incorporated into the model or, depending on the desired application, layers could also be used to represent spatial location. The ultimate goal was to produce limit theorems for the distribution of the number of occupied boxes in any given layer.

Before moving to this multiple-layer scheme for a general number of boxes, I began by studying a small finite example where there were just three boxes, to establish the behaviour of the model in this case when brute-force methods are tractable. In this way, for three boxes I was able to establish explicit, non-recursive formulae for the expectation, variance and p.g.f. for the number of occupied boxes in any given layer. It proved fairly straightforward to show that this p.g.f. has all real roots for all layers. Additionally, for three boxes I could get explicit formulae for the success probabilities of the independent Bernoulli random variables into which the number of

occupied boxes could be decomposed.

When moving to four and five boxes it can be seen how even adding one additional box adds complexity to these formulas and to the difficulty in proving the p.g.f. has all real roots. I was able to show applying a strategy first used by Vatutin and Mikhailov [3] for the scheme with N boxes that the p.g.f. for the number of occupied boxes in any given layer has all real roots. Hence, the number of occupied boxes in any layer can be written as a sum of independent (but not identically distributed) Bernoulli random variables. From this, normal and Poisson limiting results could then be established for the distribution of the number of occupied boxes.

The distributional results proven in this thesis are asymptotical but we have shown that even for any finite number of shots and boxes the number of occupied boxes can be written as a sum of independent Bernoulli random variables. We do not however have explicit formulae for the parameters of those Bernoullis and we showed that even for three boxes these take quite complex form. If a particular application involved working with a finite sample size then trying to establish exactly what the parameters of these Bernoullis would be could be beneficial but challenging. Note that this can be thought of in terms of roots of polynomials since we have shown the p.g.f. has all real roots but we have not established what the roots are exactly. Numerical approaches to determining the roots will work so long as N is not too large. Further investigation into how to check that a given choice of n and N satisfy the conditions of the limit theorems proven in this thesis would also be beneficial.

An advantage of considering a multiple-layer scheme is that it allows us to consider for example the difference in the number of occupied boxes between consecutive layers. Again, for small finite examples, I could provide explicit formulas for the expectation, variance and p.g.f. of this difference between any pair of consecutive layers. More importantly, I showed that the p.g.f. for the difference in the number of occupied boxes between consecutive layers of the scheme with a general number of boxes has all real roots. As a result, normal and Poisson limiting results followed as well as examples where Poisson approximation in particular works well.

7.2 Estimating population size

Suppose we take a small set of n mitochondrial DNA sequences which we can represent as shots in the shots and boxes model. Behind these sequences is a tree that we do not know but we can get an estimate of this using maximum likelihood. At this stage all the branches are measured in units of the average number of mutations but we can convert this to generations using the mutation rate. Then, if you look at the tree for a particular generation k and count the number of ancestral lines this is equivalent to $R^*(k)$ (number of occupied boxes in layer k of the shots and boxes model).

Now, the question is whether we can use what we know about $R^*(k)$ to construct the likelihood of N which here would represent the population size. In a setting such as this one we anticipate

that n will be much smaller than N and we have seen that in this scenario the distribution of the difference in the number of occupied boxes between consecutive layers is close to Poisson. We also know that these differences are independent so if we assume a Poisson distribution we will have a product of Poisson probabilities which can then be maximised with respect to N.
I started with a dataset containing 53 mitochondrial DNA sequences and assumed the differences in the number of occupied boxes between consecutive layers was Poisson. I then constructed the log-likelihood and found it reached its maximum when $N = 31049$. To check the plausibility of this value I simulated some trees starting with $n = 53$ and $N = 31049$ and compared them with the actual tree we are considering.

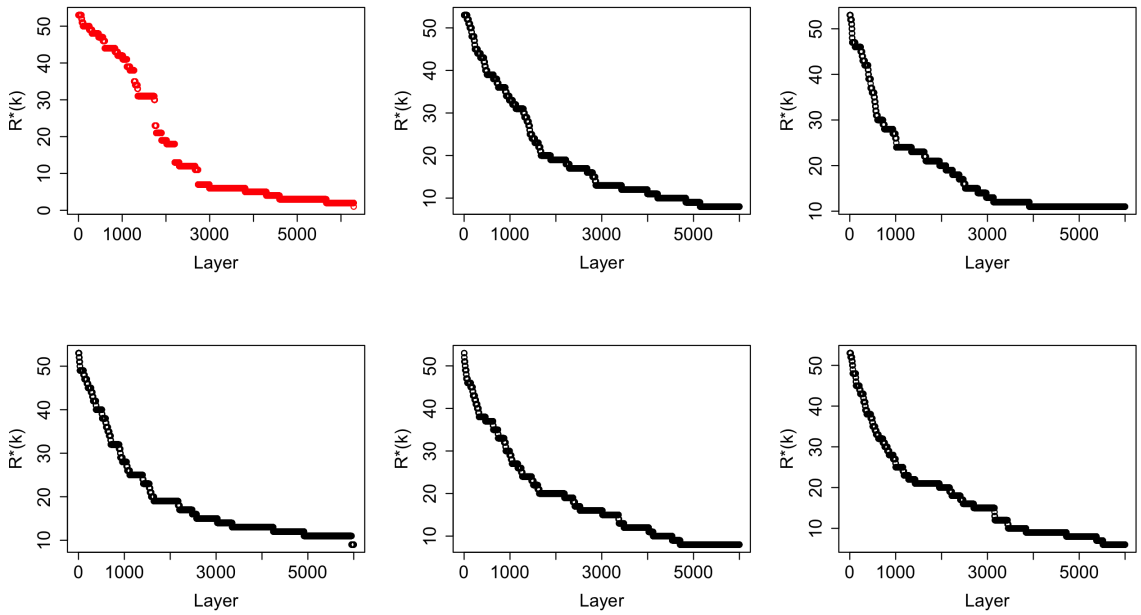

Figure 7.1: Comparing the estimated tree from the DNA data being considered (red, top-left) with simulated trees using the obtained value of N.

7.3 Limitations and future work

Due to the added complexity of a multiple-layer scheme all of the limiting results and even the finite results shown in this thesis are under the assumption that shots are allocated using the uniform distribution. In the case of a single-layer shots-and-boxes scheme most of the results still hold when considering non-uniform allocation but further assumptions were needed. It would be interesting to try to prove which (if any) of our results still hold and under what conditions. This would also help to improve applicability as it would allow for more flexibility in the model. For a single layer limiting results were first obtained by various authors for the number of empty (or occupied) boxes and then their work was extended (usually by others) to show that these same results held for the number of boxes containing exactly m shots. Hence, another natural

extension of the work in this thesis would be to persist with the assumption of a uniform allocation of shots but try to establish analogous results for the number of boxes containing exactly m shots. If it were the case that the results for the number of occupied boxes produced in this thesis also held for a non-uniform allocation of shots (under appropriate restrictions) then, as for the uniform case, it would be of interest to try and extend this to the number of boxes containing exactly m shots. In this thesis the distributional results for the number of occupied boxes are for a single layer or for a single difference between layers so that they are univariate limit theorems. A possible though challenging extension to this would be to establish multivariate asymptotic distributional results for example for the joint distribution of occupancy numbers across l layers. Another advantage of this multiple-layer scheme is that we can consider random variables that did not arise for a single layer. In this thesis I looked at the difference in the number of occupied boxes between consecutive layers but it would be worth considering, based on potential applications, other univariate quantities that could be of interest.

At the beginning of this thesis, (single-layer) schemes with an infinite number of boxes were discussed but were not explored further. This would be another possible area of direction of research based on work shown here. More recently, functional central limit theorems for these infinite schemes have been produced ([14], [15]) and this is again something not considered in this thesis but it would allow the multiple-layer scheme to be studied as a whole rather than only considering one layer at a time.

As well as considering how to extend the multiple-layer scheme presented in thesis we could also consider what would happen if, say, rather than just being interested in the number of boxes which are occupied (empty or occupied by exactly m shots) we also wanted to know where the shots were located. Additionally, for some applications it also might be of interest to label the shots from the outset. An area of recent research in the context of iterations of random functions has considered having two multiple-layer schemes running simultaneously. Here, the schemes are marginally uniform but they are dependent on one another which could be used, for example, to model DNA recombination [28].

Overall, given that the results produced under a uniform allocation in single-layer scheme extend to a multiple-layer scheme it is promising that there are many possible avenues of future research based on the foundations developed in this thesis.

Bibliography

[1] Weiss, I., (1958), Limiting distributions in some occupancy problems, *Annals of Mathematical Statistics*, **29**, 878-884.

[2] Kolchin, V., Sevast'yanov, B., Chistyakov, V., (1978), *Random Allocations*, V. H. Winston, Washington.

[3] Vatutin, V.A., Mikhailov, V.G., (1982), Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles, *Theory of Probability and Its Applications*, **27**, 734-743.

[4] Békéssy, A., (1963), On classical occupancy problems, *Magy.Tud.Akad.Mat.Kutato Int.Kozl*, **8**, 59-71.

[5] Rényi, A., (1962), Three new proofs and generalization of a theorem of Irving Weiss, *Magy.Tud.Akad.Mat.Kutato Int.Kozl*, **7**, 203-214.

[6] Elliot, W.E.Y., Valenza, R.J., (1996), And then there were none: winnowing the Shakespeare claimants, *Computers and the Humanities*, **30**, 191-245.

[7] Bunge, J., Fitzpatrick, M., Estimating the number of species: a review, *Journal of the American Statistical Association*, **88**, 364-373.

[8] Gnedin, A., Hansen, B., Pitman, J., (2007) Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws, *Probability Surveys*, **4**, 146-171.

[9] Skinner, C.J., Elliot, M.J., (2002), A measure of disclosure risk for microdata, *Journal of the Royal Statistical Society: Series B*, **64**, 855-867.

[10] Chaudhuri, S., Motwani, R., Narasayya, (1998), Random sampling for histogram construction: how much is enough?, *ACM SIGMOD Conference on Management of Data*, **1998**, 436-447.

[11] Wakeley, J., (2009), *Coalescent Theory: An Introduction*, Roberts and Company Publishers, Colorado.

[12] Dukhin, S.S., Sjoblom, J., Wasan, D.T. et al., (2001), Coalescence coupled with either coagulation or flocculation in dilute emulsions, *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, **180**, 223-234.

[13] Rani, Y., Rohil, H., (2013), A Study of Hierarchical Clustering Algorithm, *International Journal of Information and Computation Technology*, **3**, 1115-1122.

[14] Chebunin, M., Zuyev, S., (2019), Functional central limit theorems for occupancies and missing mass process in infinite urn models, arXiv:1906.10949.

[15] Chebunin, M., Kovalevskii, A., (2016), Functional central limit theorems for certain statistics in an infinite urn scheme, *Statistics and Probability Letters*, **119**, 344-348.

[16] Chistyakov, V.P., (1967), Computation of the power of the empty cell test, *Matem. Zametki*, **1**, 9-16.

[17] David, F. N., Barton, D. E., (1962), *Combinatorial Chance*, Hafner Publishing Company, London.

[18] Englund, G., (1981), A remainder term estimate for the normal approximation in classical occupancy, *Annals of Probability*, **9**, 684-692.

[19] Quine, M. P., Robinson, J., (1982), A Berry-Esseen bound for an occupancy problem, *Annals of Probability*, **10**, 663-671.

[20] Penrose, M.D., (2009), Normal approximation for isolated balls in an urn allocation model, *Electronic Journal of Probability*, **14**, 2156-2181.

[21] Goldstein, L., Penrose, M. D., (2010), Normal approximation for coverage models over binomial point processes, *Annals of Applied Probability*, **20**, 696-721.

[22] Hwang, H., Janson, S., (2008), Local limit theorems for finite and infinite urn models, *Annals of Probability*, **36**, 992-1022.

[23] Karlin, S., (1967), Central limit theorems for certain infinite urn schemes, *Journal of Mathematics and Mechanics*, **17**, 373-401.

[24] Bahadur, R. R., (1960), On the number of distinct values in a large sample from an infinite discrete distribution, *Proceedings of the National Academy of Sciences India Part A*, **26**, 67-75.

[25] Darling, D. A., (1967), Some limit theorems associated with multinomial trials, *Contributions to Probability Theory*, **5.2A**, 345-350.

[26] Dutko, M., (1989), Central limit theorems for infinite urn models, *Annals of Probability*, **17**, 1255-1263.

[27] Zubkov, A.M., Serov, A.A., (2015), Images of subset of finite set under iterations of random dependent mappings, *Discrete Mathematics and Applications*, **25**, 179-185.

[28] Serov, A.A., (2016), Images of a finite set under iterations of two random dependent mappings, *Discrete Mathematics and Applications*, **26**, 175-181.

[29] Zubkov, A.M., Serov, A.A., (2018), Estimates of the mean size of the subset image under composition of random mappings, *Discrete Mathematics and Applications*, **28**, 331-338.

[30] Kurtz, D.C., (1992), A sufficient condition for all the roots of a polynomial to be real, *American Mathematical Monthly*, **99**, 259-263.

[31] Alotaibi, I.J., (2015), *Generalised Sturm Sequences*, California State University, Los Angeles.

[32] Rahman, Q.I., Schmeisser, G., (2002), *Analytic Theory of Polynomials*, Clarendon Press, Oxford.

[33] Diaconis, P., Freedman, D., (1999), Iterated random functions, *SIAM Review*, **41**, 45-76.

[34] Wright, S., (1931), Evolution in Mendelian populations, *Genetics*, **16**, 97-159.

[35] Pongprasert, S., Chaengsisai, K., Kaewleamthong, W., Sriphrom, P., (2021), Real Root Polynomials and Real Root Preserving Transformations, *International Journal of Mathematics and Mathematical Sciences*, **2021**, 1-5.

[36] Mises, R., (1939), Uber Aufteilungs and Besetzungs-Wahrscheinlichkeiten, *Revue de la Faculté des Sciences de l'Université d'Istanbul*, **4**, 145-163.

[37] Kolchin, V., (1966), The speed of convergence to limit distributions in the classical ball problem, *Teoriya Veroyatn. i Yeye Primenen.*, **11**, 144-156.

[38] Harris, B., Park, C.J., (1971), The limiting distribution of the sample occupancy numbers from the multinomial distribution with equal cell probabilities, *Annals of the Institute of Statistical Mathematics*, **23**, 125-133.

[39] Erdos, P., Rényi, A., (1961), On a classical problem of probability theory, *Magy.Tud.Akad.Mat.Kutato Int.Kozl*, **6**, 215-220.

[40] Chistyakov, V.P., (1964) Computation of the power of the empty cell test, *Teoriya Veroyatn. i Yeye Primenen.*, **9**, 718-724.

[41] Holst, L., (1971), Limit theorems for some occupancy and sequential occupancy problems, *Annals of Mathematical Statistics*, **42**, 1671-1680.

[42] Kitabatake, S., (1958) A remark on a non-parametric test, *Japanese Journal of Mathematics*, **5**, 45-49.

[43] Sevast'yanov, B.A., (1972), A limiting Poisson law in a scheme of dependent random variables, *Teoriya Veroyatn. i Yeye Primenen.*, **17**, 733-738.

[44] Petrov, V.V., (1975), *Sums of Independent Random Variables*, Springer-Verlag, Berlin-Heidelberg.

[45] Le Cam, L., (1960), An approximation theorem for the Poisson binomial distribution, *Pacific Journal of Mathematics*, **10**, 1181-1197.

[46] R Core Team, (2020), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.

[47] Nee, S., May, R.M., Harvey, P.H., (1994), The reconstructed evolutionary process, *Philosophical Transactions: Biological Sciences*, **344**, 305-311.

[48] Feller, W., (1971), *An Introduction to Probability Theory and its Applications, Vol.2*, Wiley, New York.

[49] Chen, L.H., Xia, A., (2004), Stein's method, palm theory and Poisson process approximation, *Annals of Probability*, **32**, 2545-2569.

[50] Borcea, J., Branden, P., (2009), Polya-Schur master theorems for circular domains and their boundaries, *Annals of Mathematics*, **170**, 465-492.

[51] Dedieu, J.P, (1992), Obreschkoff's theorem revisited: what convex sets are contained in the set of hyperbolic polynomials?, *Journal of Pure and Applied Algebra*, **81**, 269-278.

[52] Branden, P., Krasikov, I., (2016), Elements of Polya-Schur theory in the finite difference setting, *Proceedings of the American Mathematical Society*, **144**, 4831-4843.

[53] Fisk, S., (2006), Polynomials, roots, and interlacing, arXiv:math/0612833.

[54] Rogers, J.W., (1983), Locations of roots of polynomials, *Society for Industrial and Applied Mathematics*, **25**, 327-342.

[55] Michelen, M., Sahasrabudhe, J., (2018), Central limit theorems from the roots of probability generating functions, arXiv:1804.07696.

[56] Dilcher, K., (1991), Real Wronskian zeros of polynomials with nonreal zeros, *Journal of Mathematical Analysis and Applications*, **154**, 164-183.

[57] Sevast'yanov, B.A., (1966), Limit theorems in a scheme for allocation of particles in cells, *Theory of Probability and Its Applications*, **11**, 614-619.

[58] Mikhailov, V.G., (1977), A Poisson limit theorem in a scheme for group allocation of particles, *Theory of Probability and Its Applications*, **22**, 152-156.

[59] Mikhailov, V.G., (1977), An estimate of the rate of convergence to the Poisson distribution in group allocation of particles, *Theory of Probability and Its Applications*, **22**, 554-562.

[60] Mikhailov, V.G., (1980), Asymptotic normality of the number of empty cells for group allocation of particles, *Theory of Probability and Its Applications*, **25**, 82-90.

[61] Mikhailov, V.G., (1981), Convergence to a multi-dimensional normal law in an equiprobable scheme for group allocation of particles, *Math. USSR-Sb*, **39**, 145-168.

[62] Harper, L., (1961), Stirling behaviour is asymptotically normal, *Annals of Mathematical Statistics*, **38**, 410-414.

[63] Barbour, A. D., Gnedin, A. V., (2009), Small counts in the infinite occupancy scheme, *Electronic Journal of Probability*, **14**, 365-384.

[64] Barbour, A. D., Holst, L., Janson, S., (1992), *Poisson Approximation*, Oxford University Press, New York.

[65] Chatterjee, S., (2008), A new method of normal approximation, *Annals of Probability*, **36**, 1584-1610.

[66] Johnson, N. L., Kotz, S., (1977), *Urn Models and their Application: An Approach to Modern Discrete Probability Theory*, John Wiley & Sons, New York.

[67] Mikhailov, V. G., (1981), The central limit theorem for a scheme of independent allocation of particles by cells, *Number Theory, Mathematical Analysis and their Applications*, **157**, 138-152.

[68] Steele, J. M., (1986), An Efron-Stein inequality for non-symmetric statistics, *Annals of Statistics*, **14**, 753-758.

[69] Feller, W., (1968), *An Introduction to Probability Theory and its Applications Vol.1 3rd ed.*, John Wiley and Sons, New York.

[70] Sachkov, V. N., (1978), *Probabilistic Methods in Combinatorial Analysis*, Nauka, Moscow.

[71] Ivanov, V. A., Ivchenko, G. I., Medvedev, Y. I., (1985), Discrete problems in probability theory, *Journal of Mathematics Sciences*, **31**, 2759-2795.

[72] Sevast'yanov, B. A., Chistyakov, V. P., (1964), Asymptotic normality in the classical ball problem, *Theory of Probability and Its Applications*, **9**, 198-211.

[73] Kolchin, V. F., (1967), Uniform local limit theorems in the classical ball problem for a case with varying lattices, *Theory of Probability and Its Applications*, **12**, 57-67.

[74] Kotz, S., Balakrishnan, N., (1997), Advances in urn models during the past two decades, *Advances in Combinatorial Methods and Applications to Probability and Statistics*, 203-257.

[75] Kesten, H., (1968), Review of Darling. Some limit theorems associated with multinomial trials, *Mathematical Reviews*, **35**, 73-78.

[76] Bogachev, L. V., Gnedin, A. V., Yakubovich, Y. V., (2006), On the variance of the number of occupied boxes, *Advances in Applied Mathematics*, **40**, 401-432.

[77] Rais, B., Jacquet, P., Szpankowski, W., (1993), Limiting distribution for the depth in PATRICIA tries, *SIAM Journal on Discrete Mathematics*, **6**, 197-213.

[78] Hitczenko, P., Louchard, G., (2001), Distinctness of compositions of an integer: A probabilistic analysis, *Random Structures Algorithms*, **19**, 407-437.

[79] Hwang, H. K., Yeh, Y. N., (1997), Measures of distinctness for random partitions and compositions of an integer, *Advances in Applied Mathematics*, **19**, 378-414.

[80] Gnedin, A., Pitman, J., Yor, M., (2006), Asymptotic laws for compositions derived from transformed subordinators, *Annals of Probability*, **34**, 468-492.