The solution of a recursive sequence arising from a combinatorial problem in botanical epidemiology

Z. AlSharawi[†][*], A. Burstein[§], M. Deadman[‡], A. Umar[†]

†Department of Mathematics and Statistics, Sultan Qaboos University
P. O. Box 36, PC 123, Al-Khod, Sultanate of Oman

§Department of Mathematics, Howard University
Washington, DC 20059, USA

‡Department of Crop Sciences, Sultan Qaboos University
P. O. Box 34, PC 123, Al-Khod, Sultanate of Oman

June 9, 2012

Abstract

One of the central problems in botanical epidemiology is whether disease spreads within crops in a regular pattern or follows a random process. In this paper, we consider a row of n plants in which m are infected. We then develop a rigorous mathematical approach to investigate the total number of ways to obtain k isolated individuals among m infected plants. We give a recurrence relation in three parameters that describes the problem, then we find a closed form solution, and give two different approaches to tackle the proof. Finally, we find interesting formulas for the expectation and variance of the random variable that represents the number of infected and isolated plants.

Keywords: Spread of disease, Recurrence relation, Binomial coefficients, Hypergeometric function.
Mathematics Subject Classification: 92D30, 65Q30, 05A05, 05A15.

1 Introduction

An epidemic, whether afflicting plants or animals, might usefully be described as disease that is concentrated in time and space. Analysis of temporal change in the amount of disease is well documented and frequently utilizes non-linear regression techniques to estimate parameters such as rates of epidemic development and asymptotic disease levels. In botanical epidemiology, the logistic model $x'(t) = \frac{r}{K}x(t)(K - x(t))$ has perhaps been the most widely used to describe and compare temporal progress of disease [8], where $x(t)$ represents the quantity of disease at time t, r is a disease

[*] Corresponding author: alsha1zm@alsharawi.info

rate parameter, and K is the maximum quantity of disease. Variations of the logistic, especially the Gompertz and monomolecular [6] have also been widely adopted by plant pathologists wishing to study change in the amount of disease over time or variations in disease levels caused by plant variety or experimental location. Precise characterization of patterns of disease increase over time is essential for timely intervention through disease management practices.

The spatial characteristics of disease progress in a plant population are just as important for disease management as are the temporal characteristics [3]. The pattern of spatial progress provides vital information about the effectiveness of management practices. Plant pathologists frequently use various classifications to describe spatial patterns of diseased plants: random or non-random (aggregated, clustered, regular).

Statistical techniques more recently adopted by plant pathologists to investigate spatial aspects of disease progress have included spatial autocorrelation analysis to measure disease aggregation [1], Lloyd's Index of Patchiness [9], variogram analysis [7] and the use of probability distributions such as beta-binomial analysis [5].

Certain plantings affected by disease represent a special case requiring a distinct approach to spatial analysis of progress. Amenity plantings and certain row crops are essentially one dimensional with little or no interference between rows. Disease progresses from isolated foci which expand via neighboring plants. Significant information can therefore be obtained by analyzing the extent of clustering of infected individuals along rows, or conversely, the extent to which infected individuals are isolated along rows. To draw some inferences about the dispersal mechanism of the pathogen and consequent actions formulated to limit the spread of disease, a contrast is required between experimental data and theoretical results concerning the random distribution of infected individuals. This paper is devoted to developing the theoretical tools necessary for achieving this objective. We consider the infected and isolated individuals to be our random variable and find explicit formulas for the total number of ways to have k isolated singles among m infected plants in a row of n cells. Then we find a simplified formula for the expectation of our random variable, and use it to find an explicit form for the variance. The theoretical results of this paper will be used in the analysis of experimental work and will be published elsewhere.

2 The number of the isolated-infected plants

Let $F(n, m, k)$ be the total number of ways to have k diseased but isolated individuals among m infected plants in a row of n plants (or cells). Before we embark on giving a formula for $F(n, m, k)$, we clarify our notation by a simple example. Suppose that we have a row of $n = 8$ cells with $m = 5$ infected, and we are looking for the total number of ways to have $k = 3$ isolated individuals. Table 1 shows that we have a total of $F(8, 5, 3) = 4$ possibilities.

In this section, we give the following main result and its proof:

●	●	○	●	○	●	○	●
●	○	●	●	○	●	○	●
●	○	●	○	●	●	○	●
●	○	●	○	●	○	●	●

Table 1: This table shows all possibilities of having 3 isolated individuals among 5 infected plants in a row of 8 cells. In this table, ● denotes an infected plant and ○ denotes a healthy one.

Theorem 2.1. *For any integers $n \geq m \geq k \geq 0$ such that $m \geq k+2$, we have*

$$F(n,m,k) = \binom{n-m+1}{k} \sum_{j\geq 0} \binom{n-m-k+1}{j+1}\binom{m-k-j-2}{j}.$$

It is worth mentioning that here and in the sequel of this paper, $\binom{n}{k}$ is given the usual meaning when $n \geq k \geq 0$, and we extend its meaning to be 0 when $n < k \neq 0$.

The proof of Theorem 2.1 can be achieved using two independent approaches. The first approach is by a combinatorial argument. The second approach is by establishing a recursive sequence, then use a triple-induction argument to solve the recursive sequence. Since each approach can be appealing to certain segment of the audience, we give both approaches. However, we alert the reader that the induction argument is long; however, we shorten it by writing the main ideas and leave some technical steps for the interested reader.

2.1 The combinatorial approach

We give the combinatorial approach for proving Theorem 2.1.

A proof of Theorem 2.1. Given a row of n plants with m infected ones in which k of the infected are isolated, we can map it to a weak composition of m with $n-m+1$ parts (a *weak composition* is an ordered partition of a nonnegative integer into nonnegative integer parts. For more information, see reference [4]). Let the first part of the composition be the number of (infected) plants to the left of the leftmost healthy plant, the last part be the number of the (infected) plants to the right of the rightmost healthy plant, and for $2 \leq i \leq n-m$, let the ith part of the composition be the number of the (infected) plants between the $(i-1)$-st and the ith healthy plant. Obviously, the sum of the parts thus obtained is m, the number of infected plants, and the number of parts is $n-m+1$, one more than the number of healthy plants.

Note that every isolated infected plant is thus mapped to a part of size 1. Therefore, we need to enumerate weak compositions of m into $n-m+1$ parts exactly k of which are 1s.

Let $C(a,b)$ be the set of weak compositions of a into b parts. Given a weak composition $\lambda \in C(m, n-m+1)$ with exactly k parts of size 1, let j be the number of parts of size at least 2 in λ.

Remove the parts of size 0 and 1 from λ to obtain a composition $\lambda' \in C(m-k, j)$ with no part of size 0 or 1. Thus, either $\lambda = \emptyset$ or (if $j \geq 1$) all parts of λ' are of size at least 2. Each composition

$\lambda' \in C(m-k,j)$ corresponds to $\binom{n-m+1}{k}\binom{n-m-k+1}{j}$ weak compositions $\lambda \in C(m, n-m+1)$ since there are that many ways to choose the positions of the k ones and $n-m-k-j+1$ zeros in λ and fill the rest of the positions with parts of λ' in order.

If $j = 0$, then $k = m$, so that

$$F(n,m,m) = \binom{n-m+1}{m},$$

If $j \geq 1$, then subtract 1 from each part of λ' to obtain a composition $\lambda'' \in C(m-k-j, j)$ with all positive parts. The number of such compositions is well known; it is $\binom{m-k-j-1}{j-1}$ (e.g., see [2, Section 3.5]). Thus, if $j \geq 1$, i.e. if $m \geq k+2$, then

$$F(n,m,k) = \sum_{j\geq 1} \binom{n-m+1}{k}\binom{n-m-k+1}{j}\binom{m-k-j-1}{j-1},$$

which is the same as the required result.

□

2.2 The recursive sequence approach

We start by listing the following facts, which are elementary or can be deduced from the definition of $F(n,m,k)$.

Proposition 2.1. *Each of the following holds true:*
(i) $F(n,0,0) = 1$ *and* $F(n,1,0) = 0$ *for all* $n \geq 1$.
(ii) $F(n,n,n) = 1$ *for* $n = 0,1$ *and* $F(n,n,n) = 0$ *for all* $n \geq 2$.
(iii) $F(n,m,k) = 0$ *whenever* $n < m$, $m < k$ *or* $n,m,k < 0$.
(iv) $F(n,n,0) = 1$ *if* $n \neq 1$ *and* $F(n,n,k) = 0$ *if* $k > 0, n > 1$.
(v) $F(n,1,1) = n$ *for all* $n \geq 1$.

Next, we give some particular cases of $F(n,m,k)$, which can give some convenience to the reader before we embark on the general case.

Proposition 2.2. *Each of the following holds true:*

(i) $F(n,2,0) = n-1$ *for all* $n \geq 1$.
(ii) $F(n,n-1,1) = 2$ *for* $n = 2$ *and all* $n \geq 4$. *For* $n = 3$, *we have* $F(3,2,1) = 0$.
(iii) $F(n,n-1,0) =$

$$\begin{cases} \frac{1}{2}(n-2)(3n-5), & n = 1,2,3 \\ n-2 & n \geq 4. \end{cases}$$

Proof. (i) Zero isolated means that the two infected plants are adjacent, and therefore, we can consider the two infected ones as one unit to be chosen from $n-1$ cells. Therefore, we obtain

$$F(n,2,0) = \binom{n-1}{1} = n-1.$$

To prove (ii), observe that the only way to obtain one isolated plant among $n-1$ infected in a row of n cells is when the infected plant is situated in the far left or far right cell, and the adjacent cell is occupied by a noninfected one. So we obtain only two cases as long as $n \neq 3$. The case $n = 3$ is an exception and obviously we obtain $F(3, 2, 1) = 0$. Finally, we prove (iii). Observe that Part (i) of Proposition 2.1 gives the formula for $n = 1, 2$. The formula for $n = 3$ follows from (i). Thus, we need to focus on $n \geq 4$, but this is simple if we think about the healthy plant rather than the $n-1$ infected ones. To obtain zero isolated plants, the healthy plant can be anywhere except in the second cell from the left or the second cell from the right. Therefore, we obtain $n-2$ cases and the proof is complete. □

Now, we give the following general recursive sequence for $F(n, m, k)$.

Lemma 2.1. *Suppose that $n > m \geq k \geq 0$, then $F(n, m, k)$ satisfies the recursive relation*

$$F(n,m,k) = F(n-1,m,k) + F(n-2,m-1,k-1) + \sum_{j=2}^{m-k} F(n-1-j, m-j, k).$$

Proof. Consider $n > m \geq k \geq 0$. Observe that $F(r, s, t) = 0$ if either $s < 0$ or $t < 0$ as given in Proposition 2.1. This fact makes the notation on the right hand side of the formula well-defined. Next, without loss of generality, we start with the cell on the far left, either it is occupied by an infected plant or not. If not occupied by an infected plant, then we are left with k isolated among the m infected in $n-1$ cells, i.e., $F(n-1, m, k)$. If the first cell is occupied by an infected one, then we proceed to the second cell. In this case, the second cell is either occupied by an infected plant or not; if not, then we are left with $k-1$ isolated singles among $m-1$ infected in $n-2$ cells, i.e., $F(n-2, m-1, k-1)$. However, if the second cell is occupied by an infected plant, then we look at the third cell. In this case, we are looking for the number of ways to have k isolated among $m-2$ infected in $n-3$ cells. By proceeding in the same way, the induction process defines the required recursive sequence. □

Next, we use Lemma 2.1 to obtain the following result:

Theorem 2.2. *For $N \geq m+1 \geq 3$, each of the following holds true for $F(N, m, k)$:*

(i) For $k \geq 1$, we have

$$F(N,m,k) = \sum_{n=m+1}^{N} F(n-2, m-1, k-1) + \sum_{n=m+1}^{N} \sum_{j=0}^{m-k-2} F(n+k-m-1+j, k+j, k).$$

(ii) For $k = 0$, we have

$$F(N,m,0) = 1 + \sum_{n=m+1}^{N} \sum_{j=0}^{m-2} F(n-m-1+j, j, 0).$$

Proof. Sum both sides of the formula in Lemma 2.1 from $n = m + 1$ to $n = N$, then use the fact that $F(m, m, k) = 0$ for all $k \geq 1$ to obtain (i). To obtain (ii), observe that $F(m, m, 0) = 1$ and $F(n - 2, m - 1, -1) = 0$. □

Before we proceed, we layout our strategy for tackling the problem. Our strategy is based on induction, and since we have three parameters $n \geq m \geq k \geq 0$, we start by establishing the "anchors" $F(n, n, k), F(n, m, m)$ and $F(n, m, 0)$. Observe that $F(n, n, k)$ is given in Part (iv) of Proposition 2.1. Next, we establish the formula for $F(n, m, m)$, but first, recall Pascal's formula [2]

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1} \tag{2.1}$$

and the following consequence of Pascal's formula

$$\sum_{j=k}^{n-2} \binom{j}{k} = \binom{n-1}{k+1}. \tag{2.2}$$

Lemma 2.2. *Let n and m be nonnegative integers. We have*

$$F(n, m, m) = \binom{n+1-m}{m}.$$

Proof. The proof is by induction on n and m. Since $n \geq m$, the "anchors" are $F(n, 0, 0)$ and $F(m, m, m)$, which are given in Proposition 2.1. Now, assume the formula is satisfied for $F(n - 1, m, m)$, then use Lemma 2.1 and Formula 2.1 to obtain the required formula for $F(n, m, m)$. □

In a similar way, we find the following result.

Proposition 2.3. *Let n and m be nonnegative integers. Each of the following holds true:*

$$(i) \quad F(n, m+1, m) = 0$$

$$(ii) \quad F(n, m+2, m) = (m+1)\binom{n-m-1}{m+1}$$

$$(iii) \quad F(n, m+3, m) = (m+1)\binom{n-m-2}{m+1}.$$

Proof. (i) This is obvious. So, we proceed to prove (ii) by induction on m and n. When $m = 0$, we obtain $F(n, 2, 0) = n - 1$ as given in Proposition 2.2, which is the same as the formula. Also, when $n = m + 2$, we have $F(m + 2, m + 2, m)$ addressed in Part (iv) of Proposition 2.1. Now, assume the formula holds for $F(n - 1, m + 2, m)$, and use it to prove the formula for $F(n, m + 2, m)$. From Theorem 2.2 and Lemma 2.2, we obtain

$$\begin{aligned} F(n, m+2, m) &= \sum_{j=m+3}^{n} F(j-2, m+1, m-1) + \sum_{j=m+3}^{n} F(j-3, m, m) \\ &= (m+1) \sum_{j=1}^{n-m-2} \binom{j}{m}. \end{aligned}$$

Next, use the formula in Eq. (2.2) to obtain the required result. Proving (iii) is similar, and therefore, the proof is omitted. □

In our induction in the previous result, we used $F(n,2,0)$ in Part (ii) and $F(n,3,0)$ in part (iii), which are easy to find. However, we need to establish the "anchor" $F(n,m,0)$ in general. The next result gives the answer.

Lemma 2.3. *For all integers $n \geq m > 1$, we have*

$$F(n,m,0) = \sum_{j\geq 0} \binom{m-2-j}{j}\binom{n-m+1}{j+1}.$$

Proof. At $m=2$ or $n=m$, the formula is satisfied. Assume the formula is satisfied for $F(n-1,m,0)$. Now, use Lemma 2.1 to obtain

$$F(n,m,0) = F(n-1,m,0) + 1 + \sum_{k=2}^{m-2} F(n-m-1+k,k,0).$$

Next, apply the induction hypothesis inside the sum to obtain

$$\begin{aligned}\sum_{k=2}^{m-2} F(n-m-1+k,k,0) &= \sum_{k=2}^{m-2}\sum_{j\geq 0}\binom{k-2-j}{j}\binom{n-m}{j+1}\\ &= \sum_{j\geq 0}\left(\binom{n-m}{j+1}\sum_{k=2}^{m-2}\binom{k-2-j}{j}\right)\\ &= -1 + \sum_{j\geq 0}\binom{n-m}{j}\binom{m-2-j}{j}.\end{aligned}$$

Apply the induction hypothesis on $F(n-1,m,0)$ and combine the sums to obtain

$$\begin{aligned}F(n,m,0) =& \sum_{j\geq 0}\binom{n-m}{j+1}\binom{m-2-j}{j} + \sum_{j\geq 0}\binom{n-m}{j}\binom{m-2-j}{j}\\ =& \sum_{j\geq 0}\binom{m-2-j}{j}\binom{n-m+1}{j+1}.\end{aligned}$$

□

Now, we have all the machinery needed for the proof of Theorem 2.1.

An alternative proof of Theorem 2.1. The proof is by induction on n, m and k. Notice that we have established the "anchors" for this formula, i.e., $F(n,2+k,0), F(n,m+2,m)$ and $F(m+2+k,m+2+k,m)$. Thus, we assume the formula holds for up to $F(n-1,m+2+k,m)$. We verify the formula for $F(n,m+2+k,m)$. From Part (i) of Theorem 2.2, we obtain

$$F(N,m+2+k,m) = \sum_{n=m+3+k}^{N} F(n-2,m+1+k,m-1) + \sum_{n=m+3+k}^{N}\sum_{q=0}^{k} F(n-k-3+q,m+q,m).$$

Observe that n can replace N, and we can apply the induction hypothesis on all terms. To handle the double summation first, we write

$$\begin{aligned}\sum_{q=0}^{k} F(n-k-3+q, m+q, m) =& F(n-k-3, m, m) + F(n-k-2, m+1, m) + \\ & \sum_{q=0}^{k-2} F(n-k-1+q, m+2+q, m) \\ =& \binom{n-k-2-m}{m} + 0 + \\ & \sum_{q=0}^{k-2}\sum_{j\geq 0} \binom{q-j}{j}\binom{m+j+1}{j+1}\binom{n-k-m-2}{m+1+j} \\ =& \sum_{j\geq 0}\binom{k-j}{j}\binom{m+j}{j}\binom{n-k-m-2}{m+j}.\end{aligned}$$

Thus, $F(N, m+2+k, m) =$

$$\begin{aligned}&\sum_{n=m+3+k}^{N} F(n-2, m+1+k, m-1) + \sum_{n=m+3+k}^{N}\sum_{j\geq 0}\binom{k-j}{j}\binom{m+j}{j}\binom{n-k-m-2}{m+j} \\ =& \sum_{n=m+3+k}^{N} F(n-2, m+1+k, m-1) + \sum_{j\geq 0}\binom{k-j}{j}\binom{m+j}{j}\binom{N-k-m-1}{m+j+1}.\end{aligned}$$

Now, we handle the first sum to obtain $F(N, m+2+k, m) =$

$$\begin{aligned}&\sum_{n=m+3+k}^{N}\sum_{j\geq 0}\binom{k-j}{j}\binom{m+j}{j+1}\binom{n-2-m-k}{m+j} + \sum_{j\geq 0}\binom{k-j}{j}\binom{m+j}{j}\binom{N-k-1-m}{m+j+1} \\ =& \sum_{j\geq 0}\binom{k-j}{j}\binom{m+j}{j+1}\binom{N-1-m-k}{m+j+1} + \sum_{j\geq 0}\binom{k-j}{j}\binom{m+j}{j}\binom{N-k-1-m}{m+j+1} \\ =& \sum_{j\geq 0}\binom{k-j}{j}\binom{m+j+1}{j+1}\binom{N-m-1-k}{m+j+1}.\end{aligned}$$

Next, use the fact that

$$\binom{m+i}{i}\binom{M}{m+i} = \binom{M-m}{i}\binom{M}{m}$$

to obtain

$$F(N, m+2+k, m) = \sum_{j\geq 0}\binom{k-j}{j}\binom{N-2m-k-1}{j+1}\binom{N-m-1-k}{m}.$$

Finally, replace N by n, $m+2+k$ by m^* and m by k^*, then delete the asterisk to obtain the form given in Theorem 2.1. □

2.3 A connection with the hypergeometric function ${}_3F_2(4)$

The hypergeometric function ${}_3F_2(z)$ is defined as

$$ {}_3F_2\left(\begin{array}{c} a_1, a_2, a_3 \\ b_1, b_2;\ z \end{array}\right) = \sum_{j=0}^{\infty} \frac{(a_1)_j (a_2)_j (a_3)_j}{(b_1)_j (b_2)_j} \frac{z^j}{j!}, \tag{2.3} $$

where z is a complex number and

$$ (a)_0 = 1,\ (a)_j = a(a+1)(a+2)\cdots(a+j-1) $$

for positive integers j.

We can use the hypergeometric function ${}_3F_2(4)$ to give a convenient representation of $F(n,m,k)$. Furthermore, the formulas found in this section can be used to give closed forms of ${}_3F_2$ in certain cases.

Proposition 2.4. *For any integers $n \geq m \geq k \geq 0$ such that $m \geq k+2$, we have*

$$ F(n,m,k) = (n-m+1)\binom{n-m}{k} {}_3F_2\left(\begin{array}{c} m-n+k, \frac{1}{2}(k-m+2), \frac{1}{2}(k-m+3) \\ 2, k-m+2;\ 4 \end{array}\right). $$

Proof. Use the result of Theorem 2.1 together with ${}_3F_2(4)$ as defined in (2.3), then simplify to obtain the required form. Observe that the condition $m \geq k+2$ makes $k-m+2 \leq 0$; however since one of the factors $(\frac{1}{2}(k-m+2))_j$ or $(\frac{1}{2}(k-m+3))_j$ vanishes before $(k-m+2)_j$ reaches zero, the expression is well defined. □

Using the fact that $\sum_{k=0}^{m} F(n,m,k) = \binom{n}{m}$, one can use the hypergeometric function to give the identity

$$ \sum_{k=0}^{m-2} (n-m+1)\binom{n-m}{k} {}_3F_2\left(\begin{array}{c} m-n+k, \frac{1}{2}(k-m+2), \frac{1}{2}(k-m+3) \\ 2, k-m+2;\ 4 \end{array}\right) = \binom{n}{m} - \binom{n-m+1}{m}. $$

Finally, it is worth mentioning that one can use some Computer Algebra Systems such as MAPLE and its hypergeometric package to compute and manipulate the obtained formulas.

3 The expectation and variance

In this section, we find a simple formula for the expectation of the infected and isolated plants, then we use the expectation to find the variance. Let X be a random variable representing the number of infected and isolated plants, the expectation $E(X)$ is defined by

$$ E(X) = \frac{1}{\binom{n}{m}} \sum_{k=0}^{m} kF(n,m,k), $$

where we assume that all configurations are equally likely.

Theorem 3.1. *The expected value (or mean) of infected and isolated plants among the m infected plants in a row of n cells is given by*

$$E(X) = \frac{1}{\binom{n}{m}} \sum_{k=1}^{m} kF(n,m,k) = \frac{m(n-m)(n+1-m)}{n(n-1)},$$

where we assume that all configurations are equally likely.

Proof. Let X_i be the number of ways in which the plant number i can be infected and isolated. Then $X = \sum_{i=1}^{n} X_i$, so

$$E(X) = \sum_{i=1}^{n} E(X_i).$$

The sample space for each X_i is $\{0,1\}$, with

$$P(X_i = 1) = \begin{cases} \binom{n-2}{m-1}/\binom{n}{m}, & i = 1, n, \\ \binom{n-3}{m-1}/\binom{n}{m}, & i = 2, \ldots, n-1. \end{cases}$$

Therefore,

$$E(X_i) = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) = P(X_i = 1) = \begin{cases} \binom{n-2}{m-1}/\binom{n}{m}, & i = 1, n, \\ \binom{n-3}{m-1}/\binom{n}{m}, & i = 2, \ldots, n-1. \end{cases}$$

Hence, noting that $(n-2)\binom{n-3}{m-1} = (n-m-1)\binom{n-2}{m-1}$, we have

$$E(X) = 2\frac{\binom{n-2}{m-1}}{\binom{n}{m}} + (n-2)\frac{\binom{n-3}{m-1}}{\binom{n}{m}} = \frac{(n-m+1)\binom{n-2}{m-1}}{\binom{n}{m}} = \frac{m(n-m)(n-m+1)}{n(n-1)}. \qquad \square$$

Remark 3.1. *It is worth mentioning here that an induction argument can be used to give an alternative approach for finding the expectation. Indeed, for $N \geq m \geq 2$, one can establish the anchors $E(N,2)$ and $E(m,m)$, then use induction on N to prove that*

$$\sum_{k=1}^{m} kF(N,m,k) = (N+1-m)\binom{N-2}{m-1} =: E(N,m),$$

which provides the formula of the expectation given in Theorem 3.1.

Next, we give a formula for the variance.

Theorem 3.2. *The variance of infected and isolated plants among the m infected plants in a row of $n \geq 4$ cells is given by*

$$\mathrm{Var}(X) = \frac{m(m-1)(n-m)(n-m+1)}{n^2(n-1)^2(n-2)(n-3)} p(n,m),$$

where

$$p(n,m) = 4n^3 - (7m+10)n^2 + (4m^2+11m+6)n - 6m^2.$$

Proof. Since $\mathrm{Var}(X) = E(X^2) - (E(X))^2$ and we know $E(X)$ from the preceding theorem, we only need to find

$$E(X^2) = \frac{1}{\binom{n}{m}} \sum_{k=0}^{m} k^2 F(n,m,k).$$

Let X_i be as in the proof of Theorem 3.1. Then $X = \sum_{i=1}^{n} X_i$, so

$$E(X^2) = E\left(\left(\sum_{i=1}^{n} X_i\right)^2\right) = \sum_{i=1}^{n} E(X_i^2) + 2\sum_{1\le i<j\le n} E(X_iX_j).$$

We will now calculate each of the two summands on the right. As in the previous theorem, the sample space of each X_i is $\{0,1\}$, so

$$E(X_i^2) = 1^2 \cdot P(X_i = 1) + 0^2 \cdot P(X_i = 0) = P(X_i = 1) = \begin{cases} \binom{n-2}{m-1}/\binom{n}{m}, & i = 1, n, \\ \binom{n-3}{m-1}/\binom{n}{m}, & i = 2, \dots, n-1. \end{cases}$$

Thus, as in the preceding proof, we have

$$\sum_{i=1}^{n} E(X_i^2) = 2\frac{\binom{n-2}{m-1}}{\binom{n}{m}} + (n-2)\frac{\binom{n-3}{m-1}}{\binom{n}{m}} = \frac{m(n-m)(n-m+1)}{n(n-1)}.$$

Likewise,

$$E(X_iX_j) = P(X_i = 1 \text{ and } X_j = 1),$$

so that if $j = i \pm 1$, we have $E(X_iX_j) = 0$ (note that there are $(n-1)$ such pairs (i,j) where $i < j$), and if $|j-i| \ge 2$, then we have

$$E(X_iX_j) = P(X_i = 1 \text{ and } X_j = 1) = \binom{n - n_{ij}}{m-2},$$

where $n_{ij} = |N_i \cup N_j|$ and $N_1 = \{1,2\}$, $N_n = \{n-1, n\}$, $N_i = \{i-1, i, i+1\}$ for $2 \le i \le n-1$.

Assume $n \ge 4$, then $n_{ij} \in \{4,5,6\}$. More precisely, if $1 \le i < j \le n$, then

- $n_{ij} = 4$ for $(i,j) \in \{(1,3), (1,n), (n-2,n)\}$ (3 cases);
- $n_{ij} = 5$ if $i = 1$ and $j \in \{4, \dots, n-1\}$, or $j = n$ and $i = \{2, \dots, n-3\}$, or $j = i+2$ and $i \in \{2, \dots, n-3\}$ (for the total of $3(n-4)$ cases);
- $n_{ij} = 6$ in the remaining $\binom{n}{2} - (n-1) - 3 - 3(n-4) = \binom{n-4}{2}$ cases.

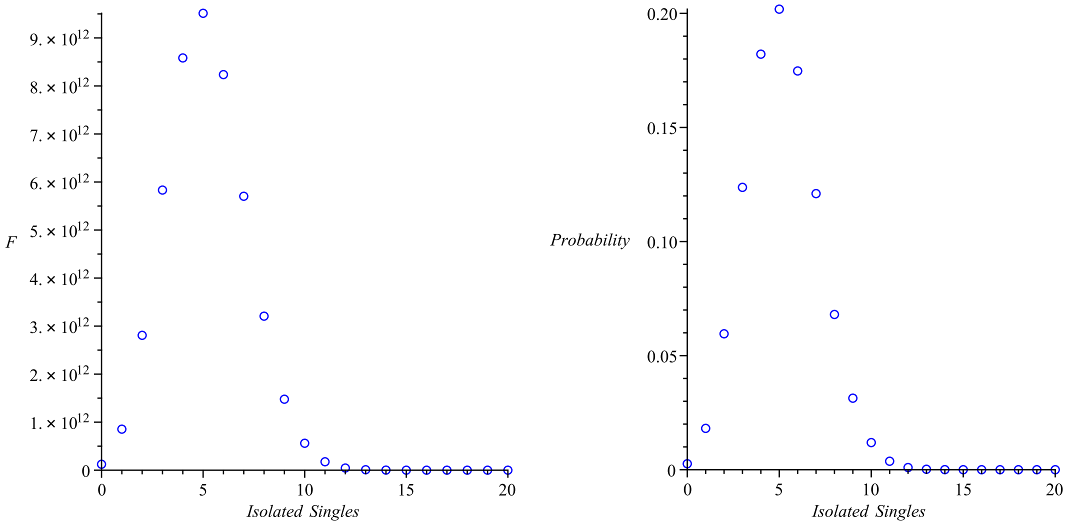
Therefore,

$$\begin{aligned} 2\sum_{1\le i<j\le n} E(X_iX_j) &= \frac{1}{\binom{n}{m}}\left(6\binom{n-4}{m-2} + 6(n-4)\binom{n-5}{m-2} + (n-4)(n-5)\binom{n-6}{m-2}\right) \\ &= \frac{\binom{n-4}{m-2}}{\binom{n}{m}}\left(6 + 6(n-m-2) + (n-m-2)(n-m-3)\right) \\ &= \frac{m(m-1)(n-m)(n-m-1)}{n(n-1)(n-2)(n-3)}(n-m)(n-m+1) \\ &= \frac{m(m-1)(n-m)^2(n-m-1)(n-m+1)}{n(n-1)(n-2)(n-3)}. \end{aligned}$$

Routine technical manipulations now yield the desired result. □

Notice that we restrict $n \geq 4$ in Theorem 3.2 to obtain a simplified formula. The cases $n = 1, 2$ and 3 are easy to handle individually. Now, we illustrate our results with the following example:

Example 3.1. *Suppose that we have a row of $n = 50$ plants in which $m = 30$ are infected. Observe that the isolated singles should be within the range 0 to 20. The values of $F(50, 30, k)$, $0 \leq k \leq 20$ are shown together with their probabilities in Figure 1.*



(a) (b)

Figure 1: The graph in (a) shows the sequence $F(50, 30, k)$, $0 \leq k \leq 20$. The graph in (b) shows the probability of having k isolated singles, $0 \leq k \leq 20$.

Notice that one can use Markov's inequality, $P(X \geq \lambda) \leq \frac{1}{\lambda}E(X)$, and the simple expectation formula above to find bounds on the probability of obtaining a certain number of isolated singles. For instance,

$$P(X > 15) = P(X \geq 16) \leq \frac{1}{16}\frac{30(50-30)(51-30)}{50(49)} = \frac{9}{28} \approx 0.32142857.$$

This can be improved using the one-tailed Chebyshev's inequality,

$$P(X - E(X) \geq \lambda) \leq \frac{\text{Var}(X)}{\text{Var}(X) + \lambda^2}.$$

We first find $E(X) = \frac{30(50-30)(51-30)}{50(49)} = \frac{36}{7}$, so in order to estimate the same probability as above, we will need $\lambda = 16 - (36/7) = 76/7$. This yields

$$P(X \geq 16) \leq \frac{\text{Var}(X)}{\text{Var}(X) + (76/7)^2} = \frac{8787}{280259} \approx 0.0313353.$$

This improves our bound by an order of magnitude. However, even the improved bound is still very rough as we can see using the exact formula given in Theorem 2.1. Indeed, we find that $P(X > 15) = 4.907252827 \cdot 10^{-7}$*, which shows the significance of dealing with exact values rather than approximations or samples.*

Acknowledgement: The first author thanks W. Abu-Dayyeh for some stimulating discussion about the problem. Also, the authors thank the anonymous referees.

References

[1] C. W. Barnes, L. L. Kinkel, J. V. Groth, *Spatial and temporal dynamics of Puccinia andropogonis on Comandra umbellata and Andropogon gerardii in a native prairie,* Canadian Journal of Botany **83** (2005), 1159–1173.

[2] R. A. Brualdi, *Introductory Combinatorics,* 5th ed., Pearson, 2010.

[3] C. L. Campbell, L. V. Madden, *Introduction to Plant Disease Epidemiology,* John Wiley & Sons, New York, 1990.

[4] S. Heubach, T. Mansour, *Combinatorics of Compositions and Words,* Chapman and Hall/CRC, 2009

[5] G. Hughes, L. V. Madden, *Using the beta-binomial distribution to describe aggregated patterns of disease incidence,* Phytopathology **83** (1993), 759–763.

[6] L. V. Madden, C. L. Campbell, *Nonlinear disease progress curves,* Pages 181-229 in: Epidemics of Plant Diseases. J. Kranz, ed. 2nd ed. Springer-Verlag, Berlin, 1990.

[7] L. Real, P. McElhany, *Spatial pattern and process in plant-pathogen interactions,* Ecology **77** (1996), 1011–1025.

[8] J. Segarra, M. J. Jeger, F. van den Bosch, *Epidemic dynamics and patterns of plant diseases,* Phytopathology **91** (2001), 1001–1010.

[9] C. L. Xiao, J. J. Hao, K. V. Subbarao, *Spatial patterns of microsclerotia of Verticillium dahliae in soil and Verticillium wilt of cauliflower,* Phytopathology **87** (1997), 325–331.