Yet Another Application of a Binomial Recurrence

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Vernon J. Rego
Purdue University, rego@cs.purdue.edu

Report Number:
88-765
YET ANOTHER APPLICATION OF A
BINOMIAL RECURRENCE
Order Statistics

Wojciech Szpankowski
Vernon Rego

CSD-TR-765
April 1988
Abstract

We investigate the moments of the maximum of a set of i.i.d geometric random variables. Computationally, the exact formula for the moments (which does not seem to be available in the literature) is inhibited by the presence of an alternating sum. A recursive expression for the moments is shown to be superior. However, the recursion can be both computationally intensive as well as subject to large round-off error when the set of random variables is large, due to the presence of factorial terms. To get around this difficulty we develop accurate asymptotic expressions for the moments and verify our results numerically.

Keywords: geometric distribution, order statistics, binomial recurrence, asymptotic approximation.

1. INTRODUCTION

In this paper, we present some results concerning the maximum order statistic \( M_n \) of \( n \) independent and identically distributed geometric random variables. We first develop a recurrence, and then an exact formula for the mean and variance of this order statistic. To the best of our knowledge, these results appear to be new (see, for example [2, 3]). The exact formula for the first two moments of this order statistic is computationally unsatisfactory, since it involves an alternative sum. The recurrence is computationally sound, but due to the presence of a factorial term, is not to be recommended for very large values of \( n \). In order to get around this barrier, we develop asymptotic forms for all the moments of this order statistic. In particular, we present extremely accurate asymptotic results for the mean and variance of \( M_n \).
The motivation for this study came from a model of a program unification technique in concurrency enhancement methods. The details of the model can be found in [6]. Briefly, consider a set of \( n \) balls placed in an urn at time \( k=0 \). At each discrete time step \( k, k \geq 1 \), we are required to pick up all the balls in the urn and toss them into the air. For each ball that is tossed, and for each value of \( k \), there is a nonzero probability \( p \) that the ball will fall out of the urn, for \( k \geq 1 \). Balls that fall out of the urn are ignored. We are interested in the number of tosses that is required to empty out the urn. The number of tosses required for any one ball (say the \( i^{th} \)) to fall out of the urn is given by a geometric random variable \( X_i \). Since the balls do not influence one another when tossed, the random variables \( X_i, 1 \leq i \leq n \), are i.i.d. It follows that the number of tosses required to empty the urn is precisely the maximum, \( M_n \), of this set of random variables. The original model in [6] investigates the situation where the number of urns is variable, and different rules may be used when choosing urns from which throws are to be made. In this paper we restrict our attention to \( M_n \) and accurate asymptotics for its moments.

In section 2.1 we present the problem formally and prove a proposition concerning the exact solution and asymptotic approximations for the first two moments of \( M_n \). In section 2.2 we develop a general solution to some binomial recurrences, and in section 2.3 we present asymptotic solutions for some alternating sums. Finally, in section 3 we briefly present some computational results demonstrating the accuracy of the asymptotics, and conclude the paper.

2. MAIN RESULTS

In this section, we present our main results. We start with a short, but formal, description of the problem, and we formulate our main results in a form of Proposition. In the next two subsections we provide the proof of the Proposition.
2.1 Problem statement

Let $X_i, i = 1, 2, \ldots, n$ be a set of i.i.d random variables distributed according to the geometric distribution with parameter $p$. That is, for every $i = 1, 2, \ldots, n$,

$$Pr(X_i = k) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots \quad (2.1a)$$

$$EX_i = p^{-1}, \quad EX_i^2 = (2-p)p^{-2} \quad (2.1b)$$

We shall investigate, in particular, the first two moments of the maximum of $X_1, X_2, \ldots, X_n$. Let

$$M_n = \max(X_1, X_2, \ldots, X_n) \quad (2.2a)$$

and define

$$M_n = EM_n, \quad M_n^2 = EM_n^2 \quad (2.2b)$$

Then, $M_n$ and $M_n^2$ satisfy the following recurrences

$$M_0 = 0 \quad M_1 = p^{-1} \quad (2.3a)$$

$$M_n = 1 + \sum_{k=0}^{n} \binom{n}{k} (1-p)^{k}p^{n-k}M_k \quad n \geq 2 \quad (2.3b)$$

and

$$M_0^2 = 0 \quad M_1^2 = (2-p)p^{-2} \quad (2.4a)$$

$$M_n^2 = 1 + 2M_n + \sum_{k=0}^{n} \binom{n}{k} (1-p)^{k}p^{n-k}M_k^2 \quad (2.4b)$$

To derive (2.3) and (2.4), we adopt a slight variation of the urn scheme discussed in the introduction. The intention is obtain an intuitive view of $M_n$. Let us assume that $n$ balls are put into $n$ distinct urns, with one ball in each urn. The action of all urns is synchronized in the sense that at the beginning of a slot time (e.g., every second) each nonempty urn attempts to get rid of the ball by tossing it into the air. For each ball tossed, the ball falls out of the urn with probability $p$, and falls back into the urn with probability $q = 1-p$. Each urn acts independently of the other urns. Note that $M_n$ defined in (2.2), is equal to the number of slots needed to
empty all urns. Therefore, after the first slot we have \( \binom{n}{k}(1-p)^{p^n-k} \) nonempty urns, and the time to empty them is equal to \( 1 + M_k \). This suggests the following recurrence for the \( l \)-th moment \( EM^l_n \) of \( M_n \)

\[
EM^l_n = \sum_{k=0}^{n} \binom{n}{k} (1-p)^k p^{n-k} E(1 + M_k)^l
\]

Solving (2.5) for \( l = 1 \) and 2, one finally obtains our recurrences (2.3) and (2.4).

Using recurrences (2.3) and (2.4), we prove our main results, which can be summarized as follows.

**PROPOSITION (i).** The exact solution for the first moment \( M_n \) of \( M_n \) is

\[
M_n = - \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{1}{1-q^k} \quad (2.6)
\]

where \( q = 1-p \), and for the second moment we obtain

\[
M^2_n = \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{2k-1}{1-q^k} - 2 \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{1}{(1-q^k)^2} \quad (2.7)
\]

(ii) For large \( n \) the average \( M_n \) of \( M_n \) is asymptotically equal to

\[
M_n = \frac{\log n}{\log Q} + \frac{\gamma}{\log Q} + \frac{1}{2} + P(n) + O(n^{-1}) \quad (2.8)
\]

where \( \log \) denotes the natural logarithm, \( Q \stackrel{\text{def}}{=} q^{-1} \), \( \gamma = 0.577 \) is the Euler constant, and \( P(n) \) is a periodic function with very small amplitude (e.g., Knuth [5] computed that \( P(0.5) < 1.725 \cdot 10^{-7} \) for \( q = 0.5 \)), which can be expressed as

\[
P(n) = \frac{1}{\log Q} \sum_{k=-\infty}^{\infty} \Gamma(2\pi n_\text{ik}/\log q) \exp[-2\pi n_\text{ik}\log q] n \quad (2.9)
\]

where \( \Gamma(x) \) is the gamma function [1].

The variance \( \text{var} \{M_n\} \) of \( M_n \) for large \( n \) can be asymptotically expressed as
\[
\text{Var } M_n = 2 \frac{\log n}{\log Q} + 2\beta + \frac{1}{4} - \frac{\gamma^2}{\log^2 Q} - \frac{2}{\log Q} + P_1(n) + \mathcal{O}(n^{-1}) \quad (2.10)
\]

where
\[
\beta = \frac{1}{h_1^2} \left\{ \frac{\pi^2}{12} + \frac{\gamma^2}{2} + \frac{3}{4} \frac{h_2^2}{h_1^2} - \frac{1}{3} \frac{h_3}{h_1} + \frac{\gamma h_2}{h_1} \right\} \quad (2.11)
\]

with \( h_j \equiv (-1)^j \log^j q, j = 1, 2, 3 \).

(iii) The \( k \)-th moment \( M_n^k \) of \( M_n \) is asymptotically equal to
\[
M_n^k = \log^k n + \mathcal{O}(\log^{k-1} n) \quad (2.12)
\]

The rest of the paper is devoted to proving the proposition and demonstrating the accuracy of the asymptotic results.

2.2 Exact solution

In this subsection, we prove part (i) of the Proposition. Note that both recurrences (2.3) and (2.4) fall into the following general recurrence: given \( x_0 \) and \( x_1 \), solve for \( n \geq 2 \)
\[
x_n = a_n + \sum_{k=0}^{n} \binom{n}{k} q^k p^{n-k} x_k \quad (2.13)
\]
where \( p + q = 1 \), and \( a_n \) is any sequence of numbers which we further call an additive term of the recurrence (2.13). In our case \( a_n = 1 \) for (2.3) and \( a_n = 1 + 2M_n \) for (2.4). This type of recurrence was extensively studied in [7, 8] (see also [5]). We quote some results from [7, 8] here.

Let us define a sequence \( \hat{a}_n \), called binomial inverse relations [5], as
\[
\hat{a}_n = \sum_{k=0}^{n} \binom{n}{k} a_k \quad a_n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \hat{a}_n \quad (2.14)
\]

Then,

Lemma 1. (i) The recurrence (2.13) possesses the following solution
\[ x_n = x_0 + \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{\hat{a}_k + kA - a_0}{1 - q^k} \]  \hspace{1cm} (2.15)

where \( A = a_1 - x_1 p - x_0 q \).

(ii) The inverse \( \hat{a}_n \) to \( x_n \) satisfies

\[ x_n = \frac{\hat{a}_n + nA - a_0}{1 - q^n} \quad n \geq 1 \]  \hspace{1cm} (2.16)

**Proof.** The details are given in [7]. Here we present only sketch of the derivation. To prove (2.15), we multiply (2.13) by \( \frac{z^n}{n!} \) and computing the exponential generation function

\[ X(z) = \sum_{n=0}^{\infty} x_n \frac{z^n}{n!} \]  

we obtain the following functional equation

\[ X(z) - X(qz)e^{(1-q)z} = A(z) - a_0 - Az \]  \hspace{1cm} (2.17)

where \( A(z) \) is the exponential generation function for \( a_n \). Introducing \( H(z) = X(z)e^{-z} \) we transform (2.17) into

\[ H(z) - H(qz) = A(z)e^{-z} - a_0 e^{-z} - A ze^{-z} \]

This functional equation is easy to solve by consecutive iterations. Noting, in addition, that \( \hat{a}_n \) has generating function \( \hat{A}(-z) = A(z)e^{-z} \) we prove, after some algebra, Eq. (2.15). The proof of (2.16) is immediate by comparing (2.15) with the definition (2.14).

\[ \square \]

Using our Lemma 1, the solution of (2.3) is simple, since one computes \( \hat{a}_n = \delta_{n,0} \), where \( \delta_{n,0} \) is the Kronecker delta [5], and \( A = 1 - M_1 p = 0 \). Hence (2.6) follows. To solve (2.4) we need the inverse \( \hat{M}_n \) of \( M_n \). But, from (2.6) and (2.16) we find \( \hat{M}_n = -(1 - q^n)^{-1} \) for \( n \geq 1 \), and then (2.7) follows from Lemma 1 and some simple algebra.
2.3 Asymptotic approximation

The exact solutions given in Proposition (i) are not attractive from a numerical viewpoint. In fact, to compute $M_n$ and $M_n^2$, one would do well to use the recurrences (2.3) and (2.4) instead of the exact solutions (2.6) and (2.7).

Nevertheless, the exact formulas allow us to obtain very sharp asymptotic approximations.

Note that formulas (2.6) and (2.7) fall into the general pattern

$$S_n = \sum_{k=1}^{n} (-1)^k \binom{n}{k} f_k$$

(2.18)

where $f_k$ is any sequence. It turns out to be useful to plug into complex analysis to obtain asymptotics of (2.18) [5]. Let us assume that $f_k$ has an analytical continuation $f(z)$, that is, $f(k) = f_k$. In [9], we have proved

Lemma 2. Under certain mild conditions on the growth of $f(z)$ at infinity (cf. [9]) one obtains

$$S_n = \frac{1}{2\pi i} \int_{-1/2 - i\infty}^{-1/2 + i\infty} \Gamma(z)f(-z)n^{-z}dz + e_n$$

(2.19)

where $\Gamma(z)$ is the gamma function [1, 4], and $e_n$ is the error function given by

$$e_n = O(n^{-1}) \int_{-1/2 - i\infty}^{-1/2 + i\infty} z\Gamma(z)f(-z)n^{-z}dz$$

(2.20)

Proof. Formula (2.19) follows from Cauchy's theorem [4], and some algebraic manipulation. For details the reader is referred to [9].

To apply Lemma 2, by Cauchy's residue theorem, we find residues of the function under the integral right to the line $(-1/2 - i\infty, 1/2 + i\infty)$. Then

$$\sum_{k=1}^{n} (-1)^k \binom{n}{k} f_k = -\sum_{k=-\infty}^{\infty} \text{res}(\Gamma(z_k)f(-z_k)n^{-z_k}) + e_n + O(n^{-M})$$

(2.21)
for any \( M > O \), and the sum is taken over all poles \( z_k, k = 0, \pm 1, \pm 2, \ldots \), of the function under the integral (2.19).

To illustrate Lemma 2, we apply it to asymptotic analysis of \( M_n \) and \( M_n^2 \). From (2.6) and (2.19) one finds

\[
M_n = -\frac{1}{2\pi i} \int_{-\frac{1+i}{2}}^{\frac{1+i}{2}} \frac{\Gamma(z)n^{-z}}{1 - q^{-z}} \, dz + e_n = \sum_{k = -\infty}^{\infty} \text{res} \left\{ \frac{\Gamma(z_k)n^{-z_k}}{1 - q^{-z_k}} \right\} + e_n
\]

But, the poles \( z_k \) are given by (i.e., roots of the denominator)

\[
z_k = \frac{2\pi ik}{\log q} \quad k = 0, \pm 1, \ldots ,
\]

It is well known that the leading factor in the asymptotics comes from \( k = 0 \). In fact, \( z_0 = 0 \) is a double pole, since zero is also a pole of the gamma function \([1, 4]\). To compute the residue at \( z_0 = 0 \), we use the following Taylor expansions

\[
\Gamma(z) = z^{-1} - \gamma + O(z)
\]

\[
n^{-z} = 1 - z \log n + O(z^2)
\]

\[
\frac{1}{1 - q^{-z}} = -\frac{z^{-1}}{\log q^{-1}} + \frac{1}{2} + O(z)
\]

Multiplying (2.23a) through (2.23c) and finding the coefficient at \( z^{-1} \) leads to the residue at \( z_0 = 0 \). That is,

\[
\text{res}_{z_0} \left\{ \frac{\Gamma(z_0)n^{-z_0}}{1 - q^{-z_0}} \right\} = \frac{\log n}{\log Q} + \frac{\gamma}{\log Q} + \frac{1}{2}
\]

where \( Q = q^{-1} \). The other poles contribute to the periodic function \( P(n) \) defined in (2.9). This function has a rather small amplitude (see \([5]\)) and can safely be ignored in practice. The error function \( e_n \) contributes \( O(n^{-1}) \). Indeed, it is enough to apply the residue theorem, and to see that the integral in (2.20) is \( O(1) \).
The evaluation of $M_n^2$ is a little more intricate. Let us split $M_n^2$ into three parts, that is,

\[ M_n^2 = S_n^1 + S_n^2 + S_n^3 \]

where

\[ S_n^1 = - \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{1}{1 - q^k} \]

\[ S_n^2 = \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{2k}{1 - q^k} \]

\[ S_n^3 = \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{1}{(1 - q^k)^2} \]

The asymptotic for $S_n^1$ and $S_n^2$ are immediately available from Lemma 2. In fact, $S_n^1 = M_n$ and, after some algebra,

\[ S_n^2 = - \frac{2}{\log Q} + F_1(n) + O(n^{-1}) \]

where

\[ F_1(n) = \frac{1}{\log Q} \sum_{k=-\infty}^{\infty} \Gamma(1 + 2\pi i k \log q) \exp[-2\pi i k \log n] \]

Now, for $S_n^3$ we immediately obtain from Lemma 2

\[ S_n^3 = \frac{1}{2\pi i} \int_{-\infty - i\infty}^{\infty - i\infty} \frac{\Gamma(z) n^{-z}}{(1 - q^{-z})^2} \ dz \]

But, this time the pole $z_0 = 0$ is a triple pole of the function under the integral. This case was extensively studied in [8], and we show there that the following expansions must be used

\[ \Gamma(z) = z^{-1} - \gamma + \left( \frac{\pi^2}{12} + \frac{\gamma^2}{2} \right) z + O(z^2) \]

\[ n^{-z} = 1 - z \log n + \frac{z^2}{2} \log^2 n + O(z^3) \]

\[ (1 - q^{-z})^2 = z^2 (b_0 + b_1 z + b_2 z^2 + O(z^3)) \]

where $b_0 = h_1^2$, $b_1 = h_1 h_2$ and $b_2 = \frac{1}{4} h_2^2 + \frac{1}{3} h_1 h_3$ and $h_n = (-1)^n \log^n q$. In [8] an algorithm is given to compute the residue in such a case (note that the last equation in (2.27) shows the Taylor expansion of $(1 - q^{-z})^2$ and not $(1 - q^{-z})^{-2}$). After some algebra, and using (2.25),
we prove

\[ M_n^2 = \frac{\log^2 n}{\log^2 Q} + 2 \frac{\log n}{\log Q} \left( \frac{\gamma}{\log Q} + \frac{3}{2} \right) + 2\beta + \frac{\gamma}{\log Q} + \frac{1}{2} - \frac{2}{\log Q} + F_2(n) \]

where \( \beta \) is given in (2.11), and \( F_2(n) \) is a periodic function. Using this and (2.8), we immediately prove (2.10) in Proposition (ii). Finally, formula (2.12) in Proposition (iii), is proved in a similar manner. In that case, we consider only the leading factor, which corresponds to \( S_n^3 \) in (2.26) with the denominator replaced by \((1 - q^{-n})^k\) for the \( k \)-th moment of \( M_n \).

3. NUMERICAL WORK

In the following table, we present the values of \( M_n \) and \( \text{Var}(M_n) \) for selected values of \( n \), \( 2 \leq n \leq 200 \). Observe that the asymptotics are accurate even for values of \( n \) below 10.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( M_n )</th>
<th>Asymptotic</th>
<th>( \text{Var}(M_n) )</th>
<th>Asymptotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5.71429</td>
<td>4.91586</td>
<td>19.75510</td>
<td>22.83865</td>
</tr>
<tr>
<td>10</td>
<td>10.68127</td>
<td>10.51036</td>
<td>33.21958</td>
<td>34.02765</td>
</tr>
<tr>
<td>20</td>
<td>13.00595</td>
<td>12.91978</td>
<td>38.42961</td>
<td>38.84650</td>
</tr>
<tr>
<td>30</td>
<td>14.38679</td>
<td>14.32920</td>
<td>41.38469</td>
<td>41.66534</td>
</tr>
<tr>
<td>40</td>
<td>15.37245</td>
<td>15.32920</td>
<td>43.45371</td>
<td>43.66534</td>
</tr>
<tr>
<td>50</td>
<td>16.13946</td>
<td>16.10486</td>
<td>45.04716</td>
<td>45.21666</td>
</tr>
<tr>
<td>60</td>
<td>16.76744</td>
<td>16.73862</td>
<td>46.34302</td>
<td>46.48418</td>
</tr>
<tr>
<td>70</td>
<td>17.25920</td>
<td>17.27446</td>
<td>47.43432</td>
<td>47.55858</td>
</tr>
<tr>
<td>80</td>
<td>17.76026</td>
<td>17.73862</td>
<td>48.37788</td>
<td>48.48418</td>
</tr>
<tr>
<td>90</td>
<td>18.16725</td>
<td>18.14804</td>
<td>49.20893</td>
<td>49.30302</td>
</tr>
<tr>
<td>100</td>
<td>18.53154</td>
<td>18.51428</td>
<td>49.95121</td>
<td>50.03550</td>
</tr>
<tr>
<td>110</td>
<td>18.86126</td>
<td>18.84559</td>
<td>50.62172</td>
<td>50.69811</td>
</tr>
<tr>
<td>120</td>
<td>19.16237</td>
<td>19.14804</td>
<td>51.23636</td>
<td>51.30302</td>
</tr>
<tr>
<td>130</td>
<td>19.43962</td>
<td>19.42628</td>
<td>51.79352</td>
<td>51.85949</td>
</tr>
<tr>
<td>140</td>
<td>19.69623</td>
<td>19.68388</td>
<td>52.31404</td>
<td>52.37469</td>
</tr>
<tr>
<td>150</td>
<td>19.93521</td>
<td>19.92370</td>
<td>52.79796</td>
<td>52.85434</td>
</tr>
<tr>
<td>160</td>
<td>20.15882</td>
<td>20.14804</td>
<td>53.25046</td>
<td>53.30302</td>
</tr>
<tr>
<td>170</td>
<td>20.36887</td>
<td>20.35878</td>
<td>53.67583</td>
<td>53.72449</td>
</tr>
<tr>
<td>180</td>
<td>20.56697</td>
<td>20.55746</td>
<td>54.07622</td>
<td>54.12186</td>
</tr>
<tr>
<td>190</td>
<td>20.75439</td>
<td>20.74540</td>
<td>54.45491</td>
<td>54.49774</td>
</tr>
<tr>
<td>200</td>
<td>20.93217</td>
<td>20.92370</td>
<td>54.81474</td>
<td>54.85434</td>
</tr>
</tbody>
</table>

The absolute error between the asymptotic value of \( M_n \) and the value given by the recurrence is
0.17 for $n = 10$. When $n = 100$, the error drops to 0.017, which is a tenth of the original error. For $n = 200$, the error is 0.008. The situation is similar in the case of the variance. When $n = 10$, the error is 0.808. However, when $n = 100$, the error falls to 0.0842, which again is about a tenth of the original error. As in the previous case, the error falls again to 0.03960, about one-half its value, when $n$ is increased to 200. Thus, we can expect the error to decrease roughly geometrically, halving for every increase of $n$ by a hundred. In our Proposition we have proved that the error is not larger than $O(n^{-1})$. Clearly, when $n$ takes on such large values, the recurrence becomes impractical, both due to the amount of time required to compute it, as well as numerical overflow caused by the computation of factorial terms.

References


