THE NUMBER OF WINNERS IN A DISCRETE GEOMETRICALLY DISTRIBUTED SAMPLE

PETER KIRSCHENHOFER AND HELMUT PRODINGER

Department of Algebra and Discrete Mathematics
Technical University of Vienna, Austria

Let the random variable G be geometrically distributed, i.e. $\mathbb{P}\{G=k\}=pq^{k-1}$, with $q=1-p$. Also, assume that n independent copies are given. Finally, let X count the number of winners, i.e. the number of random variables with highest value. In [1] the probability distribution of X, the expectation $\mathbf{E}X$ of X and the asymptotic behaviour of $\mathbb{P}\{X=1\}$ (probability of a single winner) for $n\to\infty$ was to be determined. In the solution [2] it was remarked that – surprisingly – this probability does not converge as $n\to\infty$ but rather has an oscillating behaviour. At the same time, Eisenberg, Stengle and Strang [5] discussed this problem and related topics, exhibiting the structure of the periodic fluctuation, for which an explicit Fourier expansion was given. Also about this time, Brands, Steutel and Wilms [3] came independently to roughly the same results. A recent paper due to Baryshnikov, Eisenberg and Stengle [4] deals with the existence of the limiting probability of a tie for first place.

In fact, a fluctuating behaviour in asymptotic expansions is not at all uncommon. There are numerous results of that type, e.g. in the analysis of divide–and–conquer recursions [7, 8, 16] or digital sums [6], to mention just some examples, that play a prominent role in the probabilistic analysis of algorithms.

Our aim in this paper is to some extent tutorial: The asymptotic technique that yields the Fourier expansions of the fluctuating functions very comfortably is called

1991 *Mathematics Subject Classification.* 60 C 05.
Key words and phrases. Geometric distribution, generating functions, Rice's method.

"Rice's method", see the recent survey [9]. In order to convince the reader of the advantages of this method we will rederive some results on the initially mentioned problem in the sequel, and afterwards present some new results concerning higher moments of distribution as well as the number of persons reaching a specified level beyond the winner.

Let us abbreviate $Q := 1/q$ and $L := \log Q$. Let further p_m denote the probability $p_m = \mathbb{P}\{X = m\}$, i.e. the probability to have m winners (amongst n players). Then

$$p_m = p^m \binom{n}{m} \sum_{j \geq 1} q^{(j-1)m} \left(1 - q^{j-1}\right)^{n-m}.$$

This follows from the observation that m out of n people have a (winning) value j, while the other ones have a smaller value. Now we set $N := n - m$, expand the binomial and sum over j to get the alternative

$$p_m = p^m \binom{n}{m} \sum_{k=0}^{N} \binom{N}{k} (-1)^k \frac{1}{1 - Q^{-k-m}}. \tag{1}$$

The key point in analyzing this alternating sum asymptotically is the following Lemma.

Lemma. [9] *Let $f(z)$ be a function that is analytic on $[n_0, +\infty[$. Assume that $f(z)$ is meromorphic in the whole of $\mathbb{C}$ and analytic on $\Omega = \bigcup_{j=1}^{\infty} \gamma_j$ where the γ_j are concentric circles whose radius tends to infinity. Let $f(z)$ be of polynomial growth on Ω. Then, for n large enough,*

$$\sum_{k=n_0}^{n} \binom{n}{k} (-1)^k f(k) = \sum_z \operatorname{Res}[n; z] f(z) \ , \tag{2}$$

where

$$[n; z] = \frac{(-1)^{n-1} n!}{z(z-1)\dots(z-n)} = \frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)} \quad \square$$

and the sum is extended to all poles not on $[n_0, +\infty[$.

Extending the contour of integration to the left it turns out that, under suitable growth conditions on $f(z)$ (which are fulfilled in all our examples), an asymptotic equivalent of (2) is given by the residues of $[n; z] f(z)$ at the additional poles.

In our example we only have to consider the residues of

$$[N;z]\frac{1}{1-Q^{-z-m}}$$

left of the line $\Re z = 0$. They are at $z = -m+\chi_j$, with $\chi_j = \frac{2j\pi i}{L}$. The computation of the residues is simple;

$$\mathrm{Res}_{z=-m+\chi_j} = n^{\chi_j}\frac{\Gamma(N+1)\Gamma(N+1+m-\chi_j)}{m!}\frac{1}{L},$$

whence we have (after multiplication by the factor $p^m\binom{n}{m}$ and going back to n instead of N) the asymptotic formulæ (for m fixed and $n \to \infty$)

$$p_m = \frac{1}{L}\frac{p^m}{m!}\sum_{j\in\mathbb{Z}}\frac{\Gamma(n+1)}{\Gamma(n+1-\chi_j)}\Gamma\left(m-\chi_j\right) + \text{smaller order terms.} \tag{3}$$

It is often convenient to pull out the term with index $j = 0$ (the "mean") and give a name to the remaining periodic fluctuation (of mean zero),

$$\delta_m(x) = \frac{1}{m!}\sum_{j\neq 0}\Gamma(m-\chi_j)e^{2j\pi i x}.$$

This is a continuous, periodic function of period 1, (integral) mean zero and small amplitude (for m not too large, compare the discussion in [12]), and we can thus write

$$p_m \sim \frac{1}{L}\frac{p^m}{m} + \frac{p^m}{L}\delta_m(\log_Q n) + \mathcal{O}\Big(\frac{1}{n}\Big), \quad m \text{ fixed }, n \to \infty. \tag{4}$$

It is interesting to note that the alternating sum (1) can also be rewritten using the partial fraction decomposition of the meromorphic function $\frac{1}{1-q^{z+m}}$, namely (compare [11, 7.10])

$$\frac{1}{1-q^{z+m}} = \frac{1}{2} + \frac{1}{L}\sum_{j\in\mathbb{Z}}\frac{1}{z+m+\chi_j}, \quad \text{with } \chi_j = \frac{2j\pi i}{L}. \tag{5}$$

Inserting (4) into (1) we find

$$p_m = p^m\binom{n}{m}\sum_{k=0}^{N}\binom{N}{k}(-1)^k\frac{1}{2} + \frac{1}{L}p^m\binom{n}{m}\sum_{j\in\mathbb{Z}}\sum_{k=0}^{N}\binom{N}{k}(-1)^k\frac{1}{k+m+\chi_j}.$$

Now for $x \in \mathbb{C} \setminus \{-N, \dots, 0\}$

$$\sum_{k=0}^{N} \binom{N}{k} (-1)^k \frac{1}{k+x} = \frac{\Gamma(N+1)\Gamma(x)}{\Gamma(N+1+x)} = \frac{N!}{(N+x)_{N+1}}$$

(compare [10, (5.41)]), so that (remembering $N = n - m$)

$$\begin{aligned} p_m &= \frac{p^m}{2}\delta_{n,m} + \frac{1}{L}\frac{p^m}{m!}\sum_{j\in\mathbb{Z}} \frac{n!}{(m+\chi_j)\dots(n+\chi_j)} \\ &= \frac{p^m}{2}\delta_{n,m} + \frac{1}{L}\frac{p^m}{m!}\sum_{j\in\mathbb{Z}} \Gamma(m-\chi_j)\frac{\Gamma(n+1)}{\Gamma(n+1-\chi_j)}. \end{aligned} \tag{6}$$

From (6) we may conclude that the smaller order terms from expansion (3) equal zero in fact.

Now let us consider the expectation,

$$\begin{aligned} \mathrm{E}_n &= \sum_{m\ge 1} m p_m \\ &= n \sum_{m\ge 1} \binom{n-1}{m-1} p^m \sum_{j\ge 0} q^{jm} \left(1-q^j\right)^{n-m} \\ &= \frac{np}{q} \sum_{j\ge 0} q^{j+1} \left(1-q^j\right)^{n-1} \sum_{m\ge 0} \binom{n-1}{m} \left(\frac{pq^j}{1-q^j}\right)^m \\ &= \frac{np}{q} \sum_{j\ge 1} q^j (1-q^j)^{n-1} \\ &= (Q-1)n \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k \frac{1}{Q^{k+1}-1} \\ &= Q(p_1 - p\delta_{n,1}). \end{aligned}$$

From (3) we conclude the asymptotic formula

$$E_n = \frac{p}{q}\frac{1}{L}\Big(1 + \delta_1(\log_Q n)\Big) + \mathcal{O}\big(\frac{1}{n}\big). \tag{7}$$

This computation was already reported in [2]. Let us now engage in the computation of the *variance*, which seems to be new.

For this we compute the second factorial moment M_n for $n \geq 2$:

$$\begin{aligned}
M_n &= \sum_{m\geq 2} m(m-1)p_m \\
&= n(n-1)\sum_{m\geq 2}\binom{n-2}{m-2}p^m \sum_{j\geq 0} q^{jm}\left(1-q^j\right)^{n-m} \\
&= \frac{n(n-1)p^2}{q^2}\sum_{j\geq 1} q^{2j}\left(1-q^j\right)^{n-2} \\
&= \frac{n(n-1)p^2}{q^2}\sum_{k=0}^{n-2}\binom{n-2}{k}(-1)^k\frac{1}{Q^{k+2}-1} \\
&= 2Q^2p_2 - 2\frac{p^2}{q^2}\delta_{n,2} = 2Q^2(p_2 - p^2\delta_{n,2}).
\end{aligned} \tag{8}$$

The variance is obtained in the usual way by computing $V_n = M_n + E_n - E_n^2$. Hence

$$V_n = \frac{p^2}{q^2}\frac{1}{L} + \frac{p}{q}\frac{1}{L} - \frac{p^2}{q^2}\frac{1}{L^2} + \tau(\log_Q n) + \mathcal{O}\Big(\frac{1}{n}\Big). \tag{9}$$

This time $\tau(x) = \frac{p^2}{q^2}\frac{2}{L}\delta_2(x) + \frac{p}{q}\frac{1}{L}\delta_1(x) - \frac{p^2}{q^2}\frac{1}{L^2}\delta_1^2(x)$ has (integral) mean *different* from zero! While this quantity is quite small, it can be extracted using the methods described in [13], and we find the alternative

$$\mathrm{V}_n = \frac{p}{q^2}\frac{1}{L} - \frac{p^2}{q^2}\frac{1}{L}\bigg(2\sum_{j\geq 1}\frac{Q^j}{(Q^j+1)^2} + \frac{1}{4}\bigg) + \tau_1(\log_Q n) + \mathcal{O}\Big(\frac{1}{n}\Big), \tag{10}$$

where the function $\tau_1(x)$ has now mean zero.

There is a nice way to derive the explicit forms of the expectation and the second factorial moment, using generating functions. Let the coefficient of z^k in $F_n(z)$ denote the probability that n player produce k winners. We get the following recursion

$$F_n(z) = \sum_{k=1}^{n}\binom{n}{k}p^{n-k}q^kF_k(z) + p^nz^n, \qquad n \geq 1. \tag{11}$$

It is convenient to set $F_0(z) = 1$. This recursion is almost selfexplanatory. When at a certain level the remaining players all fail, we label each of them by a "z" and leave the recursion. The expectation E_n is obtained via $F_n'(1)$; therefore

$$E_n = \sum_{k=1}^{n}\binom{n}{k}p^{n-k}q^kE_k + np^n, \qquad n \geq 1.$$

Define the exponential generating function

$$E(z) = \sum_{n \geq 0} E_n \frac{z^n}{n!},$$

then we obtain

$$E(z) = e^{pz} E(qz) + pze^{pz},$$

which will be easier by using the "Poisson transform"

$$\hat{E}(z) = e^{-z} E(z) = \sum_{n \geq 0} \hat{E}_n \frac{z^n}{n!};$$

$$\hat{E}(z) = \hat{E}(qz) + pze^{-qz}.$$

Equating coefficients we see, that, for $n \geq 1$,

$$\hat{E}_n = \frac{np}{q} (-1)^{n-1} \frac{q^n}{1 - q^n}$$

and furthermore for $n \geq 1$

$$\begin{aligned} E_n &= \sum_{k=0}^{n} \binom{n}{k} \hat{E}_k \\ &= \sum_{k=1}^{n} \binom{n}{k} \frac{kp}{q} (-1)^{k-1} \frac{1}{Q^k - 1} \\ &= \frac{np}{q} \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k \frac{1}{Q^{k+1} - 1}. \end{aligned}$$

For the second factorial moment we have to derive twice, and evaluate at $z = 1$. An almost identical computation gives the same expression that we obtained already.

This approach was used by Knuth in [15] under the name "binomial transform" and subsequently used by many people, see for instance [17].

Finally we want to produce some new results which shed some additional light on the original question about the number of winners.

Assume that the winners have reached the level j. We are interested in the number of people who reached the level $j - d$, where d is a parameter. ($d = 0$ is the case that was just considered.) Call the random variable in question X_d. Then we have

$$\begin{aligned}&\mathbb{P}\{X_d = m_d\}\\ &= \sum \sum_{j\ge d+1} \binom{n}{m_0,\ldots,m_d,\mu} \left(pq^{j-1}\right)^{m_0} \ldots \left(pq^{j-d-1}\right)^{m_d}\left(1-q^{j-d-1}\right)^{\mu},\end{aligned} \tag{12}$$

where the first sum runs over all $m_0 \ge 1$, $m_1 \ge 0$, $\ldots$, $m_{d-1} \ge 0$ and $\mu = n - m_0 - \cdots - m_d$. The expected value is just the sum of m_d times this quantity. We get

$$\begin{aligned}E_n^{(d)} &= n \sum_{j\ge d+1} pq^{j-d-1}\Big[\Big(pq^{j-1}+\cdots+pq^{j-d-1}+1-q^{j-d-1}\Big)^{n-1}\\ &\qquad - \Big(pq^{j-2}+\cdots+pq^{j-d-1}+1-q^{j-d-1}\Big)^{n-1}\Big]\\ &= pn \sum_{j\ge d+1} q^{j-d-1}\Big[\left(1-q^j\right)^{n-1} - \left(1-q^{j-1}\right)^{n-1}\Big]\\ &= \frac{p^2}{q^{d+1}} n \sum_{j\ge d} q^j \left(1-q^j\right)^{n-1} - \frac{p}{q} n \left(1-q^d\right)^{n-1}.\end{aligned}$$

This formula holds only for $d \ge 1$. Since

$$E_n^{(0)} = E_n = n\frac{p}{q}\sum_{j\ge 1} q^j\left(1-q^j\right)^{n-1},$$

we find

$$E_n^{(d)} = \frac{p}{q^d}E_n^{(0)} - \frac{p^2}{q^{d+1}} n \sum_{j=1}^{d-1} q^j\left(1-q^j\right)^{n-1} - \frac{p}{q} n\left(1-q^d\right)^{n-1}, \tag{13}$$

and, since the extra terms are exponentially small, we can use the asymptotic result for $E_n^{(0)}$ and have

$$\begin{aligned}E_n^{(d)} &= \frac{p}{q^{d+1}} p_1 \qquad (n \ge 2)\\ &= \frac{1}{L}\frac{p^2}{q^{d+1}} + \frac{1}{L}\frac{p^2}{q^{d+1}}\delta_1(\log_Q n) + \mathcal{O}\Big(\frac{1}{n}\Big) \qquad (n\to\infty).\end{aligned} \tag{14}$$

The second factorial moment $M_n^{(d)}$ is computed analogously as $(d \ge 1)$

$$\begin{aligned}M_n^{(d)} &= n(n-1)\sum_{j\ge d+1}\Big(pq^{j-d-1}\Big)^2\Big[\left(1-q^j\right)^{n-2} - \left(1-q^{j-1}\right)^{n-2}\Big]\\ &= \frac{p^3}{q^{2d+2}} n(n-1)\sum_{j\ge d} q^{2j}\left(1-q^j\right)^{n-2} - \frac{p^2}{q^2} n(n-1)\left(1-q^d\right)^{n-2}\\ &= \frac{p}{q^{2d}} + \text{exponentially small terms in } n, \qquad n\to\infty.\end{aligned} \tag{15}$$

(with $M_n^{(0)} = M_n$ from above). Altogether we find for the variance

$$V_n^{(d)} = \frac{p^2}{q^{2d+2}}\frac{1}{L}\left(1+q^{d+1}\right) - \frac{p^4}{q^{2d+2}}\frac{1}{L}\left(2\sum_{j\geq 1}\frac{Q^j}{(Q^j+1)^2}+\frac{1}{4}\right) + \tau_1^{(d)}(\log_Q n) + \mathcal{O}\left(\frac{1}{n}\right). \tag{16}$$

References

1. E3436 (Problem), American Mathematical Monthly **99** (1991), 366.
2. E3436 (Solution), American Mathematical Monthly **101** (1994), 78–80.
3. J.J.A.M.Brands, F.W.Steutel, and R.J.G.Wilms, *On the number of maxima in a discrete sample*, Statistics and Probability Letters **20** (1994), 209–217.
4. Y. Baryshnikov, B.Eisenberg, and G.Stengle, *A necessary and sufficient condition for the existence of the limiting probability of a tie for first place*, Statistics and Probability Letters (1995), 203–209.
5. B.Eisenberg, G.Stengle, and G.Strang, *The asymptotic probability of a tie for first place*, Annals of Applied Probability **3** (1993), 731–745.
6. P.Flajolet, P. Grabner, P.Kirschenhofer, H.Prodinger and R.F.Tichy, *Mellin Transforms and Asymptotics: Digital Sums*, Theoretical Computer Science **123** (1994), 291–314.
7. P.Flajolet and L.B.Richmond, *Generalized Digital Trees and Their Difference-Differential Equations*, Random Structures and Algorithms **3** (1992), 305–320.
8. P. Flajolet and R. Sedgewick, *Digital Search Trees Revisited*, SIAM J.Comput. **15** (1986), 748–767.
9. P.Flajolet and R.Sedgewick, *Mellin transforms and asymptotics: Finite differences and Rice's integrals*, Theoret. Comput. Sci. (1994) (to appear).
10. R.Graham, D.E.Knuth, and O.Patashnik, *Concrete Mathematics*, Addison-Wesley, 1989.
11. P. Henrici, *Applied and Computational Complex Analysis, vol.1*, J. Wiley, New York, 1977.
12. P. Kirschenhofer and H.Prodinger, *A Result in Order Statistics Related to Probabilistic Counting*, Computing **51** (1993), 15–27.
13. P.Kirschenhofer and H.Prodinger, *On Some Applications of Formulæof Ramanujan in the Analysis of Algorithms*, Mathematika **38** (1991), 14–33.
14. P.Kirschenhofer, H.Prodinger and W.Szpankowski, *On the Variance of the External Pathlength in a Symmetric Digital Trie*, Discrete Applied Mathematics **25** (1989), 129–143.
15. D.E.Knuth, *The Art of Computer Programming, Vol. 3*, Addison-Wesley, 1973.
16. H.M.Mahmoud, *Evolution of Random Search Trees*, Wiley-Interscience Series in Discrete Mathematics and Optimization, New York, 1992.
17. H.Prodinger and W.Szpankowski, *A note on binomial recurrences arising in the analysis of algorithms (Letter to the Editor)*, Information Processing Letters **46** (1993), 309–311.
18. W. Szpankowski, *The Evaluation of an Alternative Sum with Applications to the Analysis of Some Data Structures*, Information Processing Letters **28** (1988), 13–19.

TU Vienna
Wiedner Hauptstrasse 8-10
A-1040 Vienna
Austria