RAW MOMENTS AND ENTROPY ASSOCIATED WITH THE LAST PART IN A COMPOSITION

Meenakshi Rana
School of Mathematics, Thapar Institute of Engineering and Department, Patiala, Punjab, India
mrana@thapar.edu

Abhimanyu Kumar
Department of Electrical and Instrumentation Engineering, Thapar Institute of Engineering and Department, Patiala, Punjab, India
akumar6_be17@thapar.edu

Received: 9/23/22, Revised: 1/13/23, Accepted: 3/8/23, Published: 3/24/23

Abstract

Compositions are ordered partitions of a positive integer. They are useful in models for many computer science problems. This short article refines the statistical study of the last part in a composition. More precisely, the evaluation of raw and central moments as well as the entropy is conducted which is of deep significance to analysts.

1. Introduction

A composition of a positive integer n is an ordered partition, and total number of compositions are denoted by $c(n)$ (see [1, 12]). For instance, compositions of 3 are $\{3, 2+1, 1+2, 1+1+1\}$ so $c(3) = 4$. It can be easily seen that there are $c(n-1)$ compositions with last part 1 since deleting the last part 1 from these compositions gives the compositions of $n-1$. Similarly, the number of compositions of n with 2 as a last part is $c(n-2)$. Continuing this, we get $c(n) = c(n-1)+c(n-2)+\ldots+c(1)+1$ and $c(n) = 2^{n-1}$ for $n \geq 1$.

Hitczenko and Louchard [13] probabilistically analyzed the first empty part, the maximum part-size and the distribution of the number of distinct part sizes; later on, Louchard [16] analyzed the moments, the first full part, the maximum part size, and the distribution of the number of distinct parts of given multiplicity. Archibald et al. [3] investigated the probability of a random composition having no parts occurring exactly m times, where m belongs to a forbidden finite set of multiplicities. Considering all compositions as equiprobable, we know (see [13, 12]) that the num-

DOI:

ber of parts is asymptotically Gaussian, and that the part sizes are asymptotically distributed as $\mathsf{GEOM}(0.5)$. So, Louchard [15] asymptotically studied the properties of geometrically distributed random variables such as the limiting trajectories, the number of runs and the run length distribution, the hitting time to a length k run and the maximum run length, using a Markov chain approach and a polyomino-like description. Applying the machinery developed in [13], Louchard et al. [17] obtained the asymptotic forms of all moments. However, there are fluctuations (oscillations) involved which were given in the form of Fourier series using approximations obtained from the extreme value distribution together with the Mellin transform. Louchard and Prodinger [19] asymptotically evaluated the moments of the number of empty urns and largest non-empty urn in sequences of geometrically distributed random variables.

The longest run in a random word of length n has been studied in depth by Grabner et al. [10]. Gafni [8] examines the average and distribution of the length of the longest run of consecutive equal parts in a composition of size n. The analysis here has some similarities to the analytic treatment of compositions in [2], and the methods inspired from the book by Flajolet and Sedgewick [7]. Knopfmacher and Mansour [14] evaluate the asymptotic mean values for the number, and for the sum of positions, of records in compositions of n, where a part is a record based on whether it is greater than (or equal to) other parts of the composition. Brennan and Knopfmacher [5] studied the distribution of the number of large ascents in an integer composition. Two years later, Falah and Mansour [6] examined the small ascents. Archibald et al. [4] studied descents from maximal elements in samples of geometric random variables and determined the asymptotic expression for the mean of the greatest descent after a maximum value in a sample.

Let X be the last part in a composition of n. Then the relative frequency with which $X = k$ (for $k = 1, 2, \dots, n$) is given by

$$f_k = \frac{c(n-k)}{c(n)} = \begin{cases} 2^{-k} & : k < n \\ 2^{-n+1} & : k = n \end{cases},$$

where $\sum_{k=1}^{n} f_k = 1$. The first raw moment of X is simply given by

$$\mathbb{E}[X] = \sum_{k=1}^{n} f_k k = \frac{c(n-1) + 2c(n-2) + \ldots + nc(0)}{c(n)}$$

and the second moment is

$$\mathbb{E}[X^2] = \sum_{k=1}^{n} f_k k^2 = \frac{c(n-1) + 4c(n-2) + \ldots + n^2 c(0)}{c(n)}.$$

These formulas allow the calculation of the mean $\mu = \mathbb{E}[X]$ and the variance $\sigma^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Goyal-Rana [9] studied these quantities by approximating the summation with an integral. The arguments therein are strengthened here to obtain the exact relations. Unlike what was done before, here we also compute the entropy of X, given by

$$\mathbb{H}[X] = -\sum_{i=1}^{n} f_k \log f_k,$$

where log stands for the base-2 logarithm.

The main results are derived in Section 2, and concluded in Section 3.

2. Analysis and Discussion

Lemma 1. *We have*

$$\sum_{k=1}^{n} kr^k = \frac{r(1-r^{n+1})}{(1-r)^2} - \frac{(n+1)r^{n+1}}{1-r} \tag{1}$$

and

$$\sum_{k=1}^{n} k^2 r^k = \frac{2r^2(1-r^{n+1})}{(1-r)^3} + \frac{r(1-r^{n+1})}{(1-r)^2} - \frac{2(n+1)r^{n+2}}{(1-r)^2} - \frac{(n+1)^2 r^{n+1}}{1-r}, \tag{2}$$

where $r \neq 1$.

Proof. The proof is a straightforward calculation. Differentiate both sides of the well-known identity:

$$\sum_{k=1}^{n} r^k = \frac{1-r^{n+1}}{1-r}$$

with respect to r, and then multiply both sides by r in order to obtain Equation (1). Then replicate the process on Equation (1) in order to get Equation (2). □

Theorem 1. *We have $\mathbb{E}[X] = 2 - 2^{-n+1}$ and $\mathbb{E}[X^2] = 6 - (3+2n)2^{-n+1}$.*

Proof. Using Equation (1), we can simplify the first raw moment as

$$\begin{aligned}
\mathbb{E}[X] = \sum_{k=1}^{n} f_k k &:= 2^{-n+1}n + \sum_{k=1}^{n-1} 2^{-k}k \\
&= 2^{-n+1}n + \left(\frac{2^{-1}(1-2^{-n})}{(1-2^{-1})^2} - \frac{2^{-n}n}{1-2^{-1}} \right) \\
&= 2^{-n+1}n + \big(2(1-2^{-n}) - 2^{-n+1}n\big) \\
&= 2 - 2^{-n+1},
\end{aligned}$$

which is the final result. Similarly, using Equation (2), we have the second raw moment as

$$\begin{aligned}\mathbb{E}[X^2] = \sum_{k=1}^{n} f_k k &:= 2^{-n+1}n^2 + \sum_{k=1}^{n-1} 2^{-k}k^2 \\ &= 2n^2 2^{-n} + \big(6(1-2^{-n}) - 4n2^{-n} - 2n^2 2^{-n}\big) \\ &= 6 - (3+2n)2^{-n+1},\end{aligned}$$

which is the final result. □

The above theorem allows us to calculate the central moments, namely the mean $\mu = \mathbb{E}[X] = 2 - 2^{-n+1}$ and variance $\sigma^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 2 - (2n-1)2^{-n+1} - 2^{-2n+2}$. These exact formulas are a result of our improvements to the procedure opted in [9]. For large n, the mean of the last part of the composition is 2 since $\lim_{n\to\infty} \mathbb{E}[X] = 2$. Variance is a measure of dispersion, that is, how far a set of numbers is spread out from their mean. The variance too turns out to be 2 for large n. This observation hints at properties analogous to the Poisson distribution.

Given that entropy is relevant to information theory, statistical thermodynamics, and combinatorics; with the following result, we propose the evaluation of entropy in the study of statistics of partitions.

Theorem 2. *We have* $\mathbb{H}[X] = 2(1 - 2^{-n+1})$.

Proof. The entropy is expressible as

$$\begin{aligned}\mathbb{H}[X] &= -\sum_{k=1}^{n} f_k \log(f_k) \\ &= -2^{-n+1}\log(2^{1-n}) - \sum_{k=1}^{n-1} 2^{-k}\log(2^{-k}) \\ &= (n-1)2^{-n+1} + \sum_{k=1}^{n-1} 2^{-k}k,\end{aligned}$$

where we use Equation (1) to achieve

$$\begin{aligned}\mathbb{H}[X] &= (n-1)2^{-n+1} + \left(\frac{2^{-1}(1-2^{-n})}{(1-2^{-1})^2} - \frac{2^{-n}n}{1-2^{-1}}\right) \\ &= 2(1-2^{-n+1}),\end{aligned}$$

which is the desired result. □

In information theory, the entropy $\mathbb{H}[X]$ of a random variable X measures the average level of uncertainty in the possible outcomes of X. In our context, the entropy measures how "surprising" the average value is of the last part in a composition of n, viewed as the average outcome of a random variable. For reference, remember that entropy is zero for a double-headed coin since the outcome of coin toss is always heads so there is no uncertainty.

For interested readers, Figure 1 shows the plot of mean, variance, and entropy of the last part of a composition of n.

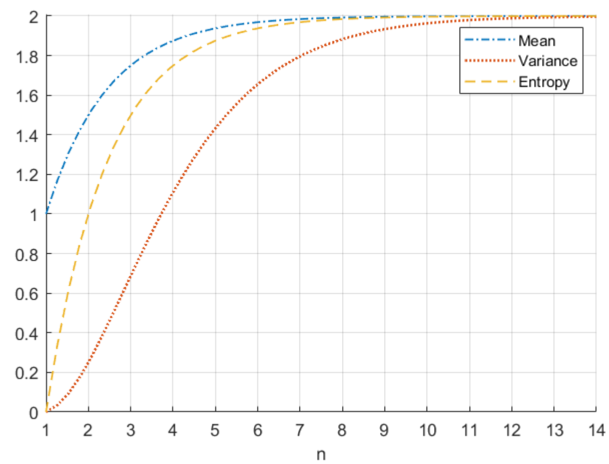

Figure 1: Plot of mean, variance, and entropy.

The mean and variance are standard central moments, but if necessary then higher-order central moments may also be calculated using raw moments. Below, Lemma 2 mentions the well-known Faulhaber's formula (see [11, Section 6.5]), which gives an expression for the sum of the pth powers of the first n positive integers as a $(p+1)$th degree polynomial in n.

Lemma 2 (Faulhaber's formula)**.** *Let us consider a non-negative integer p and the Bernoulli numbers of the second kind $\{B_j\}_{j=0,1,\ldots,p}$ where $B_0 = 1$ and $B_1 = \frac{1}{2}$. Then the following formula holds:*

$$\sum_{k=1}^{n} k^p = \frac{1}{p+1} \sum_{j=0}^{p} \binom{p+1}{j} B_j n^{p-j+1}.$$

This formula enables us to obtain the recursive formula for the raw moments, given in the theorem below.

Theorem 3. *For $m \in \mathbb{N}$, we have*

$$\mathbb{E}[X^{m+1}] = \frac{3}{2}(m+1)\mathbb{E}[X^m] - 2^{-n+1}(m+1)n^m - \sum_{j=2}^{m} \binom{m+1}{j} B_j \mathbb{E}\left[X^{m-j+1}\right].$$

Proof. Applying summation-by-parts gives

$$\begin{aligned}\sum_{k=1}^{n} f_k k^m &= f_n \sum_{k=1}^{n} k^m - \sum_{\ell=1}^{n-1} (f_{\ell+1} - f_\ell) \sum_{k=1}^{\ell} k^m \\ &= 2^{-n+1} \sum_{k=1}^{n} k^m - (2^{-n+1} - 2^{-n+1}) \sum_{k=1}^{n-1} k^m - \sum_{\ell=1}^{n-2} (2^{-\ell-1} - 2^{-\ell}) \sum_{k=1}^{\ell} k^m \\ &= 2^{-n+1} \sum_{k=1}^{n} k^m - 0 + 2^{-1} \sum_{\ell=1}^{n-2} 2^{-\ell} \sum_{k=1}^{\ell} k^m.\end{aligned}$$

Now, using Faulhaber's formula gives

$$\begin{aligned}\mathbb{E}[X^m] &= 2^{-n+1} \sum_{k=1}^{n} k^m + 2^{-1} \sum_{\ell=1}^{n-2} 2^{-\ell} \left(\frac{1}{m+1} \sum_{j=0}^{m} \binom{m+1}{j} B_j \ell^{m-j+1} \right) \\ &= 2^{-n+1} \sum_{k=1}^{n} k^m + \frac{2^{-1}}{m+1} \sum_{j=0}^{m} \binom{m+1}{j} B_j \sum_{\ell=1}^{n-2} 2^{-\ell} \ell^{m-j+1}.\end{aligned}$$

Since $\mathbb{E}\left[X^{m-j+1}\right] = 2^{-n+1} n^{m-j+1} + 2^{-n+1}(n-1)^{m-j+1} + \sum_{\ell=1}^{n-2} 2^{-\ell} \ell^{m-j+1}$, we have

$$\begin{aligned}\mathbb{E}[X^m] = 2^{-n+1} \sum_{k=1}^{n} k^m &+ \frac{2^{-1}}{m+1} \sum_{j=0}^{m} \binom{m+1}{j} B_j \mathbb{E}\left[X^{m-j+1}\right] \\ &- \frac{2^n}{m+1} \sum_{j=0}^{m} \binom{m+1}{j} B_j (n^{m-j+1} + (n-1)^{m-j+1}),\end{aligned}$$

which can be simplified using Faulhaber's formula in reverse, that is,

$$\begin{aligned}\mathbb{E}[X^m] &= 2^{-n+1} \sum_{k=1}^{n} k^m + \frac{2^{-1}}{m+1} \sum_{j=0}^{m} \binom{m+1}{j} B_j \mathbb{E}\left[X^{m-j+1}\right] \\ &\quad - 2^{-n} \left(\sum_{k=1}^{n} k^m + \sum_{k=1}^{n-1} k^m \right) \\ &= 2^{-n+1} \sum_{k=1}^{n} k^m + \frac{2^{-1}}{m+1} \sum_{j=0}^{m} \binom{m+1}{j} B_j \mathbb{E}\left[X^{m-j+1}\right] \\ &\quad - 2^{-n} \left(2 \sum_{k=1}^{n} k^m - n^m \right) \\ &= 2^{-n} n^m + \frac{2^{-1}}{m+1} \sum_{j=0}^{m} \binom{m+1}{j} B_j \mathbb{E}\left[X^{m-j+1}\right].\end{aligned}$$

The final result follows immediately upon expanding the first few terms of the summation and rearranging it to have $\mathbb{E}[X^{m+1}]$ on the left side of the equation. □

3. Conclusion

This paper obtains exact formulas for expectation and variance of the last part in a composition. If needed, skewness, kurtosis, and other central moments can also be determined likewise. Additionally, entropy is also evaluated here.

Acknowledgements. The first author is supported by Science and Engineering Research Board (SERB-MATRICS) grant MTR/2019/000123. The authors express their sincere thanks to the editor and reviewer for the constructive comments which led to improvements in content and presentation of this paper.

References

[1] G. E. Andrews, *The Theory of Partitions* (Encyclopedia of Mathematics and its Applications), Cambridge University Press, Cambridge, 1998.

[2] M. Archibald, A. Knopfmacher, and H. Prodinger, The number of distinct values in a geometrically distributed sample, *European J. Combin.* **27** (2006), 1059–1081.

[3] M. Archibald, A. Knopfmacher, and T. Mansour, Compositions and samples of geometric random variables with constrained multiplicities, *Discrete Math. Theor. Comput. Sci. Proc.* **AN** (2010), 449–460.

[4] M. Archibald, A. Blecher, C. Brennan, A. Knopfmacher, and H. Prodinger, Geometric random variables: Descents following maxima, *Statist. Probab. Lett.* **124** (2017), 140–147.

[5] C. Brennan, A. Knopfmacher, The distribution of ascents of size d or more in compositions, *Discrete Math. Theor. Comput. Sci.* **11** (2009), 1–10.

[6] M. Falah and T. Mansour, Ascents of size less than d in compositions, *Cent. Eur. J. Math.* **9** (2011), 196–203.

[7] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge Univ. Press, Cambridge, 2009.

[8] A. Gafni, Longest run of equal parts in a random integer composition, *Discrete Math.* **338(2)** (2015), 236–247.

[9] M. Goyal, M. Rana: On central tendencies of the parts with some restrictions, *Math. Rep. (Bucur.)* **20** (2018), 291–300.

[10] P. Grabner, A. Knopfmacher, and H. Prodinger, Combinatorics of geometrically distributed random variables: run statistics, *Theoret. Comput. Sci.* **297** (2003), 261–270.

[11] R. L. Graham, D. E. Knuth, O. Patashnik, *Concrete Mathematics*, 2nd ed., Addison-Wesley, Reading, MA, 1998.

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

[13] P. Hitczenko and G. Louchard, Distinctness of compositions of an integer: a probabilistic analysis, *Random Structures Algorithms* **19(3-4)** (2001), 407–437.

[14] A. Knopfmacher and T. Mansour, Record statistics in integer compositions, *Discrete Math. Theor. Comput. Sci. Proc.* **AK** (2009), 527–536.

[15] G. Louchard, Runs of geometrically distributed random variables: a probabilistic analysis, *J. Comput. Appl. Math.* **142(1)** (2002), 137–153.

[16] G. Louchard, The number of distinct part sizes of some multiplicity in compositions of an Integer: a probabilistic analysis, *Discrete Math. Theor. Comput. Sci. Proc.* **AC** (2003), 155–170.

[17] G. Louchard, H. Prodinger, and M. D. Ward, The number of distinct values of some multiplicity in sequences of geometrically distributed random variables, *Discrete Math. Theor. Comput. Sci. Proc.* **AD** (2005), 231–256.

[18] G. Louchard, H. Prodinger, Asymptotics of the moments of extreme-value related distribution functions, *Algorithmica* **46** (2006), 431–467.

[19] G. Louchard and H. Prodinger, On gaps and unoccupied urns in sequences of geometrically distributed random variables, *Discrete Math.* **308(9)** (2008), 1538–1562.

[20] H. Wilf, The distribution of run lengths in integer compositions, *Electron. J. Combin.* **18** (2011).