A JOINT CENTRAL LIMIT THEOREM FOR THE SUM-OF-DIGITS FUNCTION, AND ASYMPTOTIC DIVISIBILITY OF CATALAN-LIKE SEQUENCES

MICHAEL DRMOTA* AND CHRISTIAN KRATTENTHALER†

*Institute of Discrete Mathematics and Geometry, TU Wien,
Wiedner Hauptstraße 8–10, A-1040 Vienna, Austria.
WWW: `https://www.dmg.tuwien.ac.at/drmota`

†Fakultät für Mathematik der Universität Wien,
Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria.
WWW: `http://www.mat.univie.ac.at/~kratt`

Abstract. We prove a central limit theorem for the joint distribution of $s_q(A_j n)$, $1 \le j \le d$, where s_q denotes the sum-of-digits function in base q and the A_j's are positive integers relatively prime to q. We do this in fact within the framework of quasi-additive functions. As application, we show that most elements of "Catalan-like" sequences — by which we mean integer sequences defined by products/quotients of factorials — are divisible by any given positive integer.

1. Introduction

In [5], Burns shows that most of the ubiquitous *Catalan numbers* $C_n := \frac{1}{n+1}\binom{2n}{n}$ (cf. [17, Ex. 6.19]) are divisible by p, where p is some given prime number. Let $v_p(N)$ denote the *p-adic valuation* of the integer N, which by definition is the maximal exponent α such that p^α divides N. In view of Legendre's formula [10, p. 10] for the p-adic valuation of factorials,

$$v_p(n!) = \frac{1}{p-1}\big(n - s_p(n)\big), \tag{1.1}$$

where $s_p(N)$ denotes the *p-ary sum-of-digits function*

$$s_p(N) = \sum_{j\ge 0} \varepsilon_j(N),$$

with $\varepsilon_j(N)$ denoting the j-th digit in the p-adic representation of N, we have

$$v_p\left(\frac{1}{n+1}\binom{2n}{n}\right) = \frac{1}{p-1}\big(2s_p(n) - s_p(2n)\big) - v_p(n+1).$$

Thus, one sees that the above and many more asymptotic divisibility results — such as the divisibility of most of the Catalan numbers, or even of most of the *Fuß–Catalan*

2010 *Mathematics Subject Classification.* Primary 11K65; Secondary 05A15 11A63 11N60 60F05.

Key words and phrases. Sum-of-digits function, quasi-additive functions, central limit theorem, Catalan numbers, central binomial coefficients.

*† Research partially supported by the Austrian Science Foundation FWF, in the framework of the Special Research Program "Algorithmic and Enumerative Combinatorics", project F50-N15.

numbers (cf. [1, pp. 59–60]) by any given prime *power* — can be proved if one has sufficiently precise results on the distribution of the vector

$$(s_p(A_1 n), s_p(A_2 n), \dots, s_p(A_d n)), \qquad n < N. \tag{1.2}$$

Indeed, for $p = 2$, Schmidt [14] and Schmid [13] showed that for pairwise different positive odd integers $A_1, A_2, \dots, A_d$ the vector (1.2) satisfies a d-dimensional central limit theorem with asymptotic mean vector $(1/2, \dots, 1/2) \cdot \log_2 N$ and asymptotic covariance matrix $\Sigma \cdot \log_2 N$ with

$$\Sigma = \left(\frac{\gcd(A_i, A_j)^2}{4 A_i A_j} \right)_{1 \le i,j \le d}.$$

We want to mention that the distribution of the vector (1.2) in residue classes was intensively studied in [6]. Furthermore we also emphasize that many results on divisiblity properties of non-central binomial coefficients are available, too, see for example [16].

The first goal of the present paper is to generalize the central limit theorem by Schmidt [14] and Schmid [13] to arbitrary primes p, and even to arbitrary bases q. We do this in Theorem 1 in Section 2, by using an even more general concept, namely the concept of *q-quasi-additive functions*.

We then apply this result in Section 3 (see Theorem 4 and Corollary 5) to prove the somewhat non-intuitive fact that most elements of any sequence $\big(S(n)\big)_{n \ge 0}$ of integers given by a (non-trivial) formula

$$\frac{P(n)}{Q(n)} \frac{\prod_{i=1}^r (C_i n)!}{\prod_{i=1}^s (D_i n)!}$$

are divisible by any given prime power, and thus by any given positive integer. Here, $P(n)$ and $Q(n)$ are polynomials in n over the integers, where $Q(n)$ is a product of linear factors, and the C_i's and D_i's are positive integers with $\sum_{i=1}^r C_i = \sum_{i=1}^s D_i$. The attribute "non-trivial" means that the set of C_i's is different from the set of D_i's. As is pointed out in more detail in Section 3, numerous (mainly combinatorial) sequences that appear in the literature in various contexts are of this form.

2. A central limit theorem

Let $q \ge 2$ be a given integer. It is well known that the sum-of-digits function $s_q(n)$ satisfies a central limit theorem of the form

$$\frac{1}{N} \# \left\{ n < N : s_q(n) \le \mu_q \log_q N + t \sqrt{\sigma_q^2 \log_q N} \right\} = \Phi(t) + o(1), \tag{2.1}$$

uniformly in t, where $\mu_q = (q-1)/2$, $\sigma_q^2 = (q^2-1)/12$, and $\Phi(t)$ denotes the distribution function of the standard Gaußian distribution. This result is easy to prove since the digits $\varepsilon_j(n)$, $0 \le j < \log_q(N)$, behave almost as i.i.d. random variables if n varies between 0 and $N-1$. Actually much more is known (see for example [2]). Suppose that $P(x)$ is a polynomial of degree $D \ge 1$ with non-negative integer coefficients. Then we also have

$$\frac{1}{N} \# \left\{ n < N : s_q(P(n)) \le \mu_q \log_q P(N) + t \sqrt{\sigma_q^2 \log_q P(N)} \right\} = \Phi(t) + o(1).$$

Note that the value $P(N)$ can be replaced by N^D without changing the validity of the statement.

This result applies in particular to linear polynomials $P_j(n) = A_j n$ (with integers $A_j \ge 1$). In what follows, we will consider linear combinations of the form

$$f(n) = c_1 s_q(A_1 n) + c_2 s_q(A_2 n) + \cdots + c_d s_q(A_d n), \qquad n < N, \tag{2.2}$$

with real numbers c_j and integers $A_j \ge 1$, $1 \le j \le d$. Clearly, the central limit result of Schmidt [14] and Schmid [13] mentioned in the introduction is equivalent to the fact that $f(n)$ as in (2.2) with $q = 2$, $n < N$, satisfies a one-dimensional central limit theorem with asymptotic mean $\frac{1}{2}(c_1 + c_2 + \cdots + c_d) \cdot \log_2 N$ and asymptotic covariance $\mathbf{c}\Sigma\mathbf{c}^t \cdot \log_2 N$, where $\mathbf{c} = (c_1, c_2, \dots, c_d)$.

It is also clear that the results of [13, 14] should directly transfer to a general basis $q \ge 2$ so that we can cover general $f(n)$. We will establish this generalization, however, with a completely different (and in fact more modern) proof.

Theorem 1. *Let $q \ge 2$ be an integer, and let $A_1, A_2, \dots, A_d$ be positive integers. Then the vector*

$$(s_q(A_1 n), s_q(A_2 n), \dots, s_q(A_d n)), \qquad 0 \le n < N, \tag{2.3}$$

satisfies a d-dimensional central limit theorem with asymptotic mean vector $((q-1)/2, \dots, (q-1)/2) \cdot \log_q N$ and asymptotic covariance matrix $\Sigma \cdot \log_q N$, where Σ is positive semi-definite.

If we further assume that q is prime and that the integers $A_1, A_2, \dots, A_d$ are not divisible by q, then Σ is explicitly given by

$$\Sigma = \left(\frac{(q^2-1)}{12} \frac{\gcd(A_i, A_j)^2}{A_i A_j} \right)_{1 \le i,j \le d}. \tag{2.4}$$

For the proof we make use of the (recent) concept of quasi-additivity which is thoroughly discussed in [8]. There, a function f defined on the non-negative integers is called *q-quasi-additive*, if there exists $r \ge 0$ such that

$$f(q^{k+r} a + b) = f(a) + f(b) \qquad \text{for all } b < q^k. \tag{2.5}$$

We note that if (2.5) holds for some $r \ge 0$, then it holds as well for every larger r. This also shows that linear combinations of q-quasi-additive functions are q-quasi-additive, too. We further note that $s_q(n)$ is q-quasi-additive with parameter $r = 0$.

One of the main results of the paper [8] is that any q-quasi-additive function $f(n)$ of at most logarithmic growth satisfies a central limit theorem of the form

$$\frac{1}{N} \# \left\{ n < N : f(n) \le \mu \log_q N + t \sqrt{\sigma^2 \log_q N} \right\} = \Phi(t) + o(1),$$

for appropriate constants μ and σ^2.

Our first observation is that $f(n)$ given in (2.2) is q-quasi-additive. The logarithmic growth property is trivially satisfied since $s_q(n) \le (q-1) \log_q n$.

Lemma 2. *Let A and r be positive integers with $q^r \ge A$. Then $g(n) = s_q(An)$ is q-quasi-additive (with parameter r).*

Proof. Suppose that $b < q^k$. Then $Ab < q^{k+r}$, and consequently

$$g(q^{k+r} a + b) = s_q(q^{k+r} Aa + Ab) = s_q(Aa) + s_q(Ab) = g(a) + g(b). \qquad \square$$

Since linear combinations of q-quasi-additive functions are q-quasi-additive, it directly follows that $f(n)$, as given by (2.2), satisfies a central limit theorem of the prescribed form. This also implies that the vector (2.3) satisfies a d-dimensional central limit theorem with asymptotic mean $((q-1)/2, \ldots, (q-1)/2) \cdot \log_q N$. (This follows from the fact that a random vector $\mathbf{X} = (X_1, \ldots X_d)$ is Gaußian with mean vector $(\mu_1, \ldots, \mu_d)$ and covariance matrix Σ if and only if every projection $c_1 X_1 + \cdots + c_d X_d$ with real $c_1, \ldots, c_d$ is univariate Gaußian with mean $c_1\mu_1 + \cdots + c_d\mu_d$ and variance $(c_1, \ldots, c_d)\Sigma(c_1, \ldots, c_d)^t$).

It remains to compute the covariance matrix in the case, where q is a prime number.

Lemma 3. *Let $q \ge 2$ be a prime number, let A_1, A_2 be positive integers that are not divisible by q, and set $D = \gcd(A_1, A_2)$. Then, uniformly for $(\log N)^{1/3} \le i, j \le \log_q N - (\log N)^{1/3}$ and $a, b \in \{0, 1, \ldots, q-1\}$, we have*

$$\frac{1}{N} \# \{n < N : \varepsilon_i(A_1 n) = a,\ \varepsilon_j(A_2 n) = b\}$$
$$= \begin{cases} \frac{1}{q^2} + O\left((\log N)^{-C}\right), & \text{if } i \ne j, \\ \frac{1}{q^2} + \frac{D^2}{A_1 A_2} \sum\limits_{\ell \ne 0} \frac{1}{4\pi^2 \ell^2} \left(e\left(-\frac{\ell A_2 a}{qD}\right) - e\left(-\frac{\ell A_2 (a+1)}{qD}\right) \right) \left(e\left(\frac{\ell A_1 b}{qD}\right) - e\left(-\frac{\ell A_1 (b+1)}{qD}\right) \right) & \\ \qquad\qquad +O\left((\log N)^{-C}\right), & \text{if } i = j, \end{cases}$$

for any given $C > 0$. Here, $e(x) = e^{2\pi i x}$.

Proof. We adapt the method of [2] to the present situation. However, in order to make the presentation more transparent, we first present a slightly simplified approach. First we note that $\varepsilon_j(n) = a$ if and only if $\{nq^{-j-1}\} \in [a/q, (a+1)/q)$, where $\{x\} = x - \lfloor x \rfloor$ denotes the fractional part of x. We also note that the Fourier series of the characteristic function $\mathbf{1}_{[a/q,(a+1)/q)}(x)$ is given by

$$\mathbf{1}_{\left[\frac{a}{q}, \frac{a+1}{q}\right)}(x) = \sum_m d_m(a) e(mx) \quad \text{with} \quad d_m(a) = \begin{cases} \frac{1}{q}, & \text{if } m = 0, \\ \frac{e\left(-\frac{ma}{q}\right) - e\left(-\frac{m(a+1)}{q}\right)}{2\pi i m}, & \text{if } m \ne 0. \end{cases}$$

This Fourier series is not absolutely convergent. This is the major reason that we have to be more precise in a second round. Observe that $d_m(a) = 0$ if $m \ne 0$ and $m \equiv 0 \bmod q$.

We have

$$\# \{n < N : \varepsilon_i(A_1 n) = a,\ \varepsilon_j(A_2 n) = b\} = \sum_{n<N} \mathbf{1}_{\left[\frac{a}{q}, \frac{a+1}{q}\right)} \left(\frac{A_1 n}{q^{i+1}} \right) \mathbf{1}_{\left[\frac{b}{q}, \frac{b+1}{q}\right)} \left(\frac{A_2 n}{q^{j+1}} \right)$$
$$= \sum_{m_1, m_2} d_{m_1}(a)\, d_{m_2}(b) \sum_{n<N} e\left(\left(\frac{A_1 m_1}{q^{i+1}} + \frac{A_2 m_2}{q^{j+1}} \right) n \right).$$

Since

$$\left| \sum_{n<N} e(\alpha n) \right| \le \frac{2}{|1 - e(\alpha)|},$$

we may *neglect* all exponential sums where $\alpha = \frac{A_1 m_1}{q^{i+1}} + \frac{A_2 m_2}{q^{j+1}}$ is not an integer. At this moment, this step is not rigorous since the Fourier series is not absolutely convergent.

Next suppose that α is an integer. If $i \neq j$, the number $\frac{A_1 m_1}{q^{i+1}} + \frac{A_2 m_2}{q^{j+1}}$ can be an integer only if $m_1 = m_2 = 0$ since we also assume that A_1 and A_2 are not divisible by q. Thus we should get

$$\#\{n < N : \varepsilon_i(A_1 n) = a,\, \varepsilon_j(A_2 n) = b\} = d_0(a) d_0(b) N + o(N) = \frac{N}{q^2} + o(N).$$

If $i = j$, then the assumption $\frac{A_1 m_1}{q^{j+1}} + \frac{A_2 m_2}{q^{j+1}} = k$ for an integer k leads to $|m_1| \geq \frac{1}{2A_1}|k| q^{i+1} \geq \frac{1}{2A_1}|k| q^{(\log N)^{1/3}}$ or $|m_2| \geq \frac{1}{2A_2}|k| q^{j+1} \geq \frac{1}{2A_2}|k| q^{(\log N)^{1/3}}$ so that the corresponding terms are negligible (if the Fourier series would be absolutely convergent).

Thus we should get (again by observing that all summands for which α is an integer can be put into an error term)

$$\begin{aligned}\#\{n < N : \varepsilon_j(A_1 n) = a,\, \varepsilon_j(A_2 n) = b\} &= \sum_{\ell} d_{\ell A_2/D}(a) d_{-\ell A_1/D}(b)\, N + o(N) \\ &= \frac{N}{q^2} + \sum_{\ell \neq 0} \frac{1}{4\pi^2 \ell^2} \left(e\left(-\tfrac{\ell A_2 a}{qD}\right) - e\left(-\tfrac{\ell A_2 (a+1)}{qD}\right)\right) \left(e\left(\tfrac{\ell A_1 b}{qD}\right) - e\left(-\tfrac{\ell A_1 (b+1)}{qD}\right)\right) N \\ &\qquad + o(N).\end{aligned}$$

In order to make the above heuristics rigorous, we proceed as in [2]. We replace the characteristic function $\mathbf{1}_{[a/q,(a+1)/q)}(x)$ by a smoothed version. Let $\chi_{a,\Delta}(x)$ be defined by

$$\chi_{a,\Delta}(x) := \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} \mathbf{1}_{[\frac{a}{q}, \frac{a+1}{q}]}(\{x+z\})\, dz,$$

The Fourier coefficients of the Fourier series $\chi_{a,\Delta}(x) = \sum_{m \in \mathbb{Z}} d_{m,\Delta}(a) e(mx)$ are given by

$$d_{0,\Delta}(a) = \frac{1}{q},$$

and for $m \neq 0$ by

$$d_{m,\Delta}(a) = \frac{e\left(-\frac{ma}{q}\right) - e\left(-\frac{m(a+1)}{q}\right)}{2\pi i m} \cdot \frac{e\left(\frac{m\Delta}{2}\right) - e\left(-\frac{m\Delta}{2}\right)}{2\pi i m \Delta}.$$

Note that $d_{m,\Delta}(a) = 0$ if $m \neq 0$ and $m \equiv 0 \bmod q$, and that

$$|d_{m,\Delta}(a)| \leq \min\left(\frac{1}{\pi |m|}, \frac{1}{\Delta \pi m^2}\right).$$

By definition, we have $0 \leq \chi_{a,\Delta}(x) \leq 1$ and

$$\chi_{a,\Delta}(x) = \begin{cases} 1, & \text{if } x \in \left[\frac{a}{q} + \Delta, \frac{a+1}{q} - \Delta\right], \\ 0, & \text{if } x \in [0,1] \setminus \left[\frac{a}{q} - \Delta, \frac{a+1}{q} + \Delta\right]. \end{cases}$$

In particular, we set $\Delta = (\log N)^{-C}$ for some (sufficiently large) constant C. Of course we have to take into account all error terms. The smoothing error can be handled with the help of a discrepancy estimate (see [2]). Putting the resulting estimates together — we leave the details to the reader —, one obtains

$$\begin{aligned}\#\left\{n<N:\varepsilon_i(A_1n)=a,\,\varepsilon_j(A_2n)=b\right\}&=d_{0,\Delta}(a)d_{0,\Delta}(b)N+O\left(N(\log N)^{-C}\right)\\&=\frac{N}{q^2}+O\left(N(\log N)^{-C}\right)\end{aligned}$$

for $i\neq j$, and

$$\begin{aligned}\#\left\{n<N:\varepsilon_j(A_1n)=a,\,\varepsilon_j(A_2n)=b\right\}=\sum_\ell d_{\ell A_2/D,\Delta}(a)d_{-\ell A_1/D,\Delta}(b)\,N\\+O\left(N(\log N)^{-C}\right)\end{aligned}$$

for $i=j$, where all estimates are uniform for $(\log N)^{1/3}\le i,j\le\log_q N-(\log N)^{1/3}$. Since

$$d_{m,\Delta}(a)=d_m(a)\frac{\sin(\pi m\Delta)}{\pi m\Delta}=d_m(a)\left(1+O\left(\frac{1}{m\Delta}\right)\right)$$

for $1\le|m|\le 1/\Delta$, we obtain (with $A=\max\{A_1,A_2\}$)

$$\begin{aligned}\sum_{1\le|\ell|\le 1/(A\Delta)}d_{\ell A_2/D,\Delta}(a)d_{-\ell A_1/D,\Delta}(b)=\sum_{1\le|\ell|\le 1/(A\Delta)}d_{\ell A_2/D}(a)d_{-\ell A_1/D}(b)\\+O\left(\Delta\log(1/\Delta)\right),\end{aligned}$$

and

$$\begin{aligned}\sum_{|\ell|>1/(A\Delta)}d_{\ell A_2/D,\Delta}(a)d_{-\ell A_1/D,\Delta}(b)=O\left(\Delta\right),\\\sum_{|\ell|>1/(A\Delta)}d_{\ell A_2/D}(a)d_{-\ell A_1/D}(b)=O\left(\Delta\right).\end{aligned}$$

Thus,

$$\sum_\ell d_{\ell A_2/D,\Delta}(a)d_{-\ell A_1/D,\Delta}(b)=\sum_\ell d_{\ell A_2/D}(a)d_{-\ell A_1/D}(b)+O\left(\frac{\log\log N}{(\log N)^C}\right)$$

This completes the proof of the lemma. □

It is now not difficult to complete the computation of the covariance matrix (which also completes the Proof of Theorem 1). By definition, the covariance of $s_q(A_1n)$ and $s_q(A_2n)$ is given by

$$\mathbf{Cov}=\frac{1}{N}\sum_{n<N}s_q(A_1n)s_q(A_2n)-\frac{1}{N}\sum_{n<N}s_q(A_1n)\cdot\frac{1}{N}\sum_{n<N}s_q(A_2n).$$

In order to apply Lemma 3, we neglect the digits ε_j with $j\le(\log N)^{1/3}$ or $j\ge\log_q(N)-(\log N)^{1/3}$ and denote by $\overline{s}_q$ the sum of digits of the remaining digits ε_j with $(\log N)^{1/3}<j<\log_q N-(\log N)^{1/3}$. Then the corresponding approximate covariance $\overline{\mathbf{Cov}}$ satisfies

$$\mathbf{Cov}-\overline{\mathbf{Cov}}=O\left((\log N)^{5/6}\right),$$

which can be shown with the help of the Cauchy–Schwarz inequality. Hence, by rewriting $\overline{\mathbf{Cov}}$ with the help of the numbers $\frac{1}{N}\#\{n<N:\varepsilon_i(A_1n)=a,\,\varepsilon_j(A_2n)=b\}$ (from Lemma 3) and the numbers

$$\frac{1}{N}\#\{n<N:\varepsilon_j(A_in)=a\}=\frac{1}{q}+O\left((\log N)^{-C}\right)$$

(note that the fact that this asymptotic property holds uniformly for $(\log N)^{1/3}\le j\le \log_q N-(\log N)^{1/3}$, $a\in\{0,1,\ldots,q-1\}$, and $i=1,2$ follows from Lemma 3), we immediately get

$$\begin{aligned}
\overline{\mathbf{Cov}} &= L\frac{D^2}{A_1A_2}\sum_{a,b=0}^{q-1}ab\sum_{\ell\not\equiv 0 \bmod q}\frac{1}{4\pi^2\ell^2}\\
&\qquad\qquad\cdot\left(e\left(-\tfrac{\ell A_2 a}{qD}\right)-e\left(-\tfrac{\ell A_2(a+1)}{qD}\right)\right)\left(e\left(\tfrac{\ell A_1 b}{qD}\right)-e\left(-\tfrac{\ell A_1(b+1)}{qD}\right)\right)\\
&\qquad\qquad+O\left((\log N)^{2-C}\right)\\
&= L\frac{D^2}{A_1A_2}\frac{q^2}{4\pi^2}\sum_{\ell\not\equiv 0 \bmod q}\frac{1}{\ell^2}+O\left((\log N)^{2-C}\right)\\
&= L\frac{D^2}{A_1A_2}\frac{q^2-1}{12}+O\left((\log N)^{2-C}\right),
\end{aligned}$$

where $L=\lfloor \log_q N-2(\log N)^{1/3}\rfloor$, and where we have used the identity

$$\sum_{a=0}^{q-1}a\,e(ak/q)=\frac{q}{e(k/q)-1},$$

which is valid for integers k that are not divisible by q. We can choose C appropriately — for example $C=2$ — and finally obtain

$$\mathbf{Cov}=\frac{q^2-1}{12}\frac{\gcd(A_i,A_j)^2}{A_iA_j}\log_q N+O\left((\log N)^{5/6}\right),$$

which completes the proof of Theorem 1.

3. Asymptotic divisibility of Catalan-like integer sequences

The main result in this section concerns divisibility of "Catalan-like" sequences by prime powers.

Theorem 4. *Let p be a given prime number, α a positive integer, $P(n)$ a polynomial in n with integer coefficients, and $(C_i)_{1\le i\le r}$, $(D_i)_{1\le i\le s}$, $(E_i)_{1\le i\le t}$, $(F_i)_{1\le i\le t}$ given integer sequences with $C_i,D_i>0$ and $p\nmid\gcd(E_i,F_i)$ for all i, $\sum_{i=1}^r C_i=\sum_{i=1}^s D_i$, and $\{C_i:1\le i\le r\}\ne\{D_i:1\le i\le s\}$. If all elements of the sequence $\big(S(n)\big)_{n\ge0}$, defined by*

$$S(n):=\frac{P(n)}{\prod_{i=1}^t(E_in+F_i)}\frac{\prod_{i=1}^r(C_in)!}{\prod_{i=1}^s(D_in)!},\tag{3.1}$$

are integers, then

$$\lim_{N\to\infty}\frac{1}{N}\#\{n<N:S(n)\equiv 0 \pmod{p^\alpha}\}=1.\tag{3.2}$$

We note that (3.2) remains true if α increases slowly with N. In particular, we can choose $\alpha = \lfloor \eta \log N \rfloor$ for an appropriate $\eta > 0$. We will actually show in the proof of Theorem 4 that

$$\#\left\{n < N : v_p(S(n)) < \eta \log N\right\} = O\left(\frac{N}{\sqrt{\log N}}\right)$$

if $\eta > 0$ is sufficiently small.

Furthermore we note that the assumption $p \nmid \gcd(E_i, F_i)$ is not really necessary since we can always reduce the problem to this case by separating the factors $p^{v_p(\gcd(E_i,F_i))}$. Thus, we immediately obtain the following corollary.

Corollary 5. *Let $S(n)$ be given as in Theorem 4 (without assuming the condition $p \nmid \gcd(E_i, F_i)$). Then, for all positive integers m, we have*

$$\lim_{N\to\infty} \frac{1}{N} \#\left\{n < N : S(n) \equiv 0 \pmod m\right\} = 1. \tag{3.3}$$

We call integer sequences of the form as in (3.1) — that is, integer sequences given by a product/quotient of factorials multiplied by a rational function — "Catalan-like" since the Catalan numbers $\frac{1}{n+1}\binom{2n}{n}$ represent obviously such a sequence, but as well many other sequences that one finds in the literature (and in the On-Line Encyclopedia of Integer Sequences [15]).

Examples. All of the following sequences are "Catalan-like" in the sense of (3.1).

(1) *Binomial coefficients* such as the *central binomial coefficients* $\binom{2n}{n}$, or more generally $\binom{(a+b)n}{an}$ for positive integers a and b, including variations such as $\binom{2n}{n-1}$, etc.

(2) *Multinomial coefficients* such as $\frac{((a_1+a_2+\cdots+a_s)n)!}{(a_1n)!\,(a_2n)!\cdots(a_sn)!}$, etc.

(3) *Fuß–Catalan numbers.* These are defined by (cf. [1, pp. 59–60]) $\frac{1}{n}\binom{(m+1)n}{n-1}$, where m is a given positive integer.

(4) Gessel's [7] *super ballot numbers* (often also called *super-Catalan numbers*) $\frac{(2n)!\,(2m)!}{n!\,m!\,(m+n)!}$ for non-negative integers m, or for $m = an$ with a a positive integer.

(5) Many counting sequences in *tree* and *map enumeration* (cf. [12] for a survey) such as $\frac{m+1}{n((m-1)n+2)}\binom{mn}{n-1}$ (m-ary blossom trees with n white nodes; cf. [11, Sec. 3]), $\frac{2\cdot 3^n}{(n+2)(n+1)}\binom{2n}{n}$ (number of rooted planar maps with n edges; cf. [20]), $\frac{2}{(3n-1)(3n-2)}\binom{3n-1}{n}$ (number of rooted non-separable planar maps with n edges; cf. [4]), $\frac{2}{(3n+1)(n+1)}\binom{4n+1}{n}$ (number of rooted planar triangulations with $n+3$ vertices; cf. [18]), $\frac{1}{2(n+2)(n+1)}\binom{2n}{n}\binom{2n+2}{n+1}$ (number of rooted Hamiltonian maps with $2n$ vertices; cf. [19]), to mention just a few.

What Theorem 4 says is that, for any of these sequences, most elements (in the sense that the proportion of those in the set of all elements tends to 1) are divisible by p^α, for a given prime number p and given positive integer α.

We should at this point remind the reader of *Landau's criterion* [9], which says that

$$U(n) := \frac{\prod_{i=1}^{r}(C_i n)!}{\prod_{i=1}^{s}(D_i n)!}$$

is an integer for all n if and only if

$$\sum_{i=1}^{r} \lfloor C_i x \rfloor - \sum_{i=1}^{s} \lfloor D_i x \rfloor \geq 0 \tag{3.4}$$

for all real numbers x. (Here, we still assume that $\sum_{i=1}^{r} C_i = \sum_{i=1}^{s} D_i$.) In view of [3, Lemma 3.3], which says that if $U(n)$ is non-integral for some n then, for almost all primes p, there exists an n such that $v_p\big(U(n)\big) < 0$, this means (more or less; we do not believe that the polynomial $P(n)$ can "correct" non-integrality of $U(n)$ for all n) that (3.4) is an implicit assumption in Theorem 4.

For the proof of Theorem 4, we consider the integer interval $[0, N-1]$ as a probability space, with each integer equally likely, precisely as in Section 2. For notational convenience, the corresponding probability function will be denoted by $\mathbf{P}_N$. Functions on the non-negative integers are then interpreted as random variables X on this space by restricting them to $[0, N-1]$. The expectation of X on the space, that is, $\frac{1}{N}\sum_{i=0}^{N-1} X(i)$, will be denoted by $\mathbf{E}_N(X)$, the variance will be denoted by $\mathbf{Var}_N(X)$, and the covariance of two functions by $\mathbf{Cov}_N(X,Y)$.

We need two auxiliary lemmas. The first concerns asymptotic mean and variance for the p-adic valuation of a linear function.

Lemma 6. *Let E and F be integers, $E > 0$, not both divisible by the prime p. If $v_p(En + B)$ is considered as a random variable for n in the integer interval $[0, N-1]$, then*

$$\mathbf{E}_N\big(v_p(En+F)\big) = \begin{cases} 0, & \text{if } p \mid E, \\ \frac{1}{p-1} + o(1), & \text{if } p \nmid E, \end{cases} \qquad \text{as } N \to \infty, \tag{3.5}$$

and

$$\mathbf{Var}_N\big(v_p(En+F)\big) = \begin{cases} 0, & \text{if } p \mid E, \\ \frac{p}{(p-1)^2} + o(1), & \text{if } p \nmid E, \end{cases} \qquad \text{as } N \to \infty. \tag{3.6}$$

Proof. The first assertion in the case distinction in (3.5) is obvious since our assumptions imply that $En + F \not\equiv 0 \pmod p$ if $p \mid E$. If $p \nmid E$, then the congruence $En + F \equiv 0 \pmod{p^\alpha}$ has a unique solution for n modulo p^α for any given positive integer α. Thus, we have

$$\mathbf{E}_N\big(v_p(an+b)\big) = \frac{1}{N} \sum_{\ell=1}^{\lfloor \log_p N \rfloor} \left(\frac{N}{p^\ell} + O(1)\right) = \frac{1}{p-1} + o(1), \qquad \text{as } N \to \infty.$$

Similarly, still assuming $p \nmid E$, for the variance we have

$$\begin{aligned} \mathbf{Var}_N\big(v_p(an+b)\big) &= \frac{1}{N} \sum_{\ell=1}^{\lfloor \log_p N \rfloor} (2\ell - 1)\left(\frac{N}{p^\ell} + O(1)\right) - \Big(\mathbf{E}_N\big(v_p(an+b)\big)\Big)^2 \\ &= \frac{p+1}{(p-1)^2} - \frac{1}{(p-1)^2} + o(1), \qquad \text{as } N \to \infty, \end{aligned}$$

establishing also (3.6). □

The second auxiliary lemma provides an asymptotic upper bound on the covariance of a linear function and the sum-of-digits function of a linear function.

Lemma 7. *Let C, E, and F be integers, $C, E > 0$, and E and F not both divisible by p. If $s_p(Cn)$ and $v_p(En + B)$ are considered as random variables for n in the integer interval $[0, N-1]$, then*

$$\mathbf{Cov}_N(s_p(Cn), v_p(En + F)) = O\left(\log_p^{1/2}(N)\right), \qquad \text{as } N \to \infty. \tag{3.7}$$

Proof. By the Cauchy–Schwarz inequality in probabilistic setting, we have

$$\mathbf{Cov}_N\Big(s_p(Cn), v_p\big(En + F\big)\Big) \le \mathbf{Var}_N\big(s_p(Cn)\big)^{1/2}\,\mathbf{Var}_N\Big(v_p\big(En + F\big)\Big)^{1/2}.$$

The variance of $s_p(Cn) = s_p(Cp^{-v_p(C)}n)$ has been (implicitly) given in (2.1) (see the line below that equation; see also (2.4) with $q = p$ and $A_i = A_j = Cp^{-v_p(C)}$) and turned out to be of the order $\log_p(N)$, while the variance of $v_p\big(En+F\big)$ has been given in (3.6) and turned out to be bounded. The assertion of the lemma is hence obvious. □

Proof of Theorem 4. With $S(n)$ given in (3.1), we have

$$\begin{aligned} v_p\big(S(n)\big) &= v_p\big(P(n)\big) - \sum_{i=1}^{t} v_p(E_in + F_i) + \sum_{i=1}^{r} v_p\big((C_in)!\big) - \sum_{i=1}^{s} v_p\big((D_in)!\big) \\ &\ge -\sum_{i=1}^{t} v_p(E_in + F_i) - \frac{1}{p-1}\sum_{i=1}^{r} s_p(C_in) + \frac{1}{p-1}\sum_{i=1}^{s} s_p(D_in). \end{aligned} \tag{3.8}$$

Here, we used Legendre's formula (1.1) and the assumption that $\sum_{i=1}^{r} C_i = \sum_{i=1}^{s} D_i$.

Now, it follows from [3, Lemma 3.5 and its proof], that under the integrality and non-triviality assumption for $S(n)$ of the theorem, we have $r < s$.

The expression on the right-hand side of (3.8) is a linear combination of the functions $v_p(E_in + F_i)$, $s_p(C_in)$, and $s_p(D_in)$, which we view again as random variables on $[0, N-1]$. For convenience, let us denote the function on the right-hand side of (3.8) by $T(n)$. By Theorem 1 and (3.5), we have

$$\mathbf{E}_N\big(T(n)\big) = \Omega\big(\log_p(N)\big), \qquad \text{as } N \to \infty.$$

The reader should observe that the inequality $r < s$ is used here crucially. On the other hand, the variance of $T(n)$ is bounded above by the sum of the pairwise covariances of the involved random variables (functions). By Theorem 1, (3.6), and (3.7), we see that

$$\mathbf{Var}_N\big(T(n)\big) = O\big(\log_p(N)\big), \qquad \text{as } N \to \infty.$$

Given a random variable X, Chebyshev's inequality reads

$$\mathbf{P}\big(|X - \mathbf{E}(X)| < \varepsilon\big) > 1 - \frac{1}{\varepsilon^2}V(X). \tag{3.9}$$

Choosing $\varepsilon = \big(\log_p(n)\big)^{3/4}$ and $X = T(n)$, we get

$$\mathbf{P}_N\Big(\big|T(n) - \mathbf{E}_N\big(T(n)\big)\big| < \log_p^{3/4}(N)\Big) = 1 + O\left(\log_p^{-1/2}(N)\right), \qquad \text{as } N \to \infty.$$

Thus, for N large enough, we have

$$T(n) > \mathbf{E}_N\big(T(n)\big) - \log_p^{3/4}(N) = \Omega\big(\log_p(N)\big) > \alpha,$$

with probability $1 + O\big(\log_p^{-1/2}(N)\big)$. If we use this information in (3.8), then the assertion of the theorem follows immediately. □

References

[1] D. Armstrong, *Generalized Noncrossing Partitions and Combinatorics of Coxeter Groups*, Mem. Amer. Math. Soc., vol. 202, no. 949, Amer. Math. Soc., Providence, R.I., 2009.

[2] N. L. Bassily and I. Kátai, *Distribution of the values of q-additive functions on polynomial sequences*, Acta Math. Hung. **68** (1995), 353–361.

[3] J. W. Bober, *Factorial ratios, hypergeometric series, and a family of step functions*, J. London Math. Soc. **79** (2009), 422–444.

[4] W. G. Brown, *Enumeration of non-separable planar maps*, Can. J. Math. **15** (1963), 526–545.

[5] R. Burns, *Structure and asymptotics for Catalan numbers modulo primes using automata*, preprint, arχiv:1701.02975.

[6] C. Dartyge and G. Tenenbaum, *Sommes des chiffres de multiples d???entiers*, Ann. Inst. Fourier **55**(7) (2005), 2423–2474.

[7] I. M. Gessel, *Super ballot numbers*, J. Symbolic Comput. **14** (1992), 179–194.

[8] S. Kropf and S. Wagner, *On q-quasiadditive and q-quasimultiplicative functions*, Electron. J. Combin. **24**(1) (2017), Article 1.60, 22 pp.

[9] E. Landau, *Sur les conditions de divisibilité d'un produit de factorielles par un autre*, Nouv. Ann. Math. **19** (1900), 344–362.

[10] A. M. Legendre, *Essai sur la théorie des nombres*, 2ed., Courcier, Paris, 1808.

[11] G. Schaeffer, A combinatorial interpretation of super-Catalan numbers of order two, unpublished manuscript, 2003; available at http://www.lix.polytechnique.fr/Labo/Gilles.Schaeffer/Biblio/Sc03.ps.

[12] G. Schaeffer, *Planar maps*, in: Handbook of Enumerative Combinatorics, M. Bóna (ed.), CRC Press, Boca Raton, London, New York, 2015, pp. 335–395. 589–678.

[13] J. Schmid, *The joint distribution of the binary digits of integer multiples*, Acta Arith. **43** (1984), 391–415.

[14] W. M. Schmidt, *The joint distribution of the digits of certain integer s-tuples*, Studies in Pure Mathematics, Mem. of P. Turán (1983), 605–622.

[15] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, http://oeis.org/.

[16] L. Spiegelhofer and M. Wallner, *Divisibility of binomial coefficients by powers of two*, preprint, arχiv:1710.10884.

[17] R. P. Stanley, *Enumerative Combinatorics*, vol. 2, Cambridge University Press, Cambridge, 1999.

[18] W. T. Tutte, *A census of planar triangulations*, Canad. J. Math. **14** (1962), 21–38.

[19] W. T. Tutte, *A census of Hamiltonian polygons*, Canad. J. Math. **14** (1962), 402–417.

[20] W. T. Tutte, *A census of planar maps*, Canad. J. Math. **15** (1963), 249–271.

*Institute of Discrete Mathematics and Geometry, TU Wien, Wiedner Hauptstrasse 8–10, A-1040 Vienna, Austria. WWW: https://www.dmg.tuwien.ac.at/drmota.

†Fakultät für Mathematik, Universität Wien, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria. WWW: http://www.mat.univie.ac.at/~kratt.