Contents

1 Kolmogorov and number theory **3**

1.1 Introduction 3

1.2 New estimates for uniform order statistics 6

1.3 Number theory applications 9

Chapter 1

From Kolmogorov's theorem on empirical distribution to number theory

By ***Kevin Ford***

We describe some new estimates for the probability that an empirical distribution function of uniform-[0,1] random variables stays on one side of a given line, and give applications to number theory.

1.1 Introduction

Let $X_1, \dots, X_n$ be real-valued independent random variables, each with distribution function $F(t)$. Let

$$F_n(t) = \frac{1}{n} \#\{i : X_i \le t\}$$

be the corresponding empirical distribution function. For n, t fixed, $F_n(t)$ is a random variable. Applying the strong law of large numbers to the Bernoulli variables

$$\mathbf{1}_{\{X_n \le t\}} \quad (= 1 \text{ if } X_n \le t, 0 \quad \text{otherwise}),$$

we see that $F_n(t) \underset{n\to\infty}{\longrightarrow} F(t)$ almost surely. In 1933, Glivenko [Gli33] and (slightly later) Cantelli [Can33] proved that the convergence is uniform on the real line : $\sup \mid F_n(t) - F(t) \mid \underset{n\to\infty}{\longrightarrow} 0$ almost surely. Immediately, in his seminal paper [Kol33], Kolmogorov made a careful study of the convergence of $F_n(t)$ to $F(t)$ as $n \to \infty$: he showed that if F is continuous, then for each $\lambda > 0$, the probability $\mathbf{P}(\sup |F_n(t) - F(t)| < \lambda/\sqrt{n})$ is independent of F, and that

$$\mathbf{P}(\sup |F_n(t) - F(t)| < \lambda/\sqrt{n}) \to \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2} \quad (n \to \infty) \tag{1.1}$$

uniformly in λ.

The three papers of Glivenko, Kolmogorov and Cantelli appeared (in this order) in the same issue of the *Giornale dell Istituto Italiano degli Attuari*, all in Italian, and with almost the same title. The paper [Kol33] of Kolmogorov also appears in his Selected Works ([KolW], p. 139-146; comments p. 574-583).

Six years later, Smirnov [Smi39] studied the corresponding one-sided bounds, showing for $\lambda \geq 0$ that

$$\mathbf{P}(\sup(F_n(t) - F(t)) < \lambda/\sqrt{n}) \to 1 - e^{-2\lambda^2} \quad (n \to \infty). \tag{1.2}$$

Together, (1.1) and (1.2) form the basis for the well-known *Kolmogorov-Smirnov* goodness-of-fit tests.[(1)]

It is sometimes convenient to express probabilities of the above type in terms of the "order statistics" of $X_1, \ldots, X_n$, which is the increasing sequence $\xi_1 \leq \cdots \leq \xi_n$ obtained by ordering (each realization of) $X_1, \ldots, X_n$.

From now on, we will consider uniform distribution on $[0, 1]$, that is

$$F(t) = \begin{cases} 0 & t \leq 0 \\ t & 0 < t < 1 \\ 1 & t \geq 1. \end{cases} \tag{1.3}$$

In this case, the numbers $\xi_1, \ldots, \xi_n$ are called *uniform order statistics*. In this note, we are interested in the behavior of

$$Q_n(u, v) = \mathbf{P}\left(\forall\, i \in \{1, \ldots, n\} : \xi_i \geq \frac{i - u}{v}\right).$$

In this notation, Smirnov's theorem reads[(2)] $Q_n(\lambda\sqrt{n}, n) \to 1 - e^{-2\lambda^2}$.

Refinements to (1.2) were given later in the range $\lambda_0 \leq \lambda = O(n^{1/6})$ for a *fixed* positive λ_0 (e.g. Smirnov [Smi44], Lauwerier [Lau63]; see also Ch. 9 of [SW86]), in particular

$$Q_n(\lambda\sqrt{n}) = 1 - e^{-2\lambda^2}\left(1 - \frac{2\lambda}{3n^{1/2}} + O\left(\frac{\lambda^4 + 1}{n}\right)\right). \tag{1.4}$$

[1] Notice that applying the Central Limit Theorem to the Bernoulli variables $\mathbf{1}_{\{X_n \leq t\}}$, we have only

$$\mathbf{P}(|F_n(t) - F(t)| < \lambda/\sqrt{n}) \to \frac{1}{2\pi}\int_{-\lambda/\sigma(t)}^{\lambda/\sigma(t)} e^{-s^2/2} ds,$$

with $\sigma(t) = \sqrt{F(t)(1 - F(t))}$. In Kolmogorov's theorem, $|F_n(t) - F(t)|$ is replaced by its supremum over t, and the limit in the right-hand side is a universal (independent of F) function, of which Kolmogorov gave the first table of values.

[2] Notice that

$$F_n(t) = \begin{cases} 0 & t \in (-\infty, \xi_1) \\ i/n & t \in [\xi_i, \xi_{i+1}) \quad (1 \leq i \leq n - 1) \\ 1 & t \in [\xi_n, +\infty) \end{cases}$$

thus we see (with (1.3)) that

$$\mathbf{P}(\sup(F_n(t) - F(t)) < \lambda/\sqrt{n}) = \mathbf{P}\left(\max_i(\frac{i}{n} - \xi_i) < \lambda/\sqrt{n}\right) = Q_n(\lambda\sqrt{n}, n).$$

Let $w = u + v - n$. Trivially $Q_n(u, v) = 0$ when $w \le 0$ and $Q_n(u, v) = 1$ when $u \ge n$ (recall that $0 \le X_i \le 1$ from the choice of F). If $u \le 1$ and $w > 0$, the exact formula $Q_n(u, v) = \dfrac{w}{v}(1 + u/v)^{n-1}$ was found by Daniels [Dan45]. Estimating $Q_n(u, v)$ when $u > 1$ is much more difficult, however there is an exact formula

$$\begin{aligned} Q_n(u, v) &= \frac{w}{v^n} \sum_{0 \le j < u} \binom{n}{j} (w + n - j)^{n-j-1} (j - u)^j \\ &= 1 - \frac{w}{v^n} \sum_{u < j \le n} \binom{n}{j} (w + n - j)^{n-j-1} (j - u)^j. \end{aligned} \tag{1.5}$$

The special case $v = n$ of (1.5) is due to Smirnov [Smi44], and the general case is due to Pyke [Pyk59]. The equivalence of the two expressions for $Q_n(u, v)$ follows from one of Abel's identities ([Rio68], p. 18, (13a)). The first is more convenient when u is very small and fixed, while the second is more convenient for larger u because all summands are positive.

Smirnov [Smi44] estimated $Q_n(\lambda\sqrt{n}, n)$ using (1.5) and Stirling's formula for $k!$, and Csáki [Csa74] used similar methods to show

$$Q_n(\alpha\sqrt{n}, n + (\beta - \alpha)\sqrt{n}) \to 1 - e^{-2\alpha\beta} \qquad (n \to \infty).$$

for fixed $\alpha \ge 0$, $\beta \ge 0$. Lauwerier [Lau63] and Penkov [Pen76], by contrast, started with (1.5) and used complex analytic methods to approximate $Q_n(\lambda\sqrt{n})$. Yet another approach is based on what are called "almost sure invariance principles" or "strong approximation theorems" ([CR81], [Phi86]). The strong Komlós-Major-Tusnády theorem [KMT75] implies

$$|F_n(t) - t - n^{-1/2} B_n(t)| \ll \frac{\log n}{n} \qquad (0 \le t \le 1)$$

with probability $\ge 1 - O(1/n)$, where $B_n(t)$ is a Brownian bridge process. The order $\frac{\log n}{n}$ on the right side is also best possible [KMT75] (see also Ch. 4 of [CR81]). Since

$$\mathbf{P}\left(\sup_{0 \le t \le 1} (B_n(t) - (at + b)) \le 0 \right) = 1 - e^{-2b(a+b)},$$

the KMT theorem implies the uniform estimate

$$\begin{aligned} Q_n(u, v) &= O\left(\frac{1}{n}\right) + 1 - e^{-\frac{2(u + O(\log n))(w + O(\log n))}{n}} \\ &= 1 - e^{-2uw/n} + O\left(\frac{(u + w + \log n) \log n}{n} \right). \end{aligned} \tag{1.6}$$

This gives an asymptotic for $Q_n(u, v)$ in a wide range of u and w, but requiring $\frac{u}{\log n} \to \infty$ and $\frac{w}{\log n} \to \infty$.

For the application to number theory in [For04], we need sharper uniform bounds than (1.6). In particular, we need the bound $Q_n(u, v) = O(u/n)$ uniformly for $n \ge 1$, $w = O(1)$ and $1 \le u \le n$.

1.2 New estimates for uniform order statistics

Theorem 1.2.1. *Uniformly in $u > 0$, $w > 0$ and $n \geq 1$, we have*

$$Q_n(u,v) = 1 - e^{-2uw/n} + O\left(\frac{u+w}{n}\right),$$

i.e. $|O(\frac{u+w}{n})| \leq \mathrm{const}(\frac{u+w}{n})$ *where the constant is independent of* u, v, n.

In addition we have the following useful approximation.

Corollary 1.2.2. *Uniformly in $u > 0$, $w > 0$ and $n \geq 1$, we have*

$$Q_n(u,v) = \frac{2uw}{n}\left(1 + O\left(\frac{1}{u} + \frac{1}{w} + \frac{uw}{n}\right)\right).$$

In particular, when $uw/n \to 0$, $u \to \infty$ and $w \to \infty$ as $n \to \infty$, we see that $Q_n(u,v)$ is asymptotic to $2uw/n$. Starting with (1.5), a complicated modification of the complex analytic method of Lauwerier [Lau63] can be used to prove Theorem 1.2.1. This was carried out in the original version of [For04], and a sketch of the argument appears in [For04a].

Here we outline a new method based on the theory of random walks, full details of which appear in [For06]. Rather than work with (1.5), we reinterpret $Q_n(u,v)$ in terms of a random walk. Let $Y_1, \cdots, Y_{n+1}$ be independent random variables with exponential distribution, and let $W_k = Y_1 + \cdots + Y_k$ for $1 \leq k \leq n+1$. By a well-known theorem of Rényi [Ren53], the vectors $(\xi_1, \ldots, \xi_n)$ and $(W_1/W_{n+1}, \ldots, W_n/W_{n+1})$ have identical distributions. Similarly, given that $W_{n+1} = v$, the probability density function of the vector $(W_1/v, \ldots, W_n/v)$ is identically $n!$ on the set $\{(x_1, \ldots, x_n) : 0 \leq x_1 \leq \cdots \leq x_n \leq 1\}$. Therefore,

$$Q_n(u,v) = \mathbf{P}[\min_{1 \leq i \leq n}(W_i - i) \geq -u \mid W_{n+1} = v].$$

Put $X_i = 1 - Y_i$, so that the X_i have mean 0, variance 1 and $X_i < 1$ for all i. Let

$$S_i = X_1 + \cdots + X_i, \quad T_i = \max(0, S_1, \ldots, S_i) \qquad (i \geq 0).$$

The sequence $0, S_1, S_2, \ldots$ can be thought of as a recurrent random walk on the real line, with T_i measuring the farthest extent to the right that the walk has achieved during the first i steps. Setting

$$R_m(x,y) = \mathbf{P}[T_{m-1} < y \mid S_m = x],$$

we have

$$Q_n(u,v) = R_{n+1}(n+1-v, u). \tag{1.7}$$

If we label the point y as a barrier, then $R_m(x,y)$ is the probability of stopping after m steps at x without crossing the barrier.

In proving (1.1) in [Kol33], Kolmogorov used a relation similar to (1.7). Specifically, let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with discrete distribution

$$\mathbf{P}[Y_j = r-1] = \frac{e^{-1}}{r!} \qquad (r = 0, 1, \ldots)$$

and let $Z_j = Y_1 + \cdots + Y_j$ for $j \geq 1$. The variables Y_i have mean 0 and variance 1. Kolmogorov proved that for integers $u \geq 1$,

$$\begin{aligned}\mathbf{P}(\sup |F_n(t) - F(t)| \leq u/n) &= \frac{n!e^n}{n^n}\mathbf{P}\left(\max_{0\leq j\leq n-1} |Z_j| < u, Z_n = 0\right) \\ &= \mathbf{P}\left(\max_{0\leq j\leq n-1} |Z_j| < u \,|Z_n = 0\right).\end{aligned}$$

Small modifications to the proof yield, for *integers* $u \geq 1$ and for $n \geq 2$, that

$$Q_n(u,n) = \mathbf{P}\left(\max_{0\leq j\leq n-1} Z_j < u \,|Z_n = 0\right).$$

Let f_m be the pdf for S_m ($m = 1, 2, \ldots$). The Central Limit Theorem for densities (e.g., Theorem 1 in §46 of [GK68]) implies that for large m and $|x| \ll \sqrt{m}$, $f_m(x) \approx (2\pi m)^{-1/2}e^{-x^2/2m}$. However, there are asymmetries in the distribution for $|x| > \sqrt{m}$, which can be seen using the exact formula

$$f_m(x) = \begin{cases} \frac{(m-x)^{m-1}}{e^{m-x}(m-1)!} & x \leq m \\ 0 & x > m, \end{cases} \tag{1.8}$$

easily proved by induction on m.

Our principal tool for estimating $R_n(x, y)$ is a reccurrence formula based on the reflection principle for random walks. Suppose $y \geq 0$ and $y \geq x$. By reflecting about the point y that part of the walk beyond the first crossing of y, a recurrent random walk of n steps that crosses the point y and ends at the point x is about as likely as a random walk which ends at $2y - x$ after n steps. This of course is inexact, since the steps of a random walk may not be symmetric and the walk may not hit y exactly. The next lemma gives a precise measure of the accuracy of the reflection principle for our specific walk. For convenience, define

$$\widetilde{R}_n(x, y) = f_n(x)R_n(x, y) = \mathbf{D}[T_{n-1} < y, S_n = x],$$

where the last expression stands for $\frac{d}{dx}\mathbf{P}[T_{n-1} < y, S_n < x]$. From the reflection principle we expect that $\widetilde{R}_n(x, y) \approx f_n(x) - f_n(2y - x)$.

Lemma 1.2.3. *For a positive integer $n \geq 2$, real $y > 0$, real x, and real $a \geq 1$,*

$$\widetilde{R}_n(x, y) = f_n(x) - f_n(y+a) + \int_0^1 \sum_{k=1}^{n-1} \widetilde{R}_k(y+\xi, y)\left(f_{n-k}(a - \xi) - f_{n-k}(x - y - \xi)\right) d\xi. \tag{1.9}$$

Proof. Start with

$$\widetilde{R}_n(x, y) = f_n(x) - f_n(y + a) + f_n(y + a) - \mathbf{D}[T_{n-1} \geq y, S_n = x].$$

If $S_n = y + a$, then there is a unique k, $1 \le k \le n-1$, so that $T_{k-1} < y$ and $S_k \ge y$. Thus,

$$\begin{aligned} f_n(y+a) &= \sum_{k=1}^{n-1} \mathbf{D}[T_{k-1} < y, S_k \ge y, S_n = y+a] \\ &= \sum_{k=1}^{n-1} \int_0^1 \mathbf{D}[T_{k-1} < y, S_k = y+\xi, S_n = y+a]\, d\xi \\ &= \sum_{k=1}^{n-1} \int_0^1 \widetilde{R}_k(y+\xi, y) f_{n-k}(a-\xi)\, d\xi. \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbf{D}[T_{n-1} \ge y, S_n = x] &= \sum_{k=1}^{n-1} \mathbf{D}[T_{k-1} < y, S_k \ge y, S_n = x] \\ &= \sum_{k=1}^{n-1} \int_0^1 \widetilde{R}_k(y+\xi, y) f_{n-k}(x-y-\xi)\, d\xi. \end{aligned}$$

□

In Lemma 1.2.3, we choose $a = y - x - b(n, y-x)$, where $b = b(n,z)$ is the unique solution of $f_n(-z) = f_n(z-b)$ ith $-2 \le b \le z-1$ ($b(n,z)$ exists and is unique since $f_n(x)$ is unimodular with maximum at $x = 1$). This makes $|f_{n-k}(a-\xi) - f_{n-k}(x-y-\xi)|$ small, at least when k is small. Also, $\widetilde{R}_k(y+\xi, y)$ should be small, since it measures the liklihood of a walk staying to the left of y for $n-1$ steps and jumping over y on the n-th step. Suppose $n \ge 10$, $0 \le y \le \frac{n}{10}$, and $y \le x \le y+1$. We have $f_n(1+x) \le f_n(1-x)$ for $x \ge 0$, thus when $0 \le \xi \le 1$ and $1 \le j \le n-1$, $f_j(5-\xi) \le f_j(x-y-\xi)$. By Lemma 1.2.3 with $a = 5$,

$$\widetilde{R}_n(x,y) \le f_n(x) - f_n(y+5) = \int_x^{y+5} \frac{t-1}{n-t} f_n(t)\, dt \ll \frac{(y+1) f_n(y)}{n}.$$

Together with estimates for $|f_{n-k}(a-\xi) - f_{n-k}(x-y-\xi)|$ obtained from (1.8), the integral-sum on the right of (1.9) can be shown to be small. We conclude that, with small error,

$$R_n(x,y) \approx 1 - \frac{f_n(2y-x-b)}{f_n(x)}.$$

The desired asymptotic for $Q_n(u,v)$ now follows from (1.8) and the asymptotic $b = b(n,z) = -2 + O(\frac{(z+1)^2}{n-1})$.

We note that when the steps in a recurrent random walk have an arbitrary continuous or lattice distribution, one can define a quantity analogous to $R_n(x,y)$. The same argument provides an analogous formula to (1.9) and an analog of Theorem 1.2.1, namely

$$R_m(y-z, y) = 1 - e^{-2yz/n} + O\left(\frac{y+z+1}{n}\right) \quad (0 \le y \ll \sqrt{n}, 0 \le z \ll \sqrt{n}),$$

can be shown to hold for a very general class of distributions (see [Ford06a]).

1.3 Number theory applications

Hardy and Ramanujan initiated the study of the statistical distribution of the prime factors of integers in their ground-breaking 1917 paper [HR17], and much work has been done on this topic since then. Write an arbitrary integer $n = p_1 p_2 \cdots p_k$, where the p_i are primes and $p_1 \le \cdots \le p_k$. Roughly speaking, the quantities $g_j = \log\log p_{j+1} - \log\log p_j$ behave like independent exponentially distributed random variables. Of course the g_j have discrete distributions, but the distributions approach the exponential distribution as $j \to \infty$. It is well-known that a typical integer n has about $\log\log b - \log\log a$ prime factors in an interval $(a, b]$ (see e.g. Ch. 1 of [HT88]), and the probability that n has at least one prime factor in $(a, b]$ is approximately[3]

$$1 - \prod_{a<p\le b} (1 - 1/p) = 1 - \frac{\log a + O(1)}{\log b}.$$

One can also consider integers with a fixed number of prime factors and examine the statistics

$$(\xi_1, \cdots, \xi_m), \qquad \xi_i = \frac{\log\log p_{j+i} - \log\log p_j}{\log\log p_k - \log\log p_j}, \quad m = k - 1 - j.$$

With k and j fixed, the numbers $\xi_1, \ldots, \xi_m$ behave much like uniform order statistics. This means that for "nice" functions $f : [0,1]^m \to \mathbb{R}$, the average of $f(\xi_1, \ldots, \xi_m)$ over n which are the product of k primes is about

$$m! \int\limits_{0\le x_1 \le \cdots \le x_m \le 1} f(x_1, \ldots, x_m)\, dx_1 \cdots dx_m.$$

The approximation gets better as $j \to \infty$.

These phenomena can be explained by considering the following "model" of the integers (known as the Kubilius model). Let $\{X_p : p \text{ prime}\}$ be independent Bernoulli random variables so that $\mathbf{P}(X_p = 0) = 1 - \frac{1}{p}$ and $\mathbf{P}(X_p = 1) = \frac{1}{p}$. Thus X_p models the event that a random integer is divisible by p. By an elementary estimate,

$$\sum_{a<p\le b} \mathbf{E}(X_p) = \sum_{a<p\le b} \frac{1}{p} = \log\log b - \log\log a + O(1/\log a).$$

(The $\log\log$, rather than $\log$, are due to the fact that we sum only on primes.) For more about probabilistic number theory, the reader may consult the excellent monographs of Elliott [Ell79].

Questions about the distribution of all divisors of integers are much more difficult, since the corresponding random variables $\{X_d : d \ge 1\}$ are not at all independent (e.g., $X_6 = 1 \implies X_3 = 1$). Consider the problem of estimating

[3] p will always denote a prime number; $\prod_{a<p\le b}$ will be a product on primes, $\sum_{a<p\le b}$ a sum on primes.

$\varepsilon(y, z)$, the probability that a random integer has a divisor d satisfying $y < d \le z$. More precisely,

$$\varepsilon(y, z) = \lim_{x\to\infty} \frac{\#\{n \le x : \exists\, d|n, y < d \le z\}}{x}.$$

Similarly, let $\varepsilon_r(y, z)$ be the probability that a random integer has exactly r divisors in the interval $(y, z]$. Interest in bounding $\varepsilon(y, z)$ began in the 1930s with a paper by Besicovitch [Bes34], who proved that $\liminf_{y\to\infty} \varepsilon(y, 2y) = 0$. A year later, Erdős [Erd35] improved this to $\lim_{y\to\infty} \varepsilon(y, 2y) = 0$. Later work, especially by Erdős [Erd36], [Erd60] and Tenenbaum [Ten84], focused on determining the rate at which $\varepsilon(y, 2y) \to 0$ and on bounding $\varepsilon(y, z)$ for more general y, z. Chapter 2 of the book [HT88] contains a thorough exposition on such bounds and their applications. The main theorem of [For04] is a determination of the order of magnitude of $\varepsilon(y, z)$ for all y, z; that is, bounding $\varepsilon(y, z)$ between two constant multiples of a smooth function of y, z. In particular, we show that for some positive constants c_1 and c_2,

$$\frac{c_1}{(\log y)^{\delta}(\log\log y)^{3/2}} \le \varepsilon(y, 2y) \le \frac{c_2}{(\log y)^{\delta}(\log\log y)^{3/2}}, \tag{1.10}$$

where $\delta = 1 - \dfrac{1 + \log\log 2}{\log 2} = 0.08607\ldots$. A relatively short, complete proof of this special case is given in [For06b].

Concerning the behavior of $\varepsilon_r(y, z)$, Erdős conjectured in [Erd60] that

$$\lim_{y\to\infty} \frac{\varepsilon_1(y, 2y)}{\varepsilon(y, 2y)} = 0.$$

The ratio $\dfrac{\varepsilon_r(y, z)}{\varepsilon(y, z)}$ can be considered as the conditional probability that a random integer contains exactly r divisors in $(y, z]$ given that it has at least one such divisor. In [Ten87] a lower bound $\dfrac{\varepsilon_r(y, 2y)}{\varepsilon(y, 2y)} \ge c_3 f(y)$ was given, where $f(y) \to 0$ very slowly as $y \to \infty$. Erdős conjecture is disproved in [For04], where the order of $\varepsilon_r(y, z)$ is determined for a wide range of y, z. In particular, for any $r \ge 1$ and any constant $c > 1$,

$$\liminf_{y\to\infty} \frac{\varepsilon_r(y, cy)}{\varepsilon(y, cy)} > 0.$$

Also,

$$\frac{\varepsilon_r(y, z)}{\varepsilon(y, z)} \to 0 \qquad (z/y \to \infty),$$

confirming a conjecture of Tenenbaum [Ten87].

We now say a few words about the proofs. Let m be the product of the distinct prime factors of n which are $\le y$. First, $\varepsilon(y, 2y)$ can be estimated in terms of

$$\sum_m \frac{L(m)}{m}, \qquad L(m) = \mu\{u : \exists\, d|m, e^u < d \le 2e^u\},$$

where μ denotes Lebesgue measure. The quantity $L(m)$ is a kind of measure of the global distribution of the divisors of m. If $m = p_1 \cdots p_k$, then

$$L(m) \le \min_{0 \le h \le k} 2^{k-h} \log(2p_1 \cdots p_h).$$

Most of the time, we expect $\log(2p_1 \cdots p_h) = O(\log p_h)$, so

$$L(m) \approx O\left(2^k \exp\{\min_{1 \le h \le k}(-h \log 2 + \log\log p_h)\}\right).$$

Putting $\xi_i = \dfrac{\log\log p_i}{\log\log y}$, then $\xi_1, \ldots, \xi_k$ behave much like uniform order statistics. Thus, upper bounds for averages of $L(m)$ depend on the size of $Q_k(u,v)$ with $v = \dfrac{\log\log y}{\log 2}$. Utilizing Theorem 1.2.1 (actually, the weaker bound $Q_n(u,v) = O(\frac{(u+1)(w+1)^2}{n})$ proved in [For04] suffices) leads to the upper bound in (1.10). Furthermore, the bulk of the contribution comes from numbers n with $k = \dfrac{\log\log y}{\log 2} + O(1)$. This implies that most integers which have a divisor in $(y, 2y]$ have about $\dfrac{\log\log y}{\log 2}$ prime factors $\le y$. By contrast, most integers n have about $\log\log y$ prime factors $\le y$.

Acknowledgement

This research was supported by National Science Foundation grants DMS-0301083 and DMS-0555367.

Bibliography

[Bes34] Besicovitch, A.S.: On the density of certain sequences of integers. Math. Ann., **110**, 336–341 (1934)

[Can33] Cantelli, F.G.: Sulla determinazione empirica delle leggi di probabilità. Giorn. Ist. Ital. Attuari, **4**, 421–424 (1933)

[CR81] Csörgő, M., Révész, P.: Strong Approximations in probability and statistics. Academic Press (1981)

[Csa74] Csáki, E.: On tests based on empirical distribution functions (Hungarian). Magyar Tud. Akad. Mat. Fiz. Oszt. Közl., **23**, 239–327 (1977). English translation in: Leifman, L.J. (ed) Selected translations in mathematical statistics and probability, **15**, Amer. Math. Soc., 229–317 (1981)

[Dan45] Daniels, H.E.: The statistical theory of the strength of bundles of threads, I. Proc. Roy. Soc. London. Ser. A., **183**, 405–435 (1945)

[Ell79] Elliott, P.D.T.A.: Probabilistic number theory I, II. Grund. Math. Wissen. 239, 240. Springer, Berlin Heidelberg New York (1979, 1980)

[Erd35] Erdős, P.: Note on the sequences of integers no one of which is divisible by any other. J. London Math. Soc., **10**, 126–128 (1935)

[Erd36] Erdős, P.: A generalization of a theorem of Besicovitch. J. London Math. Soc., **11**, 92–98 (1936)

[Erd60] Erdős, P.: An asymptotic inequality in the theory of numbers (Russian). Vestnik Leningrad. Univ., **15**, 41–49 (1960)

[For04] Ford, K.: The distribution of integers with a divisor in a given interval (2004), 62 pages. preprint available at : `http://front.math.ucdavis.edu/math.NT/0401223`

[For04a] Ford, K.: Du théorème de Kolmogorov sur les distributions empiriques à la théorie des nombres. In *L'héritage de Kolmogorov en mathématiques*. Editions Belin, Paris, 111–120 (2004)

[For06] Ford, K.: Sharp probability estimates for generalized Smirnov statistics (2006), 10 pages. preprint available at : `http://front.math.ucdavis.edu/math.PR/0609224`

[Ford06a] Ford, K.: Sharp probability estimates for random walks with barriers (2006), preprint available at : `http://front.math.ucdavis.edu/math.PR/06xxxxx`

[For06b] Ford, K.: Integers with a divisor in $(y, 2y]$, 18 pages. preprint available at : `http://front.math.ucdavis.edu/math.NT/0607473`

[Gli33] Glivenko, V.: Sulla determinazione empirica delle leggi di probabilità. Giorn. Ist. Ital. Attuari, **4**, 92–99 (1933)

[GK68] Gnedenko, B. V., Kolmogorov, A. N.: Limit distributions for sums of independent random variables. (Translated from the Russian, annotated, and revised by K. L. Chung. With appendices by J. L. Doob and P. L. Hsu. Revised edition), Addison-Wesley, Reading, Mass.-London-Don Mills., Ont. (1968)

[HT88] Hall, R.R., Tenenbaum, G.: Divisors. Cambridge Tracts in Mathematics, **90**. Cambridge University Press, Cambridge, UK (1988)

[HR17] Hardy, G.H., Ramanujan, S.: The normal number of prime factors of a number n. Quart. J. Math., **158**, 76–92 (1917)

[Kol33] Kolmogorov, A.N.: Sulla determinazione empirica di una legge di distribuzione (On the empirical determination of a distribution law). Giorn. Ist. Ital. Attuar., **4**, 83–91 (1933)

[KolW] Kolmogorov, A.N.: Selected works, vol. II: Probability theory and mathematical statistics (with a preface by Aleksandrov, P.S.; translated from the Russian by Lindquist, G.; translation edited by Shiryayev, A.N.). Kluwer Academic Publishers Group, Dordrecht (1992)

[KMT75] Komlós, J., Major, P., Tusnády, G.: An approximation of partial sums of independent RV's and the sample DF. I. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, **32**, 111–131 (1975)

[Lau63] Lauwerier, H.A.: The asymptotic expansion of the statistical distribution of N. V. Smirnov (German). Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, **2**, 61–68 (1963)

[Pen76] Penkov, B.I.: Asymptotic distribution of Pyke's statistics (Russian). Teor. Verojatnost. i Primenen., **21**, 378–383 (1976). English translation in: Theory of probability and its applications, **21**, 370–374 (1976)

[Per39] Perron, O.: Über Bruwiersche Reihen. Math. Z., **45**, 127–141 (1939)

[Phi86] Philipp, W.: Invariance principles for independent and weakly dependent random variables, in *Dependence in Probability and Statistics (Oberwolfach, 1985)*, Progr. Probab. Statist. **11**, Birkhäuser Boston, Boston, MA. 225–268 (1986)

[Pyk59] Pyke, R.: The supremum and infimum of the Poisson process. Ann. Math. Statist., **30**, 568–576 (1959) AUTHOR = Rényi, A.,

[Ren53] Rényi, A.: On the theory of order statistics. Acta Math. Acad. Sci. Hung., **4**, 191–232 (1953)

[Rio68] J. Riordan: Combinatorial identities. Wiley, New York (1968)

[SW86] Shorack, G.R., Wellner, J.A.: Empirical processes with applications to statistics. Wiley Series in Probability and Mathematical Statistics. Wiley, New York (1986)

[Smi39] Smirnov, N.V.: Sur les écarts de la courbe de distribution empirique (Russian, French summary). Rec. Math. Moscou (Mat. Sbornik), **6**, 3–26 (1939)

[Smi44] Smirnov, N.V.: Approximate laws of distribution of random variables from empirical data (Russian). Uspekhi Matem. Nauk, **10**, 179–206 (1944)

[Ten84] Tenenbaum, G.: Sur la probabilité qu'un entier possède un diviseur dans un intervalle donné. Compositio Math., **51**, 243–263 (1984)

[Ten87] Tenenbaum, G.: Un problème de probabilité conditionnelle en arithmétique. Acta Arith., **49**, 165–187 (1987)