FROM KOLMOGOROV'S THEOREM ON EMPIRICAL DISTRIBUTION TO NUMBER THEORY

KEVIN FORD

ABSTRACT. We describe some new estimates for the probability that an empirical distribution function stays on one side of a given line, and give applications to number theory.

1. INTRODUCTION

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

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

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

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

we see that $F_n(u) \underset{n\to\infty}{\longrightarrow} F(u)$ almost surely. In 1933, Glivenko [10] (and, slightly later, Cantelli [2]) proved that the convergence is uniform on the real line : $\sup \mid F_n(u) - F(u) \mid \underset{n\to\infty}{\longrightarrow} 0$ almost surely. Immediately, in his seminal paper [13], Kolmogorov made a careful study of the convergence of $F_n(u)$ to $F(u)$ as $n \to \infty$: he showed that if F is continuous, then for each $\lambda > 0$, the probability $\mathbf{P}(\sup |F_n(u) - F(u)| < \lambda/\sqrt{n})$ is independent of F, and that

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

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 [13] of Kolmogorov also appears in his Selected Works ([14], p. 139-146; comments p. 574-583).

Six years later, Smirnov [21] studied the corresponding one-sided bounds, showing for $\lambda \ge 0$ that

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

Date: August 24, 2004.
2000 Mathematics Subject Classification: 62G30, 11N25.
Research supported by National Science Foundation grant DMS-0301083.

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 $X_1, \dots, X_n$, which is the increasing sequence $\xi_1 \le \cdots \le \xi_n$ obtained by ordering each realization of the initial sequences $X_1, \dots, X_n$.

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

$$F(u) = \begin{cases} 0 & u \le 0 \\ u & 0 < u < 1 \\ 1 & u \ge 1. \end{cases} \tag{1.3}$$

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

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

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

A generalization of Smirnov's theorem was given by Csáki in 1974 (see [3], Theorem 2.1). For each fixed pair (α, β) of positive real numbers,

$$Q_n(\alpha\sqrt{n}, n + (\beta-\alpha)\sqrt{n}) \to 1 - e^{-2\alpha\beta} \qquad (n \to \infty). \tag{1.4}$$

Smirnov and others gave later refinements to (1.2) (e.g. [15]). All of these estimates require that $u \ge c\sqrt{n}$ and $u + v - n \ge c\sqrt{n}$ for some fixed $c > 0$. For more information about such estimates, including connections with non-crossing probabilities for Brownian motion and Brownian bridge processes, see [20]. For an application to number theory, we will need estimates which are uniform in $n \ge 1, 0 < u < n, u+v > n$ and are particularly strong when u and $u + v - n$ are very small.

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

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

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

with $\sigma(u) = \sqrt{F(u)(1-F(u))}$. In Kolmogorov's theorem, $|F_n(u) - F(u)|$ is replaced by its supremum over u, 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

$$F_n(u) = \begin{cases} 0 & u \in (-\infty, \xi_1) \\ \frac{i}{n} & u \in [\xi_i, \xi_{i+1}) \quad (1 \le i \le n-1) \\ 1 & u \in [\xi_n, +\infty). \end{cases}$$

thus we see (with (1.3)) that

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

$Q_n(u,v) = \dfrac{w}{v}(1+u/v)^{n-1}$ was found by Daniels [4]. 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 [22], and the general case is due to Pyke [18]. The equivalence of the two expressions for $Q_n(u,v)$ follows from one of Abel's identities ([19], 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.

2. New estimates for uniform order statistics

Theorem 2.1. Uniformly in $u > 0$, $w > 0$ and $n \ge 1$, we have

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

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

Two immediate corollaries are the asymptotics (1.2) and (1.4). In addition we have the following useful approximation.

Corollary 2.1. Uniformly in $u > 0$, $w > 0$ and $n \ge 1$, 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 $\dfrac{uw}{n} \to 0$, $u \to \infty$ and $w \to \infty$ as $n \to \infty$, we see that $Q_n(u,v)$ is asymptotic to $\dfrac{2uw}{n}$.

Both Smirnov [22] and Csáki [3] approximate the second sum in (1.5) via Stirling's formula for $n!$, while we use the complex analytic approach of Lauwerier [15] (see also [16]). Full details appear in §11 of [9], and we only indicate the main ideas here. First,

$$Q_n(u,v) = 1 - \frac{n!}{v^n}T, \quad T = w\sum_{u<j\le n}\frac{(w+n-j)^{n-j-1}}{(n-j)!}\,\frac{(j-u)^j}{j!}. \tag{2.1}$$

Introduce the so-called Bruwier functions (see [17])

$$\phi(z) = w\sum_{k=0}^{\infty}\frac{(k+w)^{k-1}}{k!}z^k, \quad \psi(z) = \sum_{k>u}\frac{(k-u)^k}{k!}z^k \quad (|z| < 1/e).$$

The residue theorem then gives

$$T = \int_{|z|=r}\frac{\phi(z)\psi(z)}{z^{n+1}}\,dz \qquad (r < 1/e). \tag{2.2}$$

The functions ϕ and ψ may be analytically continued to $S = \mathbb{C}\backslash(1/e, \infty)$ by writing

$$\begin{aligned} \phi(z) &= e^{wf(z)} \\ \psi(z) &= \frac{e^{-ug(z)}}{g(z)-1} + z^u \sum_{j\in\mathbb{Z}, j\neq 0} \frac{1}{g_j(z)^u (g_j(z)-1)}, \end{aligned} \tag{2.3}$$

where $f(z)$, $g(z)$ and $g_j(z)$ $(j \in \mathbb{Z}, j \neq 0)$ are the branches of the inverse of ze^{-z} as given in [15].

We next expand the contour in (2.2) into a large circle plus two real line segments above and below the branch cut of S. Letting the radius of the circle tend to ∞, we obtain (since there is no pole outside the circle of radius $1/e$) :

$$T = \frac{1}{2\pi i} \left(\int_{D^+} - \int_{D^-} \right) \frac{\phi(z)\psi(z)}{z^{n+1}}\, dz, \tag{2.4}$$

where D^+ and D^- are the branches of the real segment $(1/e, \infty)$ lying above, respectively below, the branch cut. The contribution to the integrals from the terms in the sum over j in (2.3) is negligible (this requires some effort to prove, however). The bulk of the contribution to the integrals from the remaining part comes from z near the singularity at $1/e$. Writing $z = (1/e)(1 - re^{i\theta}/2)$ with $r \geq 0$ and $|\theta| \leq \pi$, we have

$$\begin{aligned} f(z) &= 1 - r^{1/2}e^{i\theta/2} + \frac{1}{3}re^{i\theta} + O(r^{3/2}), \\ g(z) &= 1 + r^{1/2}e^{i\theta/2} + \frac{1}{3}re^{i\theta} + O(r^{3/2}). \end{aligned}$$

Here $\theta = -\pi$ corresponds to $z \in D^+$ and $\theta = \pi$ corresponds to $z \in D^-$. Next, let $s \geq 0$ and suppose $\frac{1}{e}(1 + s^2/2) \in D^+$. Put

$$\begin{aligned} F(s) &= f(\frac{1}{e}(1+s^2/2)) = 1 + is - \frac{1}{3}s^2 \cdots \\ G(s) &= g(\frac{1}{e}(1+s^2/2)) = F(-s). \end{aligned} \tag{2.5}$$

The power series for $F(s)$ and $G(s)$ represent analytic functions for $|s| < \sqrt{2}$ and $G(s) = \overline{F(s)}$ for real s. By (2.4), for some small constant $\beta > 0$

$$T = \frac{e^{w+n-u}}{2\pi} \int_{-\beta}^{\beta} \frac{-is}{G(s)-1} \frac{e^{w(F(s)-1)-u(G(s)-1)}}{(1+s^2/2)^{n+1}}\, ds + E, \tag{2.6}$$

E being a negligible error term.

The above integral is estimated using the saddle point method, which requires good estimates for the location of the critical point $\gamma = \gamma(x, y)$ of the function

$$a(s) = a(s; x, y) = x(F(s) - 1) - y(G(s) - 1) - \log(1 + s^2/2).$$

which lies closest to the origin. Here x, y are complex numbers with $|x| \le c$, $|y| \le c$ for some small $c > 0$. Using (2.5) and the fact that f and g are inverses of ze^{-z}, we obtain

$$\begin{aligned}\gamma(x,y) &= (x+y)\left(\frac{i+O(|x+y|^2)}{1-\frac{2}{3}(y-x)}\right),\\ a(\gamma(x,y)) &= -xy+y-x+\log(1+x)+\log(1-y)+O(|x^2y|+|xy^2|),\\ a''(\gamma(x,y)) &= -1+O(|x|+|y|).\end{aligned}$$

Taking $x = \frac{w}{n+1}$ and $y = \frac{u}{n+1}$, we find that the integral in (2.6) is

$$\left(\sqrt{\frac{2\pi}{n}}+O\left(\frac{w+u}{n^{3/2}}\right)\right) e^{u-w-\frac{uw}{n}+O(\frac{uw(u+w)}{n^2})}\left(1+\frac{w}{n}\right)^n\left(1-\frac{u}{n}\right)^n. \tag{2.7}$$

Theorem 2.1 readily follows from (2.1), (2.6) and (2.7).

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 [12], and much work has been done on this topic since then. Write an arbitrary integer $n = p_1p_2\cdots p_k$, 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 $(a, b]$, 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. Let $\{X_p : p \text{ prime}\}$ be independent Bernoulli random variables so that $\mathbf{P}(X_p = 0) = 1 - \frac{1}{p}$

[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.

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 [5].

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. Consider the problem of estimating $\varepsilon(y,z)$, 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 [1], who proved that $\liminf_{y\to\infty} \varepsilon(y,2y) = 0$. A year later, Erdős [6] improved this to $\lim_{y\to\infty} \varepsilon(y,2y) = 0$. Later work, especially by Erdős [7], [8], and Tenenbaum [23], 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 [11] contains a thorough exposition on such bounds and their applications. The main theorem of [9] 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,

$$\frac{c_1}{(\log y)^\delta (\log\log y)^{3/2}} \le \varepsilon(y,2y) \le \frac{c_2}{(\log y)^\delta (\log\log y)^{3/2}}, \tag{3.1}$$

where $\delta = 1 - \frac{1+\log\log 2}{\log 2} = 0.08607\ldots$.

Concerning the behavior of $\varepsilon_r(y,z)$, Erdős conjectured in [8] 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 [24] 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 [9], 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 [24].

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

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

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 $v = \frac{\log\log y}{\log 2}$. Utilizing Theorem 2.1 leads to the upper bound in (3.1). 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 $\le y$.

References

[1] A. S. Besicovitch, *On the density of certain sequences of integers*, Math. Ann. **110** (1934), p. 336-341.

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

[3] E. Csáki, *On tests based on empirical distribution functions* (Hungarian), Magyar Tud. Akad. Mat. Fiz. Oszt. Közl. **23** (1977), no. 3-4, p. 239-327. English translation in *Selected translations in mathematical statistics and probability* **15**, ed. by L. J. Leifman, Amer. Math. Soc. (1981), p. 229-317.

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

[5] P. D. T. A. Elliott, *Probabilistic number theory. I,II*, Grund. Math. Wissen. **239-240**, Springer-Verlag, New York, 1979-1980.

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

[7] P. Erdős, *A generalization of a theorem of Besicovitch*, J. London Math. Soc. **11** (1936), p. 92-98.

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

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

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

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

[12] G. H Hardy and S. Ramanujan, *The normal number of prime factors of a number n*, Quart. J. Math. **158** (1917), p. 76-92.
[13] A. N. Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione (On the empirical determination of a distribution law)*, Giorn. Ist. Ital. Attuar. **4** (1933), p. 83-91.
[14] A. N. Kolmogorov, *Selected works, vol. II : Probability theory and mathematical statistics*, (with a preface by P. S. Aleksandrov ; translated from the Russian by G. Lindquist, translation edited by A. N. Shiryayev), Kluwer Academic Publishers Group, Dordrecht, 1992.
[15] H. A. Lauwerier, *The asymptotic expansion of the statistical distribution of N. V. Smirnov* (German), Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **2** (1963), p. 61-68.
[16] B. I. Penkov, *Asymptotic distribution of Pyke's statistics* (Russian), Teor. Verojatnost. i Primenen. **21** (1976), p. 378-383. English translation in *Theory of probability and its applications* **21** (1976), p. 370-374.
[17] O. Perron, *Über Bruwiersche Reihen*, Math. Z. **45** (1939), p. 127-141.
[18] R. Pyke, *The supremum and infimum of the Poisson process*, Ann. Math. Statist. **30** (1959), p. 568-576.
[19] J. Riordan, *Combinatorial identities*, John Wiley & Sons Inc., New York, 1968.
[20] G. R. Shorack and J. A. Wellner, *Empirical processes with applications to statistics*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986.
[21] N. V. Smirnov, *Sur les écarts de la courbe de distribution empirique* (Russian. French summary), Rec. Math. Moscou (Mat. Sbornik) **6** (1939), p. 3-26.
[22] N. V. Smirnov, *Lois de distribution approchées de variables aléatoires à partir de données empiriques* Uspekhi Matem. Nauk **10** (1944), p. 179-206.
[23] G. Tenenbaum, *Sur la probabilité qu'un entier possède un diviseur dans un intervalle donné*, Compositio Math. **51** (1984), p. 243-263.
[24] G. Tenenbaum, *Un problème de probabilité conditionnelle en arithmétique*, Acta Arith. **49** (1987), p. 165-187.

Department of Mathematics, 1409 West Green Street, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
E-mail address: ford@math.uiuc.edu