THE POISSON-DIRICHLET DISTRIBUTION AND THE FREQUENCY OF LARGE PRIME DIVISORS

J.F.C. Kingman, Isaac Newton Institute

Of the 100 integers from 2 to 101, 68 have the property of possessing a prime divisor larger than their square root. This phenomenon persists for larger integers, and the density of the set of such integers is just $\log 2 = 0.69315$.

This is a special case of a theorem of Dickman (1930), who calculated the density

$$\mathbf{D}\{n\,;p_1(n) > n^x\}\,, \tag{1}$$

where $p_1(n)$ is the largest prime divisor of n, and

$$\mathbf{D}(E) = \lim_{n\to\infty} n^{-1} \sum_{m\in E\,,m\leqslant n} 1 \tag{2}$$

is (when it exists) the usual arithmetic density of a set E of natural numbers. Dickman's expression for (1) is in general very complicated, but simplifies when $x \geqslant \frac{1}{2}$ to

$$\mathbf{D}\{n\,;p_1(n) > n^x\} = -\log x\,. \tag{3}$$

Dickman's result was rediscovered, for instance by Ramaswami (1949) and de Bruijn (1951), but it was greatly generalised by Billingsley (1972). He showed that, if

$$p_1(n) \geqslant p_2(n) \geqslant \ldots \geqslant p_m(n) \tag{4}$$

are the (not necessarily distinct) prime factors of n, and if we define $p_r(n) = 1$ for $r > m$, so that

$$n = \prod_{r=1}^{\infty} p_r(n)\,, \tag{5}$$

then the density

$$\mathbf{D}\{n\,;\log p_r(n)/\log n > x_r(r = 1, 2, \ldots, k)\} \tag{6}$$

exists for all $k \geqslant 1$ and all x_r, and he gives explicit complicated formulae for these densities.

Lloyd (1984) noticed that Billingsley's formulae were the same as those that Shepp and Lloyd (1966) had derived for the problem of cycle lengths in a random permutation, although there is no natural isomorphism between these two problems. As will be shown, his analysis holds the key to Billingsley's theorem, but it was ignored by subsequent authors (and only drawn to my attention by Simon Tavaré in 2004).

In particular, it was not cited in an otherwise definitive paper by Donnelly and Grimmett (1993). They observed that Billingsley's expressions for (6) were the marginal distributions of the Poisson-Dirichlet distribution $\mathcal{PD}(1)$. By exploiting a characterisation of this distribution as that of the order statistics of the GEM distribution, they produced a transparent and elegant proof of Billingsley's theorem.

Not, however, transparent enough, since Arratia, Barbour and Tavaré (1997, 2003) pointed out that the prime divisor problem is unusual among the many applications of the Poisson-Dirichlet distribution in being difficult to relate to other instances. The analysis of Lloyd, however, does lead to a clear explanation of the occurrence of $\mathcal{PD}(1)$ in (6).

The Poisson-Dirichlet distribution was introduced in Kingman (1975) as a limiting case of the Dirichlet distribution, relevant to problems of computer storage and of population genetics. The marginal distributions can be read off from results of Watterson (1974), and these coincide with those of Dickman and Billingsley. Kingman showed that the distribution can be derived from the jumps of the gamma subordinator of Moran (1959), and the familiar description of the jumps of a subordinator in terms of a non-homogeneous Poisson process leads at once to the modern definition of $\mathcal{PD}(\theta)$, where $\theta > 0$.

Thus let a random sequence

$$Y_1 > Y_2 > Y_3 > \ldots > 0 \tag{7}$$

form a Poisson process on $(0, \infty)$ whose mean measure has density

$$\theta y^{-1} e^{-y} \quad (y > 0)\,. \tag{8}$$

(Definitions and properties of Poisson processes as in Kingman (1993)). Campbell's theorem shows easily that

$$\sigma = \sum_{r=1}^{\infty} Y_r \tag{9}$$

is finite with probability one, and has the gamma distribution with density

$$\Gamma(\theta)^{-1} s^{\theta-1} e^{-s} \quad (s > 0)\,. \tag{10}$$

The key property of the density (8) is that the sequence

$$X_1 > X_2 > X_3 > \ldots > 0 \tag{11}$$

defined by

$$X_r = Y_r/\sigma \tag{12}$$

is independent of σ. Its distribution is the Poisson-Dirichlet distribution $\mathcal{PD}(\theta)$, which is a probability measure on the space of infinite sequences $(x_r; r \leqslant 1)$ satisfying

$$x_1 > x_2 > x_3 > \ldots > 0\,,\ \sum_{r=1}^{\infty} x_r = 1\,. \tag{13}$$

Billingsley's theorem can then be expressed by saying that the density (6) exists and is given by

$$\mathbf{P}\left\{X_r > x_r (r = 1, 2, \ldots, k)\right\} \tag{14}$$

where the sequence (X_r) has distribution $\mathcal{PD}(1)$.

In order to make this manifest, we need to find a Poisson process with mean measure defined by (8) with $\theta = 1$. The clue to this construction is in Lloyd (1984), where he considers, for any $s > 1$, a distribution $\mathbf{P}_s$ on the natural numbers with

$$\mathbf{P}_s\{n\} = \zeta(s)^{-1} n^{-s} \quad (n \geqslant 1)\,, \tag{15}$$

$\zeta(s)$ being of course the Riemann zeta function. If the unique factorisation of n into primes is written not in the form (5) but as

$$n = \prod_p p^{\alpha_p(n)}\,, \tag{16}$$

(15) becomes

$$\mathbf{P}_s\{n\} = \prod_p \left(1 - \frac{1}{p^s}\right)\left(\frac{1}{p^s}\right)^{\alpha_p(n)}\,. \tag{17}$$

This shows that the α_p are independent random variables with geometric distributions

$$\mathbf{P}_s\{\alpha_p = a\} = \left(1 - p^{-s}\right)\left(p^{-s}\right)^a \quad (a \geqslant 0)\ . \tag{18}$$

This leads easily to the following result, implicit in Lloyd (1984).

Theorem 1 *The joint distributions of the counts of the random set*

$$\{(s-1)\log p_r \quad (r = 1, 2, \ldots)\}\,, \tag{19}$$

counting multiplicities, under $\mathbf{P}_s$*, converges as* $s \to 1$ *to those of a Poisson process with density (8) with* $\theta = 1$*.*

Proof Write

$$t = (s-1)^{-1} \tag{20}$$

Then the number of points of (19), counting multiplicities, that fall in an interval $I = (A, B] \quad (0 < A < B < \infty)$ is

$$C(I) = \sum_{e^{tA} < p \leqslant e^{tB}} \alpha_p(n)\,. \tag{21}$$

Since the α_p are independent under $\mathbf{P}_s$ for distinct primes p, the $C(I)$ for disjoint I are independent, so that we have only to prove that the $\mathbf{P}_s$-distribution of $C(I)$ converges, as $s \to 1$, to a Poisson distribution with mean

$$\int_A^B y^{-1} e^{-y} dy\,. \tag{22}$$

Lloyd does this by a clever trick which lengthens and obscures the proof. Standard theorems on Poisson approximation can be used (and would be useful if error bounds were required), but for our purposes it is simpler to proceed directly. The probability generating function of $C(I)$ is, for $0 \leqslant \xi \leqslant 1$,

$$\begin{aligned}\Phi_I(\xi) &= \prod_{e^{tA} < p \leqslant e^{tB}} \left(\frac{1 - p^{-s}}{1 - p^{-s}\xi}\right) \\ &= \exp\left(-\sum_{e^{tA} < p \leqslant e^{tB}} \log\left\{1 + \frac{p^{-s}(1-\xi)}{1 - p^{-s}}\right\}\right)\ .\end{aligned}$$

It is easy to check that

$$\left|\log\left\{1+\frac{p^{-s}(1-\xi)}{1-p^{-s}}\right\}-(1-\xi)p^{-s}\right|\leqslant 2p^{-2s}\,,$$

so that

$$\begin{aligned}\left|-\log\Phi_I(\xi)-(1-\xi)\sum_{e^{tA}<p\leqslant e^{tB}}p^{-s}\right|&\\ \leqslant 2\sum_{e^{tA}<p}p^{-2s}=O(e^{-tA})&\end{aligned}$$

as $s\to 1$, $t\to\infty$. Hence the theorem is proved if we can show that

$$\lim_{s\to 1}\sum_{e^{tA}<p\leqslant e^{tB}}p^{-s}=\int_A^B y^{-1}e^{-y}dy\,. \tag{23}$$

This is the only point at which serious number theory is needed. Equation (23) is an easy consequence of the prime number theorem, but it does not require anything like the full force of that theorem. We deduce it from the much easier theorem (Theorem 427 of Hardy and Wright (1960)) that

$$S(x)=\sum_{p\leqslant x}\frac{1}{p}=\log\log x+C+o(1) \tag{24}$$

as $x\to\infty$, where C is a constant. The sum in (23) can be written as a Stieltjes integral with respect to the step function S, which can then be integrated by parts:

$$\begin{aligned}\sum_{e^{tA}<p\leqslant e^{tB}}p^{-s}&=\int_{e^{tA}}^{e^{tB}}x^{1-s}dS(x)\\ &=e^{-B}S\left(e^{tB}\right)-e^{-A}S\left(e^{tA}\right)+\int_{e^{tA}}^{e^{tB}}(s-1)x^{-s}S(x)dx\\ &=e^{-B}S\left(e^{tB}\right)-e^{-A}S\left(e^{tA}\right)+\int_A^B e^{-y}S\left(e^{ty}\right)dy\,.\end{aligned}$$

By (24) this last expression is equal to

$$\begin{aligned} & e^{-B}(\log tB + C) - e^{-A}(\log tA + C) \\ & \qquad + \int_A^B e^{-y}(\log ty + C)\, dy + o(1) \\ & \qquad = e^{-B}\log B - e^{-A}\log A + \int_A^B e^{-y}\log y dy + o(1) \\ & \qquad = \int_A^B y^{-1}e^{-y}dy + o(1)\,. \end{aligned}$$

This proves (23), and thus the theorem.

Note, that, although (19) may contain multiple points, the limiting Poisson process does not, since its mean measure is non-atomic. Thus the theorem remains true even if multiple points are only counted once.

The following theorem is an almost immediate consequence of Theorem 1.

Theorem 2 *Under* $\mathbf{P}_s$, *the joint distributions of the sequence*

$$\log p_r / \log n \quad (r = 1, 2, 3, \ldots) \tag{25}$$

converge to those of $\mathcal{PD}(1)$ *as* $s \to 1$.

Proof Theorem 1 shows that the joint distributions of the sequence

$$Y_r = (s-1)\log p_r \quad (r = 1, 2, 3, \ldots) \tag{26}$$

converge to those of the points of a Poisson process with mean measure having density $y^{-1}e^{-y}$. If we can prove that the distributions of

$$Y_1, Y_2, Y_3, \ldots, \ \sigma = Y_1 + Y_2 + \ldots \tag{27}$$

converge to those of the points of a Poisson process and their sum, the conclusion of the theorem follows. This is not quite obvious, since σ is only a lower semicontinuous function of the Y_r. However,

$$\sigma = (s-1)\log n\,, \tag{28}$$

and it is elementary from (15) that the distribution of σ converges, as $s \to 1$, to a unit exponential distribution. Since this is same as the distribution ((10) with $\theta = 1$) in the Poisson limit, Theorem 1 does extend to the extended sequence (27). Thus the distributions of

$$X_r = Y_r/\sigma = \log p_r / \log n \tag{29}$$

converge to those in the Poisson limit, which are the marginals of $\mathcal{PD}(1)$. This completes the proof.

It is important to note that this argument does not prove Billingsley's (or even Dickman's) theorem. These make statements about the arithmetic densities $\mathbf{D}(E)$ of certain subsets E of the natural numbers. Theorem 2 makes the same statements about

$$\mathbf{HD}(E) = \lim_{s \to 1} \sum_{E} \zeta(s)^{-1} n^{-s} \,. \tag{30}$$

We use the symbol **HD** because, as Lloyd points out, (30) is equivalent to the harmonic density

$$\mathbf{HD}(E) = \lim_{m \to \infty} (\log m)^{-1} \sum_{n \in E, n \leqslant m} n^{-1} \,. \tag{31}$$

If $\mathbf{D}(E)$ exists, so does $\mathbf{HD}(E)$, and they are equal.

The converse however is false; a set E can have $\mathbf{HD}(E)$ without $\mathbf{D}(E)$ existing. A vivid example is the set E of natural numbers whose decimal expansion begins with 1. This has

$$\mathbf{HD}(E) = \log 2 / \log 10 \,,$$

but has upper and lower densities $5/9$ and $1/9$.

Thus for a self-contained proof of the Dickman-Billingsley theorem, the best approach remains that of Donnelly and Grimmett. If however we appeal to general results, such as those of Levin and Faĭnleĭt (1967), to establish the existence of $\mathbf{D}(E)$, we can use Theorem 2 and avoid the computations of joint distributions in the Donnelly-Grimmett proof.

Does the present argument make the inevitability of $\mathcal{PD}(1)$ clear? The identification of the Poisson process in Theorem 1 does help, but the computation of the mean measure is perhaps still a little unsatisfactory. The characteristic of Poisson processes with densities of the form (8) is the independence of

$$\sigma = \sum_{r=1}^{\infty} Y_r \qquad \text{and} \qquad (Y_r/\sigma \,;\; r = 1, 2, \ldots) \,.$$

This is actually a stumbling block to Arratia, Barbour and Tavaré (1997), since the $p_r(n)$ are determined by n, so how can $\log p_r(n)/\log n$ be independent of n?

The answer to this mild paradox is that the probabilities are averaged over numbers near n. What it is saying is that the statistics of the sequence

$$\log p_r(n)/\log n \qquad (r = 1, 2, \ldots)$$

for n in some distant interval do not depend on the choice of that interval.

If this were obvious, the occurrence of $\mathcal{PD}(\theta)$ would be explained. The particular value $\theta = 1$ is easier to understand, since it relates to the asymptotic exponential distribution for

$$(s-1)\log n$$

under the natural distribution $\mathbf{P}_s$.

References

R. Arratia, A.D. Barbour & S. Tavaré (1997) Random combinatorial structures and prime factorizations, *Notices Amer. Math. Soc.* **44**, 903–10.

R. Arratia, A.D. Barbour & S. Tavaré (2003) *Logarithmic Combinatorial Structures: a Probabilistic Approach*, European Mathematical Society, Zurich.

P. Billingsley (1972) On the distribution of large prime divisors, *Period. Math. Hung.* **2**, 283–9.

N.G. de Bruijn (1951) On the number of positive integers $\leqslant x$ and free of prime factors $> y$, *Nederl. Akad. Wetensch. Proc. Ser A* **54**, 50–60.

K. Dickman (1930) On the frequency of numbers containing prime divisors of a certain relative magnitude, *Ark. Math. Astronomi och Fysik* **22**, 1–14.

P.J. Donnelly & G.R. Grimmett (1993) On the asymptotic distribution of large prime factors, *J. Lond. Math. Soc.* **47**, 395–404.

G.H. Hardy & E.M. Wright (1960) *The Theory of Numbers*, Oxford.

J.F.C. Kingman (1975) Random discrete distributions, *J. Roy. Statist. Soc. B,* **37**, 1–22.

J.F.C. Kingman (1993) *Poisson Processes*, Oxford.

B.V. Levin & A.S. Faĭnleĭb (1967) Application of certain integral equations to questions of the theory of numbers, *Usephi Mat. Nauk* **22**, 119–197.

S.P. Lloyd (1984) Ordered prime divisors of a random integer, *Ann. Prob.* **12**, 1206–12.

P.A.P. Moran (1959) *The Theory of Storage*, Methuen, London.

V. Ramaswami (1949) The number of positive integers $\leqslant x$ and free of prime divisors $> y^c$, and a problem of S.S. Pillai, *Duke Math. J.* **16**, 99–109.

L.A. Shepp & S.P. Lloyd (1966) Ordered cycle lengths in a random permutation, *Trans. Amer. Math. Soc.* **121**, 340–57.

G.A. Watterson (1974) The sampling theory of selectively neutral alleles, *Adv. Appl. Prob.* **6**, 463–88.

POSTSCRIPT (19.8.04)

Construct a sequence $X_1 > X_2 > \dots$ with distribution $\mathcal{PD}(\theta)$ by equations (7)–(12). Let Z be independent of the Y_r and therefore of the X_r, with gamma distribution (10). Then the joint distribution of

$$X_1, X_2, \dots, \sigma$$

is the same as that of

$$X_1, X_2, \dots, Z,$$

and hence the joint distribution of

$$\sigma X_1, \sigma X_2, \dots$$

is the same as that of

$$ZX_1, ZX_2, \dots .$$

This proves that, if (X_r) has distribution $\mathcal{PD}(\theta)$ and the independent random variable Z has distribution $\Gamma(\theta)$, then

$$ZX_r (r = 1, 2, \dots)$$

are the points of a Poisson process with density (8). Combining this result for $\theta = 1$ with the Dickman-Billingsley theorem, we have the following analogue of Lloyd's theorem.

Theorem 3 *Let $p_1 \geqslant p_2 \geqslant \dots \geqslant p_m$ be the prime factors of a random integer n, uniformly distributed on $\{1, 2, \dots, N\}$. Let Z_N be independent of n, having a negative exponential distribution with mean $(\log N)^{-1}$. Then the joint distributions of the random sequence*

$$Z_N \log p_r \qquad (r = 1, 2, \dots)$$

converge as $N \to \infty$ to those of the points (in descending order) of a Poisson process with density $x^{-1}e^{-x}$.

Proof The Dickman-Billingsley theorem states, in effect, that the sequence

$$\log p_r / \log n \qquad (r = 1, 2, \dots)$$

converges in distribution to $\mathcal{PD}(1)$ as $N \to \infty$. Since $Z_N \log N$ has distribution $\Gamma(1)$, this shows that the sequence

$$Z_N \log N \log p_r / \log n$$

converges in distribution to the Poisson process with density (8) and $\theta = 1$. It is therefore only necessary to show that

$$\log n / \log N \to 1$$

in probability, and this is clear since

$$\begin{aligned} \mathbb{E}\left\{ \frac{\log n}{\log N} \right\} &= \frac{1}{N \log N} \sum_{n=1}^{N} \log n \\ &= \frac{1}{N \log N} \left(N \log N - N + O(\log N) \right) \to 1 \,. \end{aligned}$$

The random scaling by Z_N is essential for the result; it is easy to see that Z_N cannot be replaced by any deterministic sequence.

A direct proof of Theorem 3 would give a proof of the Dickman-Billingsley theorem. By Rényi arguments, it would be enough to prove that the distribution of the number of r for which $Z_N \log p_r$ lies in a finite union E of intervals converges, as $N \to \infty$, to the Poisson distribution with mean

$$\int_E x^{-1} e^{-x} dx \,.$$