A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 56–61

On the distribution of the greatest common divisor

Persi Diaconis[1] **and Paul Erdös**[1]

Stanford University

Abstract: For two integers chosen independently at random from $\{1, 2, \ldots, x\}$, we give expansions for the distribution and the moments of their greatest common divisor and the least common multiple, with explicit error rates. The expansion involves Riemann's zeta function. Application to a statistical question is briefly discussed.

1. Introduction and statement of main results

Let M and N be random intergers chosen uniformly and independently from $\{1, 2, \ldots, x\}$. Throughout (M, N) will denote the greatest common divisor and $[M, N]$ the least common multiple. Cesàro (1885) studied the moments of (M, N) and $[M, N]$. Theorems 1 and 2 extend his work by providing explicit error terms. The distribution of (M, N) and $[M, N]$ is given by:

Theorem 1.

$$P_x\big\{[M,N] \le tx^2 \quad \textit{and} \quad (M,N) = k\big\} = \frac{6}{\pi^2}\ \frac{1}{k^2}\big\{kt(1-\log kt)\big\} + O_{k,t}\left(\frac{\log x}{x}\right) \tag{1.1}$$

$$P_x\big\{(M,N) = k\big\} = \frac{6}{\pi^2}\frac{1}{k^2} + O\left(\frac{\log(\frac{x}{k})}{xk}\right) \tag{1.2}$$

$$P_x\big\{[M,N] \le tx^2\big\} = 1 + \frac{6}{\pi^2}\cdot\sum_{j=1}^{[1/t]}\big\{jt(1-\log jt) - 1\big\} + O_t\left(\frac{\log x}{x}\right). \tag{1.3}$$

Where $[x]$ denotes the greatest integer less than or equal to x. Christopher (1956) gave a weaker form of (1.2).

(1.2) easily yields an estimate for the expected value of (M, N):

$$E_x\big\{(M,N)\big\} = \frac{1}{x^2}\sum_{i,j\le x}(i,j) = \sum_{k\le x} k\ P_x\big\{(M,N)=k\big\} = \frac{6}{\pi^2}\log x + O(1).$$

(1.2) does not lead to an estimate for higher moments of (M, N). Similarly the form of (1.3) makes direct computation of moments of $[M, N]$ unwieldy. Using elementary arguments we will show:

Theorem 2.

$$E_x\big\{(M,N)\big\} = \frac{6}{\pi^2}\log x + C + O\left(\frac{\log x}{\sqrt{x}}\right) \tag{1.4}$$

[1]Department of Statistics, Stanford University, Stanford, 94305-4065, CA USA. e-mail: diaconis@math.stanford.edu

Keywords and phrases: Euler constant, gcd, inversion, lcm, moment, random, zeta function.
AMS 2000 subject classifications: 11N37, 11A25, 60E05.

where C is an explicitly calculated constant.

$$\text{for} \qquad k \geq 2,\ E_x\big\{(M,N)^k\big\} = \frac{x^{k-1}}{k+1}\left\{\frac{2\zeta(k)}{\zeta(k+1)} - 1\right\} + O\big(x^{k-2}\log x\big). \tag{1.5}$$

where $\zeta(z)$ is Riemann's zeta function,

$$\text{for} \qquad k \geq 1, E_x\big\{[M,N]^k\big\} = \frac{\zeta(k+2)}{\zeta(2)(k+1)^2}x^{2k} + O\big(x^{2k-1}\log x\big). \tag{1.6}$$

Section two of this paper contains proofs while section three contains remarks, further references and an application to the statistical problem of reconstructing the sample size given a table of rounded percentages.

2. Proofs of main theorems

Throughout we use the elementary estimate

$$\Phi(x) = \sum_{1\leq k\leq x}\varphi(k) = \frac{3}{\pi^2}x^2 + R(x) \tag{2.1}$$

where $R(x) = O(x\log x)$.

See, for example, Hardy and Wright (1960) Theorem 330. Since $\#\ \{m,n \leq x : (m,n) = 1\} = 2\Phi(x) + O(1)$ and $(m,n) = k$ if and only if $k|m$, $k|n$ and $(\frac{m}{k},\frac{n}{k}) = 1$, we see that $\#\ \{m,n \leq x : (m,n) = k\} = 2\Phi(\frac{x}{k}) + O(1)$. This proves (1.2). To prove (1.1) and (1.3) we need a preparatory lemma.

Lemma 1. *If $F_x(t) = \#\{m,n \leq x : mn \leq tx^2$ and $(m,n) = 1\}$, then*

$$F_x(t) = \frac{6}{\pi^2}\ t(1-\log t)x^2 + O_t(x\log x).$$

Proof. Consider the number of lattice points in the region $R_x(t) = \{m,n \leq x : mn \leq tx^2\}$. It is easy to see that there are $t(1-\log t)x^2 + O_t(x) = N_x(t)$ such points. Also, the pair $\langle m,n\rangle \in R_x(t)$ and $(m,n) = k$ if and only if $\langle\frac{m}{k},\frac{n}{k}\rangle \in R_{x/k}(t)$ and $(\frac{m}{k},\frac{n}{k}) = 1$. Thus $N_x(t) = \sum_{1\leq d\leq x}F_{x/d}(t)$. The standard inversion formula says

$$F_x(t) = \sum_{1\leq d\leq x}\mu(d)N_{x/d}(t) = \frac{6}{\pi^2}\ t(1-\log t)x^2 + O_t(x\log x).$$

Lemma 1 immediately implies that the product of 2 random integers is independent of their greatest common divisor:

Corollary 1.

$$P_x\big\{MN \leq tx^2 | (M,N) = k\big\} = t(1-\log t) + O_{t,k}\left(\frac{\log x}{x}\right).$$

To prove (1) note that

$$\begin{aligned} P_x\big\{[M,N] &\leq tx^2 \quad \text{and} \quad (M,N) = k\big\} \\ &= P_x\big\{[M,N] \leq tx^2 | (M,N) = k\big\}\cdot P_x\big\{(M,N) = k\big\} \\ &= P_x\left\{MN \leq \frac{t}{k}\ x^2 | (M,N) = k\right\}\cdot P_x\big\{(M,N) = k\big\}. \end{aligned}$$

Use of (1.2) and Corollary 1 completes the proof of (1.1). To prove (1.3) note that

$$P_x\{[M,N] \le tx^2\} = P_x\left\{(M,N) > \left[\frac{1}{t}\right]\right\} + \sum_{k=1}^{[1/t]} P_x\{[M,N] \le tx^2 | (M,N)=k\} \cdot P_x\{(M,N)=k\}.$$

Using (1.2) and Corollary 1 as before completes the proof of Theorem 1.

To prove Theorem 2, write, for $k \ge 1$,

$$\begin{aligned}\sum_{m,n\le x} (m,n)^k &= 2\sum_{1\le m\le x}\ \sum_{1\le n\le m} (m,n)^k - \sum_{1\le i\le x} i^k \\ &= 2\sum_{1\le m\le x} f_k(m) - \frac{x^{k+1}}{k+1} + O(x^k) \qquad (2.2)\end{aligned}$$

where $f_k(m) = \sum_{d|m} d^k \varphi(\frac{n}{d})$. Dirichlet's Hyperbole argument (see, e.g., Saffari (1970)) yields for any t,

$$\sum_{1\le m\le x} f_k(m) = \sum_{1\le i\le t} i^k \Phi\left(\frac{x}{i}\right) + \sum_{1\le i\le x/t} \varphi(i) I_k\left(\frac{x}{i}\right) - I_k(t)\Phi\left(\frac{x}{t}\right) \qquad (2.3)$$

where

$$I_k(t) = \sum_{1\le i\le t} i^k = \frac{t^{k+1}}{k+t} + O(t^k).$$

When $k=1$, we proceed as follows: Choose $t=\sqrt{x}$. The first sum on the right side of (2.3) is

$$\begin{aligned}&\sum_{1\le k\le\sqrt{x}} \left\{\frac{3}{\pi^2}\left(\frac{x}{k}\right)^2 + O\left(\frac{x}{k}\log\frac{x}{k}\right)\right\} \\ &\quad = \frac{3}{\pi^2}x^2\left\{\log\sqrt{x} + \gamma + O\left(\frac{1}{\sqrt{x}}\right)\right\} + O(x^{3/2}\log x). \qquad (2.4)\end{aligned}$$

The second sum in (2.3) is

$$\sum_{1\le k\le\sqrt{x}} \varphi(k)\left\{\frac{1}{2}\left(\frac{x}{k}\right)^2 + O\left(\frac{x}{k}\right)\right\} = \frac{x^2}{2}\sum_{1\le k\le\sqrt{x}} \frac{\varphi(k)}{k^2} + O(x^{3/2}). \qquad (2.5)$$

Now

$$\begin{aligned}\sum_{1\le k\le\sqrt{x}} \frac{\varphi(k)}{k^2} &= \sum_{1\le k\le\sqrt{x}} \frac{2k+1}{(k(k+1))^2}\Phi(k) + \frac{\Phi(\sqrt{x})}{[x]} \\ &= 2\sum_{1\le k\le\sqrt{x}} \frac{1}{k(k+1)^2}\left\{\frac{3}{\pi^2}k^2 + R(k)\right\} + \sum_{1\le k\le\sqrt{x}} \frac{\Phi(k)}{k^2(k+1)^2} + \frac{3}{\pi^2} + O\left(\frac{\log x}{\sqrt{x}}\right) \\ &= \frac{6}{\pi^2}\sum_{1\le k\le\sqrt{x}} \frac{k}{(k+1)^2} + 2\sum_{k=1}^{\infty}\frac{R(k)}{k(k+1)^2} + \sum_{k=1}^{\infty}\frac{\Phi(k)}{k^2(k+1)^2} + \frac{3}{\pi^2} + O\left(\frac{\log x}{\sqrt{x}}\right) \\ &= \frac{3}{\pi^2}\log x + d + O\left(\frac{\log x}{\sqrt{x}}\right)\end{aligned}$$

where

$$d = \sum_{k=1}^{\infty} \left\{ \Phi(k) + 2kR(k) - \frac{6}{\pi^2} k(2k+1) \right\} / \big(k(k+1)\big)^2 + \frac{6}{\pi^2}\left(\gamma + \frac{1}{2}\right) \tag{2.6}$$

and γ is Euler's constant. Using this in equation (2.5) yields that the second sum in (2.3) is

$$\frac{3x^2}{2\pi^2}\log x + \frac{d}{2}x^2 + O\big(x^{3/2}\log x\big). \tag{2.7}$$

The third term in (2.3) is

$$\frac{1}{2}\frac{3}{\pi^2}x^2 + O\big(x^{3/2}\log x\big). \tag{2.8}$$

Combining (2.8), (2.7) and (2.4) in (2.3) and using this in (2.2) yields:

$$\sum_{m,n\le x}(m,n) = \frac{6}{\pi^2}x^2\log x + \left(d + \frac{6}{\pi^2}\left(\gamma+\frac{1}{2}\right) - \frac{1}{2}\right)x^2 + O\big(x^{3/2}\log x\big),$$

where d is defined in (2.6).

When $k \ge 2$, the best choice of t in (2.3) is $t = 1$. A calculation very similar to the case of $k = 1$ leads to (1.3).

We now prove (1.6). Consider the sum

$$\begin{aligned}
\sum_{i,j\le x}[i,j]^k &= 2\sum_{i\le x}\sum_{j\le i}[i,j]^k + O\big(x^{k+1}\big) \\
&= 2\sum_{i\le x}\sum_{d|i}\sum_{j\le i}\left(\frac{ij}{d}\right)^k + O\big(x^{k+1}\big) \\
&= 2\sum_{i\le x} i^k \sum_{d|i} f_k\left(\frac{i}{d}\right) + O\big(x^{k+1}\big) \\
&= 2\sum_{d=1}^{x} d^k \sum_{j\le x/d} j^k f_k(j) + O\big(x^{k+1}\big).
\end{aligned} \tag{2.9}$$

Where

$$f_k(n) = \sum_{\substack{j\le n \\ (j,n)=1}} j^k.$$

We may derive another expression for $f_k(n)$ by considering the sum

$$\sum_{i=1}^{n} i^k = \frac{n^{k+1}}{k+1} + R_k(n) = n^k \sum_{d|n}\frac{f_k(d)}{d^k}. \tag{2.10}$$

Dividing (2.10) by n^k and inverting yields

$$\frac{f_k(n)}{n^k} = \frac{1}{k+1}\sum_{d|n}\mu\left(\frac{n}{d}\right)d + \sum_{d|n}\mu\left(\frac{n}{d}\right)\frac{R_k(d)}{d^k}$$

or

$$f_k(n) = \frac{n^k}{k+1}\varphi(n) + \sum_{d|n}\mu\left(\frac{n}{d}\right)\left(\frac{n}{d}\right)^k R_k(d) = \frac{n^k\varphi(n)}{k+1} + E(n).$$

When we substitute this expression for $f_k(j)$ in (2.9) we must evaluate:

$$\begin{aligned} S_1(y) &= \sum_{j\le y} j^k E(j) = \sum_{j\le y} j^k \sum_{d|j} \mu\left(\frac{j}{d}\right)\left(\frac{j}{d}\right)^k R_k(d) \\ &= \sum_{i\le y} \mu(i) i^{2k} \sum_{d\le y/i} R_k(d) d^k. \end{aligned}$$

Now $R_k(d)$ is a polynomial in d of degree k. Thus,

$$|S_1(y)| \le \sum_{i\le y} i^{2k}\left(\frac{y}{i}\right)^{2k+1} = O\big(y^{2k+1}\log y\big).$$

We must also evaluate

$$\begin{aligned} S_2(y) &= \frac{1}{k+1}\sum_{j\le y} j^k \varphi(j) \\ &= \frac{1}{k+1}\left\{2k\sum_{j\le y} -j^{2k-1}\Phi(j) + O\left(\sum_{j\le y} j^{2k-2}\Phi(j)\right) + \Phi(y)y^{2k}\right\} \\ &\quad -\frac{6}{\pi^2}\frac{k}{(k+1)}\ \frac{y^{2k+2}}{(2k+2)} + \frac{3}{\pi^2}\ \frac{1}{(k+1)}\ y^{2k+2} + O\big(y^{2k+1}\log y\big) \\ &= \frac{6}{\pi^2(k+1)}\left(\frac{1}{2} - \frac{k}{2k+2}\right)y^{2k+2} + O\big(y^{2k+1}\log y\big) \\ &= \frac{3}{\pi^2}\frac{1}{(k+1)^2}y^{2k+2} + O\big(y^{2k+1}\log y\big). \end{aligned}$$

Substituting in the right side of (2.9) we have

$$\begin{aligned} \sum_{i,j\le x} [i,j]^k &= 2\sum_{d=1}^{x} d^k\left\{S_1\left(\frac{x}{d}\right) + S_2\left(\frac{x}{d}\right)\right\} + O\big(x^{k+1}\big) \\ &= \frac{6}{\pi^2}\frac{1}{(k+1)^2}x^{2k+2}\sum_{d=1}^{x}\frac{1}{d^{k+2}} + O\big(x^{2k+1}\log x\big) \\ &= \frac{\zeta(k+2)}{\zeta(2)}\ \frac{x^{2k+2}}{(k+1)^2} + O\big(x^{2k+1}\log x\big). \end{aligned}$$

□

3. Miscellaneous remarks

1. If $M_1, M_2, \ldots, M_k$ are random integers chosen uniformly at random then the results stated in Christopher (1956) (see also Cohen (1960), Herzog and Stewart (1971), and Neymann (1972)) imply that

$$P_x\big\{(M_1, M_2, \ldots, M_k) = j\big\} = \frac{1}{\zeta(k)}\ \frac{1}{j^k} + O\left(\frac{1}{x\ j^{k-1}}\right) k \ge 3. \tag{3.1}$$

We have not tried to extend theorems 1 and 2 to the k-dimensional case.

(3.1) has an application to a problem in applied statistics. Suppose a population of n individuals is distributed into k categories with n individuals in category i. Often only the proportions $p_i = n_i/n$ are reported. A method for estimating n given p_i, $1 \le i \le k$ is described in Wallis and Roberts (1956), pp. 184–189. Briefly,

let $m = \min |\sum_{i=1}^{k} p_i b_i|$ where the minimum is taken over all k tuples $(b_1, b_2, \ldots, b_k)$, with $b_i \epsilon \{0, \pm 1, \pm 2, \ldots\}$ not all b_i equal zero. An estimate for n is $[1/m]$. This method works if the p_i are reported with enough precision and the n_i are relatively prime for then the Euclidean algorithm implies there are integers $\{b_i\}_{i=1}^{k}$ such that $\sum b_i n_i = 1$. These b_i give the minimum $m = \frac{1}{n}$. If it is reasonable to approximate the n_i as random integers then (3.1) implies that $\text{Prob}((n_1, n_2, \ldots, n_k) = 1) \doteq \frac{1}{\zeta(k)}$ and, as expected, as k increases this probability goes to 1. For example, $\frac{1}{\zeta(5)} \doteq .964$, $\frac{1}{\zeta(7)} \doteq .992$, $\frac{1}{\zeta(9)} \doteq .998$. This suggests the method has a good chance of working with a small number of categories. Wallace and Roberts (1956) give several examples and further details about practical implementation.

2. The best result we know for $R(x)$ defined in (2.1) is due to Saltykov (1960). He shows that

$$R(x) = O\big(x(\log x)^{2/3}(\log\log x)^{1+\epsilon}\big).$$

Use of this throughout leads to a slight improvement in the bounds of theorems 1 and 2.

3. The functions (M, N) and $[M, N]$ are both multiplicative in the sense of Delange (1969, 1970). It would be of interest to derive results similar to Theorems 1 and 2 for more general multiplicative functions.

References

[1] Cesàro, E. (1885). Etude Moyenne du plus grand Commun Diviseur de deux nombres, *Ann. Nat. Pura. Appl.* **13**(2), 233–268.

[2] Christopher, J. (1956). The asymptotic density of some k dimensional sets, *Amer. Math. Monthly* **63**, 399-401. MR97363

[3] Cohen, E. (1960). Arithmetical Functions of a Greatest Common Divisor I, *Proc. Amer. Math. Soc.* **11**, 164–171. MR111713

[4] Delange, H. (1969). Sur Les Fonctions De Plusiurs Entiers Strictement Positifs, *Enseignement Math.* **15**, 77–88. MR245538

[5] Delange, H. (1970). Sur Les Fonctions Multiplicative de Plusiurs Entiers, *Enseignement Math.* **16**, 219–246. MR294275

[6] Hardy, G. H. and Wright, E. M. (1960). *The Theory of Numbers*, Oxford University Press.

[7] Herzog, F. and Stewart, B. (1971). Patterns of Visible and non Visible Lattices, *Amer. Math. Monthly* **78**, 487–496. MR284403

[8] Neymann, J. E. (1972). On the Probability that k Positive Integers are Relatively Prime, *Jour. Number Th.* **4**, 469–473. MR304343

[9] Saffari, B. (1968). Sur quelques Applications de la "Méthode de l'hyperbole" de Dirichlet à la Théorie des Numbres Premiers, *Enseignement Math.* **14**, 205–224. MR268138

[10] Saltykov, A. I. (1960), On Eulers Function, *Vestnik Maskov Univ. Ser. I Mat. Meh.* **6**, 34–50. MR125088

[11] Wallis, W. A. and Roberts, H. V. (1956). *Statistics a New Approach*. New York, Free Press. MR88841