Number Theory

(Hardy & Ramanujan 1920)

Thm Let $\nu(x) = \#\{p \in \mathbb{P} : p \mid x\}$. If $f(n) \to \infty$ as $n \to \infty$ (arbitrarily slowly) then

$$\#\left\{\omega \in [n] : |\nu(\omega) - \ln\ln n| > f(n)\sqrt{\ln\ln n}\right\} = o(n)$$

i.e. "almost all" n have "very close to" $\ln\ln$ prime factors

Pf: (Turán 1934) Let $([n], P)$ be a p. space with $P(\omega) = \frac{1}{n}$ $\forall \omega \in [n]$.

$$X_p(\omega) \overset{\text{def}}{=} \begin{cases} 1 & \text{if } p \mid \omega \\ 0 & \text{o.w.} \end{cases}, \quad X \overset{\text{def}}{=} \sum_{p \in \mathbb{P}_{\le n}} X_p \quad \text{and} \quad L_n \overset{\text{def}}{=} f(n)\sqrt{\ln\ln n}$$

$$\mathcal{E}X = \sum_{p \in \mathbb{P}_{\le n}} \frac{\lfloor \frac{n}{p} \rfloor}{n} \le \sum_{p \in \mathbb{P}_{\le n}} \frac{1}{p} = \ln\ln n + A + O\left(\frac{1}{\ln n}\right)$$

(see Apostol: Introduction to Analytic Number Theory p90, Thm 4.12)

$\pi(n) = |\mathbb{P}_{\le n}|$

$$\sum_{\substack{p \ne q \\ p,q \in \mathbb{P}_{\le n}}} \mathrm{Cov}(X_p, X_q) = \sum \left(\frac{\lfloor \frac{n}{pq} \rfloor}{n} - \frac{\lfloor \frac{n}{p} \rfloor}{n} \frac{\lfloor \frac{n}{q} \rfloor}{n} \right) \le \frac{1}{n} \sum_{\substack{p \ne q \\ p,q \in \mathbb{P}_{\le n}}} \left(\frac{1}{p} + \frac{1}{q} \right) = \frac{2\pi(n)}{n} \sum_{p \in \mathbb{P}_{\le n}} \frac{1}{p} = o(1)$$

$$\frac{1}{pq} - \frac{(\frac{n}{p}-1)}{n}\frac{(\frac{n}{q}-1)}{n}$$

prime Number Theorem $\pi(x) \sim \frac{x}{\ln x}$

$$\mathrm{LHS}/n = P_r(|X - \ln\ln n| > L_n) \le P(|X - \mathcal{E}X| + |\mathcal{E}X - \ln\ln n| > L_n)$$

$$\le P_r\left(|X - \mathcal{E}X| > \frac{1}{2} L_n\right) \text{ as n sufficiently}$$

$\because \mathcal{E}X = \ln\ln n + A + o(1)$ and $L_n \to \infty$ as $n \to \infty$

$$\le 4\frac{\mathrm{Var}X}{L_n^2} \le 4\frac{\mathcal{E}(\sum X_p) + \sum \mathrm{Cov}(X_p, X_q)}{L_n^2} = 4\frac{\mathcal{E}X + o(1)}{L_n^2} \le 4\frac{\ln\ln n + A + o(1)}{f^2(n)\ln\ln n} = o(1)$$

QED

The second moment method is an effective tool in number theory. Let $\nu(n)$ denote the number of primes p dividing n. (We do not count multiplicity though it would make little difference.) The folllowing result says, roughly, that "almost all" n have "very close to" $\ln \ln n$ prime factors. This was first shown by Hardy and Ramanujan in 1920 by a quite complicated argument. We give a remarkably simple proof of Paul Turan [1934], a proof that played a key role in the development of probabilistic methods in number theory.

Theorem 2.1 Let $\omega(n) \to \infty$ arbitrarily slowly. Then number of x in $\{1, \ldots, n\}$ such that

$$|\nu(x) - \ln \ln n| > \omega(n)\sqrt{\ln \ln n}$$

is $o(n)$.

Proof. Let x be randomly chosen from $\{1, \ldots, n\}$. For p prime set

$$X_p = \begin{cases} 1 & \text{if } p|x \\ 0 & \text{otherwise} \end{cases}$$

Set $M = n^{1/10}$ and set $X = \sum X_p$, the summation over all primes $p \le M$. As no $x \le n$ can have more than ten prime factors larger than M we have $\nu(x) - 10 \le X(x) \le \nu(x)$ so that large deviation bounds on X will translate into asymptotically similar bounds for ν. (Here 10 could be any large constant.) Now

$$E[X_p] = \frac{\lfloor n/p \rfloor}{n}$$

As $y - 1 < \lfloor y \rfloor \le y$

$$E[X_p] = 1/p + O(1/n)$$

By linearity of expectation

$$E[X] = \sum_{p \le M} \frac{1}{p} + O(\frac{1}{n}) = \ln \ln n + O(1)$$

Now we find an asymptotic expression for $Var[X] = \sum_{p \le M} Var[X_p] + \sum_{p \ne q} Cov[X_p, X_q]$. As $Var[X_p] = \frac{1}{p}(1 - \frac{1}{p}) + O(\frac{1}{n})$,

$$\sum_{p \le M} Var[X_p] = \sum_{p \le M} \frac{1}{p} + O(1) = \ln \ln n + O(1)$$

With p, q distinct primes, $X_pX_q = 1$ if and only if $p|x$ and $q|x$ which occurs if and only if $pq|x$. Hence

$$\begin{aligned}
&Cov[X_p, X_q] = E[X_p]E[X_q] - E[X_pX_q] \\
&= \tfrac{\lfloor n/pq \rfloor}{n} - \tfrac{\lfloor n/p \rfloor}{n}\tfrac{\lfloor n/q \rfloor}{n} \\
&\le \tfrac{1}{pq} - (\tfrac{1}{p} - \tfrac{1}{n})(\tfrac{1}{q} - \tfrac{1}{n}) \\
&\le \tfrac{1}{n}(\tfrac{1}{p} + \tfrac{1}{q})
\end{aligned}$$

Thus

$$\sum_{p \neq q} Cov[X_p, X_q] \le \frac{1}{n} \sum_{p \neq q} \frac{1}{p} + \frac{1}{q} \le \frac{2M}{n} \sum \frac{1}{p}$$

Thus

$$\sum_{p \neq q} Cov[X_p, X_q] = O(n^{-9/10} \ln \ln n) = o(1)$$

That is, the covariances do not affect the variance, $Var[X] = \ln \ln n + O(1)$ and Chebyschev's Inequality actually gives

$$\Pr[|X - \ln \ln n| > \lambda \sqrt{\ln \ln n}] < \lambda^{-2} + o(1)$$

for any constant λ. As $|X - \nu| \le 10$ the same holds for ν. $\square$

In a classic paper Paul Erdős and Marc Kac [1940] showed, essentially, that ν does behave like a normal distribution with mean and variance $\ln \ln n$. Here is their precise result.

Theorem 2.2. Let λ be fixed, positive, negative or zero. Then

$$\lim_{n \to \infty} \frac{1}{n} |\{x : 1 \le x \le n, \nu(x) \ge \ln \ln n + \lambda \sqrt{\ln \ln n}\}| = \int_{\lambda}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

We outline the argument, emphasizing the similarities to Turan's proof. Fix a function $s(n)$ with $s(n) \to \infty$ and $s(n) = o((\ln \ln n)^{1/2})$ - e. g. $s(n) = \ln \ln \ln n$. Set $M = n^{1/s(n)}$. Set $X = \sum X_p$, the summation over all primes $p \le M$. As no $x \le n$ can have more than $s(n)$ prime factors greater than M we have $\nu(x) - s(n) \le X(x) \le \nu(x)$ so that it suffices to show Theorem 2.2 with ν replaced by X. Let Y_p be independent random variables with $\Pr[Y_p = 1] = p^{-1}$, $\Pr[Y_p = 0] = 1 - p^{-1}$ and set $Y = \sum Y_p$, the summation over all primes $p \le M$. This Y represents an idealized version of X. Set

$$\mu = E[Y] = \sum_{p \le M} p^{-1} = \ln \ln n + o((\ln \ln n)^{1/2})$$

and

$$\sigma^2 = Var[Y] = \sum_{p \leq M} p^{-1}(1-p^{-1}) \sim \ln\ln n$$

and define the normalized $\tilde{Y} = (Y-\mu)/\sigma$. From the Central Limit Theorem (well, an appropriately powerful form of it!) $\tilde{Y}$ approaches the standard normal N and $E[\tilde{Y}^k] \to E[N^k]$ for every positive integer k. Set $\tilde{X} = (X - \mu)/\sigma$. We compare $\tilde{X}, \tilde{Y}$.

For any distinct primes $p_1, \ldots, p_s \leq M$

$$E[X_{p_1} \cdots X_{p_s}] - E[Y_{p_1} \cdots Y_{p_s}] = \frac{\lfloor \frac{n}{p_1 \cdots p_s} \rfloor}{n} - \frac{1}{p_1 \cdots p_s} = O(n^{-1})$$

We let k be an arbitrary fixed positive integer and compare $E[\tilde{X}^k]$ and $E[\tilde{Y}^k]$. Expanding, $\tilde{X}^k$ is a polynomial in X with coefficients $n^{o(1)}$. Further expanding each $X^j = (\sum X_p)^j$ - always reducing X_p^a to X_p when $a \geq 2$ - gives the sum of $O(M^k) = n^{o(1)}$ terms of the form $X_{p_1} \cdots X_{p_s}$. The same expansion applies to $\tilde{Y}$. As the corresponding terms have expectations within $O(n^{-1})$ the total difference

$$E[\tilde{X}^k] - E[\tilde{Y}^k] = n^{-1+o(1)} = o(1)$$

Hence each moment of $\tilde{X}$ approach that of the standard normal N. A standard, though nontrivial, theorem in probability theorem gives that $\tilde{X}$ must therefore approach N in distribution. □

We recall the famous quotation of G. H. Hardy:

> 317 is a prime, not because we think so, or because our minds are shaped in one way rather than another, but *because it is so*, because mathematical reality is built that way.

How ironic - though not contradictory - that the methods of probability theory can lead to a greater understanding of the prime factorization of integers.