

Two Proofs of the Central Limit Theorem

Yuval Filmus

January/February 2010

In this lecture, we describe two proofs of a central theorem of mathematics, namely the central limit theorem. One will be using cumulants, and the other using moments. Actually, our proofs won't be entirely formal, but we will explain how to make them formal.

1 Central Limit Theorem

What is the central limit theorem? The theorem says that under rather general circumstances, if you sum *independent* random variables and normalize them accordingly, then at the limit (when you sum lots of them) you'll get a normal distribution.

For reference, here is the density of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2 :

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We now state a very weak form of the central limit theorem. *Suppose that X_i are independent, identically distributed random variables with zero mean and variance σ^2 . Then*

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \longrightarrow \mathcal{N}(0, \sigma^2).$$

Note that if the variables do not have zero mean, we can always normalize them by subtracting the expectation from them.

The meaning of $Y_n \longrightarrow Y$ is as follows: for each interval $[a, b]$,

$$\Pr[a \leq Y_n \leq b] \longrightarrow \Pr[a \leq Y \leq b].$$

This mode of convergence is called *convergence in distribution*.

The exact form of convergence is not just a technical nicety — the normalized sums do *not* converge *uniformly* to a normal distribution. This means that the tails of the distribution converge more slowly than its center. Estimates for the speed of convergence are given by the Berry-Esséen theorem and Chernoff's bound.

The central limit theorem is true under wider conditions. We will be able to prove it for independent variables with bounded moments, and even more general versions are available. For example, limited dependency can be tolerated (we will give a number-theoretic example). Moreover, random variables not having moments (i.e. $\mathbb{E}[X^n]$ doesn't converge for all n) are sometimes well-behaved enough to induce convergence. Other problematical random variable will converge, under a different normalization, to an α -stable distribution (look it up!).

2 Normal Distribution and Meaning of CLT

The normal distribution satisfies a nice convolution identity:

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2) \implies X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Moreover, we can scale a normally distributed variable:

$$X \sim \mathcal{N}(\mu, \sigma^2) \implies cX \sim \mathcal{N}(c\mu, c^2\sigma^2).$$

Even more exciting, we can recover the normal distribution from these properties. The equation $\mathcal{N}(0, 1) + \mathcal{N}(0, 1) = \sqrt{2}\mathcal{N}(0, 1)$ in essence defines $\mathcal{N}(0, 1)$ (up to scaling), from which the entire ensemble can be recovered.

These properties point at why we should expect the normalized sums in the central limit theorem to converge to a normal variable. Indeed, suppose the convergence is to a hypothetical distribution \mathcal{D} . From the equations

$$\begin{aligned} \frac{X_1 + \cdots + X_n}{\sqrt{n}} &\longrightarrow \mathcal{D} \\ \frac{X_1 + \cdots + X_{2n}}{\sqrt{2n}} &\longrightarrow \mathcal{D} \end{aligned}$$

we would expect $\mathcal{D} + \mathcal{D} = \sqrt{2}\mathcal{D}$, so \mathcal{D} must be normal. Therefore the real content of the central limit theorem is that convergence *does* take place. The

exact form of the *basin of attraction* is deducible beforehand — the only question is whether summing up lots of independent variables and normalizing them accordingly would get us closer and closer to the only possible limit, a normal distribution with the limiting mean and variance.

3 Moment Generating Function

The main tool we are going to use is the so-called *moment generating function*, defined as follows for a random variable X :

$$M_X(t) = \mathbb{E}[e^{tX}].$$

Expanding the Taylor series of e^{tX} , we discover the reason it's called the moment generating function:

$$M_X(t) = \sum_{n=0}^{\infty} \frac{\mathbb{E}[X^n]}{n!} t^n.$$

The moment generating function is thus just the exponential generating function for the moments of X . In particular,

$$M_X^{(n)}(0) = \mathbb{E}[X^n].$$

So far we've assumed that the moment generating function exists, i.e. the implied integral $\mathbb{E}[e^{tX}]$ actually converges for some $t \neq 0$. Later on (on the section on characteristic functions) we will discuss what can be done otherwise. For now, we will simply assume that the random variable X has moments of all orders, and it follows that $M_X(t)$ is well-defined (the diligent reader will prove this using monotonicity of the p -norm $\|\cdot\|_p$).

The moment generating function satisfies the following very useful identities, concerning convolution (sum of independent variables) and scaling (multiplication by a constant):

$$\begin{aligned} M_{X+Y}(t) &= \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX} e^{tY}] = M_X(t)M_Y(t), \\ M_{cX}(t) &= \mathbb{E}[e^{tcX}] = M_X(ct). \end{aligned}$$

For the first identity, X and Y must be independent of course.

4 Example: Bernoulli and Poisson

A Bernoulli random variable $\text{Ber}(p)$ is 1 with probability p and 0 otherwise. A binomial random variable $\text{Bin}(n, p)$ is the sum of n independent $\text{Ber}(p)$ variables.

Let us find the moment generating functions of $\text{Ber}(p)$ and $\text{Bin}(n, p)$. For a Bernoulli random variable, it is very simple:

$$M_{\text{Ber}(p)} = (1 - p) + pe^t = 1 + (e^t - 1)p.$$

A binomial random variable is just the sum of many Bernoulli variables, and so

$$M_{\text{Bin}(n,p)} = (1 + (e^t - 1)p)^n.$$

Now suppose $p = \lambda/n$, and consider the asymptotic behavior of $\text{Bin}(n, p)$:

$$M_{\text{Bin}(n,\lambda/n)} = \left(1 + \frac{(e^t - 1)\lambda}{n}\right)^n \longrightarrow e^{\lambda(e^t - 1)}.$$

As the reader might know, $\text{Bin}(n, p) \longrightarrow \text{Poisson}(\lambda)$, where the Poisson random variable is defined by

$$\Pr[\text{Poisson}(\lambda) = n] = e^{-\lambda} \frac{\lambda^n}{n!}.$$

Let us calculate the moment generating function of $\text{Poisson}(\lambda)$:

$$M_{\text{Poisson}(\lambda)}(t) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n e^{tn}}{n!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}.$$

This is hardly surprising. In the section about characteristic functions we show how to transform this calculation into a bona fide proof (we comment that this result is also easy to prove directly using Stirling's formula).

5 Cumulants

We are now almost ready to present our first proof. We first define the cumulant generating function of a random variable X :

$$K_X(t) = \log M_X(t).$$

This somewhat strange definition makes more sense once we notice that $M_X(t) = 1 + O(t)$, so that it makes sense to take its logarithm. In fact, using the Taylor series of $\log(1 + x)$,

$$\log(1 + t) = t - \frac{t^2}{2} + \dots$$

we can expand $K_X(t)$ as a power series, which begins as follows:

$$\begin{aligned} K_X(t) &= \left(\mathbb{E}[X]t + \frac{\mathbb{E}[X^2]}{2}t^2 + \dots \right) - \frac{(\mathbb{E}[X]t + \dots)^2}{2} + \dots \\ &= \mathbb{E}[X]t + \frac{\mathbb{E}[X^2] - \mathbb{E}[X]^2}{2}t^2 + \dots \end{aligned}$$

Hence the first two coefficients of $K_X(t)$ (as an exponential generating function, that is disregarding the $1/n!$ factors) are the expectation and the variance. We call these coefficients *cumulants*. Formally, we can define the n th cumulant $K_n[X]$ as follows:

$$K_n[X] = K_X^{(n)}(0).$$

In particular, we have just shown that

$$K_0[X] = 0, \quad K_1[X] = \mathbb{E}[X], \quad K_2[X] = \mathbb{V}[X].$$

In general, using the Taylor series of $\log(1 + x)$, we can express $K_n[X]$ as a polynomial in the moments. Conversely, using the Taylor series of e^x we can express the moments as polynomials in the cumulants. This provides an example of Moebius inversion (in lattices!), which alas we do not explain here. The moral is that we can rephrase the proof below completely in terms of moments, although it wouldn't make as much sense!

We are finally ready to give the proof, which is extremely simple. First notice that the formulas for scaling and convolution extend to cumulant generating functions as follows:

$$K_{X+Y}(t) = K_X(t) + K_Y(t), \quad K_{cX}(t) = K_X(ct).$$

Now suppose X_1, \dots are independent random variables with zero mean. Hence

$$K_{\frac{X_1 + \dots + X_n}{\sqrt{n}}}(t) = K_{X_1}\left(\frac{t}{\sqrt{n}}\right) + \dots + K_{X_n}\left(\frac{t}{\sqrt{n}}\right).$$

Rephrased in terms of the cumulants,

$$K_m \left[\frac{X_1 + \cdots + X_n}{\sqrt{n}} \right] = \frac{K_m[X_1] + \cdots + K_m[X_n]}{n^{m/2}}.$$

Note that $K_1[X_k] = 0$, so the first cumulant doesn't blow up. The second cumulant, the variance, is simply averaged. What happens to all the higher cumulants? If the cumulants are bounded by some constant C , then for $m > 2$,

$$K_m \left[\frac{X_1 + \cdots + X_n}{\sqrt{n}} \right] \leq \frac{nC}{n^{m/2}} \longrightarrow 0.$$

In other words, all the higher cumulants disappear in the limit! Thus the *cumulants* of the normalized sums tend to the cumulants of some fixed distribution, which must be the normal distribution!

In order to get convinced that the limit cumulant generating function, which is of the form $\frac{\sigma^2}{2}t^2$, indeed corresponds to a normal distribution, we can explicitly calculate the cumulant generating function of a normal variable (this is a simple exercise). Conversely, note that the limiting distribution does not depend on the distributions we started with. In particular, if we start with normally distributed X_i , notice that $(X_1 + \cdots + X_n)/\sqrt{n}$ will always be normal. This argument shows that $\frac{\sigma^2}{2}t^2$ *must* be the cumulant generating function of $\mathcal{N}(0, \sigma^2)$!

Let's see what we proved and what's missing. We proved that the cumulant generating function of the normalized sum tends to the cumulant generating function of a normal distribution with zero mean and the correct (limiting) variance, all under the assumption that the cumulants are bounded. This is satisfied whenever the moments are bounded, for example when all variables are identically distributed. However, we're really interested in proving convergence *in distribution*. The missing ingredient is Lévy's continuity theorem, which will be explained (without proof) in the next section.

6 Characteristic Functions

In this section we both indicate how to complete the proof of the central limit theorem, and explain what to do when the moment generating function is not well defined.

The moment generating function is not always defined, since the implicit integral $\mathbb{E}[e^{tX}]$ need not converge, in general. However, the following trick

will ensure that the integral will always converge: simply make t imaginary! This prompts the definition of the characteristic function

$$\varphi_X(t) = \mathbb{E}[e^{itX}].$$

Here the integrand is bounded so the integral always converges (since $\mathbb{E}[1] = 1$). The astute reader will notice that φ_X is just the Fourier transform of X (in an appropriate sense). Therefore, we would expect that convergence in terms of characteristic functions implies convergence in distribution, since the inverse Fourier transform is continuous. This is just the contents of Lévy's continuity theorem! Hence, to make our proof completely formal, all we need to do is make the argument t imaginary instead of real.

The classical proof of the central limit theorem in terms of characteristic functions argues directly using the characteristic function, i.e. without taking logarithms. Suppose that the independent random variables X_i with zero mean and variance σ^2 have bounded third moments. Thus

$$\varphi_{X_i}(t) = 1 - \frac{\sigma^2}{2}t^2 + O(t^3).$$

Using the identities for the moment generating function,

$$\varphi_{\frac{X_1 + \dots + X_n}{\sqrt{n}}} = \left(1 - \frac{\sigma^2}{2n}t^2 + O\left(\frac{t^3}{n^{3/2}}\right)\right)^n \longrightarrow e^{-\frac{\sigma^2}{2}t^2}.$$

The righthand side is just the characteristic function of a normal variable, so the proof is concluded with an application of Lévy's continuity theorem.

7 Moments of the Normal Distribution

The next proof we are going to describe has the advantage of providing a combinatorial explanation for the values of the moments of a normal distribution. In this section, we will calculate these very moments.

Calculating the moments of a normal distribution is easy. The only thing needed is integration by parts. We will concentrate on the case of zero mean and unit variance. Notice that

$$\begin{aligned} \int_{-\infty}^{\infty} x^n e^{-\frac{x^2}{2}} dx &= \frac{x^{n+1}}{n+1} e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{x^{n+1}}{n+1} \left(-x e^{-\frac{x^2}{2}}\right) dx \\ &= \frac{1}{n+1} \int_{-\infty}^{\infty} x^{n+2} e^{-\frac{x^2}{2}} dx. \end{aligned}$$

In terms of moments, we get the recurrence relation

$$M_n = \frac{M_{n+2}}{n+1} \implies M_{n+2} = (n+1)M_n.$$

Since $M_0 = 1$ and $M_1 = 0$ we get that all odd moments are zero (this happens because the distribution is symmetric), and the even moments are

$$M_n = (n-1)M_{n-2} = \cdots = (n-1)(n-3)\cdots 1.$$

The next proof will explain what these numbers stand for.

8 Proof using Moments

In this proof we will bravely compute the limit of the moments of $Y_n = (X_1 + \cdots + X_n)/\sqrt{n}$. For simplicity, we assume that the variables X_i are independent with zero mean, unit variance and bounded moments. The proof can be adapted to the case of varying variances.

It is easy to see that $M_1[Y_n] = 0$. What about the second moment?

$$\begin{aligned} M_2[Y_n] &= \mathbb{E} \left[\frac{(X_1 + \cdots + X_n)^2}{n} \right] \\ &= \frac{\sum_i \mathbb{E}[X_i^2]}{n} + \frac{\sum_{i \neq j} \mathbb{E}[X_i X_j]}{n} = 1. \end{aligned}$$

Calculation was easy since $\mathbb{E}[X_i^2] = 1$ whereas $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i] \mathbb{E}[X_j] = 0$. Now let's try the third moment, assuming $M_3[X_i] \leq C_3$:

$$\begin{aligned} M_3[Y_n] &= \mathbb{E} \left[\frac{(X_1 + \cdots + X_n)^3}{n^{3/2}} \right] \\ &= \frac{\sum_i \mathbb{E}[X_i^3]}{n^{3/2}} + 3 \frac{\sum_{i \neq j} \mathbb{E}[X_i^2 X_j]}{n^{3/2}} + \frac{\sum_{i \neq j \neq k} \mathbb{E}[X_i X_j X_k]}{n^{3/2}} \leq \frac{nC_3}{n^{3/2}} = \frac{C_3}{\sqrt{n}}. \end{aligned}$$

Thus the third moment tends to zero.

The fourth moment brings forth a more interesting calculation:

$$\begin{aligned}
M_4[Y_n] &= \mathbb{E} \left[\frac{(X_1 + \cdots + X_n)^4}{n^2} \right] \\
&= \frac{\sum_i \mathbb{E}[X_i^4]}{n^2} + 4 \frac{\sum_{i \neq j} \mathbb{E}[X_i^3 X_j]}{n^2} + 3 \frac{\sum_{i \neq j} \mathbb{E}[X_i^2 X_j^2]}{n^2} \\
&\quad + 6 \frac{\sum_{i \neq j \neq k} \mathbb{E}[X_i^2 X_j X_k]}{n^2} + \frac{\sum_{i \neq j \neq k \neq l} \mathbb{E}[X_i X_j X_k X_l]}{n^2} \\
&= O(n^{-2}) + 3 \frac{n(n-1)}{n^2} \longrightarrow 3.
\end{aligned}$$

What a mess! In fact, what we were doing is classifying all terms of length 4. These come in several kinds (replacing X_i with i):

$$\begin{array}{cccc}
i & i & i & i \\
i & i & i & j \quad (\times 4) \\
i & i & j & j \quad (\times 3) \\
i & i & j & k \quad (\times 6) \\
i & j & k & l
\end{array}$$

For example, the interesting term ijj comes in these varieties:

$$\begin{array}{cccc}
i & i & j & j \\
i & j & i & j \\
i & j & j & i
\end{array}$$

All terms which contain a singleton (a variable appearing only once) equal zero and can be dropped. Of these, the term corresponding to X_i^4 is asymptotically nil, and the term corresponding to $X_i^2 X_j^2$ is asymptotically 3, since the extra condition $i \neq j$ becomes insignificant in the limit.

We can now explain the general case. In the calculation of the m th moment, we need to deal with terms of length m . We can identify each term with a “form” similar to the ones given earlier, as follows. We go over all the factors, and let the r th *unique* factor get the appellation r (this is just what we did before, with i, j, k, l for 1, 2, 3, 4). Each term containing a singleton is identically zero. The contribution of a term with t variables with multiplicities m_1, \dots, m_t is at most

$$\frac{n^{m_1 + \cdots + m_t}}{n^{m/2}} C_{m_1} \cdots C_{m_t},$$

where C_s is a bound on $\mathbb{E}[X_i^s]$. Thus the term is asymptotically nil if $m_1 + \dots + m_t < m/2$. If $m_1 + \dots + m_t \geq m/2$, then since $m_i \geq 2$ (otherwise the term is identically nil) we see that $t = m/2$ and $m_i = 2$. In that case, the contribution of the term is

$$\frac{n(n-1)\cdots(n-m/2+1)}{n^{m/2}} \longrightarrow 1,$$

since the random variables have unit variance. Thus the m th moment converges to the number of such terms! (note that the number of asymptotically nil terms is finite)

If m is odd then clearly, there are no such terms, hence the moment is asymptotically zero. If m is even then the number of terms can be evaluated combinatorially. The term must begin with X_1 . There are $m-1$ possible positions of the other X_1 . Given its position, the first available slot must be X_2 . Its counterpart must be in one of the $m-3$ available positions, and so on. Thus the number of possible terms is

$$(m-1)(m-3)\cdots 1,$$

which is just the formula we obtained in the previous section!

9 Bonus: Number of Prime Divisors

In this section we adapt the proof in the previous subsection to a number-theoretic setting, by calculating the average number of prime factors of a number.

Let $c(n)$ be the number of prime factors of n (if n is divisible by a prime power p^d then we count it only once; counting it d times will lead to very similar results). We will show that $c(n)$ is asymptotically normally distributed. By that we mean that if X_n is a uniformly random integer between 1 and n , then $c(X_n)$ is close to a normal distribution, with parameters that we will calculate.

The quantity $c(n)$ can be written in the following form, where $[P]$ is the Iverson bracket (equal to 1 if P holds):

$$c(n) = \sum_p [p \mid n].$$

The sum is taken over all *primes*. Thus, if we denote $I_p = [p \mid X_n]$, we get

$$c(X_n) = \sum_{p \leq n} I_p.$$

The random variables I_p are *almost* independent. Bounding their dependence will allow us to conclude that $c(X_n)$ is asymptotically normally distributed.

Notice that the I_p are all Bernoulli variables. Thus for $m_i > 0$ we have

$$\mathbb{E}[I_{p_1}^{m_1} \cdots I_{p_t}^{m_t}] = \frac{\lfloor n/p_1 \cdots p_t \rfloor}{n} = \frac{1}{p_1 \cdots p_t} + O(1/n).$$

We now begin calculating the moments of $c(X_n)$. The expectation $\mathbb{E}[c(X_n)]$ is given by

$$\begin{aligned} \sum_{p \leq n} \mathbb{E}[I_p] &= \sum_{p \leq n} \frac{1}{n} \left\lfloor \frac{n}{p} \right\rfloor \\ &= \sum_{p \leq n} \frac{1}{p} + O\left(\frac{1}{\log n}\right) = \log \log n + B_1 + O\left(\frac{1}{\log n}\right). \end{aligned}$$

We have used the well known estimate for $\sum_{p \leq n} 1/p = \log \log n + B_1 + O\left(\frac{1}{\log n}\right)$, which can be recovered (non-rigorously and without the constant and the error term) by using the approximate parametrization $p_t = t \log t$. The constant $B_1 \approx 0.2615$ is Mertens' constant.

The second moment is given by

$$\begin{aligned} \sum_{p \leq n} \mathbb{E}[I_p^2] + \sum_{p \neq q \leq n} \mathbb{E}[I_p I_q] &= \sum_{p \leq n} \frac{1}{p} + \sum_{p \neq q \leq n} \frac{1}{pq} \\ &= \sum_{p \leq n} \frac{1}{p} \left(1 + \sum_{q \leq n} \frac{1}{q} - \frac{1}{p}\right) \\ &= \mathbb{E}[c(X_n)]^2 + \sum_{p \leq n} \frac{1}{p} - \sum_{p \leq n} \frac{1}{p^2}. \end{aligned}$$

Since the series $1/n^2$ converges, we get that the variance is

$$\mathbb{V}[c(X_n)] = \mathbb{E}[c(X_n)] + O(1) = \log \log n + O(1).$$

In fact, if we want we can be more accurate:

$$\mathbb{V}[c(X_n)] = \log \log n + B_1 - \sum_{p=1}^{\infty} \frac{1}{p^2} + O\left(\frac{1}{\log n}\right),$$

since the error in computing the infinite sum $\sum p^{-2}$ is only $O(1/n)$.

Before we calculate any further moments, we would like to normalize the indicators. Let $J_p = I_p - \mathbb{E}[I_p]$. Expanding the expectation of the product, with k_i being the power of $\mathbb{E}[I_i]$, we get

$$\begin{aligned} \mathbb{E}\left[\prod_{s=1}^t J_{p_s}^{m_s}\right] &= \sum_{k_1=0}^{m_1} \cdots \sum_{k_t=0}^{m_t} (-1)^{\sum_{s=1}^t k_s} \prod_{s=1}^t \binom{m_s}{k_s} \prod_{s=1}^t \mathbb{E}[I_{p_s}]^{k_s} \mathbb{E}\left[\prod_{s=1}^t I_{p_s}^{m_s - k_s}\right] \\ &= \sum_{k_1=0}^{m_1} \cdots \sum_{k_t=0}^{m_t} (-1)^{\sum_{s=1}^t k_s} \prod_{s=1}^t \binom{m_s}{k_s} \prod_{s=1}^t p_s^{-k_s} \prod_{k_s < m_s} p_s^{-1} + O(1/n). \end{aligned}$$

If $m_s = 1$ then the two summands cancel exactly, and so the entire expectation is $O(1/n)$. Therefore we can assume that all $m_s \geq 2$.

When calculating the moments, we are going to sum over all t -tuples of different primes smaller than n . Letting $m = \sum_{s=1}^t m_s$, we are going to divide by $\mathbb{V}[c(X_n)]^{m/2}$, so every term that is asymptotically smaller is negligible. Therefore, a natural next step is to estimate the following sum:

$$\sum_{p_1 \neq \dots \neq p_t} \prod_{s=1}^t p_i^{-\mu_i} \leq \sum_{p_1, \dots, p_t} \prod_{s=1}^t p_i^{-\mu_i} = O(\mathbb{E}[c(X_n)]^{\{i: \mu_i=1\}}).$$

This is negligible unless the number of μ_i s which equal 1 is at least $m/2$.

Considering now the previous expression for the expectation of a product of J_{p_s} , unless $t \geq m/2$ the term is negligible. If $t \geq m/2$ then since $m_s \geq 2$, we see that in fact $t = m/2$ and $m_s = 2$; in particular, m must be odd. By considering the $m = 2$ case, we get that the contribution of any such term tends to $\mathbb{V}[c(X_n)]^{m/2}$ (the loss due to the requirement that the primes be different tends to zero). As in the proof of the central limit theorem in the previous section, there are $(m-1)(m-3)\cdots 1$ such terms, and so all the moments tend to the moments of a normal distribution with stated expectation and variance. Hence the number of prime factors is asymptotically normal.

We leave the reader to verify that even if we count prime factors according to multiplicity, the result is still asymptotically normal, albeit with a slightly larger expectation and variance. The main point is that the total expected contribution of the higher powers to the sum is $O(1)$.

10 Sources

Both proofs are taken from the nice book *The Semicircle Law, Free Random Variables, and Entropy*. This book is actually about free probability, which is an example of non-commutative probability. What can be non-commutative about probability?

To answer this intriguing question, let us notice that so far, we have virtually identified a distribution with its moments. Now suppose X is a random matrix. The expectations $\mathbb{E}[\text{Tr } X^n]$ give the empirical distribution of the eigenvalues of X , since denoting the eigenvalues by λ_i ,

$$\mathbb{E}[\text{Tr } X^n] = \sum \lambda_i^n.$$

Let us now define expectation to always include the trace. We identify a distribution with the moments $\mathbb{E}[X^n]$ of a random variable having this distribution. Independent random matrices X, Y do not necessarily satisfy $\mathbb{E}[XY] = \mathbb{E}[YX]$. However, if X and Y are very large then we do asymptotically have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. This, along with similar facts, is expressed by saying that independent random matrices are *asymptotically free*. By presenting a large random matrix as a sum of many random matrices (using properties like the additivity of the normal distribution, according to which the entries might be distributed), we get, using very similar methods, the semicircle law, which gives the asymptotic empirical distribution of eigenvalues; the limiting density is a semicircle!

The counterpart of the cumulant generating function is called the R-transform. Similar relations between moments and cumulants are true also in the free setting, but they are different, having something to do with non-crossing partitions (in our case, the relevant lattice is that of all partitions). For all this and more, consult the book or the works of Alexandru Nica and Roland Speicher.

We comment that old-style proofs of the semicircle law massage an exact expression for the density of the eigenvalues. This corresponds to proving the central limit theorem for binomial random variables by using the exact probabilities and Stirling's formula,

Finally, the number-theoretic example was suggested by a lecture of Balázs Szegedy (the theorem itself appears in classic books on number theory like Hardy and Wright).