

The ubiquitous Gaussian limit law in analytic combinatorics

HSIEN-KUEI HWANG

Institute of Statistical Science, Institute of Information Science
Academia Sinica, Taipei 115, Taiwan

April 22, 2012

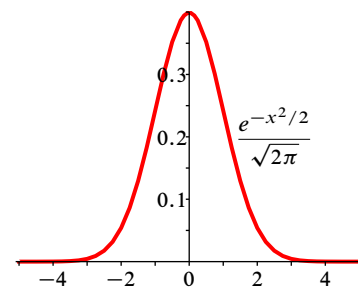
*"...tous les effets de la nature, ne sont que
les résultats mathématiques d'un petit
nombre de lois immuables."
--- P. S. Laplace*

Data is everywhere, notably in this information explosion era. Beyond the first-order summary of a sample by its average value or its median, the bell-shaped Gaussian (or normal) curve has long been observed in the histograms of many data samples since the early history of probability, statistics and related fields. The Gaussian distribution, known commonly under the name of *Law of Frequency of Errors*, first appeared in 1733 in de Moivre's works (see his *Doctrine of Chance*, second edition, first published in 1738, and [9]) as the limit of suitably normalized binomial distributions; see Figure 1. The following intuitive arguments of G. Galilei (see [9, Ch. 1]) may be regarded as the first-level description of the Law:

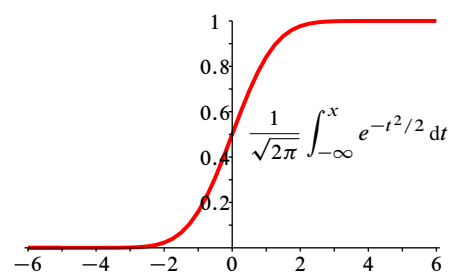
"...random errors are inevitable in instrumental observations, that measurements are equally prone to err in one direction or the other, and that the majority of observations tend to center around the mean value."

While intensive recent attention has been paid on complex networks having *the power law* as one of the omnipresent characteristics, the importance and power of the Gaussian distribution in modeling and predicting theoretical and practical situations should not be underestimated. Indeed, as Francis Galton put it (see [4]):

"The law would have been personified by the Greeks and deified, if they had known of it."



The Gaussian (normal) density.



The Gaussian distribution function.

The prototype source of the normal distribution is the classical central limit theorem (CLT), first given by Laplace in his *Théorie Analytique des Probabilités* in the beginning of the nineteenth century.

Classical CLT: *Given a sequence of independent and identically distributed random variables $\{X_j\}_{j \geq 1}$, if $0 < \sigma^2 := \mathbb{V}(X_1) < \infty$ exists, then the distributions of the sums $S_n = \sum_{1 \leq j \leq n} X_j$, when centered and normalized, are asymptotically Gaussian (or normal) in the following sense*

$$\mathbb{P} \left(\frac{S_n - \mu n}{\sigma \sqrt{n}} < x \right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt, \quad (1)$$

as $n \rightarrow \infty$, for $x \in \mathbb{R}$, where $\mu := \mathbb{E}(X_1)$.

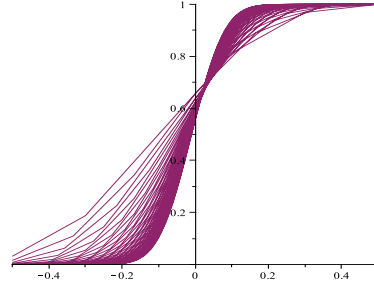


Figure 1: Binomial to Gaussian.

Laplace’s analytic realization via characteristic functions of this classical “Law of Frequency of Errors” represents (see [11, p. 283]) “the crowning achievement of his persistent efforts, extending over a period of more than twenty years” (almost forty years, according to [3, p. 18]).

The early two-century history (from 1730’s to 1930’s) of the central limit theorem is tersely and amusingly summarized by Le Cam (see [8]):

In the beginning there was de Moivre, Laplace, and many Bernoullis, and they beget limit theorems, and the wise men saw that it was good and they called it by the name of Gauss. Then there were new generations and they said that it had experimental vigor but lacked in rigor. Then came Chebyshev, Liapounov, and Markov and they beget a proof and Pólya saw that it was momentous and he said that its name shall be called the Central Limit Theorem. . . .

For more detailed information on the historical aspects of the CLT, see the books [1, 3, 5, 9, 10].

The final form of the *classical* CLT, from its first primitive version by Laplace in 1810 to its final iff-version for sums of independent random variables by Lindeberg (sufficiency in 1922), Feller and Lévy (necessity in 1935), also marks the beginning of the modern dominant *measure-theoretic approach* in probability theory and related fields. In contrast, the *classical analytic approach*, based either on characteristic functions or on the method of moments, has become almost obsolete, and appeared only sporadically in the literature. It is from this methodological perspective that Flajolet’s works stand out in the modern literature on random discrete structures, notably through his persistent use and developments of *general analytic schemes* (see below for more details). We aim here to highlight, through a brief analysis and a simple classification of his works on limit theorems, his contribution in random combinatorial structures and related areas, focusing on one of the central aspects—the Gaussian limit law in analytic combinatorics.

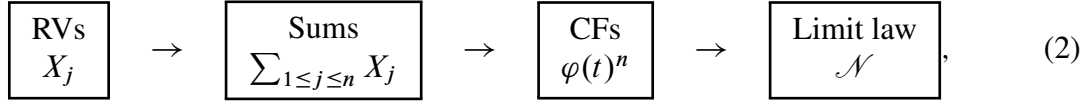
One standard proof of (1) follows from the line of *characteristic functions*, first introduced by Laplace and then popularized by Lyapunov: Define $\varphi(t) := \mathbb{E}(e^{itX_1})$. Then $\mathbb{E}(e^{itS_n}) = \varphi^n(t)$. The asymptotic relation (1) results from a Taylor expansion of second order

$$\mathbb{E} \left(e^{it(S_n - \mu n)/(\sigma \sqrt{n})} \right) = e^{-\mu \sqrt{nit}/\sigma} \varphi \left(\frac{t}{\sigma \sqrt{n}} \right)^n = e^{-t^2/2(1+o(1))},$$

and Lévy’s continuity theorem:

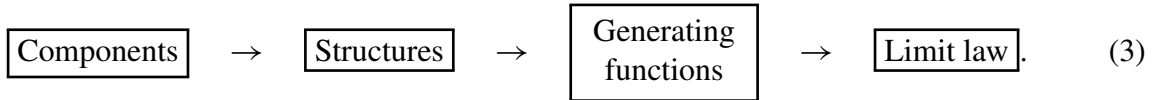
If the sequence of characteristic functions $\varphi_n(t) = \mathbb{E}(e^{itX_n})$ tends to $\varphi(t) = \mathbb{E}(e^{itX})$ as $n \rightarrow \infty$, and if $\varphi(t)$ is continuous at $t = 0$, then the distribution function of X_n converges to that of X .

We can schematize the above approach as follows.



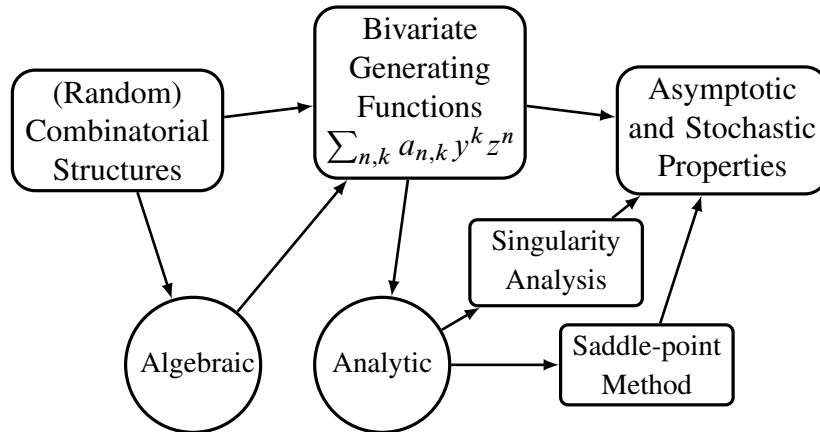
which, although simple enough, leads to fruitful extensions of many combinatorial distributions. Here RV denotes “random variable” and CF is the abbreviation of “characteristic function”.

For example, an intuitive “translation” of the above flow pattern (2) to combinatorial structures goes as follows.



Roughly, we expect that each component (in generic sense) “mimics” the rôle of an individual random variable and the parameter or characteristic of interest in the structures that of the “sums”. (Note that the components are generally not independent.) Since the Gaussian law is pervasive in the situations of large sums of small RVs, one expects that such a “normal” law has a similar central rôle in analytic combinatorics.

More precisely, this translation leads indeed to a paradigm for random combinatorial structures, briefly summarized in the following diagram.

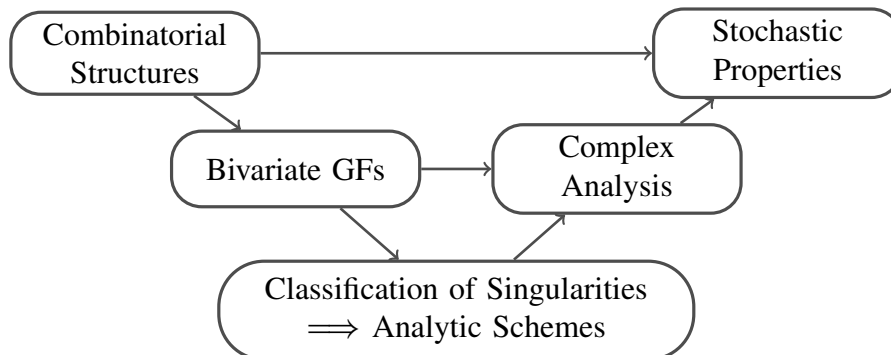


Such a simple diagram is thoroughly discussed in Chapter IX of his authoritative book, jointly written with Sedgewick, “Analytic Combinatorics” [PF201]. The general question of interest here is roughly as follows. To each double-indexed nonnegative sequence $\{a_{n,k}\}$, which often enumerates the number (weighted or unweighted) of combinatorial structures of size n with k “components” (or other parameters), we associate the sequence of random variables

$$\mathbb{P}(X_n = k) := \frac{a_{n,k}}{\sum_j a_{n,j}},$$

and study the stochastic properties of X_n . A general analytic approach consists in considering the bivariate generating function $f = f(z, y) := \sum_{n,k} a_{n,k} y^k z^n$ and deducing the required asymptotic properties by analytic properties of the function f , regarded as an analytic function in the complex (z, y) -planes.

Once this paradigm between the connection of random combinatorial structures and their stochastic behaviors becomes clear through the use of generating functions, one can then focus on classifying the generating functions, which are themselves analytic functions in almost all cases of interest, leading to the systematic study of *schemas* (or *analytic schemes*) fully explored in [PF201, Ch. IX]; see the following diagram.



In particular, in the case of asymptotic normality, nine different schemas of diverse nature and degree of generality (see Table 3 below) are proposed and discussed with great length, largely extending many previous works including particularly his own papers [PF088], [PF112], [PF117] and building firmly a stochastic theory via analytic combinatorics.

As described in [PF201, P. 12]:

A parameter of a combinatorial class is fully determined by a bivariate generating function, which is a deformation of the basic counting generating function of the class ... Then, the asymptotic distribution of a parameter of interest is characterized by a collection of surfaces, each having its own singularities. The way the singularities' locations move or their nature changes under deformation encodes all the necessary information regarding the distribution of the parameter under consideration. Limit laws for combinatorial parameters can then be obtained and the corresponding phenomena can be organized into broad categories, called schemas. It would be inconceivable to attain such a far-reaching classification of metric properties of combinatorial structures by elementary real analysis alone.

It is from this methodological perspective and his systematic developments that we see how Flajolet succeeded in building up a modern theory of limit theorems based on classical analytic tools. While the notion of analytic schemes is not new and finds its roots in many early papers on analytic number theory and integer partitions, his unprecedented thoroughgoing investigation using powerful analytic tools institutes the foundation for analytic combinatorics.

Among all Flajolet's published papers, there are about two dozens where Gaussian limit law explicitly appears or is proved, and diverse techniques are developed or theorized. The majority of random variables studied follow an asymptotically Gaussian distribution of the form $\mathcal{N}(cn, c'n)$ with linear mean and linear variance, conforming to some extent the analogy

Paper	Parameters of Structures	Equation	Tool
[PF045]	size of trees of given height	$f_{k+1}(z) = \text{Poly}(z, f_k(z))$	SPM
[PF047]	wagons in trains	$g(z)h(z)^k$	SPM
[PF076]	local discrepancy of seqs.	$\sum_{0 \leq j \leq n} X_{n,j}$	MM
[PF088]	components in set (multiset)	$(1 - z/\rho)^{-y} g(z, y)$	SA
[PF096]	depth of increasing trees	$g(z)^y \int_0^z g(t)^{1-y} dt$	SA
–	leaves of increasing trees	$\int_0^f \frac{dt}{(y-1)\phi_0 + \sum_j \phi_j t^j} = z$	SA
[PF107]	coeffs. of polynomials	$\frac{\log \frac{1}{1-z}}{z(1-y \log \frac{1}{1-z})}$	SA
[PF112]	algebraic-logarithmic	$\left(\frac{1}{1-yg(z)}\right)^\alpha \left(\log \frac{1}{1-yg(z)}\right)^k$	SA
[PF105] [PF115]	cost of mergesort	$X_n \stackrel{d}{=} X_{\lfloor n/2 \rfloor} + X_{\lceil n/2 \rceil} + T_n$	$\sum \text{RV}_i$
[PF117]	depth of quadrees	$(z(1-z)\mathbb{D})^d (f-g) = 2^d yf$	SA
–	a DE schema	$\sum_{0 \leq j \leq r} \frac{\varphi_j}{(1-z)^j} f^{(r-j)} = 0$	SA
[PF135]	patterns in BSTs	$f' = f^2 + g$	SA
[PF142]	cost of hashing	$g(z, y)^{cn}$	SPM
[PF149]	parameters of non-crossing structures	$\sum_{0 \leq j \leq r} \varphi_j f^j = 0$	SA

Table 1: *Gaussian limit laws in Flajolet’s published papers before 2000. Here MM denotes “method of moments”, and BST “binary search tree”. For convenience, f is the abbreviation of $f(z, y)$ and the φ_j ’s are analytic functions of z and y .*

between the two concept patterns (2) and (3); see also Table 3 for a list of more general schemas. In words, we may say that *the standard combinatorial constructions (such as sequence, set, etc.), tend to build structures that have regularly behaved components, generating the Gaussian law.*

Instead of classifying the diverse CLTs by the order of the mean and that of the variance, we provide a different means based on the major analytic techniques used: singularity analysis (SA), saddle-point method (SPM) and other approaches; see Tables 1 and 2 and Figure 2. Note that this simple classification is completely general and not limited to Flajolet’s works.

Tables 1 and 2 summarize all Gaussian limit laws appearing in Flajolet’s published papers, where we see not only the diversity of the random structures studied (from algorithmics and combinatorics), but also, through the wide spectrum of analytic forms, the power of his analytic knowledge in solving these problems and in establishing an “analytic universe”. Note that the classification by the major techniques used is not unique and serves mostly for a better synopsis of his works. Indeed, many analytic problems solvable by SA can be changed into a problem suitable for SPM by a change of variables, or particularly Lagrange’s inversion formula.

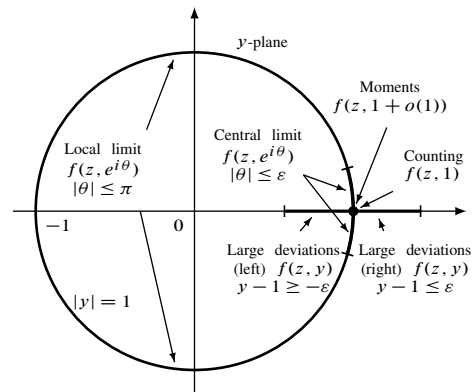
An important notion to be mentioned here is the *quasi-power approximation* (another term coined by Flajolet; see [PF201], [6]), which refers to the property that the moment probability generating function satisfies asymptotically $g(s)e^{\lambda_n h(s)}$, where $\lambda_n \rightarrow \infty$ (not necessarily integers) with n , and g, h are analytic functions for sufficiently small $s \in \mathbb{C}$. The major characteristic of such an estimate, as mostly resulted from SA or SPM in the case of large powers of

Paper	Parameters of Structures	Equation	Tool(s)
[PF155]	crossings in chord diagrams	$\sum_{k=-n}^n \frac{2^n n! (-1)^k y^{k(k-1)/2}}{(n+k)!(n-k)!(1-y)^n}$	SPM
[PF159]	cost of radix selection	$f = be^{yz(b-1)/b} f\left(\frac{yz}{b}, y\right) + z(1-y)$	QP
–	cost of distributive sort	$g(z, y)^n$	SD
–	cost of radix sort	$f = f^b\left(\frac{yz}{b}, y\right) + z(1-y)$	SD
[PF152] [PF160]	composition scheme	$g(yh(z))$ or $h^k(z)$	SD
[PF168]	final altitude of meander	$\frac{g(z, y)}{1-yh(z)}$	SA
[PF151] [PF174]	motifs in text	$\frac{g(z, y)}{\det(I_{m \times m} - zh(y))}$	SG
[PF176] [PF193]	probabilistic counters	$\frac{1}{m} \sum_{1 \leq j \leq m} C_{m, j}$	$\sum RV_i$
[PF183] [PF186]	balls in urns	$(1 - sz y^{b+s}) f'_z + (y^{b+s+1} - y^{1-a}) f'_u = t_0 y^{b+s} f$	PDE SA
[PF164] [PF191]	hidden pattern in texts	$\sum_j X_{n, j}$	MM
–	(fully constrained)	$\mathbf{g}(y)^t \mathbf{h}(y)^{n-c} \mathbf{1}$	QP
[PF198]	symmetries in phylogenetic trees	$f = z + \frac{f^2}{2} + (y - \frac{1}{2}) f(z^2, y^2)$	SA

Table 2: *Gaussian limit laws in Flajolet’s published papers after 2000. Here PDE denotes “partial differential equation” and QP “quasi-power theorem”. Again, f without parameters denotes $f(z, y)$.*

generating functions, is that the asymptotic approximation holds uniformly for s lying in some neighborhood of the origin, implying more analytic techniques can be applied for obtaining the properties of interest.

Flajolet was one of the few who really managed to prove the local limit theorem, which provides a theoretical justification closer to what one often draws for discrete distributions. Papers where local limit theorems are proved include [PF047] (wagons in trains), [PF135] (patterns in BSTs), [PF151], [PF174] (motifs in texts), [PF160] (planar maps), [PF164], [PF191] (texts), [PF198] (phylogenetic trees). An application of the SPM is needed for which the hard part always lies in obtaining a uniform estimate of $|f(z, e^{i\theta})|$ (or that of $|\mathbb{E}(e^{X_n i\theta})|$) for $z \sim 1$ and $|\theta| \leq \pi$.



The correspondence between the stochastic properties of the random variables in question and the corresponding bivariate GF; this is essentially Figure IX.9 (p. 649) of [PF201].

While most analytic schemes seem centering around the use of SA, one should note that

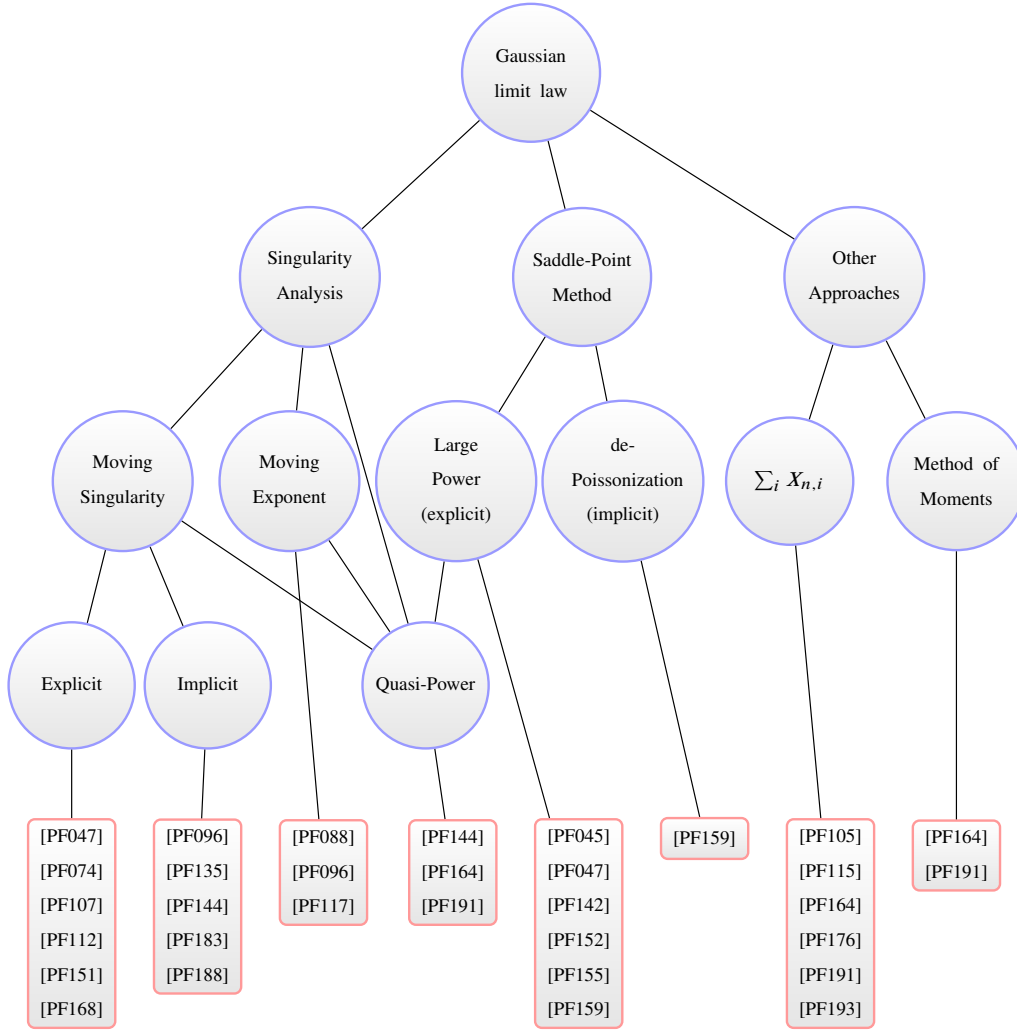


Figure 2: A classification of all Gaussian limit laws appearing in Flajolet's published papers. Following the treatment in [PF201, Ch. IX], all CLTs established by SA are divided into two subclasses: moving singularity and moving exponent. Moving singularity refers to a singularity of the type $g(y)(1 - z/\rho(y))^{-\alpha}$, implying an asymptotic estimate of the form $\tilde{g}(y)(\rho(y)/\rho(1))^{-n}$ for y near unity, which then leads to the asymptotic normality $\mathcal{N}(cn, c'n)$. On the other hand, moving exponent is connected to the local expansion $g(y)(1 - z/\rho)^{-\alpha(y)}$ near the dominant singularity, yielding the asymptotic estimate $\tilde{g}(y)n^{\alpha(y)-1}$ for $y \sim 1$, which in turn entails $\mathcal{N}(c \log n, c' \log n)$. In the case of CLT via SPM, we can further distinguish between large powers of generating functions and de-Poissonization, the functions in question in the former class being often explicit while those in the latter implicit.

the application of SPM consists itself of a process of Gaussian approximation (if the order of the saddle-point is two). SPM is applied in the papers [PF036], [PF037], [PF043], [PF045], [PF047], [PF069], [PF074], [PF078], [PF079], [PF080], [PF091], [PF094], [PF097], [PF121], [PF123], [PF124], [PF125], [PF135], [PF142], [PF156], [PF160], [PF162], [PF173], [PF179], [PF183], [PF191], [PF195], [PF197], [PF198], [PF206], [PF207], where they can be further

classified into large powers of functions, Mellin integrals and product forms, details are omitted here.

Other tools used for proving Gaussian limit laws in Flajolet’s papers include *sums of independent random variables* and *method of moments*. In addition to [PF164] and [PF191], the method of moments was also employed to establish several other non-normal limit laws: [PF033] (height of random trees), [PF114], [PF132] (lattice reduction), [PF142] (linear probing hashing), [PF165] (Airy), and [PF175] (hashing).

To further illustrate the ubiquity of Gaussian law in analytic combinatorics, we summarize in Table 3 all Gaussian schemas thoroughly expounded in [PF201, Ch. IX].

Theorem or Proposition	Page(s)	Schema	Analytic form	$\mathcal{N}(\mu_n, \sigma_n^2)$
Thm IX.8	645	Quasi-Power theorem	$\sim g(y)h(y)^{\alpha n}$	$\mathcal{N}(c\alpha_n, c'\alpha_n)$
Prop IX.6	650–651	Supercritical composition	$g(yh(z))$	$\mathcal{N}(cn, c'n)$
Thm IX.9	656	Meromorphic schema	$\frac{h(z,y)}{g(z,y)}$	$\mathcal{N}(cn, c'n)$
Thm IX.10	665	Systems of functional equations	linear systems	$\mathcal{N}(cn, c'n)$
Thm IX.11	669	Variable exponent perturbation	$\varphi_1 + \varphi_2 g(z)^{-\alpha(y)}$	$\mathcal{N}(cn, c'n)$
Thm IX.12	676	Algebraic singularity schema	$\varphi_1 + \varphi_2 g(z, y)^{-\alpha}$	$\mathcal{N}(cn, c'n)$
Prop IX.17	682	Perturbation of algebraic functions	$f = \text{Poly}(z, y; f)$	$\mathcal{N}(cn, c'n)$
Prop IX.18	685–686	Linear DEs	$\sum_{0 \leq j \leq r} \frac{\varphi_j f^{(r-j)}}{(\rho-z)^j} = 0$	$\mathcal{N}(c \log n, c' \log n)$
Thm IX.13	690	Generalized quasi-power theorem	$\sim e^{g_n(y)}$	$\mathcal{N}(g'_n(1), g'_n(1) + g''_n(1))$

Table 3: All analytic schemas in [PF201, Ch. IX] for Gaussian limit laws proved by singularity analysis and related techniques. Here φ_j ’s are analytic functions of z and y .

The seven papers grouped in this chapter are all covered in the above tables; their contents can be briefly summarized as follows.

- [PF045]: motivated by the study of the height of trees, Flajolet and Odlyzko studied the polynomial iteration $f_{k+1}(z) = \text{Poly}(z, f_k(z))$, and derived very precise estimates for $[z^n]f_k(z)$ when n is near cd^k for some structural constant c and d is the degree of the polynomial. The crucial estimate is the *uniform* quasi-power approximation $f_k(z) \sim g(z)h(z)^{d^k}$ for z lying in certain region.
- [PF088]: the $\exp(\log)$ -schema was proposed and examined in details with a large number of examples of the form $\mathcal{N}(c \log n, c' \log n)$. This paper has many follow-ups, including particularly the book [2] on logarithmic combinatorial structures; see also [7].
- [PF096]: increasing trees were proposed and systematically studied, centering on combinatorial, asymptotic and stochastic properties. An important feature is the contrast between

the order $\log n$ for the depth of increasing trees (with enumerating generating functions of the form $f' = \Psi(f)$) and \sqrt{n} for that of Meir and Moon's simply-generated family of trees (with enumerating generating functions of the form $f = z\Psi(f)$).

- [PF112]: the alg-log schema (see Table 2) was developed by SA, with a wide range of applications.
- [PF144]: a highly inspiring survey paper; continued fraction algorithms (including Euclidean algorithm and the lattice reduction algorithm of Gauss) are examined through the persistent use of transfer operators, yielding either stronger estimates or simpler proofs.
- [PF159]: different selection and sorting algorithms using bucketing techniques are analyzed.
- [PF198]: what is the probability that two randomly chosen phylogenetic trees of the same size are isomorphic? Very precise asymptotic expansion is derived by analytic-combinatorial techniques; as a byproduct, the number of symmetrical nodes in random phylogenetic trees is showed to follow not only a CLT but also an LLT.

Laplace, in his *Théorie Analytique des Probabilités*, wrote

... la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul ... ;

and, according to Fischer [3],

Laplace's accomplishments in probability theory ... as he was working entirely within the framework of classical probability theory to develop stochastics into a universal method to which all scientific fields could be made accessible.

...

His "analytical" probability theory already transcended the range of its applications due to the relevance of its mathematical methods.

While the CLT is generally regarded as a bridge linking classical and modern probability theory (see [3, Ch. 8]), we see, in Flajolet's works, that classical analytic techniques, as he largely adopted, revived and promoted, served as an effective platform on which diverse aspects of Combinatorics, Algorithmics, and Probability were clarified and theorized. In a sense, Flajolet modernized Laplacian Mathematics.

We conclude this introduction with three histograms in Figure 3, showing further the ubiquity of the Gaussian law.

"If you can specify it, you can analyse it"
--- Philippe Flajolet

References

- [1] W. J. Adams. *The Life and Times of the Central Limit Theorem*, volume 35 of *History of Mathematics*. American Mathematical Society, Providence, RI, second edition, 2009. Including papers by W. Feller and L. Le Cam and comments and a rejoinder by H. F. Trotter, J. L. Doob, David Pollard and Le Cam, With an appendix containing four fundamental papers by A. M. Lyapunov.

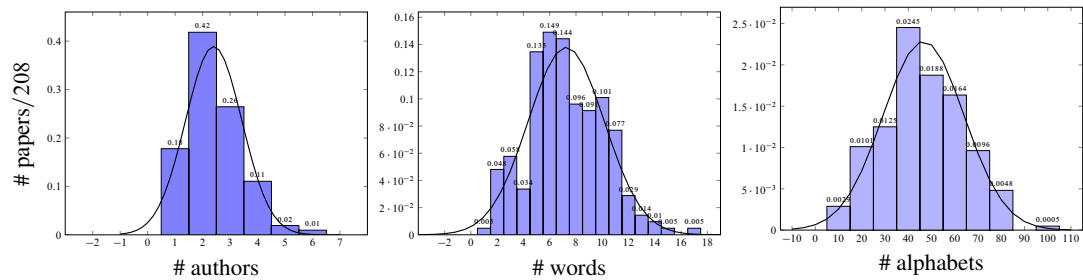


Figure 3: A few statistics of Flajolet's 208 items in his list of publications: the number of coauthors (left), the number of words in titles (middle) and the number of alphabets (right).

- [2] R. Arratia, A. D. Barbour, and S. Tavaré. *Logarithmic Combinatorial Structures: a Probabilistic Approach*. EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich, 2003.
- [3] H. Fischer. *A History of the Central Limit Theorem. Sources and Studies in the History of Mathematics and Physical Sciences*. Springer, New York, 2011. From classical to modern probability theory.
- [4] F. Galton. *Natural Inheritance*. Macmillan, London, 1889.
- [5] A. Hald. *A History of Mathematical Statistics from 1750 to 1930*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.
- [6] H.-K. Hwang. *Théorèmes Limites pour les Structures Combinatoires et les Fonctions Arithmétiques*. PhD thesis, LIX, École Polytechnique, Palaiseau, France, December 1994.
- [7] J. Knopfmacher and W.-B. Zhang. *Number Theory Arising from Finite Fields*, volume 241 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker Inc., New York, 2001. Analytic and probabilistic theory.
- [8] L. Le Cam. The central limit theorem around 1935. *Statist. Sci.*, 1(1):78–96, 1986. With comments, and a rejoinder by the author.
- [9] J. K. Patel and C. B. Read. *Handbook of the Normal Distribution*, volume 40 of *Statistics: Textbooks and Monographs*. Marcel Dekker Inc., New York, second edition, 1996.
- [10] S. M. Stigler. *The History of Statistics*. The Belknap Press of Harvard University Press, Cambridge, MA, 1990. The measurement of uncertainty before 1900, Reprint of the 1986 original.
- [11] J. V. Uspensky. *Introduction to Mathematical Probability*. McGraw-Hill, New York, London, 1937.