General combinatorial schemas with Gaussian limit distributions and exponential tails

Philippe Flajolet, Michèle Soria

INRIA

UNITE DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tel. (1) 39 63 55 11

Rapports de Recherche

N° 1002

Programme 1

GENERAL COMBINATORIAL SCHEMAS WITH GAUSSIAN LIMIT DISTRIBUTIONS AND EXPONENTIAL TAILS

Philippe FLAJOLET
Michèle SORIA

Mars 1989

General Combinatorial Schemas with Gaussian Limit Distributions and Exponential Tails

*Philippe Flajolet** and *Michèle Soria***

ABSTRACT. Under general conditions, the number of components in combinatorial structures defined as sequences, cycles or sets of components, has a Gaussian limit distribution with an exponential tail. The results are valid under general analytic conditions on the generating functions of the combinatorial structures. The proofs depend on continuity theorem for characteristic functions, Laplace transforms and techniques of singularity analysis fitted to algebraic and logarithmic singularities. Several combinatorial examples of application are given, in the fields of graphs, permutations, random mappings, ordered partitions and polynomial factorizations.

Schemas combinatoires généraux avec distributions limites gaussiennes et queues exponentielles

RÉSUMÉ. Sous des conditions générales, le nombre de composantes dans des structures combinatoires définies comme suites, cycles ou ensembles de composantes, a une distribution limite gaussienne avec queues exponentielles. Ces résultats sont valides sous des conditions analytiques générales concernant les fonctions génératrices associées aux structures combinatoires. Les preuves reposent sur l'utilisation du théorème de continuité pour les fonctions caractéristiques, de transformées de Laplace, et de techniques d'analyse de singularités adaptées aux singularités de nature algébrique et logarithmique. On donne quelques exemples en combinatoire, concernant les graphes, les permutations, les graphes fonctionnels, les surjections et les factorisations de polynômes.

* *INRIA, Rocquencourt 78153–Le Chesnay (France)*
** *LRI Université Paris-Sud 91405–Orsay*

General Combinatorial Schemas with Gaussian Limit Distributions and Exponential Tails

*Philippe Flajolet** and *Michèle Soria***

ABSTRACT. Under general conditions, the number of components in combinatorial structures defined as sequences, cycles or sets of components, has a Gaussian limit distribution with an exponential tail. The results are valid under general analytic conditions on the generating functions of the combinatorial structures. The proofs depend on continuity theorem for characteristic functions, Laplace transforms and techniques of singularity analysis fitted to algebraic and logarithmic singularities. Several combinatorial examples of application are given, in the fields of graphs, permutations, random mappings, ordered partitions and polynomial factorizations.

keywords : combinatorial constructions, generating functions, singularity analysis, limit distributions, exponential tails.

1. Introduction

In this paper we investigate the correspondence between combinatorial schemas and analytic patterns, revealing direct relations between structural definitions of combinatorial objects and statistical properties of characteristic parameters (cf. [Flajolet 1985]). This is also of interest to many problems in analysis of algorithms (cf. [Knuth 1973]), since a precise analysis of time or space complexity is often based upon counting combinatorial structures and studying parameters of these structures.

The method consists in two stages.

1. First use combinatorial enumeration methods to *associate* to a class of combinatorial structures a bivariate generating function that counts the number of structures of size n with value k for a given parameter, and systematically *translate* structural specifications into functional equations over generating functions.
2. Then use techniques of complex analysis to extract *asymptotic information* on the coefficients of these generating functions by considering both the location and nature of their singularities in the complex plane.

Our purpose is to find general conditions for combinatorial distributions to converge weakly (i.e. in the sense of distribution functions) to a Gaussian limiting distribution,

* *INRI , Rocquencourt 78153–Le Chesnay (France)*
** *LRI Université Paris-Sud 91405–Orsay*

and also show that the tails of these distributions decrease at an exponential rate. These two types of results are complementary: the existence of a limit distribution provides information on distributions near the mean value, whereas exponential tail results state that large deviations from the expected value are extremely unlikely.

We consider some classical combinatorial constructions, namely the Sequence, Cycle and Set constructions, such that each element of the composed class $\mathcal{P}$ may be uniquely decomposed into elements of the component class $\mathcal{C}$: the combinatorial schema

$$\mathcal{P} = \text{SequenceOf } \mathcal{C} \tag{1.1}$$

means that an element of $\mathcal{P}$ is a sequence of an arbitrary number of elements of $\mathcal{C}$. And we similarly study the constructions

$$\mathcal{P} = \text{SetOf } \mathcal{C} \tag{1.2}$$

and

$$\mathcal{P} = \text{CycleOf } \mathcal{C} \ . \tag{1.3}$$

We provide sufficient conditions leading to an asymptotic Gaussian distribution together with exponential tail estimate, for the number of components in a structure constructed according to each of these schemas. The results are obtained by means of characteristic functions and Laplace transforms. Limit distributions are a consequence of Lévy's continuity theorem for characteristic functions, and exponential tails come from uniformly bounding Laplace transforms. The characteristic function and the Laplace transform of a combinatorial distribution are both expressed in terms of *bivariate* counting generating function of the combinatorial structures, so that we are left with the problem of finding asymptotic information on the coefficients of analytic functions of two complex variables.

The Darboux–Pólya method [Henrici 1977] provides, under suitable analytic conditions, asymptotic expansions of the coefficients of an analytic function by examining its singularities. This is a well–known tool in combinatorial enumeration, often used for asymptotics in Pólya's theory of counting (see e.g. [Pólya 1937], [Harary, Palmer 1973]). Our proofs are based on a variant of that method developed in [Flajolet, Odlyzko 1988].

We first used these methods in [Flajolet, Soria 1988] for Set constructions. In this paper we extend the result of Gaussian limit distribution to a general analytical schema which contains the cases of Sequence and Cycle constructions, and we add exponential tail results for all the situations.

Our work takes place within the general framework of limit theorems for combinatorial enumerations (see e.g. [Sachkov 1978]): Bender [1973] obtains a Gaussian limit distribution for the number of components in a Sequence construction, by means of the continuity theorem for characteristic functions and analysis of coefficients of meromorphic functions. Using the saddle point method and the continuity theorem for characteristic functions, Canfield [1977] shows that, under natural conditions, entire exponential generating functions $C(z)$ lead to Gaussian distributions for the number of components in a Set construction. Also close in scope to our work, Compton [1987] uses real analysis to study distribution results related to the Set construction.

Section 2 contains a precise description of the methods: we first give the generating functions associated to the combinatorial schemas under consideration, and then develop the principles of the analytical methods being used. The limit distribution theorems appear in Section 3: The first result deals with the Sequence and Cycle constructions, and the second one concerns the Set construction. Section 4 is devoted to exponential tail estimates: all the schemas for which we established Gaussian limit distributions do have exponential tails. Section 5 presents some possible extensions of the results, with several combinatorial examples, in the fields of graphs, permutations, random mappings, ordered partitions and polynomial factorizations.

2. Methods

ALGEBRAIC TECHNIQUES. Symbolic methods (see e.g. [Stanley 1978], [Goulden, Jackson 1983], [Flajolet 1988]) allow us to translate a large collection of combinatorial constructions into generating function operations†.

1. The Sequence construction (1.1) translates into quasi-inverse for exponential generating functions, i.e. the analytic schema

$$\hat{P}(z) = \frac{1}{1-\hat{C}(z)}. \tag{2.1}$$

When studying parameters over structures, we use bivariate generating functions: for example, if variable u marks the number of components in a sequence, and $P_{n,k}$ is the number of $\mathcal{P}$-structures of size n having k components, the preceding equation extends to

$$\hat{P}(z,u) \equiv \sum_{n,k\geq 0} P_{n,k} u^k \frac{z^n}{n!} = \frac{1}{1-u\hat{C}(z)}. \tag{2.1'}$$

2. The Set construction translates into exponentials on generating functions: In the case of labelled structures, counted by exponential generating functions, an element of $\mathcal{P}$ is formed by taking a multiset of (labelled) elements of $\mathcal{C}$ and performing all consistent relabellings, so that Schema (1.2) gives

$$\hat{P}(z) = \exp(\hat{C}(z)). \tag{2.2}$$

And if u marks the number of elements in a set, one similarly gets

$$\hat{P}(z,u) \equiv \sum_{n,k\geq 0} P_{n,k} u^k \frac{z^n}{n!} = \exp(u\hat{C}(z)). \tag{2.2'}$$

† For a class of structures $\mathcal{S}$, we shall consistently denote by the same letter: the subclass $\mathcal{S}_n$ of $\mathcal{S}$ formed with elements of size n; S_n the cardinality of $\mathcal{S}_n$; $S(z) = \sum_n S_n z^n$, the ordinary generating function of $\mathcal{S}$; $\hat{S}(z) = \sum_{n\geq 0} S_n \frac{z^n}{n!}$, the corresponding exponential generating function.

3. In the Cycle construction, an element of $\mathcal{P}$ is a cycle of an arbitrary number of elements of $\mathcal{C}$, and this construction translates to logarithms on generating functions. In the case of labelled structures counted by exponential generating functions, Schema (1.3) gives

$$\hat{P}(z) = \log \frac{1}{1-\hat{C}(z)}. \tag{2.3}$$

If u marks the number of elements in a cycle, one also gets the bivariate generating function

$$\hat{P}(z,u) \equiv \sum_{n,k\geq 0} P_{n,k} u^k \frac{z^n}{n!} = \log \frac{1}{1-u\hat{C}(z)}. \tag{2.3'}$$

4. Ordinary generating functions are more appropriate for counting unlabelled structures (see e.g [Stanley 1978]). In the case of sequences both labelled and unlabelled cases lead to the same equation and we have

$$P(z,u) \equiv \sum_{n,k\geq 0} P_{n,k} u^k z^n = \frac{1}{1-uC(z)}.$$

The Set and Cycle constructions on unlabelled structures translate into operations that belong to Pólya's theory of counting. Two Set constructions are to be considered on unlabelled structures: The multiset construction where elements of $\mathcal{P}$ are obtained by taking arbitrary sets of elements of $\mathcal{C}$ with repetition allowed; and the power–set construction, where no component appears more than once. The multiset construction translates into the classical relation for bivariate ordinary generating functions [Pólya 1937]:

$$P(z,u) \equiv \sum_{n\geq 0} P_n u^k z^n = \exp\left(uC(z) + \frac{u^2}{2}C(z^2) + \frac{u^3}{3}C(z^3) + \cdots\right),$$

where u denotes the number of components in the multiset. In the case of the power–set construction, the generating function is the same except that the sum is alternating, i.e.

$$P(z,u) = \exp\Big(\sum_{l\geq 1} \frac{(-u)^l}{l} C(z^l)\Big).$$

Similarly the cycle construction on unlabelled structures counted by ordinary generating functions translates into

$$P(z,u) = \sum_{l\geq 1} \frac{\phi(l)}{l} \log \frac{1}{1-u^l C(z^l)},$$

where $\phi(l)$ is the Euler totient function [Read 1961].

ANALYTIC TECHNIQUES. Let Ω_n denote the random variable representing the number of components in a random $\mathcal{P}$–structure of size n. The probability that Ω_n equals k is $P_{n,k}/P_n$. Setting $p_n(u) = \sum_k P_{n,k} u^k$, we have (see [Feller 1965]):

- the *probability generating function* of Ω_n is $p_n(u)/P_n$,
- its *characteristic function* $\phi_{\Omega_n}(\theta)$ is $p_n(e^{i\theta})/P_n$,
- and its *Laplace transform* $M_{\Omega_n}(\theta)$ is $p_n(e^{\theta})/P_n$.

The mean value μ_n and the variance σ_n^2 of Ω_n are easily computed by differentiation from the probability generating function:

$$\mu_n = \frac{p_n'(1)}{p_n(1)}; \qquad \sigma_n^2 = \frac{p_n''(1)}{p_n(1)} - \frac{p_n'^2(1)}{p_n^2(1)} + \frac{p_n'(1)}{p_n(1)}.$$

As far as limiting distribution and exponential tail estimates are concerned, the problem is to find asymptotic information, as n becomes large, on the coefficients of $P(z,u)$. To establish the Gaussian limit distributions, we consider the normalized variables

$$\Omega_n^* = \frac{\Omega_n - \mu_n}{\sigma_n}$$

and prove the pointwise convergence of the characteristic functions of the Ω_n^* to the characteristic function of a Gaussian variable with mean 0 and variance 1,

$$\phi_{\Omega_n^*}(\theta) \to e^{-\theta^2/2}.$$

Hence, by Levy's continuity theorem for characteristic functions, we deduce the convergence of the distribution functions, in other words, for any two real constants $a < b$,

$$\Pr\left(a < \frac{\Omega_n - \mu_n}{\sigma_n} < b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt. \tag{2.4}$$

As we shall see in Section 4, a sufficient condition for the sequence of normalized random variables Ω_n^* to have exponential tails is that the Laplace transforms are uniformly bounded by a constant K, for all θ in a fixed real neighbourhood of 0, i.e.

$$\exists K, \quad \forall n, \qquad M_{\Omega_n^*}(\theta) < K.$$

Therefore the main technical problem lies in the estimation of $\phi_{\Omega_n^*}(\theta)$, and $M_{\Omega_n^*}(\theta)$, which are expressed as

$$\phi_{\Omega_n^*}(\theta) = E(e^{i\Omega_n^*\theta}) = e^{-i\theta\mu_n/\sigma_n} \frac{p_n(e^{i\theta/\sigma_n})}{P_n}, \tag{2.5}$$

and

$$M_{\Omega_n^*}(\theta) = E(e^{\Omega_n^*\theta}) = e^{-\theta\mu_n/\sigma_n} \frac{p_n(e^{\theta/\sigma_n})}{P_n}. \tag{2.6}$$

To compute the value of $p_n(u)$, we use the Cauchy coefficient formula for analytic functions

$$[z^n]f(z) = \frac{1}{2i\pi} \oint f(z) \frac{dz}{z^{n+1}},$$

with integration on a simple closed contour around the origin ($[z^n]f(z)$ denotes the coefficient of z^n in $f(z)$). For example in the case of ordinary generating functions, we have

$$p_n(u) = \frac{1}{2i\pi} \oint P(z,u) \frac{dz}{z^{n+1}}. \tag{2.6}$$

According to the analytic nature of $P(z,u)$, different asymptotic techniques may be applied to evaluate the Cauchy integral.

If $P(z,u)$ is entire with exponential growth , or has essential singularities, the saddle–point method will apply, the contour is a circle crossing the saddle–point and the main contribution to the integral comes from a small neighbourhood of the saddle–point.

If $P(z,u)$ has poles, algebraic or logarithmic singularities then various extensions of the Darboux–Pólya method will apply [Flajolet,Odlyzko 1988], [Wong, Wyman 1974]. In the case of meromorphic functions, the contour can be extended to a circle of large radius taking into account the residues of the integrand. For functions with algebraic and logarithmic singularities, a well–suited contour is a Hankel contour that comes close to the dominant singularity of the integrand.

The rest of the paper is devoted to functions with isolated algebraic and logarithmic singularities on their circle of convergence. Considering variable u as a parameter, we evaluate the Cauchy integral, the estimation being uniform for u sufficiently close to 1. Expanding $e^{i\theta/\sigma_n}$ in (2.5), and e^{θ/σ_n} in (2.6), when n tends to infinity, leads to Gaussian limiting distributions in the first case, and exponential tails in the second case.

3. Gaussian Limit Distributions

We show in this section that functions with algebraic and logarithmic singularities admit Gaussian limiting distributions. We use the method of *singularity analysis* developed in [Flajolet, Odlyzko 1988]: the Cauchy coefficient formula is evaluated with a Hankel-like contour of integration with main contribution to the integral coming from a small portion of the contour around the singularity.

The first result deals with the Sequence and Cycle constructions: it states that if the generating function $\hat{C}(z)$ of the component class reaches 1 before it becomes singular, then the number of components in any analytical schema of type

$$\frac{1}{(1-u\hat{C}(z))^{\alpha}}\left(\log\frac{1}{1-u\hat{C}(z)}\right)^{k}, \tag{3.1}$$

with k being an integer, and α a real number not in $\{-1,-2,\cdots\}$, has a Gaussian limiting distribution, with linear mean and variance.

The second result is about Set constructions: if the generating function $\hat{C}(z)$ of the component class has a dominant singularity of logarithmic type, then the number of components in any analytical schema

$$\exp(u\hat{C}(z)), \tag{3.2}$$

has an asymptotic distribution which is Gaussian, with logarithmic mean and variance. We give here a synthetic proof of this result which was first proved in [Flajolet, Soria 1988].

SEQUENCE AND CYCLE CONSTRUCTIONS. The first theorem deals with an extension of the analytical schemas corresponding to the Sequence and to the Cycle constructions. From this theorem it is deduced that the number of components in a Sequence or in a Cycle has a Gaussian limiting distribution, with mean and variance of order n, the size of the

structure, provided that the generating function of the components satisfies the condition that $C(z)$ reaches 1 before becoming singular.

More precisely, the condition on $C(z)$ means that $C(\rho) > 1$ where $\rho \leq +\infty$ is the radius of convergence of $C(z)$. Note that since $\hat{C}(z)$ is a function with positive coefficients, this condition is always satisfied if $\hat{C}(z)$ tends to infinity when z approaches the singularity.

THEOREM 1. *Let $\mathcal{P}$ and $\mathcal{C}$ be two classes of combinatorial structures, such that:*

$$\hat{P}(z,u) = \frac{1}{(1-u\hat{C}(z))^{\alpha}} \left(\log \frac{1}{1-u\hat{C}(z)}\right)^{k},$$

with k being an integer, and α a real number $\notin \{-1,-2,\cdots\}$, and suppose that $\hat{C}(z)$ reaches 1 before it becomes singular. Then the random variable Ω_n, that counts the number of components in a random $\mathcal{P}$-structure of size n, has mean μ_n and variance σ_n^2 satisfing, as $n \to \infty$,

$$\mu_n = c_1 n + O(1) \qquad \text{and} \qquad \sigma_n^2 = c_2 n + O(1)\cdot$$

Moreover Ω_n, once normalized, converges weakly to a limiting Gaussian distribution:

$$\Pr\left(a < \frac{\Omega_n - \mu_n}{\sigma_n} < b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt.$$

PROOF. 1. We start by determining the statistics of the P_n. Using a Taylor expansion of $\hat{C}(z)$ around ρ such that $\hat{C}(\rho) = 1$, we get

$$\frac{1}{(1-\hat{C}(z))^{\alpha}} = \frac{1}{(1-z/\rho)^{\alpha}} \frac{1}{\rho^{\alpha}\hat{C}'^{\alpha}(\rho)} \left(1 + \alpha\rho \frac{\hat{C}''(\rho)}{2\hat{C}'(\rho)} (1-z/\rho) + O(1-z/\rho)^2\right). \quad (3.3)$$

For some polynomial S, with real coefficients, of the form $S(x) = x^k + \sum_{p=0}^{k-1} \lambda_p x^p$, we thus have

$$\hat{P}(z) = \frac{1}{\rho^{\alpha}\hat{C}'^{\alpha}(\rho)} \frac{1}{(1-z/\rho)^{\alpha}} S\Big(\log \frac{1}{1-z/\rho}\Big) + O\left(\frac{1}{(1-z/\rho)^{\alpha-1}} \log^k \frac{1}{1-z/\rho}\right). \quad (3.4)$$

Hence the coefficient of z^n is extracted, using singularity analysis [Flajolet, Odlyzko 1988],

$$P_n \equiv p_n(1) \sim \frac{n!}{\Gamma(\alpha)\rho^{\alpha}\hat{C}'^{\alpha}(\rho)} \rho^{-n} n^{\alpha-1} \left(Q_0 + \frac{Q_1}{n} + \frac{Q_2}{n^2} + \cdots\right),$$

where the Q_i are polynomials of degree k in $\log n$.

2. We now compute the mean value of the distribution. Let $\hat{P}'_u(z,1)$ denote the derivative of $\hat{P}(z,u)$ with respect to u, taken at $u = 1$. It is easy to show that

$$\hat{P}'_u(z,1) = \frac{\hat{C}(z)}{\hat{C}'(z)} \hat{P}'(z).$$

Thus using a Taylor expansion of $\hat{C}(z)$ around ρ, we get

$$\hat{P}'_u(z,1) = \frac{1}{\hat{C}'(\rho)}\,\hat{P}'(z)(1 + K(1 - z/\rho) + O((1-z/\rho)^2)\,,$$

where K is expressible in terms of ρ, $\hat{C}'(\rho)$, and $\hat{C}''(\rho)$. Hence

$$[z^n]\,\hat{P}'_u(z,1) = P_{n+1}\,\frac{1}{\hat{C}'(\rho)}(1 + O(\frac{1}{n})) \tag{3.5}$$

Obviously $Q_0(\log(n+1)) = Q_0(\log n) + Q'_0(\log n)/n\,\left(1 + O(1/n\log n)\right)$, so that

$$P_{n+1} \sim \frac{(n+1)!}{\Gamma(\alpha)\rho^\alpha \hat{C}'^\alpha(\rho)}\,\rho^{-(n+1)} n^{\alpha-1}(Q_0 + \frac{R_1}{n} + \frac{R_2}{n^2} + \cdots)\,,$$

where the R_i are still polynomials of degree k in $\log n$. Consequently we have

$$\mu_n \equiv [z^n]\,\frac{\hat{P}'_u(z,1)}{P_n} = \frac{n}{\rho\hat{C}'(\rho)}\frac{Q_0 + \frac{1}{n}R_1 + \frac{1}{n^2}R_2 + \cdots}{Q_0 + \frac{1}{n}Q_1 + \frac{1}{n^2}Q_2 + \cdots}\,(1 + O(\frac{1}{n}))\cdot$$

Since the Q_i are polynomials of degree k, we have $Q_i/Q_0 = cste + O(1/\log n)$, and the denominator is $Q_0 \cdot (1 + O(1/n))$, so that finally

$$\mu_n = \frac{1}{\rho\hat{C}'(\rho)}\,n\,(1 + O(\frac{1}{n}))\cdot \tag{3.6}$$

Note that looking more carefully at the expansions, the constant term in μ_n explicitly evaluates to $1/(\rho\hat{C}'(\rho)) + (\hat{C}''(\rho)/\hat{C}'^2(\rho)) - 1$.

3. For the computation of the variance $\sigma_n^2 \equiv [z^n]\,(\hat{P}''_u(z,1)/P_n) - \mu_n^2 + \mu_n$, we use

$$\hat{P}''_u(z,1) = \frac{\hat{C}^2(z)}{\hat{C}'^2(z)}\,\hat{P}''(z) - \frac{\hat{C}^2(z)\hat{C}''(z)}{\hat{C}'^3(z)}\,\hat{P}'(z)\cdot$$

Proceeding as above, it can be shown that

$$[z^n]\frac{\hat{C}^2(z)}{\hat{C}'^2(z)}\,\hat{P}''(z) = \frac{n^2}{\rho^2\hat{C}'^2(\rho)}\,(1 + O(\frac{1}{n}))\cdot$$

The term of order n^2 cancels with the term coming from the square of the mean value, and the order of growth of the variance is linear. More precise computations show that

$$\sigma_n^2 = c_2 n\,(1 + O(\frac{1}{n})) \qquad \text{where} \qquad c_2 = \frac{1}{\rho^2\hat{C}'^2(\rho)} + \frac{\hat{C}''(\rho)}{\rho\hat{C}'^3(\rho)} - \frac{1}{\rho\hat{C}'(\rho)}\,\cdot \tag{3.7}$$

4. For the limit distribution, we have to evaluate $p_n(e^{i\theta/\sigma_n})/P_n$. Using the same techniques as above for the evaluation of $p_n(u)$, coefficient of z^n in $\hat{P}(z,u)$, we get

$$p_n(u) = \frac{n!}{\Gamma(\alpha)\rho^\alpha(u)C'^\alpha(\rho(u))}\,\rho(u)^{-n}n^{\alpha-1}\log^k n\;\left(1+O(\frac{1}{\log n})\right)\,, \tag{3.8}$$

uniformly in u in a small neighbourhood of 1, where function $\rho(u)$ is analytic at $u=1$. Thus

$$\frac{p_n(e^s)}{P_n} = \exp\left(-n\log\frac{\rho(e^s)}{\rho(1)}\right)\;(1+o(1))\,, \tag{3.9}$$

where the implied constant in the o–estimate is uniform for s sufficiently close to 0. Moreover, since function $\rho(e^s)$ admits a full asymptotic expansion around $s=0$, we have

$$\log\frac{\rho(e^s)}{\rho(1)} = s\frac{\rho'(1)}{\rho(1)} + \frac{s^2}{2}\left(\frac{\rho''(1)}{\rho(1)} - \frac{\rho'^2(1)}{\rho^2(1)}\right) + O\left(\frac{s^3}{3}\right)\cdot \tag{3.10}$$

Instantiating with $s = i\theta/\sigma_n$, we thus have for the characteristic function $\phi_{\Omega_n^*}(\theta)$ as defined in equation (2.6):

$$\phi_{\Omega_n^*}(\theta) \sim \exp\left(-i\theta\frac{\mu_n}{\sigma_n} - n\frac{i\theta}{\sigma_n}\frac{\rho'(1)}{\rho(1)} + \frac{n\theta^2}{2\sigma_n^2}\left(\frac{\rho''(1)}{\rho(1)} - \frac{\rho'^2(1)}{\rho^2(1)}\right) + O(\frac{\theta^3}{\sigma_n^3})\right)\cdot \tag{3.11}$$

Differentiating function C with respect to s gives

$$\rho'(1) = -\frac{1}{C'(\rho(1))} \text{ and } \rho''(1) = -\rho'(1) - \frac{C''(\rho(1))}{C'^3(\rho(1))}\,, \tag{3.12}$$

so that finally replacing μ_n and σ_n by their values leads to

$$\phi_{\Omega_n^*}(\theta) \to e^{-\theta^2/2}.$$

Thus the sequence $\{\Omega_n^*\}$ weakly converges to a Gaussian limit distribution. ∎

SET CONSTRUCTIONS. For the second theorem, we need a precise statement of the conditions for the generating function of components to be of logarithmic type.

We let $\Delta_0(\rho,\eta)$, with $\rho>0$ and $\eta>0$, denote the domain

$$\Delta_0(\rho,\eta) = \{\, z \mid |z| \le \rho+\eta \qquad \text{and} \qquad z \notin [\rho,\rho+\eta]\,\}\cdot$$

Let $G(z)$ be a generating function which is analytic at 0 and has a unique dominant singularity ρ on its circle of convergence. We say that $G(z)$ is a *logarithmic function with multiplier a and constant K* if near this singularity

$$G(z) = a\log\frac{1}{1-z/\rho} + R(z) \tag{3.13}$$

where a is a positive real number and $R(z)$ is analytic in Δ_0 and satisfies $R(z) = K + o(1)$ when z tends to ρ in Δ_0.

THEOREM 2. [Flajolet, Soria 1988] *Let $\mathcal{P}$ and $\mathcal{C}$ be two classes of combinatorial structures such that $\mathcal{P} = \text{SetOf}\,\mathcal{C}$, so that*

$$\hat{P}(z,u) = \exp(u\hat{C}(z)).$$

Let Ω_n be the number of components in a random $\mathcal{P}$-structure of size n. ssume that $\hat{C}(z)$ is a logarithmic function with multiplier a. Then Ω_n, once normalized, converges weakly to a limiting Gaussian distribution,

$$\Pr\left(a < \frac{\Omega_n - \mu_n}{\sigma_n} < b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt$$

where the mean μ_n and variance σ_n^2 of Ω_n satisfy, as $n \to \infty$

$$\mu_n = a \log n + O(1) \qquad \text{and} \qquad \sigma_n^2 = a \log n + O(1).$$

PROOF. This proof is slightly different from the original one given in [Flajolet, Soria 1988] .

We have $P_n = n!\rho^{-n} n^{a-1} e^K/\Gamma(a)\ (1 + O(1/n))$. Furthermore the method used in that first proof leads to

$$p_n(u) = n!\, \frac{\rho^{-n} n^{au-1} e^K}{\Gamma(au)}\ \left(1 + O(\frac{1}{n})\right), \tag{3.14}$$

uniformly, for u sufficiently close to 1. Thus we have

$$\frac{p_n(e^s)}{P_n} = n^{a(e^s-1)} \frac{\Gamma(a)}{\Gamma(ae^s)}.$$

Now $i\theta/\sigma_n$ tends to 0 when n tends to infinity, so

$$\frac{p_n(e^{i\theta/\sigma_n})}{p_n(1)} = \exp\left(a \log n\ \left(\frac{i\theta}{\sigma_n} - \frac{\theta^2}{2\sigma_n^2} + O(\theta^3)\right)\right)\ (1 + o(1)). \tag{3.15}$$

Substituting the values of μ_n and σ_n^2, we get

$$\phi_{\Omega_n^*}(\theta) \equiv e^{-i\theta\mu_n/\sigma_n} \frac{p_n(e^{i\theta/\sigma_n})}{P_n} \to e^{-\theta^2/2},$$

which implies the weak convergence of $\{\Omega_n^*\}$ to a Gaussian limit distribution. ∎

4. Exponential Tails

Weak convergence of probability distributions to a limit provides information on distributions near their center (whence the denomination of 'central limit theorems'). Such results are thus useful for characterizing relatively frequent cases. However, for applications to statistics, combinatorics or analysis of algorithms, it is often useful to characterize the rarity of extreme cases, or in other words, find information on possible 'large deviations' from the average. An important concept in this area is that of distributions with an 'exponential tail'. It turns out that distributions considered in this paper all have exponential tails, so that large deviations are extremely unlikely, and have lower probability of occurrence than would be predicted from Tchebycheff inequality of arbitrary order.

Let Y be a normalized random variable with mean 0 and standard deviation 1. We say that Y has an *exponential tail with parameter* $\alpha < 1$ if

$$\exists C > 0, \quad \forall k > 0, \quad Pr(|Y| > k) < C\alpha^k. \tag{5.1}$$

Generalizing this definition, if $\{Y_n\}_{n\geq 0}$ is a sequence of normalized random variables, we say that $\{Y_n\}$ has an *exponential tail with parameter* $\alpha < 1$ if

$$\exists C > 0, \quad \forall k > 0, \quad \forall n, \quad Pr(|Y_n| > k) < C\alpha^k. \tag{5.2}$$

This last definition is therefore a uniform version of the first one. Variables with an exponential tail have exponentially vanishing probability of large deviations from expected values. Observe first that weak convergence of a sequence $\{Y_n\}$ to a limit Y with an exponential tail (e.g. a Gaussian distribution) does not entail that the Y_n themselves have an exponential tail†. Thus exponential tail estimates are a useful complement to weak convergence results.

For a single random variable Y, it is well known (see e.g. [Billingsley 1986]) that if its Laplace transform (also known as moment generating function) $M(\theta) \equiv E(e^{\theta Y})$ is defined for θ in an interval $I = [\theta_0, \theta_1]$ enclosing 0, then Y has an exponential tail in the sense of (5.1), with

$$C = \inf_{\theta \in I} M(\theta) \qquad \text{and} \qquad \alpha = e^{-\max(\theta_1, \theta_2)}.$$

Similarly, for a sequence $\{Y_n\}$ with Laplace transforms $M_n(\theta)$, if we have

$$\exists K > 0, \quad \forall n, \quad M_n(\theta) < K \tag{5.3}$$

for all θ in some finite interval around 0, then $\{Y_n\}$ has an exponential tail in the sense of (5.2).

A nice consequence of analytic estimates derived in Section 3 is that we get with little additional work exponential tail results for combinatorial distributions that admit a Gaussian limiting law.

---

† Consider for example a probability distribution with mass $1/n$ concentrated at point $x = \sqrt{n}$, and everywhere else with a Gaussian density normalized by a factor of $1 - 1/n$.

THEOREM 3. *Let $p_n(u)$ be defined by*

$$\sum_n p_n(u)\frac{z^n}{n!} = \frac{1}{(1-u\hat{C}(z))^\alpha}\left(\log\frac{1}{1-u\hat{C}(z)}\right)^k,$$

k is an integer, α a real number $\notin \{-1,-2,\cdots\}$), and assume that $\hat{C}(z)$ reaches 1 before it becomes singular. Let Ω_n be the random variable with probability generating function $p_n(u)/p_n(1)$, mean μ_n and variance σ_n^2. Then the sequence of normalized random variables $\Omega_n^* = (\Omega_n - \mu_n)/\sigma_n$ admits an exponential tail.

PROOF. Let $M_{\Omega_n^*}(\theta)$ denote the Laplace transform of Ω_n^*,

$$M_{\Omega_n^*}(\theta) = e^{-\theta\mu_n/\sigma_n}E(e^{\Omega_n\theta/\sigma_n}) = e^{-\theta\mu_n/\sigma_n}\frac{p_n(e^{\theta/\sigma_n})}{P_n}. \tag{5.4}$$

Using the same estimate as in proof of Theorem 1 (cf. Equation (3.9)), we find

$$M_{\Omega_n^*}(\theta) = e^{-\theta\mu_n/\sigma_n}\left(\frac{\rho(e^{\theta/\sigma_n})}{\rho(1)}\right)^{-n}(1+o(1)), \tag{5.5}$$

the estimation being uniform for θ lying in a fixed real neighbourhood I of 0. Expanding function ρ around 1 like in (3.10), we get

$$M_{\Omega_n^*}(\theta) \sim \exp\left(-\theta\frac{\mu_n}{\sigma_n} - \frac{n\theta}{\sigma_n}\frac{\rho'(1)}{\rho(1)} + O(\frac{n\theta^2}{2\sigma_n^2})\right). \tag{5.6}$$

Since σ_n^2 is of order n, and $\mu_n = -n\rho'(1)/\rho(1) + O(1)$, cf. Equations (3.6), (3.7), (3.12), we conclude that $M_{\Omega_n^*}(\theta)$ is uniformly bounded for θ staying in the fixed interval I. ∎

Along the same principle of proof, we can add an exponential tail result to Theorem 2. Using Equations (3.14) and (3.15), we have

$$\frac{p_n(e^{\theta/\sigma_n})}{p_n(1)} = \exp\left(a\log n\,(\frac{\theta}{\sigma_n} + O(\frac{\theta^2}{2\sigma_n^2}))\right)\,(1+o(1)).$$

The proof now relies on the fact that the error terms of $\mu_n - a\log n$ are much smaller than σ_n.

THEOREM 4. *Let $p_n(u)$ be defined by*

$$\sum_n p_n(u)\frac{z^n}{n!} = \exp(u\hat{C}(z))$$

where $\hat{C}(z)$ is a logarithmic function. Then the sequence of normalized random variables Ω_n^*, defined as in Theorem 3, admits an exponential tail.

As a conclusion, notice that it is also possible to derive superexponential bounds with the same methods. An alternative approach to the problem could be to consider asymptotic estimates for densities ('local limit theorems'), in the style of [Bender 1973]. This may involve, however, rather delicate estimates away from the center. Exponential tail results should prove sufficient for many practical applications. Notice for instance, that the first non trivial upper bound on the height of binary search trees was obtained by Robson [1979] using exponential tail properties of Stirling numbers of the first kind (in that case explicit generating functions are available and the analysis is therefore easier).

5. Examples and Extensions

In this section, we offer a few combinatorial examples illustrating cases of applications of the theorems presented in Sections 3 and 4, and also discuss simple extensions of our results.

EXAMPLE 1. *Ordered partitions* of an n–set are described by the bivariate generating function

$$\frac{1}{1-u(e^z-1)},$$

where u marks the number of blocks. The corresponding distribution is asymptotically normal, with exponential tails. Thus the quantities $\{k!\, S_{n,k}\}_{k=0}^{n}$, where $S_{n,k}$ is a Stirling number of the second kind, are asymptotically normal, with exponential tails.

This example is well known and derives from Bender's results. If we now consider instead *cyclic partitions* of an n–set, we are lead to generating function

$$\log \frac{1}{1-u(e^z-1)},$$

which does not fall into Bender's class. From Theorem 1, the distribution of the number of blocks will again be asymptotically Gaussian, and from Theorem 3 it has exponential tails.

More generally, Theorem 1, and its companion Theorem 3 express asymptotic properties for objects obtained by 'composing' a class of structures having a generating function with an algebraico–logarithmic singularity (e.g. cycles, trees) and a class of components $\mathcal{C}$ satisfying the conditions of Theorem 1. As typical instances, Gaussian distributions and exponential tails will be present in the two bivariate schemas

$$\lambda(u\beta(z)) \qquad \text{and} \qquad \beta(u\lambda(z)),$$

where

$$\lambda(z) = \log \frac{1}{1-z} \qquad \text{and} \qquad \beta(z) = \frac{1-\sqrt{1-2z}}{z},$$

corresponding to cycles of trees and trees of cycles. (There, trees are binary, labelled and non–plane: $\beta(z) = 1 + z\beta^2(z)/2$.) ∎

EXAMPLE 2. Several examples of application of Theorem 2 have been given in [Flajolet, Soria 1988], and will not be duplicated here. Let us just say that prototypes of application are the functions

$$\exp\left(u \log \frac{1}{1-z}\right) \qquad \text{and} \qquad \exp\left(\frac{u}{2}(\log \frac{1}{1-z} - z - \frac{z^2}{2})\right),$$

corresponding to the distribution of cycles in *permutations* and of connected components in *2–regular graphs*. ∎

Thus an easy qualitative analysis of generating functions provides, in a large number of cases, direct proofs of Gaussian approximations and exponential tails estimates for combinatorial enumerations. The method is also quite robust and we proceed to indicate a few direct extensions whose proofs follow the same path.

1. *Variations on analytic conditions.* Situations where multiple dominant singularities (of the proper type) appear can be treated by our methods, just using composite Hankel contours. The net result, valid for Theorems 1 to 4, is again the occurrence of Gaussian limit distributions and exponential tails.

EXAMPLE 3. Consider the distribution of the number of cycles in a permutation where *all cycles* are restricted to have *odd length*. The analytic form is

$$\exp\left(\frac{u}{2}(\log\frac{1+z}{1-z})\right).$$

We now have singularities at $z = \pm 1$, but combining local expansions at ± 1, we can prove that the Gaussian property is preserved. ∎

Another interesting extension of Theorem 1 is when the component generating function $\hat{C}(z)$ becomes singular at ρ with $C(\rho) = 1$, the singularity at ρ being of an algebraic nature.

EXAMPLE 4. Consider the class of labelled, rooted, non–plane *trees* —as considered by Cayley— with generating function

$$t(z) = ze^{t(z)}, \qquad \text{so that} \qquad t(z) = \sum_{n\geq 1} n^{n-1}\frac{z^n}{n!}.$$

It is well known (see e.g. [Meir, Moon 1978]) that $t(z)$ has an algebraic singularity at $z = e^{-1}$:

$$t(z) = 1 - \sqrt{2(1-ez)} + \sum_{k\geq 2} c_k(1-ez)^{k/2}.$$

The bivariate function

$$\frac{1}{1-ut(z)}$$

describes the number of trees in a random (ordered) *forest* and leads to a Gaussian distribution with exponential tails. The same result will hold true for cycles of trees, i.e. the scheme

$$\log\frac{1}{1-ut(z)},$$

which provides also the number of cyclic elements in a random *connected mapping*. ∎

Observe, in contrast, that Theorem 2 has to be modified in the context of functions that become singular with finite value: For example the number of components in a random unordered forest corresponding to

$$\exp(ut(z))$$

obeys instead a limiting Poisson distribution.

2. *Variations on combinatorial/analytic schemas.* One class of applications concerns 'complex' structures with structural definitions of the type

$$\mathcal{P} = \mathcal{F} * \text{SequenceOf}(\mathcal{C})$$

as well as their Set or Cycle counterparts. There '$*$' denotes a partitional product (for labelled structures) or a plain cartesian product (for unlabelled structures). The generating function translation then becomes

$$P(z,u) = f(z)\frac{1}{1-uC(z)}.$$

If the generating function $f(z)$ of $\mathcal{F}$ is regular at the dominant singularity of $P(z,1)$ or if it has there only a dominant algebraico–logarithmic singularity, it only plays the rôle of a small perturbation, and again distributions remain Gaussian in the limit, with exponential tails.

EXAMPLE 5. This is the case for the distribution of the number of *cycles of odd length* in general permutations, described by the generating function

$$\frac{1}{\sqrt{1-z^2}}\exp\left(\frac{u}{2}(\log\frac{1+z}{1-z})\right).$$

Asymptotic properties of the distribution are preserved. ∎

Along slightly different lines, we can also go "upwards" in the category of singular behaviours, and our Theorem 2 can be extended to schemas like

$$\exp\left(u(\log^k\frac{1}{1-z/\rho}+R(z))\right).$$

The corresponding derivations then become intermediate between singularity analysis and saddle point methods.

3. *Unlabelled structures.* The analytic schemes corresponding to Set and Multiset constructions for unlabelled structures, namely

$$\exp\left(\sum_{l\geq 1}\pm\frac{u^l}{l}C(z^l)\right),$$

strongly resemble cases of application of Theorems 2 and 4. Indeed, as is well known from classical graphical enumerations [Harary, Palmer 1973], the terms $C(z^p)$ with $p \geq 2$ only tend to modify the implied constants in singular expansions of generating functions, *provided that* $C(z)$ *has radius of convergence stricly less than* 1. In that case, and if $C(z)$

is of logarithmic type, the Gaussian and exponential tail properties hold true for the Set and Multiset constructions.

Similarly, the (unlabelled) cycle construction leads to

$$P(z,u) = \log \frac{1}{1-uC(z)} + \sum_{l\geq 2} \frac{\phi(l)}{l} \log \frac{1}{1-uC(z^l)},$$

a function which is driven by its first term, provided again that $C(z)$ reaches value 1 before becoming singular. Under this condition, an analogue of Theorem 1 is valid.

As a consequence of these extensions to Pólya theory, the authors derived in [Flajolet, Soria 1988] an analogue of the Erdős–Kac theorem for polynomials over finite fields: *The number of irreducible factors in a random monic polynomial of large degree over $GF(q)$ tends to a limiting Gaussian distribution.* An exponential tail property will also hold in such a case.

References.

E. BENDER [1973]. "Central and Local Limit Theorems Applied to Asymptotic Enumeration", *J. Combinatorial Theory Series* **15**, 1973, 91-111.

P. BILLINGSLEY [1986]. *Probability and Measure*, Academic Press, 1986.

E. R. CANFIELD [1977]. "Central and Local Limit Theorems for Coefficients of Binomial Type", *J. Combinatorial Theory Series* **23**, 1977, 275-290.

K. J. COMPTON [1987]. "Some methods for computing component distribution probabilities in relational structures", *Discrete Mathematics* **66**, 1987, 59-77.

W. FELLER [1965]. *n Introduction to Probability Theory and Its pplications*, 2 Volumes, Wiley, New York, 1965.

P. FLAJOLET [1985]. "Elements of a general theory of combinatorial structures", in *Proc. FCT Conf., Lecture Notes in Comp. Sc*, Springer Verlag, 1985, 112-127.

P. FLAJOLET [1988]. "Mathematical Methods in the Analysis of Algorithms and Data Structures," in *Trends in Theoretical Computer Science*, E Börger Editor, Computer Science Press, 1988.

P. FLAJOLET AND A. M. ODLYZKO [1988]. "Singularity Analysis of Generating Functions", preprint, 1988, to appear in *SI M J. on Discrete Maths*

P. FLAJOLET AND M. SORIA [1988]. "Gaussian Limiting Distributions for the Number of Components in Combinatorial Structures", preprint, 1988, to appear in *J. Combinatorial Theory Series*

I. GOULDEN AND D. JACKSON [1983]. *Combinatorial Enumerations.* Wiley, New York, 1983.

F. HARARY AND E. PALMER [1973]. *Graphical Enumerations*, Academic Press, New–York, 1973.

P. HENRICI [1977]. *pplied and Computational Complex nalysis.* Three Volumes. Wiley, New York, 1977.

D. E. KNUTH [1973]. *The rt of Computer Programming.* Volume 3: *Searching and Sorting.* Addison-Wesley, Reading, MA, 1973.

A. MEIR AND J.W. MOON [1978]. "On the altitude of nodes in random trees", *Canadian Journal of Mathematics* **30**, 1978, 997-1015.

G. PÓLYA [1937]. "Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen", *cta Mathematica* **68**, 1937, 145-254. Translated in: G. Pólya and R. C. Read, *Combinatorial Enumeration of Groups, Graphs and Chemical Compounds*, Springer, New-York, 1987.

R. C. READ [1961]. "A note on the number of functional digraphs", *Math. nn.* **143**, 1961, 109-110.

J.M. ROBSON [1979]. "The height of binary search trees," *ustralian Computer Journal* **11**, 1979, 151-153.

V.N. SACHKOV [1978]. *Verojatnosnie Metody v Kombinatornom nalize,* Nauka, Moscow, 1978.

R. P. STANLEY [1978]. "Generating Functions," in *Studies in Combinatorics,* edited by G-C. Rota, M. A. A. Monographs, 1978.

R. WONG, M. WYMAN [1974]. "The Method of Darboux," *Journal of pproximation Theory* **10**, 1974, 159-171.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique