APPLICATION OF THE METHOD OF MOMENTS IN PROBABILITY AND STATISTICS

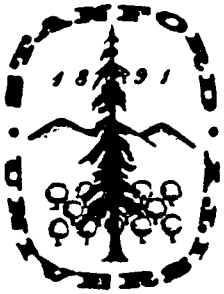
BY

PERSI DIACONIS

TECHNICAL REPORT NO. 262
NOVEMBER 1986

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS86-00235

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA



APPLICATION OF THE METHOD OF MOMENTS IN PROBABILITY AND STATISTICS

BY

PERSI DIACONIS

TECHNICAL REPORT NO. 262
NOVEMBER 1986

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT DMS86-00235

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

Application of the Method of Moments in Probability and Statistics

PERSI DIACONIS[1]

1. Introduction. The method of moments is a versatile tool, day-to-day, in probability and statistics. Probabilists use moment methods to prove limit theorems in non-standard problems, to characterize and manipulate measures, and as a source of still challenging mathematical problems.

Statisticians use moments as a basis of curve-fitting algorithms (Pearson curves through "projection pursuit"), for estimating parameters (method of moment estimators), and as theoretical tools to prove theorems in their subject.

My aim is to bring this subject to life through examples. These are drawn from my own research in an attempt to explain how moments arise in applied problems.

The first section is about probability; the basic method of moments is illustrated for Buffon's problem with a long needle. A convenient bookkeeping device — "cumulants" — is described. Natural examples show that distributions can match in many moments and still not be equal. Multivariate moment calculations are shown useful for certain limit problems in statistical mechanics.

The second section is about statistics. The method of moments is illustrated. Its demise at the hands of robustness is documented. A recent rebirth through projection pursuit regression is described. Some characterization theorems show the theoretical applicability of the method of moments.

To read further, a beginner might try Billingsley [5, Sect. 30], Uspensky [6, Appendix II] for applications in probability. Akhiezer [1], Krein [39], or Shohat and Tamarkin [56] are rich sources of exposition and literature. A clear modern account is in [40]. Berg, Christensen, and Ressel [4] show how the method of moments fuses with contemporary mathematics.

1980 *Mathematics Subject Classification.* Primary 47A20
[1] Research partly supported by NSF grant DMS86-00235.

2. The method of moments in probability.

2.1 Basics. In applied problems one often encounters a sequence μ_n of probability measures on a fixed space, say $\mathbf{R}$ the real numbers. As n increases, the measures converge in some sense to a limit. The law of large numbers and the central limit theorem are early instances.

It is widely acknowledged that the first proof of the central limit theorem up to modern standards of rigor was given by Chebyshev in 1887. His argument introduced the method of moments. Seneta [57] gives a careful history of Russian contributions at this time. Lecam [41] presents an interesting defense of the claim that Laplace proved the theorem. Laplace used transforms, but the first uniqueness theorems suitable for probability are due to Levy in the 20^{th} century.

In modern notation, the method of moments can be stated thus: Let μ be a probability measure on the Borel sets of $\mathbf{R}$. Define the k^{th} moment of μ as

$$\mu(x^k) \equiv \int x^k \mu(dx).$$

THEOREM 1. *Let μ_n be a sequence of probability measures with moments of all orders. Suppose that for each k, $\mu_n(x^k)$ converges to a number μ_k. Then, there is a measure μ with $\mu(x^k) = \mu_k$. If μ is determinate, i.e., is uniquely determined by its moments, then for every bounded continuous function*

$$\int f d\mu_n \to \int d\mu. \tag{2.1}$$

Theorem 1 is proved and applied in Billingsley [5, pp. 342-353]. Convergence in the sense (2.1) is often called weak star convergence. It is equivalent to $\mu_n(-\infty, t] \to \mu(-\infty, t]$ for every t that is a continuity point of μ.

In applications, it is often possible to get our hands on the average of simple functions such as powers x^k. The theorem says that convergence of powers yields convergence for any bounded continuous function when μ is determinate. For measures with a common compact support, this is straightforward: any continuous function can be uniformly approximated by a polynomial. For measures with unbounded support things are more subtle.

Billingsley shows how to derive the usual central limit theorem from Theorem 1 à la Chebyshev. He also works out the central limit theorem in some non-standard situations: sampling without replacement, and the number of prime divisors of an interger chosen uniformly from $\{1, 2, ..., N\}$.

EXAMPLE 1. *Buffons problem with a long needle.* Suppose a needle of length l is thrown onto a plane ruled by parallel lines distance d apart with $l > d$. We derive the probability $p(i)$ of exactly i intersections. This problem arose in the defense area: the planar grid is a grid of detection lines (e.g., a light or laser shining on a photoelectric cell). The needle might be a stream of polluting material laid down at random by a ship or plane.

It is straightforward to give exact expressions for $p(i)$. Clearly the answer only depends on $a \equiv l/d$. The number of interactions can range between 0 and $M \equiv [a]+1$, where $[a]$ denotes greatest integer. Let the angle θ_i, $0 \leqslant \theta \leqslant \pi/2$, be determined by $\cos\theta_i = i/a$, with $0 \leqslant i \leqslant M-1$. Let $\delta_i \equiv 2a\sin(\theta_i/\pi) - (2i\theta_i/\pi)$.

LEMMA 1. *Suppose* $[a] \geqslant 2$. *Then*

$$p(0) = \delta_1 + 1 - (2a/\pi)$$

$$p(i) = \delta_{i-1} + \delta_{i+1} - 2\delta_i \quad \text{for } 1 \leqslant i \leqslant M-2,$$

$$p(M-1) = \delta_{M-2} - 2\delta_{M-1}; \quad p(M) = \delta_{M-1}.$$

For $[a]=1$, *the results for* $p(0)$ *and* $p(M)$ *above hold and* $p(1)$ *is determined by subtraction.*

Lemma 1 follows from results in [35]. As is often the case with exact formulas, it is not easy to get a feel for the answer from direct inspection. In the application, a is large and we expect a fraction of a crossings — what fraction (1/2 ?) and how tightly peaked about the middle is the number of crossings (e.g., is most of the mass within $\sqrt{a}$ of the middle)? Moments are useful indicators of such behavior. The exact form suggests some neat approximation is possible. It is straightforward to calculate

$$\mu_a(x^k) \equiv \sum_{i=0}^{m} i^k p_i = c_k a^k + 0(a^{k-3/2}) \tag{2.2}$$

with $c_k = \Gamma((k+1)/2)/\Gamma((k+2)/2)\sqrt{\pi}$.

Thus $c_1 = 2/\pi$ and $c_2 = 1/2$. The mean is $2a/\pi$, $(\approx .63a)$. The standard deviation $\mu_a(x^2) - \mu_a(x)^2$ is $(1/2 - (2/\pi)^2)a$, $(\approx .1a)$. Here the usual rule of thumb, "90% of the mass is concentrated within two standard deviations of the mean", does not give a very useful approximation.

The neat form of the moments suggests rescaling, by dividing by a. This gives a new measure $\nu_a(0, t] \equiv \mu_a(0, ta]$. Now (2.2) says $\nu_a(x^k) \to c_k$. The form of the c_k's suggest beta integrals and a bit of fooling around yields

$$c_k = \frac{2}{\pi}\int_0^1 x^k \frac{dx}{(1-x^2)^{1/2}}.$$

Now, the method of moments shows that for large values of a, the measures ν_a are well approximated by the measure ν which is absolutely continuous on $[0, 1]$ with density given by $2/\pi(1-x^2)^{-1/2}$. This is a very different concentration of mass from the usual bell-shapes.

The point here is to record a typical instance of moment theory as it occurred. After the fact, a geometric argument was found to "explain" the limiting arc-sine density. Further details can be found in Diaconis (1976). See also Bell

(1977) who improved the error in (2.2) to $O(a^{k-2})$.

A number of less trivial examples received their first proofs using the method of moments. For example, Markov (1884) proved the first central limit theorem for Markov chains this way. Hoeffding (1951) proved his versatile combinatorial limit theorem this way; briefly, if a_{ij} is a square array of side n with row sums zero, and $\Sigma_i a_{ij}^2 = 1$, then, the "random diagonal" $\Sigma_i a_{i\pi(i)}$, where π is a randomly chosen permutation, has an approximate standard normal distribution if the a_{ij} do not vary too much in size. Finally Mark Kac's (1953) first proof of what is known as the Feynman-Kac Formula involved lengthy moment calculations.

All of the above now have more elegant and insightful proofs. Indeed, many young probabilists now shun the method of moments as restricted and heavy handed. There are some however who realize that moments generally get the job done without taking five years off to develop special theory. Peter Major's work on Wiener-Ito integrals [48] or Svante Janson's [30] book on random graphs provide splendid recent examples.

2.2 Quality of approximation. The method of moments is a limit theorem, yielding an approximation valid "at infinity". It is natural to inquire about the quality of the approximation, and search for correction terms. Some of the deepest and most elegant work on the moment problem is devoted to these questions. Much is known, but still the theory is not up to the demands of applications.

The basic idea begins with Chebyshev's upper and lower bounds for $\mu(0, t]$ when $\mu(x^k)$, $1 \leqslant k \leqslant n$, are given. Chebyshev stated these results in 1874. They were proved independently by Markov (1884) and Stieltjes (1884) using the analytic theory of continued fractions.

A modern approach developed by Krein [39], or Karlin and Shapley [34] considers the set of all measures with n prescribed moments as a convex set. Extreme points can be determined, and these yield upper and lower bounds. Kemperman's lecture in this volume explains this approach in detail.

Unfortunately, the numerical determination involves computation of zeros of associated orthogonal polynomials. This is feasible for a small number of moments, but appears to be quite difficult in general cases.

Here is an example in the most thoroughly studied special case: the normal distribution.

EXAMPLE 2. *Random rotations.* The following problem arose in studying a method for encrypting speech over telephones. One useful method due to Aaron Wyner requires a source of $n \times n$ orthogonal matrices chosen according to Haar measure on the orthogonal groups $\mathcal{O}(n)$. Sloane [59] contains a readable account of this problem.

It turns out that large random orthogonal matrices are expensive to generate — all usable algorithms require $O(n^3)$ operations. If $n = 256$, this is of order 10^{16} which is too slow for repeated real-time use.

To understand a method of *approximating* a random rotation by k random reflections it was important to understand the behavior of the trace of a random rotation.

Thus consider the following problem: choose an orthogonal matrix Γ from Haar measure on $\mathcal{O}(n)$. What is the distribution of $\mathrm{Tr}(\Gamma)$, the trace of Γ, when n is large? Intuitively, $\mathrm{Tr}(\Gamma)$ is the sum of a lot of little things; if all goes well it should have an approximate Gaussian distribution (the bell-shaped curve). Colin Mallows and I [13] calculated the moments and showed

LEMMA. *For* $0 \leqslant k \leqslant 2n+1$,

$$\int_{\mathcal{O}(n)} (\mathrm{Tr}\Gamma)^k d\Gamma = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^k e^{-x^2/2} dx. \tag{2.3}$$

REMARKS.

1. Here, the first $2n+1$ moments equal the corresponding moments of the standard normal distribution. Now, the method of moments shows that the measure associated to the trace on $\mathcal{O}(n)$ converges to a limiting standard normal distribution. This seems similar to Hoeffding's result described above, but the connections have not been worked out.

2. The proof of (2.3) involves a bit of character theory. Both sides vanish when k is odd. For k even, the left side of (2.3) can be interpreted as the inner product of the character of the tensor product of the usual n-dimensional representation of $\mathcal{O}(n)$ with itself k times, and the trivial character. Brauer [6] determined the decomposition of this representation, and so the multiplicity of the trivial representation, as $k!/2^{k/2}(k/2)!$. But this is just the well-known value of the right-hand side (integrating by parts).

This matching up of moments seems remarkable and suggests that the distribution of the trace might be very well approximated by the corresponding normal distribution. In this case, Chebyshev's bounds have been carefully developed. The following theorem is available.

THEOREM 2. *If* μ *is a probability measure with first* n *moments equal to the moments of a standard normal measure:*

$$\mu(x^k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^k e^{-x^2/2}\, dx, \quad 1 \leqslant k \leqslant n,$$

then, for every $t \in (-\infty, \infty)$,

$$|\mu(0,t] - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}| \leqslant \sqrt{\frac{\pi}{2n}}.$$

REMARK. Theorem 2 was first proved by Markov, improving an earlier result of Chebyshev who gave the result with a non-uniform error. A very accessible

proof from first principles is given by Uspensky [66, Appendix II]. It is known that this result is sharp in the sense that there exist measures μ where the difference is of order $1/\sqrt{n}$.

In the case at hand, there is good reason to suspect that the degree of approximation is much better. It has been generally found by applied workers that the bounds given by Chebyshev's inequalities are very broad. The bounds *are* achieved, but at discrete measures having $n+1$ atoms. There have been several attempts to tighten the bounds using additional assumptions. Royden (1953) uses restriction on the support. Mallows (1956a,b) gives improvements using smoothness and unimodality. Both authors offer useful improvements, but for day-to-day use, much further work needs to be done.

To explain this last point, I observe that the characters of naturally occurring representations of other groups have moments equal to the moments of the appropriate limiting measures. For example, if $\mathcal{U}(n)$ is the n-dimensional unitary group, the trace of a random matrix M converges to a standard complex normal and the moments match up:

$$\int_{\mathcal{U}(n)} (\operatorname{Tr} M)^a (\operatorname{Tr} \overline{M})^b dM = \frac{1}{4\pi} \int\int (x+iy)^a (x-iy)^b e^{-x^2-y^2} dx\, dy$$

for positive integers $a+b \leqslant n$. I know of no available theory to give error terms.

For a discrete example, let π be chosen uniformly in the symmetric group S_n. The trace of the permutation matrix associated to π is the number of fixed points of $\pi \equiv \mathrm{FP}(\pi)$. It is well-known [17, p. 107] that the number of fixed points has as limiting distribution the Poisson distribution with parameter 1. It is not hard to show that the first n moments are equal: for $0 \leqslant i \leqslant n$,

$$\frac{1}{n!} \sum_{\pi \in S_n} [\mathrm{FP}(\pi)]^i = \frac{1}{e} \sum_{j=0}^{\infty} j^i / j!.$$

It is also not hard to argue that the distributions are extremely close. We have shown [13] that for every $A \subset \{0, 1, 2, \cdots\}$.

$$\left| \frac{\#\{\pi \in S_n : \mathrm{FP}(\pi) \in A\}}{n!} - \frac{1}{e} \sum_{j \in A} \frac{1}{j!} \right| \leqslant \frac{2^n}{n!}.$$

It would be instructive to compare this with the extremal bounds. Alas, at present writing this seems to require days of further work.

2.3 The Hausdorff moment problem and deFinetti's theorem. In 1920, Felix Hausdorff considered the moment problem on $[0, 1]$. Among other things, he gave an elegant condition for a sequence $c_0 = 1, c_1, c_2, \cdots$ to be the moments of a probability meausre μ on $[0, 1]$:

$$c_k = \int_0^1 x^k \mu(dx). \tag{2.4}$$

To understand Hausdorff's condition, observe that such a representation implies c_k decreasing, or $c_{k+1} - c_k < 0$. Let $\Delta^1 c_k = c_{k+1} - c_k$, $\Delta^2 c_k = \Delta(\Delta c_k) = c_{k+2} - 2c_{k+1} + c_k$. Observe $\Delta^2 c_k = \int x^k (x-1)^2 \mu(dx) \geqslant 0$. Similarly,

$$(-1)^r \Delta^r c_k = \int x^k (1-x)^r \mu(dx) \geqslant 0. \tag{2.5}$$

Hausdorff proved

THEOREM 3. *A sequence $c_0 = 1, c_1, c_2, \cdots$ can be represented as a moment sequence* (2.4) *if and only if $(-1)^r \Delta^r c_k \geqslant 0$ for every k and r.*

Hausdorff's condition is appealing because it only involves the moments c_i, not general linear combinations $\Sigma a_i c_i$, or quadratic forms $\Sigma c_i c_j a_{i+j}$.

It turns out that Hausdorff's theorem is completely equivalent to a fundamental theorem in Bayesian statistics called deFinetti's theorem. The purpose of this section is to explain the connection. DeFinetti's theorem has seen vast generalization. This suggests a wealth of novel moment problems which seem untouched.

The theorem deals with a probability measure P on the space $\mathbf{Z}_2^\infty$ of infinite sequences of zeros and ones with the usual product topology. Call P *exchangeable* if it is invariant under permuting coordinates:

$$P\{e_1, e_2, ..., e_k\} = P\{e_{\pi(1)}, e_{\pi(2)}, ..., e_{\pi(k)}\}$$

for every sequence $e_1, ..., e_k \in \{0, 1\}$, all permutations π, and all k.

THEOREM 4 (deFinetti). *P is exchangeable if and only if for all n*

$$P(e_1 \cdots e_n) = \int_0^1 x^r (1-x)^{n-r} \mu(dx) \tag{2.6}$$

where $r = e_1 + \cdots + e_n$, and μ is a unique probability measure on $[0, 1]$ *which does not depend on n or $e_1, ..., e_n$.*

In probability language, P is exchangeable if and only if it is a mixture of coin tossing measures. In the language of convex sets, the set of all exchangeable measures on $\mathbf{Z}_2^\infty$ is a convex simplex with extreme points given by the product of identical factors.

DeFinetti's theorem is important in the foundations of probability. Classical Bayesian statisticians such as Bayes, Laplace and Gauss analyzed repeated phenomena using expressions like the right-hand side of (2.6), where $x^r(1-x)^{n-r}$ is called the likelihood, x is the parameter, and μ is the prior. Modern subjective Bayesians such as deFinetti, Savage, (and the present author) prefer not to talk about unobservable parameters. They prefer

assigning probability to observables such as "the next 3 flips of the coin are 0, 1, 0". This is the left side of (2.6). Theorem 4 shows that if the probability assignment is exchangeable, then the two approaches are equivalent. See deFinetti [20] for a readable discussion of these philosophical issues.

Hausdorff's and deFinetti's theorems are equivalent. To see this, define an exchangeable probability on $\mathbf{Z}_2^\infty$ by setting $c_k = P(1, \ldots, 1)$, a sequence of k ones, and using the laws of probability, $P(1) = P(1,0) + P(1,1)$, so

$$P(1,0) = P(1) - P(1,1) = c_1 - c_2 \geqslant 0, \text{ or}$$

$$P(1,0,0) = P(1,0) - P(1,0,1) = [P(1) - P(1,1)] - [P(1,1) - P(1,1,1)]$$

$$= P(1) - 2P(1,1) + P(1,1,1) = \Delta^2 c_3 .$$

More generally, let $P_{kn} = P(1, .., 1, 0, .., 0)$ for a sequence of length n beginning with k ones. So $P_{nn} = c_n$, $P_{n-1,n} = P_{n-1,n-1} - P_{n,n} = -\Delta c_{n-1}$. By induction,

$$P_{k,n} = P_{k,n-1} - P_{k+1,n} = (-1)^{n-k} \Delta^{n-k} c_k . \tag{2.7}$$

Now, if c_k has all these differences positive, the numbers $P_{k,n}$ satisfy the consistency requirements to be a probability, and from Hausdorff's theorem we get deFinettis. Conversely, if Hausdorff's condition is satisfied, deFinetti's theorem gives $c_k = \int_0^1 x^k \, \mu(dx)$. This argument was first given by deFinetti [20]. It is nicely presented by Feller [18, p. 225]. DeFinetti's theorem has seen sweeping generalization to more values than 2 (indeed to abstract spaces) to other notions of invariance (every rotationally invariant measure on $\mathbf{R}^\infty$ is a scale mixture of mean zero normals). The most general versions involve a notion of partial exchangeability which characterizes things like mixtures of Markov chains. When sufficiently abstracted, the general form of deFinetti's theorem becomes identical with the work of Dobrushin, Lanford and Ruelle on Gibbs ensembles [58, Chapt. 1].

Useful surveys of this work appear in Aldous (1985), Diaconis and Freedman (1984), Dynkin (1978), Kingman (1978) or Lauritzan (1984).

Recently, Ressel (1985) has made important connections between these generalizations of exchangeability and semi-group theory as developed by Berg, Christensen, and Ressell [4]. Representation theorems are presented as integrals over the dual semi-group. Since many duals have been classified, this gives a host of new results. Usually the results come out in a highly analytic form, and much work remains to be done in translating things back to probabilistic language.

Many of the generalizations have translation into unorthodox moment problems, where only certain moments are given. It would be interesting to see the machinery of the method of moments turned loose on these problems. A reasonable place to start might be deFinetti's theorem for Markov chains as developed in [10].

Let us return for a moment to Hausdorff's original papers. His interest lay in yet another direction. He considered lower-triangular regular summability methods A that commute with the usual Cesaro method C_1: $AC_1 = C_1A$, as a matrix equation. Hausdorff argued that $C_1A = AC_1$ if and only if $a_{nr} = \binom{n}{r}(-1)^{n-r}\Delta^{n-r}a_{rr}$. He used the integral characterization of positive a_{nr} to prove interesting summability results.

It would be instructive to have a direct probabilistic proof of Hausdorff's result. The equation $AC_1 = C_1A$ has a simple probability interpretation; why does this imply that the rows of A are mixtures of coin tossing?

2.4 Cumulants, K-statistics, and self-similar processes. For an outsider, moment theory has a messy feel to it. Consider a measure μ on $\mathbf{R}^n$. The moments of the sum of coordinates are

$$\int (x_1 + \cdots + x_n)^k \, \mu(dx_1 \cdots dx_n).$$

Expanding and approximating the cross terms seems like hard work.

Moment theorists have developed a remarkable array of notational machinery to expedite bookkeeping. By now, the machinery has taken on an elegant life of its own, closely connected to combinatorics (Möbius inversion on the partition lattice) and orthogonal polynomials. This section serves as a brief introduction.

To begin with, the moment-generating function of μ on $\mathbf{R}^n$ is defined as

$$M(\theta_1, \ldots, \theta_n) = \int e^{\theta_1 x_1 + \cdots + \theta_n x_n} \, \mu(dx_1 \cdots dx_n) = \Sigma_r \mu(x^r) \frac{\theta^r}{r!}$$

where $\theta^r = \theta_1^{r_1} \cdots \theta_n^{r_n}$, $r! = r_1! \cdots r_n!$, and

$$\mu(x^r) = \int x_1^{r_1} \cdots x_n^{r_n} \, \mu(dx_1 \cdots dx_n)$$

summed over $r_1 \geqslant 0, \ldots, r_n \geqslant 0$.

The cumulants $\{\kappa_r\}$ are defined by the identity

$$\log M(\theta_1, \ldots, \theta_n) = \sum_r \kappa_r \frac{\theta^r}{r!}. \tag{2.8}$$

Thus, for $n = 1$, $\kappa_0 = 0$, $\kappa_1 = \mu(x)$, $\kappa_2 = \sigma^2 = \mu(x^2) - \mu(x)^2$, $\cdots$. For μ a Gaussian measure, $M(\theta) = e^{\mu\theta + \sigma^2\theta^2/2}$, so all cumulants with $r > 2$ vanish. Similarly for μ an n-dimensional Gaussian measure, all teams higher than quadratic vanish. This makes cumulants for Gaussian variables particularly easy to work with. In proving limit theorems one must show only that higher order cumulants tend to zero. It also forms the basis of a theory of testing for Gaussianity developed by Brillinger and Tukey; see [7] for details.

The cumulants are polynomials in the moments. This is most easily expressed using the language of partitions. Let σ be a partition of a finite set S; for example, $\sigma = \{(1, 2), (3), (4)\}$ is a partition of $S = \{1, 2, 3, 4\}$ into *blocks* $\sigma_1 = (1, 2)$, $\sigma_2 = (3)$, $\sigma_3 = (4)$. For a partition $\sigma = \sigma_1 \sigma_2 \cdots \sigma_b$ of $\{1, \ldots, n\}$,

write

$$\kappa_\sigma = \prod_{a=1}^{b} \kappa_{r(\sigma(a))}$$

where $r(\sigma(a)) = (r_1 \cdots r_n)$ is defined by $r_i = 1$ if $i \in \sigma(a)$ and zero otherwise; for example, $\sigma = \{(1,2),(3),(4)\}$ has $\kappa_\sigma = \kappa_{1100}\kappa_{0010}\kappa_{0001}$. The relation between cumulants and moments can then be expressed

$$\kappa_\tau = \sum_\sigma \hat{\mu}(\sigma, \tau) \prod_{a=1}^{b(\sigma)} \mu\left(\prod_{i \in \sigma(a)} x_i\right) \tag{2.9}$$

where $\hat{\mu}(\sigma, \tau)$ is the Möbius function of the partition lattice under the partial order of refinement [62]. Formula (2.9) can be proved by exponentiating both sides of (2.8) and comparing coefficients.

Formula (2.9) is brilliantly presented and developed by Speed (1983) who uses this combinatorial approach to prove all of the standard facts about cumulants in a unified way.

These properties include vanishing of high-order cumulants if some set of coordinates is independent of the others, an appropriate multilinearity of cumulants, and an intriguing reduction of the cumulants of polynomials to products of cumulants of individual coordinates due to Leonov and Shiryaev (1959). It would take us too far afield to develop this here. It is elegantly derived by Speed.

A host of statistics problems involve finding a symmetric function K of observables $X_1, X_2, \ldots, X_n$ which averages to the corresponding cumulant κ:

$$\int K(x_1 \cdots x_n)\, \mu(dx_1 \cdots dx_n) = \kappa.$$

For example, the sample mean $(X_1 + \cdots + X_n)/n$ does the job for the first cumulant for independent coordinates. In the independent case this is the theory of K-statistics developed by R. A. Fisher and co-workers. There has been an active development in higher dimensional cases to cover situations like Tukey's "polykays". This is now closely related to the modern theory of symmetric functions and representation theory. The best access is Speed (1986) who gives an elegant self-contained treatment and extensive references to the literature.

There is another line of work which applies moment theory to construct remarkable examples of processes which are self-similar in the sense popularized by Mandelbrojt. To explain the idea, recall that the sum of independent Gaussian variables is again Gaussian. If x_j are Gaussian, one can consider sums like $\sum_{j=1}^{[Nt]} H_m(x_j)$, $0 \leqslant t \leqslant 1$, with H_m the m^{th} Hermite polynomial, and x_j having mean zero, and covariance $r_k = E(x_j x_{j+k})$ behaving as $k^{-\nu}L(k)$ as $k \to \infty$, where $0 < \nu < 1$ and L is slowly varying. It can be shown that when $\nu > 1/m$ the sum, adequately normalized, converges to Brownian motion. When $0 < \nu < 1/m$ the sum converges to a non-normal self-similar random process.

Such processes are currently being used to model many different types of real-world phenomena. Taqqu [65] contains a survey of these matters.

The situation is much richer in several dimensions, e.g., for averages of variables $X_{(i,j)}$ on a lattice. The possible limits are under active study by mathematical physicists using the language of renormalization. Major [48] and Sinai [58] contain accounts from this point of view.

The point of bringing these results into the present discussion is this: very intricate moment/cumulant calculations lie at the heart of many of these results. The crucial tool used is Dobrushin's "diagram formula". This is explained by Major [48, Chapt. 5] or Fox and Taqqu (1985) for example. This formula is very similar to the Leonov-Shiryaev computations in the cumulant world. For some reason there seems to be no contact between these two closely related areas.

3. Moments in statistics.

3.1 Introduction. Applied workers have long used the mean and standard deviation as numerical summaries of a bunch of numbers. The mean serves as a surrogate for the typical or central value, the standard deviation serves as a measure of how variable the numbers are about the mean.

Perhaps because of the success of the method of moments as a theoretical tool, statisticians started using higher order moments to measure "skewness" or asymmetry about the mean and "kurtosis" to measure the relative size of the extremes of the distribution.

Karl Pearson systematized these ideas, and associated a natural 4-parameter family of measures, now called Pearson curves. These are parameterized by their first 4 moments. Given some data, one computes the moments of the data and finds the Pearson curve with matching moments.

An early success of these methods was "Student's" determination of the sampling distribution of the t-statistic. He computed the moments empirically, using samples randomly drawn from numbers on slips of paper. He fit the matching Pearson curve and observed that this fit the distribution very well. It turns out that the t-distribution is actually a member of Pearson's family of curves, so the approximation was exactly correct. This was rigorously proved 10 years later by R. A. Fisher (1915).

A modern application of this approach is given by Solomon and Stephens (1980). They consider a variety of problems in geometric probability where a few moments can be computed theoretically. They find a simple distribution which matches these moments and use it to answer other questions. For example, if many random lines are dropped in the plane they form polygons. One may inquire about the average area of such a polygon. A technical description involves a Poisson field of lines with intensity parameter τ; see e.g., [60, Chapt. 3].

Miles (1973) gave the first 3 moments of the area as

$$\mu(A) = \frac{\pi}{\tau^2}, \quad \mu(A^2) = \frac{\pi^2}{2\tau^4}, \quad \mu(A^3) = \frac{4}{7}\frac{\pi^7}{\tau^6}.$$

From these, and the known lower bound on the left end point (namely zero) it is possible to fit a Pearson curve, and use it to give approximate results for the proportion of polygons with area smaller than τx. Solomon and Stephens showed this compared quite well with a Monte-Carlo investigation.

The literature on Pearson curves (and other families) is vast. We refer to [36, Chapt. 6] or Solomon and Stephens (1980) and the references therein.

There is a much more general principle known in statistics as the "method of moments". It applies when a sample is thought to come from one of a family of measures known up to a few parameters. Often, one can find simple functions of the observations whose averages are known functions of these parameters. Using sample averages, one gets a suitable number of equations and solves them for the known parameters. References and examples can be found in [38].

This is a handy first-pass method. It suffers from two flaws. The first was pointed out by R. A. Fisher (1922): moment estimators are usually not the most accurate from the point of view of mean-squared error. Usually there are a host of more accurate estimators (e.g., the method of maximum likelihood). The second problem involves what statisticians nowadays call robustness: if a small change in a few of the observations makes a big change in the final estimate, one must be wary. Moment estimators, even such standbys as the mean and sample standard deviation, are notoriously sensitive. A careful discussion of these issues is in [28].

These two problems have cut down on the use of moment methods. However, there has been a recent revival; the idea is to transform data to a bounded scale (say [0, 1]) by a known monotone function. Perform the estimation on the transformed data, and then (if needed) transform back.

A successful use of this idea can be found in Jerry Friedman's (1985) new algorithm for finding "interesting" or "structured" projections of high-dimensional data. Theory developed by Diaconis and Freedman (1984b) or Huber (1985) says that most projections of most data sets will appear like the bell shape curve. Thus the interesting projections are far from normal. Jones (1986) suggested using the L^2 distance of a projection from the normal density as an index of "interest". This approach failed because of robustness: a few stray values made the index go wild. Friedman (1985) transformed things to the unit interval and approximated the density by an expansion in Legendre polynomials. This leads to an index that can be easily computed in terms of the first few moments of the sample. This quick computation is crucial since one is faced with finding the best direction to project in 10 or 20 dimensions. The moment-based methods work wonderfully. Here, there is no need to transform back — the data are graphed in the interesting direction on the original scale.

The next example gives an application of the kinds of upper and lower bounds derived by Chebyshev to a very practical problem.

EXAMPLE 3. *Panel study data.* Sometimes data consist of many short series, say $x_1, x_2, \ldots, x_{5000}$, where $x_i \in \mathbf{Z}_2^{12}$ is a binary 12-tuple. Such data arise, for example, in economic investigations where the i^{th} person is followed for a year, and each month one records 1 if employed and 0 if not. In analyzing such data it is natural to consider simple models such as: x_i has the same distribution as flipping a coin with chance of heads p_i depending on the i^{th} person. For coin flips, the order doesn't matter, so the only information available about p_i is $S_i \equiv \#$ ones in x_i. Clearly $P\{S_i = j\} = p_i^j(1-p_i)^{12-j}$. Of course, we don't know the p_i's, only the observed S_i's. Still, one can hope to get some information about the p_i's as follows. If the people are a sample from a larger population, and μ represents the distribution of the p_i's in this larger population, then for a newly chosen person

$$P\{S = j\} = \int p^j(1-p)^{12-j}\,\mu(dp).$$

From the observed sample, one easily gets unbiased estimates of the first 12 moments (by the method of moments!) as $\hat{\mu}(x^j) = \Sigma \binom{12-j}{k} C_{j+k}$ with C_j the proportion of observed vectors with j ones.

From these estimates one can use moment theory to give upper and lower bounds on the underlying measure μ.

The above approach is classical. Burt Singer has applied the ideas in more complex cases, for instance where each person is allowed to be its own separate Markov chain. Then the moment problems alluded to at the end of the discussion on deFinetti's theorem (which are still unsolved) come into focus.

The ideas sketched out above have many further ramifications. One direction is Herbert Robbins' theory of empirical Bayes theory. The story is fascinating but too long to go into here. Robbins (1986) is a nice survey.

A second direction makes use of many results from the geometry of moment spaces. This is work by Bruce Lindsay (1983a,b). He is concerned with the estimation of mixing measures μ for more general families than binomial. Thus one observes $x_1 \cdots x_n$ from the mixture density

$$\int f_\theta(x)\, Q(d\theta),$$

where $f_\theta(x)$ is known. One wants to estimate Q. Lindsay (1986a,b) continues this work, making even more direct use of the tools of moment theory.

As a final example, here is an application of moment methods in the theory of statistics. Following Goldstein (1975), Diaconis and Ylvisaker (1985) studied the following problem: Let X and Y be independent random variables. Assume that there exist constants a and b such that

$$E(Y \mid X+Y) = a(X+Y) + b. \tag{3.1}$$

In the application, Y represents a parameter, and X represents measurement error. One observes $X+Y$. Then, the best guess at Y (using the mean-squared error criterion) is $E(Y \mid X+Y)$. The problem was to show that for a given underlying (known) measure μ for X there is a unique measure for Y resulting in the linear regression (3.1).

The uniqueness result is false without assumptions. However, it is true if the distribution of X has moments of all order that satisfy the Carleman condition, $\sum_{n=1}^{\infty} \mu(x^{2n})^{-1/2n} = \infty$, which is sufficient for μ to be a determinate measure.

It is natural to inquire if the Carleman condition is really needed or if it is enough that μ be determinate. This question produced some fascinating esoterica (it is open). With the reader's indulgence, we take a few steps down this Garden path. Any unproved assertions can be found in Diaconis and Ylvisaker (1985).

Suppose that μ is determinate and (3.1) holds. One easily shows [33, Lemma 1.1.1] that the linearity (3.1) gives the following conditions on the Fourier transforms:

$$\hat{\mu}_Y(t) = \hat{\mu}_X(t)^{\frac{a}{1-a}} \quad \text{for } t \text{ in a neighborhood of } 0.$$

Observe that if $a = 1/n$ for $n \geqslant 2$ an integer this becomes

$$\hat{\mu}_X(t) = \hat{\mu}_Y(t)^{n-1}.$$

This shows that Y has moments of all orders, and allows them to be determined from the moments of X. It shows more: by a fundamental theorem in moment theory , if μ_Y is not determinate, then, for every real s there is a probability ν_s with the same moments as μ_Y and an atom at s, so $\nu_s^{*(n-1)}$ has an atom at $(n-1)s$. Since s is arbitrary, there is a probability with the same moments as μ_X and an atom at any preselected point. This contradicts determinateness of μ_X.

If it were true that being determined by moments was inherited by convolutions then (at least) general rational values of a could be handled, for $\hat{\mu}_X^n = \hat{\mu}_Y^m$ in a neighborhood of zero, and X determinate, would then imply that Y is determinate. Alas, Christian Berg (1985) provided a probability μ which is determinate but $\mu * \mu$ not!

There is really no end to the problems and applications of the method of moments in these fields. I hope the wealth of applications convinces some readers that there is need for much further theory in practice.

References

1. N. I. Akhiezer, *The classical moment problem*, Hafner, New York, 1965.
2. D. Aldous, *Exchangeability and related topics*, Lecture Notes in Math. 1117, Springer-Verlag, New York-Heidelberg-Berlin, 1985.

3. C. Berg, *On the preservation of determinacy under convolution*, Proc. Amer. Math. Soc. **93** (1985), 351-357.
4. C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic analysis on semigroups*, Springer-Verlag, New York, 1984.
5. P. Billingsley, *Probability and measure*, Wiley, New York, 1979.
6. R. Brauer, *On algebras which are connected with the semisimple continuous groups*, Ann. Math. **38** (1937), 857-872.
7. D. Brillinger, *Time series data, analysis and theory*, Holt, Rinehart, and Winston, New York, 1975.
8. P. L. Chebyshev, *Oeuvres de P. L. Tchebychef*, St. Petersburg, 1889.
9. P. Diaconis, *Buffon's problem with a long needle*, J. Appl. Prob. **13** (1976), 614-618.
10. P. Diaconis and D. Freedman, *DeFinetti's theorem for Markov chains*, Ann. Prob. **8** (1980), 115-130.
11. P. Diaconis and D. Freedman, *Partial exchangeability and sufficiency*, Proc. Indian Statistical Institute Golden Jubilee International Conference on Statistics, (J. K. Ghosh and J. Roy, eds.), Indian Statistical Institute, Calcutta (1984a), pp. 205-236.
12. P. Diaconis and D. Freedman, *Asymptotics of graphical projection pursuit*, Ann. Stat. **12** (1984b) 793-815.
13. P. Diaconis and C. L. Mallows, *On the trace of random matrices*, unpublished manuscript.
14. P. Diaconis and M. Shahshahani, *The subgroup algorithm for generating uniform random variables*, Tecnical report, Dept. of Statistics, Stanford University, Stanford, California; Comm. in operations research (1986), to appear.
15. P. Diaconis and D. Ylvisaker, *Updating subjective probability*, Bayesian Statistics **2**, (J. Bernardo et al., eds.), North Holland, Amsterdam, 1985, pp. 133-156.
16. E. Dynkin, *Sufficient statistics and extreme points*, Ann. Prob. **6** (1978), 705-730.
17. W. Feller, *An introduction to probability theory and its applications, Vol. 1*, 3rd ed., Wiley, New York, 1968.
18. W. Feller, *An introduction to probability theory and its applications, Vol. 2*, 2nd ed., Wiley, New York, 1971.
19. B. deFinetti, *Sur la condition d'équivalence partielle*, Actualités Scientifiques et Industrielles, No. 739, Hermann and Cie, Paris, 1938; translated in Studies in Inductive Logic and Proability (R. C. Jeffrey, ed.), University of California Press, Berkeley, California, 1980, pp. 193-206.
20. B. deFinetti, *Foresight: its logical laws, its subjective sources*, in Studies in Subjective Probability (H. Kyburg and H. Smokler, eds.), Wiley, New York, 1964, pp. 95-158.
21. R. A. Fisher, *Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population*, Biometrika **9** (1915), 507-521.
22. R. A. Fisher, *On the mathematical foundations of theoretical statistics*, Phil. Trans. Roy. Soc. London A **222** (1922), 309-368.
23. R. Fox and M. Taqqu, *Noncentral limit theorems for quadratic forms in random variables having long-range dependence*, Ann. Prob. **13** (1985), 428-446.
24. J. Friedman, *Exploratory projection pursuit*, Technical report, Dept. of Statistics, Stanford University, Stanford, California; J. Amer. Stat. Soc. (1985), to appear.
25. M. Goldstein, *Uniqueness relations for linear posterior expectations*, J. Roy. Stat. Soc. B. 37 (1975), 402-405.
26. F. Hausdorff, *Summationsmethoden und Momentfolgen*, Math. Zeit. **9** (1921), 74-109; 281-299.
27. W. Hoeffding, *A combinatorial central limit theorem*, Ann. Math Stat. **22** (1951), 558-566.
28. P. Huber, *Robust Statistics*, Wiley, New York, 1981.
29. P. Huber, *Projection pursuit*, Ann. Stat. **13** (1985), 435-525.
30. S. Janson, *Normal convergence by higher semi-invariants with application to sums of dependent random variables and random graphs*, Technical report, Dept. Math., Uppsala University, Uppsala, Sweden, 1985.
31. M. C. Jones and R. Sibson, *What is projection pursuit*? J. Roy. Stat. Soc. B (1986), to appear.
32. M. Kac, *On some connections between probability theory and differential and integral equations*, Proc. 2nd Berkeley Sympos. on Math Stat. and Prob., Univ. California Press, Berkeley, California, 1953, pp. 189-215.

33. D. M. Kagan, Yu. Linnik, and C. R. Rao, *Characterization problems in mathematical statistics*, Wiley, New York, 1973.
34. S. Karlin and L. S. Shapley, *Geometry of moment spaces*, Memoirs Amer. Math. Soc. **12** (1953), Amer. Math. Soc., Providence, Rhode Island.
35. M. Kendall and P. A. P. Moran, *Geometrical probability*, Griffin, London, 1963.
36. M. Kendall and A. Stuart, *The advanced theory of statistics, Vol. 1*, 4th ed., Griffin, London, 1977.
37. J. Kingman, *Uses of exchangeability*, Ann. Prob. **6** (1978), 183-197.
38. S. Kotz and N. L. Johnson, *Encyclopedia of statistical sciences, Vols. I-VII*, Wiley, New York, 1986.
39. M. G. Krein, *The ideas of P. L. Chebyshev and A. A. Markov in the theory of limiting values of integrals and their further development*, Amer. Math Soc. Transl. **12** (1959), 1-121.
40. H. J. Landau, *Maximum entropy and the moment problem*, Bull. Amer. Math. Soc. **16** (1987), 1-31.
41. L. Lecam, *The central limit theorem around 1935*, Stat. Science **1** (1986), 78-96.
42. S. Lauritzan, *Extreme point models in statistics*, Scand. J. Stat. **11** (1984), 1-88.
43. V. P. Leonov and A. N. Shiryaev, *On a method of calculation of semi-invariants*, Theor. Prob. Appl. **4** (1959), 319-329.
44. B. G. Lindsay, *The geometry of mixture likelihoods*, Ann. Stat. **11** (1983), 86-94.
45. B. G. Lindsay, *The geometry of mixture likelihoods II; the exponential family*, Ann. Stat. **11** (1983), 783-792.
46. B. G. Lindsay, *On the determinants of moment matrices*, Technical report, Dept. of Statistics, Penn Stat University, University Park, Penn., 1986.
47. B. G. Lindsay, *Moment matrices: applications in mixtures*, Technical report, Dept. of Statistics, Penn State University, University Park, Penn., 1986.
48. P. Major, *Multiple Wiener-Ito integrals*, Lecture Notes in Math. 849, Springer-Verlag, New York, 1981.
49. A. A. Markov, *Démonstration de certaines inégalités de Tchebycheff*, Math. Annalen. **24** (1884), 172-180.
50. R. E. Miles, *The various aggregates of random polygons determined by random lines in a plane*, Adv. Math. **10** (1973), 256-290.
51. C. L. Mallows, *Generalization of Tchebychef's inequalities*, J. Roy. Stat. Soc. B. **18** (1956), 139-143.
52. C. L. Mallows, *Note on the moment problem for unimmodal distributions when one or both terminals are known*, Biometrika **43** (1956), 224-230.
53. P. Ressel, *DeFinetti-style theorems: an analytic approach*, Ann. Prob. **13** (1985), 898-922.
54. H. Robbins, *Estimating many variances*, in Statistical Decision Theory and Related Topics III (S. Gupta and J. Berger, eds.) Vol. 2, p. 251-262, Academic Press, New York, 1982.
55. H. Royden, *Bounds on a distribution function when its first n moments are given*, Ann. Math. Stat. **24** (1953), 361-368.
56. J. A. Shohat and J. D. Tamarkin, *The problem of moments*, Math. Surveys 1, Amer. Math. Soc., New York, 1943.
57. E. Seneta, *The central limit problem and linear least squares in pre-revolutionary Russia: The background*, Math. Scientist **9** (1984), 37-77.
58. Ya. G. Sinai, *Theory of phase transitions: rigorous results*, Pergamon, Oxford, 1982.
59. N. J. A. Sloane, *Encrypting by random rotations*, Cryptography: Lecture notes in Computer Science **149** (T. Beth, ed.), Springer-Verlag, Berlin, 1983.
60. H. Solomon, *Geometric probability*, SIAM, Philadelphia, Penn., 1978.
61. H. Solomon and M. A. Stephens, *Approximations to densities in geometric probability*, J. Appl. Prob. **17** (1980), 145-153.
62. T. P. Speed, *Cumulants and partition lattices*, Austral. J. Stat. **25** (1983), 378-388.
63. T. P. Speed, *Cumulants and partition lattices II: generalized k-statistics*, J. Austral. Math Soc. A **40** (1986), 34-53.
64. T. Stieltjes, *Quelques remarques sur la variation de la densite dans l'intérieur de la terre*, Arch. Néerland. Sci. Exactes Nat. **19** (1884), 435-460.
65. M. Taqqu, *Self-similar processes*, Kotz and Johnson, loc. cit.

66. J. V. Uspensky, *Introduction to mathematical probability*, McGraw-Hill, New York, 1937.

DEPARTMENT OF STATISTICS, STANFORD UNIVERSITY, STANFORD CALIFORNIA, 94305.