Institute of Mathematical Statistics

LECTURE NOTES–MONOGRAPH SERIES
Shanti S. Gupta, Series Editor

Volume 11

Group Representations in Probability and Statistics

Persi Diaconis
Harvard University

Institute of Mathematical Statistics
Hayward, California

Institute of Mathematical Statistics

Lecture Notes–Monograph Series

Series Editor, Shanti S. Gupta, Purdue University

The production of the *IMS Lecture Notes–Monograph Series* is managed by the IMS Business Office: Jessica Utts, IMS Treasurer, and Jose L. Gonzalez, IMS Business Manager.

Library of Congress Catalog Card Number: 88-82779

International Standard Book Number 0-940600-14-5

Printed in the United States of America

Table of Contents

Preface

This monograph is an expanded version of lecture notes I have used over the past eight years. I first taught this subject at Harvard's Department of Statistics 1981–82 when a version of these notes were issued. I've subsequently taught the subject at Stanford in 1983 and 1986. I've also delivered lecture series on this material at Ohio State and at St. Flour.

This means that I've had the benefit of dozens of critics and proofreaders— the graduate students and faculty who sat in. Jim Fill, Arunas Rudvalis and Hansmartin Zeuner were particularly helpful.

Four students went on to write theses in the subject — Douglas Critchlow, Peter Matthews, Andy Greenhalgh and Dan Rockmore. Their ideas have certainly enriched the present version.

I've benefited from being able to quote from unpublished thesis work of Peter Fortini, Arthur Silverberg and Joe Verducci. Andre Broder and Jim Reeds have generously shared card shuffling ideas which appear here for the first time.

Brad Efron and Charles Stein help keep me aligned in the delicate balance between honest application and honest proof. Richard Stanley seems ever willing to translate from algebraic combinatorics into English.

My co-authors David Freedman, Ron Graham, Colin Mallows and Laurie Smith helped debug numerous arguments and kept writing "our" papers while I was finishing this project.

David Aldous and I have been talking about random walk on groups for a long time. Our ideas are so intermingled, that I've found it impossible to give him his fair share of credit.

My largest debt is to Mehrdad Shahshahani who taught me group representations over innumerable cups of coffee. Our conversations have been woven into this book. I hope some of his patience, enthusiasm, and love of mathematics comes through.

Shanti Gupta kept patiently prodding and praising this work, and finally sees it's finished. Marie Sheenan typed the first version. My secretary Karola Decleve has done such a great job of taking care of me and this manuscript that words fail me. Norma Lucas 'TEX-ed' this final version beautifully.

A major limitation of the present version is that it doesn't develop the statistical end of things through a large complex example. I have done this in my Wald lectures Diaconis (1989). Thoughts like this kept delaying things. As the reader will see, there are endless places where "someone should develop a theory that makes sense of this" or try it out, or at least state an honest theorem, or

. . . . It's time to stop. After all, they're only lecture notes.

PERSI DIACONIS
Stanford, February 1987

Chapter 1. Introduction

This monograph delves into the uses of group theory, particularly non-commutative Fourier analysis, in probability and statistics. It presents useful tools for applied problems and develops familiarity with one of the most active areas in modern mathematics.

Groups arise naturally in applied problems. For instance, consider 500 people asked to rank 5 brands of chocolate chip cookies. The rankings can be treated as permutations of 5 objects, leading to a function on the group of permutations S_5 (how many people choose ranking π). Group theorists have developed natural bases for the functions on the permutation group. Data can be analyzed in these bases. The "low order" coefficients have simple interpretations such as "how many people ranked item i in position j." Higher order terms also have interpretations and the benefit of being orthogonal to lower order terms. The theory developed includes the usual spectral analysis of time series and the analysis of variance under one umbrella.

The second half of this monograph develops such techniques and applies them to partially ranked data, and data with values in homogeneous spaces such as the circle and sphere. Three classes of techniques are suggested — techniques based on metrics (Chapter 6), techniques based on direct examination of the coefficients in a Fourier expansion (spectral analysis, Chapter 8), and techniques based on building probability models (Chapter 9).

All of these techniques lean heavily on the tools and language of group representations. These tools are developed from first principles in Chapter 2. Fortunately, there is a lovely accessible little book — Serre's *Linear Representations of Finite Groups* — to lean on. The first third of this may be read while learning the material.

Classically, probability precedes statistics, a path followed here. Chapters 3 and 4 are devoted to concrete probability problems. These serve as motivation for the group theory and as a challenging research area. Many of the problems have the following flavor: how many times must a deck of cards be shuffled to bring it close to random? Repeated shuffling is modeled as repeatedly convolving a fixed probability on the symmetric group. As usual, the Fourier transform turns the analysis of convolutions into the analysis of products. This can lead to very explicit results as described in Chapter 3. Chapter 4 develops some "pure probability" tools - the methods of coupling and stopping times - for random walk problems.

Both card shuffling and data analysis of permutations require detailed knowledge of the representation theory of the symmetric group. This is developed in Chapter 7. Again, a friendly, short book is available: G. D. James' *Representation*

Theory of the Symmetric Group. This is also must reading for a full appreciation of the issues encountered.

Most of the chapters begin with examples and a self-contained introduction. In particular, it is possible to read the statistically oriented Chapters 5 and 6 as a lead in to the theory of Chapters 2 and 7.

A BRIEF ANNOTATED BIBLIOGRAPHY

Group representations is one of the most active areas of modern mathematics. There is a vast literature. Basic supplements are:

J. P. Serre (1977). *Linear Representation of Finite Groups.* Springer-Verlag: New York.

G. D. James (1978). *Representation Theory of the Symmetric Groups.* Springer Lecture Notes in Mathematics 682, Springer-Verlag: New York.

There is however the inevitable tendency to browse. I have found the following sources particularly interesting. Journal articles are referenced in the body of the text as needed.

ELEMENTARY GROUP THEORY

Herstein, I. N. (1975). *Topics in Algebra*, 2nd edition. Wiley: New York.
- The classic, best undergraduate text.

Rotman, J. (1973). *The Theory of Groups: An Introduction*, 2nd edition. Allyn and Bacon: Boston.
- Contains much hard to find at this level; the extension problem, generators and relations, and the word problem.

Suzuki, M. (1982). *Group Theory I, II.* Springer-Verlag: New York.
- Very complete, readable treatise on group theory.

BACKGROUND, HISTORY, CONNECTIONS WITH LIFE AND THE REST OF MATHEMATICS.

Weyl, H. (1950). *Symmetry.* Princeton University Press: New Jersey.
- A wonderful introduction to symmetry.

Mackey, G. (1978). *Unitary Group Representations in Physics, Probability, and Number Theory.* Benjamin/Cummings.

Mackey, G. (1980). Harmonic analysis as the exploitation of symmetry. *Bull.*

Amer. Math. Soc. **3**, 543–697.
- A historical survey.

GENERAL REFERENCES

Hewitt, E. and Ross, K. A. (1963, 1970). *Abstract Harmonic Analysis,* Vols. I, II. Springer-Verlag.
- These are encyclopedias on representation theory of abelian and compact groups. The authors are analysts. These books contain hundreds of carefully worked out examples.

Kirillov, A. A. (1976). *Elements of the Theory of Representations.*
- A fancy, very well done introduction to all of the tools of the theory. Basically a set of terrific, hard exercises. See, for example, Problem 4, Part 1, Section 2; Problem 8, Part 1, Section 3; Example 16.1, Part 3. Part 2 is readable on its own and filled with nice examples.

Pontrijagin, K. S. (1966). *Topological Groups.* Gordon and Breach.
- A chatty, detailed, friendly introduction to infinite groups. Particularly nice introduction to Lie theory.

Naimark, M. A. and Stern, D. I. (1982). *Theory of Group Representations.* Springer-Verlag: New York. Similar to Serre (1977) but also does continuous groups.

GROUP THEORY IN PROBABILITY AND STATISTICS.

Grenander, U. (1963). *Probability on Algebraic Structures.* Wiley: New York.
- Fine, readable introduction in "our language." Lots of interesting examples.

Hannen, E. J. (1965). *Group Representations and Applied Probability.* Methuen. Also in *Jour. Appl. Prob.* **2**, 1–68.
- A pioneering work, full of interesting ideas.

Heyer, H. (1977). *Probability Measures on Locally Compact Groups.* Springer-Verlag: Berlin.

Heyer has also edited splendid symposia on probability on groups. These are a fine way to find out what the latest research is. The last 3 are in *Springer Lecture Notes in Math* nos. 928, 1064, 1210.

SPECIFIC GROUPS.

Curtis, C. W. and Reiner, I. (1982). *Representation of Finite Groups and Asso-*

ciative Algebra, 2nd edition. Wiley: New York.
- The best book on the subject. Friendly, complete, long.

Littlewood, D. E. (1958). *The Theory of Group Characters*, 2nd edition. Oxford.
- "Old-fashioned" representation theory of the symmetric group.

James, G. D. and Kerber, A. (1981). *The Representation Theory of the Symmetric Group*. Addison-Wesley, Reading, Massachusetts.
- A much longer version of our basic text. Contains much else of interest.

Chapter 2. Basics of Representations and Characters

A. Definitions and examples.

We start with the notion of a *group*: a set G with an associative multiplication $s, t \to st$, an identity id, and inverses s^{-1}. A *representation* ρ of G assigns an invertible matrix $\rho(s)$ to each $s \in G$ in such a way that the matrix assigned to the product of two elements is the product of the matrices assigned to each element: $\rho(st) = \rho(s)\rho(t)$. This implies that $\rho(id) = I$, $\rho(s^{-1}) = \rho(s)^{-1}$. The matrices we work with are all invertible and are considered over the real or complex numbers. We thus regard ρ as a homomorphism from G to $GL(V)$ — the linear maps on a vector space V. The dimension of V is denoted d_ρ and called the *dimension of* ρ.

If W is a subspace of V stable under G (i.e., $\rho(s)W \subset W$ for all $s \in G$), then ρ restricted to W gives a *subrepresentation*. Of course the zero subspace and the subspace $W = V$ are trivial subrepresentations. If the representation ρ admits no non-trivial subrepresentation, then ρ is called *irreducible*. Before going on, let us consider an example.

Example. S_n the permutation group on n letters.

This is the group S_n of $1-1$ mappings from a finite set into itself; we will use the notation $\left[\begin{smallmatrix} 1 & 2 & \cdots & n \\ \pi(1) & \pi(2) & \cdots & \pi(n) \end{smallmatrix}\right]$. Here are three different representations. There are others.

(a) The *trivial representation* is 1-dimensional. It assigns each permutation to the identity map $\rho(\pi)x = x$.

(b) The *alternating representation* is also 1-dimensional. To define it, recall the sign of a permutation π is $+1$ if π can be written as a product or an even number of transpositions $\pi = \overbrace{(ab)(cd)\ldots(ef)}^{\text{even \# of factors}}$. The sign of π is -1 if π can be written as an odd number of transpositions. Elementary books on group theory show that $\operatorname{sgn}(\pi)$ is well defined and that $\operatorname{sgn}(\pi_1\pi_2) = \operatorname{sgn}(\pi_1)\operatorname{sgn}(\pi_2)$. It follows that $x \to \operatorname{sgn}(\pi) \cdot x$ is a 1-dimensional representation.

(c) The *permutation representation* is an n-dimensional representation. To define it, consider the standard basis $e_1, \ldots, e_n$ of $\mathbb{R}^n$. It is only necessary to define the linear map $\rho(\pi)$ on the basis vectors. Define $\rho(\pi)e_j = e_{\pi(j)}$. The matrix of a linear map L is defined by $L(e_j) = \Sigma L_{ij}e_i$. With this convention, $\rho(\pi)_{ij}$ is zero or one. It is one if and only if $\pi(j) = i$, so $\rho(\pi)_{ij} = \delta_{i\pi(j)}$. I will write permutations right to left. Thus $\pi_2\pi_1$ means first perform π_1 and then perform π_2.

We will also be using cycle notation for permutations, $(a_1a_2\ldots a_k)$ means $a_1 \to a_2$, $a_2 \to a_3 \ldots a_k \to a_1$. Thus (1 2)(2 3) = (1 2 3) (and *not* (1 3 2)).

Under the permutation representation this last equation transforms into

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Observe that the permutation representation has subspaces that are sent into themselves under the action of the group: the 1-dimensional space spanned by $e_1 + \cdots + e_n$, and its complement $W = \{\underset{\sim}{x} \in \mathbf{R}^n : \ \Sigma x_i = 0\}$ both have this property. A representation ρ is *irreducible* if there is no non-trivial subspace $W \subset V$ with $\rho(s)W \subset W$ for all $s \in G$. Irreducible representations are the basic building blocks of any representation, in the sense that any representation can be decomposed into irreducible representations (Theorem 2 below). It turns out (Exercise 2.6 in Serre or "a useful fact" in 7-A below) that the restriction of the permutation representation to W is an irreducible $n-1$-dimensional representation. For S_3, there are only three irreducible representations; the trivial, alternating, and 2-dimensional representation (Corollary 2 of Proposition 5 below).

EXPLICIT COMPUTATION OF THE 2-DIMENSIONAL REPRESENTATION OF S_3

Let $W = \{x \in \mathbf{R}^3 : \ x_1 + x_2 + x_3 = 0\}$. Let $w_1 = e_1 - e_2$, $w_2 = e_2 - e_3$. Clearly $w_i \in W$. They form a basis for W, for if $v = xe_1 + ye_2 + ze_3 \in W$, then $v = xe_1 + ye_2 + (-x-y)e_3 = x(e_1 - e_2) + (x+y)(e_2 - e_3)$. In this case, it is easy to argue that the restriction of the permutation representation to W is irreducible. Let (x, y, z) be nonzero in W (suppose, say $x \neq 0$) and let W_1 be the span of this vector. We want to show that W_1 is not a subrepresentation. Suppose it were. Then, we would have $(1, y', z')$ and so $(y', 1, z')$ and so $(1 - y', y' - 1, 0)$ in W_1. If $y' \neq 1$, then $e_1 - e_2$ and so $e_2 - e_3$ and $e_1 - e_2$ are in W_1. So $W_1 = W$. If $y' = 1$, then $(1, 1, -2) \in W_1$. Permuting the last two coordinates and subtracting shows $e_2 - e_3$ and so $e_1 - e_2$ are in W_1, so $W_1 = W$.

Next consider the action of π on this basis

π	$\rho(\pi)w_1$	$\rho(\pi)w_2$	$\rho(\pi)$
id	w_1	w_2	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
(1 2)	$-w_1$	$w_1 + w_2$	$\begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}$
(2 3)	$w_1 + w_2$	$-w_2$	$\begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}$
(1 3)	$-w_2$	$-w_1$	$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$
(1 2 3)	w_2	$-(w_1 + w_2)$	$\begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$
(1 3 2)	$-(w_1 + w_2)$	w_1	$\begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}$

CONVOLUTIONS AND FOURIER TRANSFORMS

Throughout we will use the notion of convolution and the Fourier transform. Suppose P and Q are probabilities on a finite group G. Thus $P(s) \geq 0, \Sigma_s P(s) = 1$. By the *convolution* $P * Q$ we mean the probability $P * Q(s) = \Sigma_t P(st^{-1})Q(t)$: "first pick t from Q, then independently pick u from P and form the product ut." Note that in general $P * Q \neq Q * P$. Let the order of G be denoted $|G|$. The *uniform* distribution on G is $U(s) = 1/|G|$ for all $s \in G$. Observe that $U * U = U$ but this does not characterize U-the uniform distribution on any subgroup satisfies this as well. However, $U * P = U$ for any P and this characterizes U.

Let P be a probability on G. The *Fourier transform* of P at the representation ρ is the matrix

$$\hat{P}(\rho) = \Sigma_s P(s)\rho(s).$$

The same definitions works for any function P. In Proposition 11, we will show that as ρ ranges over irreducible representations, the matrices $\hat{P}(\rho)$ determine P.

EXERCISE 1. Let ρ be any representation. Show $\widehat{P * Q}(\rho) = \hat{P}(\rho)\hat{Q}(\rho)$.

EXERCISE 2. Consider the following probability (random transpositions) on S_3

$$P(\text{id}) = p, \; P(12) = P(13) = P(23) = (1 - p)/3.$$

Compute $\hat{T}(\rho)$ for the three irreducible representations of S_3. (You'll learn something.)

B. THE BASIC THEOREMS.

This section follows Serre quite closely. In particular, the theorems are numbered to match Serre.

Theorem 1. *Let $\rho : G \to GL(V)$ be a linear representation of G in V and let W be a subspace of V stable under G. Then there is a complement W^0 (so $V = W + W^0$, $W \cap W^0 = 0$) stable under G.*

Proof. Let $<,>_1$ be a scalar product on V. Define a new inner product by $< u, v >= \Sigma_s < \rho(s)u,\ \rho(s)v >_1$. Then $<,>$ is invariant: $< \rho(s)u,\ \rho(s)v > = < u, v >$. The orthogonal complement of W in V serves as W^0. □

Remark 1. We will say that the representation V splits into the *direct sum* of W and W^0 and write $V = W \oplus W^0$. The importance of this decomposition cannot be overemphasized. It means we can study the action of G on V by separately studying the action of G on W and W^0.

Remark 2. We have already seen a simple example: the decomposition of the permutation representation of S_n. Here is a second example. Let S_n act on $\mathbf{R}^2$ by $\rho(\pi)(x, y) = \mathrm{sgn}(\pi)(x, y)$. The subspace $W = \{(x, y) :\ x = y\}$ is invariant. Its complement, under the usual inner product, is $W^0 = \{(x, y) :\ x = -y\}$ is also invariant. Here, the complement is not unique. For example, $W^{00} = \{(x, y) : 2x = -y\}$ is also an invariant complement.

Remark 3. The proof of Theorem 1 uses the "averaging trick;" it is the standard way to make a function of several variables invariant. The second most widely used approach, defining $< u, v >_2 = \max_g < \rho(g)u,\ \rho(g)v >_1$, doesn't work here since $<,>_2$ is not still an inner product.

Remark 4. The invariance of the scalar product $<,>$ means that if e_i is chosen as an orthonormal basis with respect to $<,>$, then $< \rho(s)e_i,\ \rho(s)e_j > = \delta_{ij}$. It follows that the matrices $\rho(s)$ are unitary. Thus, if ever we need to, we may assume our representations are unitary.

Remark 5. Theorem 1 is true for compact groups. It can fail for noncompact groups. For example, take $G = \mathbf{R}$ under addition. Take V as the set of linear polynomials $ax + b$. Define $\rho(t)f(x) = f(x + t)$. The constants form a non-trivial subspace with no invariant complement. Theorem 1 can also fail over a finite field.

Return to the setting of Theorem 1 by induction we get:

Theorem 2. *Every representation is a direct sum of irreducible representations.*

There are two ways of taking two representations (ρ, V) and (η, W) of the same group and making a new representation. The *direct sum* constructs the vector space $V \oplus W$ consisting of all pairs (v, w), $v \in V$, $w \in W$. The direct sum representation $\rho \oplus \eta(s)(v, w) = (\rho(s)v, \eta(s)w)$. This has dimension $d_\rho + d_\eta$ and clearly contains invariant subspaces equivalent to V and W.

The *tensor product* constructs a new vector space $V \otimes W$ of dimension $d_\rho d_\eta$ which can be defined as the set of formal linear combinations $v \otimes w$ subject to the rules $(av_1 + bv_2) \otimes w = a(v_1 \otimes w) + b(v_2 \otimes w)$ (and symmetrically). If $v_1, \ldots, v_a$ and $w_1, \ldots, w_b$ are a basis for V and W, then $v_i \otimes w_j$ is a basis for

$V \otimes W$. Alternatively, $V \otimes W$ can be regarded as the set of a by b matrices were $v \otimes w$ has ij entry $\lambda_i \mu_j$ if $v = \Sigma \lambda_i v_i$, $w = \Sigma \mu_j w_j$. The representation operates as $\rho \otimes \eta(s)(v \otimes w) = \rho(s)v \otimes \eta(s)w$.

The explicit decomposition of tensor products into direct sums is a booming business. New irreducible representations can be constructed from known ones by tensoring and decomposing.

The notion of the *character* of a representation is extraordinarily useful. If ρ is a representation, define $\chi_\rho(s) = \operatorname{Tr} \rho(s)$. This doesn't depend on the basis chosen for V because the trace is basis free.

PROPOSITION 1. *If χ is the character of a representation ρ of degree d then*

$$(1)\ \chi(\mathrm{id}) = \mathrm{d};\quad (2)\ \chi(\mathrm{s}^{-1}) = \chi(\mathrm{s})^*;\quad (3)\ \chi(\mathrm{tst}^{-1}) = \chi(\mathrm{s}).$$

Proof. (1) $\rho(\mathrm{id}) = \mathrm{id}$. (2) First $\rho(s^a) = I$ for a large enough. It follows that the eigenvalues λ_i of $\rho(s)$ are roots of unity. Then, with * complex conjugation,

$$\chi(s)^* = \operatorname{Tr} \rho(s)^* = \Sigma \lambda_i^* = \Sigma 1/\lambda_i = \operatorname{Tr} \rho(s)^{-1} = \operatorname{Tr} \rho(s^{-1}) = \chi(s^{-1}).$$

(3) $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$. □

PROPOSITION 2. *Let $\rho_1 : G \to GL(V_1)$ and $\rho_2 : G \to GL(V_2)$ be representations with characters χ_1 and χ_2. Then* (1) the character of $\rho_1 \oplus \rho_2$ is $\chi_1 + \chi_2$ and (2) the character of $\rho_1 \otimes \rho_2$ is $\chi_1 \cdot \chi_2$.

Proof. (1) Choose a basis so the matrix of $\rho_1 \oplus \rho_2$ is given as $\left(\begin{smallmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{smallmatrix}\right)$. (2) The matrix of the linear map $\rho_1(s) \otimes \rho_2(s)$ is the tensor product of the matrices $\rho_1(s)$ and $\rho_2(s)$. This has diagonal entries $\rho_1^{i_1 i_1}(s) \rho_2^{j_2 j_2}(s)$. □

Consider two representations ρ based on V and τ based on W. They are called *equivalent* if there is a 1-1 linear map f from V onto W such that $\tau_s \circ f = f \circ \rho_s$. For example, consider the following two representations of the symmetric group: ρ, the 1-dimensional trivial representation (so $V = \mathbb{R}$ and $\rho(\pi)x = x$) and τ, the restriction of the n-dimensional permutation representation to the subspace W spanned by the vector $e_1 + \cdots + e_n$. Here $\tau(\pi)x(e_1 + \cdots + e_n) = x(e_1 + \cdots + e_n)$. The isomorphism can be taken as $f(x) = x(e_1 + \cdots + e_n)$.

The following "lemma" is one of the most used elementary tools.

SCHUR'S LEMMA

Let $\rho^1 : G \to GL(V_1)$ and $\rho^2 : G \to GL(V_2)$ be two irreducible representations of G, and let f be a linear map of V_1 into V_2 such that

$$\rho_s^2 \circ f = f \circ \rho_s^1 \text{ for all } s \in G.$$

Then

(1) If ρ^1 and ρ^2 are not equivalent, we have $f = 0$.

(2) If $V_1 = V_2$ and $\rho^1 = \rho^2$, f is a constant times the identity.

Proof. Observe that the kernel and image of f are both invariant subspaces. For the kernel, if $f(v) = 0$, then $f\rho_s^1(v) = \rho_s^2 f(v) = 0$, so $\rho_s^1(v)$ is in the kernel. For the image, if $w = f(v)$, then $\rho_s^2(w) = f\rho_s^1(v)$ is in the image too. By irreducibility, both kernel and image are trivial or the whole space. To prove (1) suppose $f \neq 0$. Then Ker $= 0$, image $= V_2$ and f is an isomorphism. To prove (2) suppose $f \neq 0$ (if $f = 0$ the result is true). Then f has a non-zero eigenvalue λ. The map $f^1 = f - \lambda I$ satisfies $\rho_s^2 f^1 = f^1 \rho_s^1$ and has a non-trivial kernel, so $f^1 \equiv 0$. □

EXERCISE 3. Recall that the uniform distribution is defined by $U(s) = 1/|G|$, where $|G|$ is the order of the group G. Then at the trivial representation $\hat{U}(\rho) = 1$ and at any non-trivial irreducible representation $\hat{U}(\rho) = 0$.

There are a number of useful ways of rewriting Schur's lemma. Let $|G|$ be the order of G.

COROLLARY 1. *Let h be any linear map of V_1 into V_2. Let*

$$h^0 = \frac{1}{|G|}\Sigma(\rho_t^2)^{-1} h \rho_t^1.$$

Then
(1) *If ρ^1 and ρ^2 are not equivalent, $h^0 = 0$.*
(2) *If $V_1 = V_2$ and $\rho^1 = \rho^2$, then h^0 is a constant times the identity, the constant being Tr h/d_ρ.*

Proof. For any s, $\rho_{s^{-1}}^2 h^0 \rho_s^1 = \frac{1}{|G|}\Sigma \rho_{s^{-1}t^{-1}}^2 h \rho_{ts}^1 = \frac{1}{|G|}\Sigma(\rho_{ts}^2)^{-1} h \rho_{ts}^1 = h^0$. If ρ^1 and ρ^2 are not isomorphic then $h^0 = 0$ by part (1) of Schur's lemma. If $V_1 = V_2$, $\rho_1 = \rho_2 = \rho$, then by part (2), $h^0 = cI$. Take the trace of both sides and solve for c. □

The object of the next rewriting of Schur's lemma is to show that the matrix entries of the irreducible representations form an orthogonal basis for all functions on the group G. For compact groups, this sometimes is called the Peter-Weyl theorem.

Suppose ρ^1 and ρ^2 are given in matrix form

$$\rho_t^1 = (r_{i_1 j_1}(t)), \quad \rho_t^2 = (r_{i_2 j_2}(t)).$$

The linear maps h and h^0 are defined by matrices $x_{i_2 i_1}$ and $x^0_{i_2 i_1}$. We have

$$x^0_{i_2 i_1} = \frac{1}{|G|} \sum_{t j_1 j_2} r_{i_2 j_2}(t^{-1}) x_{j_2 j_1} r_{j_1 i_1}(t).$$

In case (1), $h^0 \equiv 0$ for *all* choices of h. This can only happen if the coefficients of $x_{j_2 j_1}$ are all zero. This gives

COROLLARY 2. *In case* (1)

$$\frac{1}{|G|}\sum_{t\in G} r_{i_2 j_2}(t^{-1}) r_{j_1 i_1}(t) = 0 \quad \text{for all } i_1,\ i_2,\ j_1,\ j_2.$$

COROLLARY 3. *In case* (2)

$$\frac{1}{|G|}\sum_{t\in G} r_{i_2 j_2}(t^{-1}) r_{j_1 i_1}(t) = \begin{cases} \frac{1}{d_\rho} & \text{if } i_1 = i_2 \text{ and } j_1 = j_2 \\ 0 & \text{otherwise.} \end{cases}$$

Proof. In case (2), $h^0 = \lambda I$, or $x^0_{i_2 i_1} = \lambda \delta_{i_2 i_1}$, with $\lambda = \frac{1}{d_\rho}\Sigma \delta_{j_2 j_1} x_{j_2 j_1}$. This gives

$$\frac{1}{|G|}\sum_{t j_1 j_2} r_{i_2 j_2}(t^{-1}) x_{j_2 j_1} r_{j_1 i_1}(t) = \frac{\delta_{i_1 i_2}}{d_\rho}\sum_{j_1 j_2} \delta_{j_1 j_2} x_{j_1 j_2}.$$

Since h is arbitrary, we get to equate coefficients of $x_{j_2 j_1}$. □

ORTHOGONALITY RELATIONS FOR CHARACTERS.

Corollaries 2 and 3 above assume a neat form if the representations involved are unitary, so that $r(s)^* = r(s^{-1})$ where * indicates conjugate transpose. Remark 4 to Theorem 1 implies this can always be assumed without loss of generality. Introduce the usual inner product on functions

$$(\phi|\psi) = \frac{1}{|G|}\Sigma \phi(t)\psi(t)^*.$$

With this inner product, Corollaries 2 and 3 say that the matrix entries of the unitary irreducible representations are orthogonal as functions from G into C.

Theorem 3. *The characters of irreducible representations are orthonormal.*

Proof. Let ρ be irreducible with character χ and given in matrix form by $\rho_t = r_{ij}(t)$. So $\chi(t) = \Sigma r_{ii}(t)$, $(\chi|\chi) = \Sigma_{i,j}(r_{ii}|r_{jj})$. From Corollary 3 above $(r_{ii}|r_{jj}) = \frac{1}{d_\rho}\delta_{ij}$. If χ, χ' are characters of non-equivalent representations, then in obvious notation

$$(\chi|\chi') = \sum_{ij}(r_{ii}|r'_{jj}).$$

Corollary 2 shows each term $(r_{ii}|r'_{jj}) = 0$. □

Theorem 4. *Let ρ, V be a representation of G with character ϕ. Suppose V decomposes into a direct sum of irreducible representations:*

$$V = W_1 \oplus \cdots \oplus W_k.$$

Then, if W is an irreducible representation with character χ, the number of W_i equivalent to W equals $(\phi|\chi)$.

Proof. Let χ_i be the character of W_i. By Proposition 2, $\phi = \chi_1 + \cdots + \chi_k$, and $(\chi_i|\chi)$ is 0 or 1 as W_i is not, or is, equivalent to W. □

COROLLARY 1. *The number of W_i isomorphic to W does not depend on the decomposition (e.g., the basis chosen).*

Proof. $(\phi|\chi)$ does not depend on the decomposition. □

COROLLARY 2. *Two representations with the same character are equivalent.*

Proof. They each contain the same irreducible representations the same number of times. □

We often write $V = m_1W_1 \oplus \cdots \oplus m_nW_n$ to denote that V contains W_i m_i times. Observe that $(\phi|\phi) = \Sigma m_i^2$. This sum equals 1 if and only if ϕ is the character of an irreducible representation.

Theorem 5. *If ϕ is the character of a representation then $(\phi|\phi)$ is a positive integer and equals 1 if and only if the representation is irreducible.*

EXERCISE 4. Do exercises 2.5 and 2.6 in Serre. Use 2.6 to prove that the $n-1$-dimensional part of the n-dimensional permutation representation is irreducible. (Another proof follows from "A useful fact" in Chapter 7-A.)

C. DECOMPOSITION OF THE REGULAR REPRESENTATION AND FOURIER INVERSION.

Let the irreducible characters be labelled χ_i. Suppose their degrees are d_i. The *regular representation* is based on a vector space with basis $\{e_s\}$, $s \in G$. Define $\rho_s(e_t) = e_{st}$. Observe that the underlying vector space can be identified with the set of all functions on G.

PROPOSITION 5. *The character r_G of the regular representation is given by*

$$r_G(1) = |G|$$
$$r_G(s) = 0, \quad s \neq 1.$$

Proof. $\rho_1(e_s) = e_s$ so Tr $\rho_1 = |G|$. For $s \neq 1$, $\rho_s e_t = e_{st} \neq e_t$ so all diagonal entries of the matrix for ρ_s are zero. □

COROLLARY 1. *Every irreducible representation W_i is contained in the regular representation with multiplicity equal to its degree.*

Proof. The number in question is

$$(r_G|\chi_i) = \frac{1}{|G|}\sum_{s\in G} r_G(s)\chi_i^*(s) = \chi_i^*(1) = d_i. \qquad \square$$

Remark. Thus, in particular, there are only finitely many irreducible representations.

COROLLARY 2.
(a) *The degrees d_i satisfy $\Sigma d_i^2 = |G|$.*
(b) *If $s \in G$ is different from 1, $\Sigma d_i \chi_i(s) = 0$.*

Proof. By Corollary 1, $r_G(s) = \Sigma d_i \chi_i(s)$. For (a) take $s = 1$, for (b) take any other s. □

In light of remark 4 to Theorem 1, we may always choose a basis so the matrices $r_{ij}(s)$ are unitary.

COROLLARY 3. *The matrix entries of the unitary irreducible representations form an orthogonal basis for the set of all functions on G.*

Proof. We already know the matrix entries are all orthogonal as functions. There are $\Sigma d_i^2 = |G|$ of them, and this is the dimension of the vector space of all functions. □

In practice it is useful to have an explicit formula expressing a function in this basis. The following two results will be in constant use.

PROPOSITION.
(a) *Fourier Inversion Theorem. Let f be a function on G, then*

$$f(s) = \frac{1}{|G|} \Sigma d_i \operatorname{Tr}(\rho_i(s^{-1}) \hat{f}(\rho_i)).$$

(b) *Plancherel Formula. Let f and h be functions on G, then*

$$\Sigma f(s^{-1}) h(s) = \frac{1}{|G|} \Sigma d_i \operatorname{Tr}(\hat{f}(\rho_i) \hat{h}(\rho_i)).$$

Proof. Part (a). Both sides are linear in f so it is sufficient to check the formula for $f(s) = \delta_{st}$. Then $\hat{f}(\rho_i) = \rho_i(t)$, and the right side equals

$$\frac{1}{|G|} \Sigma d_i \chi_i(s^{-1} t).$$

The result follows from Corollary 2.

Part (b). Both sides are linear in f; taking $f(s) = \delta_{st}$, we must show

$$h(t^{-1}) = \frac{1}{|G|} \Sigma d_i \operatorname{Tr}(\rho_i(t) \hat{h}(\rho_i)).$$

This was proved in part (a). □

Remark 1. The inversion theorem shows that the transforms of f at the irreducible representations determine f. It reduces to the well known discrete Fourier inversion theorem when $G = Z_n$.

Remark 2. The right hand side of the inversion theorem gives an explicit recipe for expressing a function f as a linear combination of the basis functions of Corollary 3. The right hand side being precisely the required linear combination as can be seen by expanding out the trace.

Remark 3. The Plancherel Formula says, as usual, that the inner product of two functions equals the "inner product" of their transforms. For real functions and unitary representations it can be rewritten as $\Sigma f(s)h(s) = \frac{1}{|G|}\Sigma d_i \operatorname{Tr}(\hat{h}(\rho_i)\hat{f}(\rho_i)^*)$. The theorem is surprisingly useful.

EXERCISE 5. The following problem comes up in investigating the distribution of how close two randomly chosen group elements are. Let P be a probability on G. Define $\overline{P}(s) = P(s^{-1})$. Show that $U = P * \overline{P}$ if and only if P is uniform.

EXERCISE 6. Let H be the eight element group of quarternions $\{\pm 1, \pm i, \pm j, \pm k\}$ with $i^2 = j^2 = k^2 = -1$ and multiplication given by $\begin{array}{ccc} & i & \\ \nearrow & & \searrow \\ k & \longleftarrow & j \end{array}$ so $ij = k$, $ji = -k$, etc. How many irreducible representations are there? What are their degrees? Give an explicit construction of all of them. Show that if P is a probability on H such that $P * P = U$, then $P = U$. Hint: See Diaconis and Shahshahani (1986b).

D. NUMBER OF IRREDUCIBLE REPRESENTATIONS.

Conjugacy is a useful equivalence relation on groups: s and t are called *conjugate* if $usu^{-1} = t$ for some u. This is an equivalence relation and splits the group into conjugacy classes. In an Abelian group, each class has only one element. In non-Abelian groups, the definition lumps together sizable numbers of elements. For matrix groups, the classification of matrices up to conjugacy is the problem of "canonical forms." For the permutation group, S_n, there is one conjugacy class for each partition of n: thus the identity forms a class (always), the transpositions $\{(ij)\}$ form a class, the 3 cycles $\{(ijk)\}$, products of 2–2 cycles $\{(ij)(k\ell)\}$, and so on. The reason is the following formula for computing the conjugate: if η, written in cycle notation is $(a \ldots b)(c \ldots d) \ldots (e \ldots f)$, then $\pi\eta\pi^{-1} = (\pi(a) \ldots \pi(b))(\pi(c) \ldots \pi(d)) \ldots (\pi(e) \ldots \pi(f))$. It follows that two permutations with the same cycle lengths are conjugate, so there is one conjugacy class for each partition of n.

A function f on G that is constant on conjugacy classes is called a *class function.*

PROPOSITION 6. *Let f be a class function on G. Let $\rho : G \to GL(V)$ be an irreducible representation of G. Then $\hat{f}(\rho) = \lambda I$ with*

$$\lambda = \frac{1}{d_\rho}\Sigma f(t)\chi_\rho(t) = \frac{|G|}{d_\rho}(f|\chi_\rho^*).$$

Proof. $\rho_s \hat{f}(\rho)\rho_s^{-1} = \Sigma f(t)\rho(s)\rho(t)\rho(s^{-1}) = \Sigma f(t)\rho(\mathrm{sts}^{-1}) = \hat{f}(\rho)$. So, by part 2 of Schur's lemma $\hat{f}(\rho) = \lambda I$. Take traces of both sides and solve for λ. □

Remark. Sometimes in random walk problems, the probability used is constant on conjugacy classes. An example is the walk generated by random transpositions: this puts mass $1/n$ on the class of $\{\mathrm{id}\}$ and $2/n^2$ on $\{(\mathrm{id})\}$. Proposition 6 says that the Fourier transform $\hat{f}(\rho)$ is a constant times the identity. So $\hat{P}^{*k}(\rho) = \lambda^k I$ and there is every possibility of a careful analysis of the rate of convergence. See Chapter 3-D.

EXERCISE 7. Show that the convolution of two class functions is again a class function. Show that f is a class function if and only if $f * h = h * f$ for all functions h.

Theorem 6. *The characters of the irreducible representations: $\chi_1, \ldots, \chi_h$ form an orthonormal basis for the class functions.*

Proof. Proposition 1 shows that characters are class functions and Theorem 3 shows that they are orthonormal. It remains to show there are enough. Suppose $(f|\chi_i^*) = 0$, for f a class function. Then Proposition 6 gives $\hat{f}(\rho) = 0$ for every irreducible ρ and the inversion theorem gives $f = 0$. □

Theorem 7. *The number of irreducible representations equals the number of conjugacy classes.*

Proof. Theorem 6 gives the number h of irreducible representations as the dimension of the space of class functions. Clearly, a class function can be defined to have an arbitrary value on each conjugacy class, so the dimension of the class function equals the number of classes. □

Theorem 8. *The following properties are equivalent*
(1) *G is Abelian.*
(2) *All irreducible representations of G have degree 1.*

Proof. We have $\Sigma d_\rho^2 = |G|$. If G is Abelian, then there are $|G|$ conjugacy classes, and so G terms in the sum, each of which must be 1. If all $d_\rho = 1$, then there must be $|G|$ conjugacy classes, so for each $s, t, sts^{-1} = t$, or G is Abelian. □

Example. The irreducible representations of Z_n — the integers mod n.

This is an Abelian group, so all irreducible representations have degree 1. Any ρ is determined by the image of 1: $\rho(k) = \rho(1)^k$, and $\rho(1)^n = 1$, so $\rho(1)$ must be an n^{th} root of unity. There are n such: $e^{2\pi ij/n}$. Each gives an irreducible representation: $\rho_j(k) = e^{2\pi ijk/n}$ (any 1-dimensional representation is irreducible). They are in-equivalent, since the characters are all distinct (not allowed) or $\rho^1(k) = \rho^2(k)$. The Fourier transform is the well known discrete Fourier transform and the inversion theorem translates to the familiar result: If f is a function on Z_n, and $\hat{f}(j) = \Sigma_k f(k)e^{2\pi ijk/n}$, then $f(k) = \frac{1}{n}\Sigma_j \hat{f}(j)e^{-2\pi ijk/n}$.

E. Product of Groups.

If G_1 and G_2 are groups, their *product* is the set of pairs (g_1, g_2) with multiplication defined coordinate-wise. The following considerations show that the representation theory of the product is determined by the representation theory of each factor.

Let $\rho^1 : G_1 \to GL(V_1)$ and $\rho^2 : G_2 \to GL(V_2)$ be representations. Define $\rho^1 \otimes \rho^2 : G_1 \times G_2 \to GL(V_1 \otimes V_2)$ by

$$\rho^1 \otimes \rho^2_{(s,t)}(v_1 \otimes v_2) = \rho^1_s(v_1) \otimes \rho^2_t(v_2).$$

This is a representation with character $\chi_1(s) \cdot \chi_2(t)$.

Theorem 9.
(1) *If ρ^1 and ρ^2 are irreducible, then $\rho^1 \otimes \rho^2$ is irreducible.*
(2) *Each irreducible representation of $G_1 \times G_2$ is equivalent to a representation $\rho^1 \otimes \rho^2$ where ρ^i is an irreducible representation of G_i.*

Proof.
(1) $(\chi_1|\chi_1) = (\chi_2|\chi_2) = 1$, but the norm of the character of $\rho_1 \otimes \rho_2$ is $\frac{1}{|G_1||G_2|}\Sigma\chi_1(s)\chi_2(t)\chi_1(s)^*\chi_2(t)^* = (\chi_1|\chi_1) \cdot (\chi_2|\chi_2) = 1$. So Theorem 5 gives irreducibility.
(2) The characters of the product representation are of the form $\chi_1 \cdot \chi_2$. It is enough to show these form a basis for the class functions on $G_1 \times G_2$. Since they are all characters of irreducible representations, they are orthonormal, so it must be proved that they are it all of the possible characters. If $f(s,t)$ is a class function orthogonal to all $\chi_1(s)\chi_2(t)$, then

$$\Sigma f(s,t)\chi_1(s)^*\chi_2(t)^* = 0.$$

Then for each t, $\Sigma f(s,t)\chi_1(s)^* = 0$, so $f(s,t) = 0$ for each t. □

Exercise 8. Compute all the irreducible representations of Z_2^k, explicitly.

We now leave Serre to get to applications, omitting the very important topic of induced representations. The most relevant material is Section 3.3, Chapter 7, and Sections 8.1, 8.2. A bit of it is developed here in Chapter 3-F.

Chapter 3. Random Walks on Groups

A. Examples

A fair number of real world problems lead to random walks on groups. This section contains examples. It is followed by more explicit mathematical formulations and computations.

1. Random walk on the circle and random number generation

Think of Z_p (the integers mod p) as p points wrapped around a discrete circle. The simplest random walk is a particle that moves left or right, each with probability $\frac{1}{2}$. We can ask: how many steps does it take the particle to reach a given site? How many steps does it take the particle to hit every site? After how many steps is the distribution of the particle close to random? In Section C, we show that the answer to all of these questions is about p^2.

A class of related problems arises in computer generation of pseudo random numbers based on the recurrence $X_{k+1} = aX_k + b (\text{mod } p)$ where p is a fixed number (often 2^{32} or the prime $2^{31} - 1$) and a and b are chosen so that the sequence $X_0 = 0,\ X_1,\ X_2, \ldots$, has properties resembling a random sequence. An extensive discussion of these matters is in Knuth (1981).

Of course, the sequence X_k is deterministic and exhibits many regular aspects. To increase randomness several different generators may be combined or "shuffled." One way of shuffling is based on the recurrence $X_{k+1} = a_k X_k + b_k$ (mod p) where (a_k, b_k) might be the output of another generator or might be the result of a "true random" source as produced by electrical or radioactive noise. We will study how a small amount of randomness for a and b spreads out to randomness for the sequence X_k.

If $a_k \equiv 1$ and b_k takes values ± 1 with probability $\frac{1}{2}$, we have a simple random walk. If $a_k \neq 1$ is fixed but nonrandom, the resulting process can be analyzed by using Fourier analysis on Z_p. In Section C we show that if $a_k \equiv 2$, then about $\log p \log\log p$ steps are enough to force the distribution of X_k to be close to uniform (with b_k taking values $0,\ \pm 1$ uniformly). This is a great deal faster than the p^2 steps required when $a_k \equiv 1$. If $a_k \equiv 3$, then $\log p$ steps are enough.

What if a_k is random? Then it is natural to study the problem as a random walk on A_p - the affine group mod p. This is the set of pairs (a, b) with $a, b \in Z_p,\ a \neq 0,\ gcd(a, p) = 1$. Multiplication is defined by

$$(a, b)(c, d) = (ac,\ ad + b).$$

Some results are in Example 4 of Section C, but many simple variants are unsolved.

A different group arises when considering the second order recurrence $X_{k+1} = a_k X_k + b_k X_{k-1} \pmod p$ with a and b random. It is natural to define $Y_k = \binom{X_k}{X_{k-1}}$, then

$$Y_{k+1} = \begin{pmatrix} a_k & b_k \\ 1 & 0 \end{pmatrix} Y_k = \left[\Pi \begin{pmatrix} a_i & b_i \\ 1 & 0 \end{pmatrix}\right] Y_0, \text{ with say } Y_0 = \binom{1}{0}.$$

This leads to considering a product of random matrices, and so to a random walk on $GL_2(Z_p)$. See Diaconis and Shahshahani (1986a) for some results.

2. Card Shuffling

How many times must a deck of cards be shuffled until it is close to random? Historically, this was a fairly early application of probability. Markov treated it as one of his basic examples of a Markov chain (for years, the only other example he had was the vowel/consonant patterns in Eugene Onegin). Poincare devoted an appendix of his 1912 book on probability to the problem, developing methods similar to those in Section C. The books by Doob (1935) and Feller (1968) each discuss the problem and treat it by Markov chain techniques.

All of these authors show that any reasonable method of shuffling will *eventually* result in a random deck. The methods developed here allow explicit rates that depend on the deck size. As will be explained, these are much more accurate than the rates obtained by using bounds derived from the second largest eigenvalue of the associated transition matrix.

Some examples of specific shuffles that will be treated below:

a) *Random transpositions.* Imagine n cards in a row on a table. The cards start in order, card 1 at the left, card 2 next to it,..., and card n at the right of the row. Pairs of cards are randomly transposed as follows: the left had touches a random card, and the right hand touches a random card (so left = right with probability $\frac{1}{n}$). The two cards touched are interchanged. A mathematical model for this process is the following probability distribution on the symmetric group:

$$\begin{aligned} T(\text{id}) &= \frac{1}{n} \\ T(\tau) &= \frac{2}{n^2} \text{ for } \tau \text{ any transposition} \\ T(\pi) &= 0 \quad \text{otherwise.} \end{aligned}$$

Repeatedly transposing cards is equivalent to repeatedly convolving T with itself. It will be shown that the deck is well mixed after $\frac{1}{2} n \log n + cn$ shuffles.

Some variants will also be discussed: repeatedly transposing a random card with the top card (la Librairie de la Marguerite), or repeatedly interchanging a card with one of its neighbors.

b) *Borel's shuffle.* In a book on the mathematics of Bridge, Borel and Cheron (1955) discuss the mathematics of shuffling cards at length. They suggest several open problems; including the following shuffle: The top card of a deck is removed and inserted at a random position, then the bottom card is removed and inserted at a random position. This is repeated k times. We will analyze such procedures

in Chapter 4, showing that $k = n \log n + cn$ "moves" are enough. The same techniques give similar rates for the shuffle that repeatedly puts a random card on top, or the shuffle that repeatedly removes a card at random and replaces it at random.

c) *Riffle shuffles.* This is the usual way that card players shuffle cards, cutting off about half the pack and riffling the two packets together. In Chapter 4 we will analyze a model for such shuffles due to Gilbert, Shannon, and Reeds. We will also analyze records of real riffle shuffles. The analysis suggests that 7 shuffles are required for 52 cards.

d) *Overhand shuffles.* This is another popular way of shuffling cards. The following mathematical model seems reasonable: the deck starts face down in the hand. Imagine random zeros and ones between every pair of cards with a zero under the bottom card of the deck. Lift off all the cards up to the first zero and place them on the table. Lift off all the cards up to the second zero and place this packet on top of the first removed packet. Continue until no cards remain. This is a single shuffle. It is to be repeated k times. Robin Pemantle (1988) has shown that about 2500 shuffles are required for 52 cards.

3. Random walk on the d-cube Z_2^d

Regard Z_2^d as the vertices of a cube in d dimensions. The usual random walk starts at a point and moves to one of the d neighbors with probability $\frac{1}{d}$. This is repeated k times. This is a nice problem on its own. It has a surprising connection with a classical problem in statistical mechanics: in the Ehrenfest's urn model, d balls are distributed in two urns. A ball is chosen at random and moved to the other urn. This is repeated k times and the problem is to describe the limiting distribution of the process. For a fascinating description of the classical approach see M. Kac (1947). Kac derives the eigenvalues and eigenvectors of the associated transition matrix by a tour de force. The following approach due to Siegert (1949) suggests much further research:

Let the state of the system be described by a binary vector of length d, with a 1 in the ith place denoting that ball i is in the right hand urn. The transition mechanism translates precisely to a random walk on the d cube! Indeed, the state changes by picking a coordinate and changing to its opposite mod 2. This changes the problem into analyzing the behavior of a random walk on an Abelian group. As we will see, this is straightforward; Fourier analysis gives all the eigenvalues and eigenvectors of the associated Markov chain.

Originally the state of the system in the Ehrenfest's urn was the number of balls in the right hand urn. The problem was "lifted" to a random walk on a group. That is, there was a group G (here Z_2^d) and a probability P on G (here move to the nearest neighbor) and a function $L: G \to$ state space (here the number of ones) such that the image law under L of the random walk was the given Markov chain. There has been some study of the problem of when the image of a Markov chain is Markov. Heller (1965) contains much of interest here. Mark Kac was fascinated with this approach and asked: When can a Markov chain be lifted to a random walk on a group? Diaconis and Shahshahani (1987b) give results for "Gelfand Pairs." The following exercise comes out of discussions with

Mehrdad Shahshahani.

EXERCISE 1. Let P be a probability on the symmetric group S_n. Think of the random walk generated by P as the result of repeatedly mixing a deck of n cards. For a permutation π, let $L(\pi) = \pi(1)$. The values of L are the result of following only the position of card 1. Show that the random walk induces a Markov chain for L. Show that this chain has a doubly stochastic transition matrix. Conversely, show that for any doubly stochastic matrix, there is a probability P on S_n which yields the given matrix for L.

Remark. It would be of real interest to get analogs of this result more generally. For example: find conditions on a Markov chain to lift to a random walk on an Abelian group. Find conditions on a Markov chain to lift to a random walk with a probability P that is constant on conjugacy classes. When can a Markov chain on the ordinary sphere be lifted to a random walk on the orthogonal group O_3?

Returning to the cube, David Aldous (1983b) has applied results from random walk on the d cube to solve problems in the theory of algorithms. Eric Lander (1986) gives a very clear class of problems in DNA gene mapping which really involves this process. Diaconis and Smith (1986) develop much of the fluctuation theory of coin-tossing for the cube. There is a lot going on, even in this simple example.

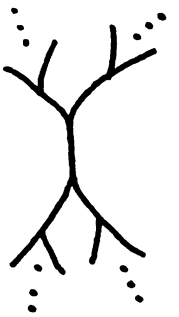
4. INFINITE GROUPS

For the most part, these notes treat problems involving finite groups. However, the techniques and questions are of interest in solving applied problems involving groups like the orthogonal group and p-adic matrix groups. Here is a brief description.

1. *The "Grand Tour" and a walk on O_n.* Statisticians often inspect high-dimensional data by looking at low-dimensional projections. To give a specific example, let $x_1, \ldots, x_{500} \in \mathbb{R}^{20}$ represent data on the Fortune 500 companies. Here x_1, the data for company 1, might have coordinates x_{11} = total value, x_{12} = number of women employed, etc. For $\gamma \in \mathbb{R}^{20}$, the projection in direction γ would be a plot (say a histogram) of the 500 numbers $\gamma \cdot x_1, \ldots, \gamma \cdot x_{500}$. Similarly the data would be projected onto various two-dimensional spaces and viewed as a scatterplot. Such inspection is often done interactively at a computer's display screen, and various algorithms exist for changing the projection every few seconds so that a scientist interested in the data can hunt for structured views.

Such algorithms are discussed by D. Asimov (1983). In one such, the direction γ changes by a small, random rotation. Thus, one of a finite collection Γ_i of 20×20 orthogonal matrices would be chosen at random, and the old view is rotated by Γ_i. This leads to obvious questions such as, how long do we have to wait until the views we have seen come within a prescribed distance (say 5 degrees) of any other view. A good deal of progress on this problem has been made by Peter Matthews in his Stanford Ph.D. thesis. Matthews (1988a) uses Fourier analysis on the orthogonal group and diffusion approximations to get useful numerical and theoretical results.

2. *Salmon fishing and* $GL_2(Q_2)$. Consider a binary tree



Let us describe a homogeneous random walk on such a tree. A particle starts at a fixed vertex. An integer distance is picked from a fixed measure, and the particle moves to one of the finite sets of vertices at this distance at random. The particle continues to move in this way. Questions involving recurrence (does the particle ever get back to where it started?) and the distribution of the distance of the walk from its starting position were raised by population geneticists studying life along a river system.

Sawyer (1978) gives background and much elegant analysis. It turns out that the tree is a coset space (homogeneous space) of the 2×2 matrices with entries in the 2-adic rationals, with respect to the subgroup of matrices with 2-adic integer entries. Number theorists have worked out enough of the representation theory to allow a dedicated probabilist to get elegant formulas and approximations.

3. *Other groups.* There is of course vast literature on random walks on $\mathbb{R}^n$. This is summarized in Feller (1971) or in Spitzer (1964). Much of this material has been generalized to non-commutative groups. Heyer (1977) contains a thorough survey. Recently there has been a really successful attack on random walk problems on Lie groups. The work of Furstenberg and Guivarclh is beautifully summarized in Bougerol-Lacroix (1985).

B. The basic setup

We now formally define what we mean by "close to random" and introduce an inequality that allows a good bound on the distance to uniformity in terms of Fourier transforms. Let G be a finite group. Let P and Q be probability distributions on G. Define the *variation distance* between P and Q as

$$||P - Q|| = \max_{A \subset G} |P(A) - Q(A)|.$$

Because we will use it heavily, we pause to discuss some basic properties.

Exercise 2. Prove that

$$||P - Q|| = \frac{1}{2} \sum_s |P(s) - Q(s)| = \frac{1}{2} \max_{||f|| \le 1} |P(f) - Q(f)|, \tag{1}$$

where, in the last expression, f is a function from G to $\mathbb{R}$ with $|f(s)| \leq 1$, and $P(f) = \Sigma_s P(s)f(s)$ is the expected value of f under P. Also, prove the validity of the following interpretation of variation distance suggested by Paul Switzer: Given a single observation, coming from P or Q with probability $\frac{1}{2}$, you are to guess which. Your chance of being correct is $\frac{1}{2} + \frac{1}{2}||P - Q||$.

EXERCISE 3. Show that if U is uniform, and $h: G \to G$ is $1-1$, then

$$||P - U|| = ||Ph^{-1} - U|| \text{ where } Ph^{-1}(A) = P(h^{-1}(A)).$$

EXERCISE 4. Let $G = S_n$. Part (a): let P be defined by "card 1 is on top, all the rest are random." Thus, $P(\pi) = 0$ if $\pi(1) \neq 1$ and $P(\pi) = 1/(n-1)!$ otherwise. What is $||P-U||$? Part (b): suppose P is defined by "card 1 is randomly distributed in a fixed set A of positions, all the other cards are random?" What is $||P - U||$?

Further properties of the variation distance are given in the following remarks and in lemma 4 of Chapter 3, lemma 4 of Chapter 4 and lemma 5 of Chapter 4.

Remark 1. The variation distance can be defined for any measurable group. It makes the measures on G into a Banach space. For G compact, the measures are the dual of the bounded continuous functions and $||\ ||$ is the dual norm. For continuous groups, the variation distance is often not suitable, since the distance between a discrete and continuous probability is 1. In this case, one picks a natural metric on G, and uses this to metrize the weak-star topology. Of course, for finite groups, all topologies are equivalent and the variation distance is chosen because of the natural interpretation given by (1): if two probabilities are close in variation distance, they give practically the same answer for any question.

Remark 2. Consider a random walk on S_n. In the language of shuffling cards, it might be thought that the following notion would be a more suitable definition of when cards are close to uniformly well shuffled: suppose the cards are turned face up one at a time and we try to guess at the value of each card before it is shown. For the uniform distribution, as in Diaconis and Graham (1977), we expect to get $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$ right on average. If the deck is not well mixed, the increase in the number of cards we can guess correctly seems like a useful measure of how far we are from uniform. Formally, one may define a guessing strategy for each possible history. Its value on a given permutation π defines a function $f(\pi)$ and (1) shows that, on average, $|P(f) - H_n| < n||P - U||$ no matter what guessing strategy is used. This may serve as a guide for how small a distance $||P - U||$ to aim for.

Remark 3. The variation distance is closely related to a variety of other metrics. For example, two other widely used measures of distance between probabilities

are

$$d_H(P,Q) = \sum_s (P(s)^{\frac{1}{2}} - Q(s)^{\frac{1}{2}})^2 - \text{ Hellinger distance}$$

$$I(P,Q) = \sum_s P(s) \log[P(s)/Q(s)] - \text{ Kullback-Leibler separation.}$$

These satisfy

$$\frac{d_H}{2} \leq \| \quad \| \leq \sqrt{d_H(1 - d_H/4)} \leq \sqrt{d_H}$$

$$\sqrt{2}\| \quad \| \leq \sqrt{I}.$$

It follows that when d_H or I are small, the variation distance is small. The converse can be shown to hold under regularity conditions.

Metrics on probabilities are discussed by Dudley (1968), Zolatorev (1983) and Diaconis and Zabell (1982). Rachev (1986) is a recent survey.

THE BASIC PROBLEM.

We can now ask a sharp mathematical question: Let P be a probability on G. Given $\varepsilon > 0$, how large should k be so that $\|P^{*k} - U\| < \varepsilon$?

It is not hard to show that P^{*k} tends to uniform if P is not concentrated on a subgroup or a coset of a subgroup. Here is a version of the theorem due to Koss (1959):

Theorem 1. *Let G be a compact group. Let P be a probability on G such that for some n_0 and $c, 0 < c < 1$, for all $n > n_0$,*

$$(*) \qquad P^{*n}(A) > cU(A) \text{ for all open sets } A.$$

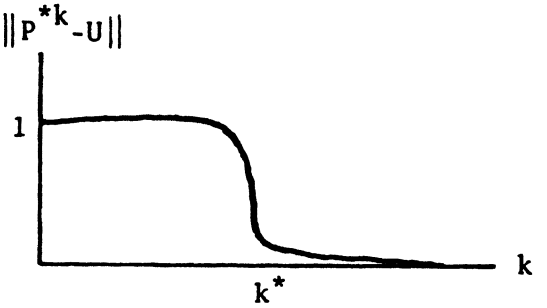
Then, for all k,

$$\|P^{*k} - U\| \leq (1 - c)^{[k/n_0]}.$$

Remarks. Condition * rules out periodicity. The conclusion shows that eventually the variation distance tends to zero exponentially fast. The result seems quantitative, but it's hard to use it to get bounds in concrete problems: as an example, consider simple random walk on Z_n. How large must k be to have the distance to uniform less than $\frac{1}{10}$? To answer, we must determine a suitable n_0 and c. This seems difficult. A short proof of the theorem is given here in Chapter 4.

There is a huge literature relating to this theorem. Heyer (1977) contains an extensive survey. Bhattacharya (1972) and Major and Shlossman (1979) contain quantitative versions which are more directly useable. Csiszar (1962) gives a proof which indicates "why it is true": briefly, convolving increases entropy and the maximum entropy distribution is the uniform. Bondesson (1983) discusses repeated convolutions of different measures.

Remark. The following "cut off" phenomena occurs in most cases where the computations can be done: the variation distance, as a function of k, is essentially 1 for a time and then rapidly becomes tiny and tends to zero exponentially fast past the cut off. Thus a graph might appear



We will determine these cut off points k^* in most of the examples discussed in Theorem 1. In such a case we will say that k^* steps suffice.

One purpose of this chapter is to discuss several ways of approximating the variation distance that give sharp non-asymptotic results. The basic tool used in the analytical approach of this section is the following inequality first used in Diaconis and Shahshahani (1981):

LEMMA 1. *(Upper bound lemma). Let Q be a probability on the finite group G. Then*

$$||Q - U||^2 \leq \frac{1}{4}\Sigma^* d_\rho \ Tr(\hat{Q}(\rho)\hat{Q}(\rho)^*)$$

where the sum is over all non-trivial irreducible representations.

Proof. From (1),

$$\begin{aligned} 4||Q - U||^2 &= \{\Sigma_s |Q(s) - U(s)|\}^2 \leq |G|\Sigma|Q(s) - U(s)|^2 \\ &= \Sigma^* d_\rho \ \mathrm{Tr}(\hat{Q}(\rho)\hat{Q}(\rho)^*). \end{aligned}$$

The inequality is Cauchy-Schwarz. The final equality is the Plancherel Theorem, and $\hat{Q}(\rho) = 1$ for ρ trivial, $\hat{U}(\rho) = 0$ for ρ non-trivial. □

Remark 1. The Cauchy-Schwarz inequality is not as crude as it may first appear. It is applied when Q is close to uniform, so $|Q(s) - U(s)|$ is roughly constant. In the examples of Section II below, and in all other "real" examples, the lemma gives the right answer in the sense that the upper bound matches a lower bound to one or two terms. The following exercise gives a lower bound of similar form. For some groups it shows the rate is off by at worst $\log|G|$. Exercise 14 gives a natural example, and Exercise 6 a contrived example, where this occurs.

EXERCISE 5. With the notation of the upper bound lemma, show that

$$||Q - U|| \geq \frac{1}{2|G|}\Sigma^* d_\rho \ \mathrm{Tr}(\hat{Q}(\rho)\hat{Q}(\rho)^*).$$

Also show

$$||Q - U||^2 \geq \frac{1}{4|G|}\Sigma^* d_\rho \ \mathrm{Tr}(\hat{Q}(\rho)\hat{Q}(\rho)^*).$$

EXERCISE 6. Let G be a finite group. Define a probability P on G by

$$P(\mathrm{id}) = 1 - \frac{\varepsilon}{2},\ P(s) = \frac{\varepsilon}{2(|G|-1)} \text{ for } s \neq \mathrm{id},\ 0 \leq \varepsilon \leq 2.$$

Show that

$$P^{*k}(\mathrm{id}) = \frac{1}{|G|} + \frac{|G|-1}{|G|}\left(1 - \frac{\varepsilon}{2}\,\frac{|G|}{|G|-1}\right)^k$$

$$P^{*k}(s) = \frac{1}{|G|} - \frac{1}{|G|}\left(1 - \frac{\varepsilon}{2}\,\frac{|G|}{|G|-1}\right)^k.$$

Using this, show that $||P^{*k} - U|| = \frac{|G|-1}{|G|}\,|1 - \frac{\varepsilon}{2}\frac{|G|}{|G|-1}|^k$. Show that

$$\Sigma^* d_\rho\ \mathrm{Tr}(\hat{P}(\rho)^k \hat{P}) = (|G|-1)\left(1 - \frac{\varepsilon}{2}\frac{|G|}{|G|-1}\right)^{2k}.$$

Remark 2. Lower bounds can be found by choosing a set $A \subset G$ and using $|Q(A) - U(A)| \leq ||Q - U||$. Often A can be chosen so that it is possible to calculate, or approximate, both $Q(A)$ and $U(A)$, *and* show that the distance between them is large. Several examples are given in the next section.

Remark 3. Total variation is used almost exclusively for the next two chapters. It is natural to inquire about the utility of the mathematically tractable L^2 norm

$$||P - U||_2^2 = \Sigma\left(P(s) - \frac{1}{|G|}\right)^2.$$

This has a fatal flaw: Suppose $|G|$ is even, and consider P uniformly distributed over half the points and zero on the others. $||P - U||_2 = \frac{1}{\sqrt{|G|}}$ is close to zero for $|G|$ large. Thus the interpretability of the L^2 norm depends on the size of the group. This makes it difficult to compare rates as the size of the group increases.

The norm $|G|^{\frac{1}{2}}||P - U||_2$ corrects for this. It seems somewhat artificial, and in light of the upper bound lemma and exercise 5, it is essentially the same as the variation distance.

C. SOME EXPLICIT COMPUTATIONS

Example 1. Simple random walk on the circle. Consider Z_p, the additive group of integers mod p. Define $P(1) = P(-1) = \frac{1}{2}$, $P(j) = 0$ otherwise. The following result shows that somewhat more than p^2 steps are required to get close to uniform.

Theorem 2. *For $n \geq p^2$, with p odd and greater than 7,*

$$||P^{*n} - U|| \leq e^{-\alpha n/p^2} \text{ with } \alpha = \pi^2/2.$$

Conversely, for $p \geq 7$ and any n

$$||P^{*n} - U|| \geq \frac{1}{2} e^{-\alpha n/p^2 - \beta n/p^4} \alpha = \pi^2/2, \ \beta = \pi^4/11.$$

Proof. The Fourier transform of P is

$$\hat{P}(j) = \frac{1}{2}\left(e^{\frac{2\pi i j}{p}} + e^{\frac{-2\pi i j}{p}}\right) = \cos(2\pi j/p).$$

The upper bound lemma yields

$$||P^{*n} - U||^2 \leq \frac{1}{4} \sum_{j=1}^{p-1} \cos(2\pi j/p)^{2n} = \frac{1}{2} \sum_{j=1}^{(p-1)/2} \cos(\pi j/p)^{2n}.$$

To bound this sum, use properties of cosine. One neat way to proceed was suggested by Hansmartin Zeuner: use the fact

$$\cos x \leq e^{-x^2/2} \quad \text{for } x \in [0, \pi/2].$$

This follows from $h(x) = \log(e^{x^2/2} \cos x), h'(x) = x - \tan x \leq 0$; so $h(x) \leq h(0) = 0$, for $x \in [0, \pi/2]$.

This gives

$$\begin{aligned} ||P^{*n} - U||^2 &\leq \frac{1}{2} \sum_{j=1}^{(p-2)/2} e^{-\pi^2 j^2 n/p^2} \leq \frac{1}{2} e^{-\pi^2 n/p^2} \sum_{j=1}^{\infty} e^{-\pi^2 (j^2-1) n/p^2} \\ &\leq \frac{1}{2} e^{-\pi^2 n/p^2} \sum_{j=0}^{\infty} e^{-3\pi^2 j n/p^2} \\ &= \frac{1}{2} \frac{e^{-\pi^2 n/p^2}}{1 - e^{-3\pi^2 n/p^2}}. \end{aligned}$$

This works for any n and any odd p. If $n \geq p^2$, $[2(1 - e^{-3\pi^2})]^{-1} < 1$ and thus we have proved more than we claimed for an upper bound.

Observe that the sum in the upper bound is dominated by the term with $k = \frac{p-1}{2}$. This suggests using the function $\cos(2\pi kj/p)$ alone to give a function bounded by 1 which has expected value zero under the uniform distribution. Using the symmetry of P,

$$\Sigma P^{*n}(j) \cos(2\pi jk/p) = \widehat{P^{*n}}(k) = \cos(2\pi k/p)^n = (-1)^n \cos(\pi/p)^n.$$

Now, (1) in section B above yields

$$2||P^{*n} - U|| \geq |\cos(\pi/p)^n|.$$

If $x \leq \frac{1}{2}$, $\cos x \geq e^{-x^2/2-x^4/11}$ say. This yields the lower bound with no conditions on n, for $p \geq 7$. □

Remark 1. The same techniques work to give the same rate of convergence (modulo constants) for other simple measures P such as $P(0) = P(1) = P(-1) = \frac{1}{3}$ or P uniform on $|j| \leq a$. Use of primitive roots and the Chinese remainder theorem gives rates for the multiplicative problem $X_n = a_n X_{n-1} \pmod{p}$ where $X_0 = 1$, and a_i are i.i.d. random variables taking values mod p. For example, suppose p is a prime and a is a primitive root mod p. Then the multiplicative random walk taking values $a, 1$ or $a^{-1} \pmod{p}$, each with probability $1/3$, takes $c(p)p^2$ steps to become random on the non-zero residues (mod p).

Remark 2. If n and p tend to infinity so $n/p^2 \to c$, the sum in the upper bound lemma approaches a theta function, so

$$||P^{*n} - U||^2 \leq \frac{1}{2}\sum_{j=1}^{\infty} e^{-\pi^2 j^2 c} + o(1).$$

Spitzer (1964), pg. 244) gives a similar result. Diaconis and Graham (1985b) show a similar theta function is asymptotically equal to the variation distance.

Remark 3. There are two other approaches to finding a lower bound in Theorem 1. Both result in a set having the wrong probability if not enough steps are taken.

Approach 1. For any set A, $||P^{*n} - U|| \geq |P^{*n}(A) - U(A)|$. Take $A = \{j: |j| \leq p/4\}$. Use the inversion theorem directly to calculate (and then approximate) $P^{*n}(A)$.

Approach 2. Consider a random walk on the integers Z taking steps ± 1 with probability $\frac{1}{2}$. Let S_n be the partial sum. The process considered in Theorem 1 is $S_n \pmod{p}$. Using the central limit theorem, if n is small compared to p^2, S_n has only a small chance to be outside $\{j: |j| \leq p/4\}$. This can be made rigorous using the Berry-Esseen theorem.

EXERCISE 7. Write out an honest proof, with explicit constants, for one of the two approaches suggested above. Show $||P^{*n} - U|| \to 1$ if $n = c(p)p^2, c(p) \to 0$.

Remark 4. The random walk based on $P(j) = P(-j) = \frac{1}{2}$ where $(j, p) = 1$ converges at the same rate as when $j = 1$ because of the invariance of variation distance under $1-1$ transformations (Exercise 3 above). Andrew Greenhalgh has shown that it is definitely possible to put $2k+1$ points down carefully, so that the random walk based on $P(j_1) = \ldots = P(j_{2k+1}) = 1/(2k+1)$ converges much faster ($c(p)p^{1/k}$ steps needed) than the random walk based on $P(j) = 1/(2k+1)$ for $|j| \leq k$.

It would be of interest to know the rate of convergence for "most" choices of k points.

The following exercises give other results connected to random walk on Z_p.

EXERCISE 8. Consider the random walk on Z_p generated by $P(1) = P(-1) = \frac{1}{2}$. It is intuitively clear (and not hard to prove) that the walk visits every point. There must be a point which is the last point visited (the last virginal point). Prove that this last point is *uniform* among the $n - 1$ non-starting points.

I do not know how to generalize this elegant result. Avrim Blum and Ernesto Ramos produced computation-free proofs of this result. Both showed that it fails for simple random walk on the cube Z_2^3.

EXERCISE 9. (Fan Chung). Prove that the convolution of symmetric unimodal distributions on Z_n is again symmetric unimodal.

EXERCISE 10. Let n be odd. Consider the random walk on Z_n generated by $P(1) = P(-1) = \frac{1}{2}$. Prove that after an even number of steps, the walk is most likely to be at zero. More generally, show that the walk is monotone in the sense that $P^{*2n}(j) \geq P^{*2n}(j + 2i)$ where $0 \leq j \leq j + 2i \leq n/2$.

This exercise originated in a statistical problem posed by Tom Ferguson. A natural way to test if an X taking values in Z_p is drawn from a uniform distribution is to look in a neighborhood of a prespecified point and reject uniformity if the point falls in that neighborhood. Consider the alternative H_1: $P = P^{*n}$ for simple random walk starting at the prespecified point. The exercise, combined with the Neyman-Pearson lemma implies classical optimality properties for this test.

Ron Graham and I showed that the same type of result holds for nearest neighbor walk on Z_2^n, but fails for nearest neighbor walks on a discrete torus like Z_{13}^3. Monotonicity also fails for the walk originally suggested by Ferguson, namely random transpositions in the symmetric group (see Section D of this chapter) with neighborhoods given by Cayley's distance — the minimum number of transpositions required to bring one permutation into another (see Chapter 6-B).

Example 2. Nearest neighbor walk on Z_2^d. Define $P(0) = P(0\ldots01) = P(0\ldots10) = \ldots = P(10\ldots0) = \frac{1}{d+1}$. The random walk generated by P corresponds to staying where you are, or moving to one of the d nearest neighbors, each with chance $\frac{1}{(d+1)}$. The following result is presented as a clear example of a useful lower bound technique.

Theorem 3. *For P as defined above, let $k = \frac{1}{4}(d+1)[\log d + c]$,*

$$||P^{*k} - U||^2 \leq \frac{1}{2}(e^{e^{-c}} - 1). \tag{1}$$

As $d \to \infty$, for any $\varepsilon > 0$ there is a $C < 0$ such that $c < C$ and k as above imply

$$||P^{*k} - U|| \geq 1 - \varepsilon. \tag{2}$$

Proof. There is a 1-dimensional representation associated to each $x \in Z_2^d$; $\hat{P}(x) = \sum_y (-1)^{x \cdot y} P(y) = 1 - \frac{2\omega(x)}{d+1}$ where $\omega(x)$ is the number of ones (or weight) of x.

The upper bound lemma gives

$$||P^{*k} - U||^2 \leq \frac{1}{4}\sum_{x\neq 0}(\hat{P}(x))^{2k} = \frac{1}{4}\sum_{j=1}^{d}\binom{d}{j}(1-\frac{2j}{d+1})^{2k}$$
$$\leq \frac{2}{4}\sum_{j=1}^{d/2}\frac{d^j}{j!}e^{-j\log d - jc} = \frac{1}{2}\sum_{j=1}^{d/2}\frac{e^{-jc}}{j!} \leq \frac{1}{2}(e^{e^{-c}}-1).$$

For the lower bound observe that the dominating terms in the upper bound come from x of weight 1. Define a random variable $Z: Z_2^d \to \mathbb{R}$ by $Z(x) = \sum_{i=1}^{d}(-1)^{x_i} = d - 2\omega(x)$. Under the uniform distribution, x_i are independent fair coin tosses so $E_U Z = 0$, $\mathrm{Var}_U(Z) = d$, and Z has an approximate normal distribution. Under the distribution P^{*k}, arguing as in Example 1 shows

$$E_k(Z) = d(1-\frac{2}{d+1})^k, \quad E_k(Z^2) = d + d(d-1)(1-\frac{4}{d+1})^k.$$
$$\mathrm{Var}_k(Z) = d + d(d-1)(1-\frac{4}{d+1})^k - d^2(1-\frac{2}{d+1})^{2k}.$$

With $k = \frac{1}{4}((d+1)\log d + cd + c)$, as $d \to \infty$,

$$E_k(Z) = \sqrt{d}e^{-c/2}(1+0(\frac{\log d}{d})),$$
$$d(d-1)(1-\frac{4}{d+1})^k = (d-1)e^{-c}(1+0(\frac{\log d}{d}))$$
$$d^2(1-\frac{2}{d+1})^{2k} = de^{-c}(1+0(\frac{\log d}{d})).$$

So $\mathrm{Var}_k(Z) = d + 0(e^{-c}\log d)$ uniformly for $c = o(\log d)$. Note that asymptotically $\mathrm{Var}_k(Z) \sim d$, independent of c of order $0(\log d)$. This is crucial to what follows.

For the lower bound, take $A = \{x: |Z(x)| \leq \beta\sqrt{d}\}$. Then we have

$$||P^{*k} - U|| \geq |P^{*k}(A) - U(A)|.$$

From (3) and Chebychev,

$$U(A) \geq 1 - \frac{1}{\beta^2}$$

$$P^{*k}(A) \leq P^{*k}\{|Z - E_k(Z)| \geq E_k(Z) - \beta\sqrt{d}\} \leq \frac{\mathrm{Var}_k(Z)}{(E_k(Z) - \beta\sqrt{d})^2}$$
$$\sim \frac{1}{(e^{-c/2} - \beta)^2} \text{ as } d \to \infty.$$

Choosing β large, and c suitably large (and negative) results in $||P^{*k} - U|| \to 1$. □

Remark 1. In this example, the set A has a natural interpretation as the set of all binary vectors with weight close to $\frac{d}{2}$. Since the random walk starts at 0, if it doesn't run long enough, it will tend to be too close to zero.

Remark 2. It is somewhat surprising that $\frac{1}{4}d \log d$ steps are enough: It takes $\frac{1}{2}d \log d$ steps to have the right chance of reaching the opposite corner $(11\ldots1)$.

Example 3. Simple random walk with randomness multiplier. Let p be an odd number. Define a process on Z_p by $X_0 = 0$, $X_n = 2X_{n-1} + \varepsilon_n (\text{mod } p)$ where ε_i are independent and identically distributed taking values 0, ± 1 with probability $\frac{1}{3}$. Let P_n be the probability distribution of X_n. In joint work with Fan Chung and R. L. Graham it was shown that $n = c \log p \text{ loglog } p$ steps are enough to get close to uniform. Note that X_n is based on the same amount of random input as simple random walk discussed in Example 1. The deterministic doubling serves as a randomness multiplier speeding convergence from p^2 to $\log p \text{ loglog } p$ steps.

Theorem 4. *For P_n defined above, if*

$$n \geq \log_2 p[\frac{\text{loglog}_2 p}{\log 9} + s],$$

then

$$||P_n - U||^2 \leq \frac{1}{2}(e^{9^{-s}} - 1).$$

Proof. Since $X_0 = 0$, $X_1 = \varepsilon_1$, $X_2 = 2\varepsilon_1 + \varepsilon_2, \ldots, X_n = 2^{n-1}\varepsilon_1 + \ldots + \varepsilon_n (\text{mod } p)$. This reduces the problem to a computation involving independent random variables. The Fourier transform of P_n at frequency j is

$$\prod_{a=0}^{n-1} (\frac{1}{3} + \frac{2}{3} \cos \frac{2\pi 2^a j}{p}).$$

Since

$$(\frac{1}{3} + \frac{2}{3} \cos(2\pi x))^2 \leq h(x) \stackrel{\text{d}}{=} \begin{cases} \frac{1}{9} & \text{if } x \in [\frac{1}{4}, \frac{3}{4}) \\ 1 & \text{otherwise.} \end{cases}$$

It will be enough to bound

$$\prod_{a=0}^{n-1} h(\{\frac{2^a j}{p}\}),$$

where $\{\cdot\}$ denotes fractional part. Observe that if the (terminating) binary expansion of x is $x = \cdot\alpha_1\alpha_2\alpha_3\ldots$, then

$$h(x) = \begin{cases} \frac{1}{9} & \text{if } \alpha_1 \neq \alpha_2 \\ 1 & \text{if } \alpha_1 = \alpha_2. \end{cases}$$

Let $A(x,n)$ denote the number of alternations in the first n binary digits of x: $A(x,n) = |\{1 \leq i < n : \alpha_i \neq \alpha_{i+1}\}|$. Successively multiplying by powers of 2 just shifts the bits to the left. It follows that

$$\prod_{a=0}^{n-1} h(\{\frac{2^a j}{p}\}) \leq 9^{-A(j/p,n)}.$$

Suppose first that p is of the special form $p = 2^t - 1$. The fractions j/p become

$$\begin{aligned} 1/p &= \overbrace{00\ldots01}^{t}\ \overbrace{00\ldots01}^{t}\ \ldots \\ 2/p &= 00\ldots10\ \ 00\ldots10\ \ldots \\ 3/p &= 00\ldots11\ \ 00\ldots11\ \ldots \\ p-1/p &= 11\ldots10\ \ 11\ldots10\ \ldots \end{aligned}$$

If $n = rt$, the number of alternations in the first n places of row j/p is no smaller than r times the number of alternations in the first t places of j/p. It follows that

$$\begin{aligned} \sum_{j=1}^{p-1} \prod_{a=0}^{n-1} h\{\frac{2^a j}{p}\} &\leq \sum_{j=1}^{p-1} 9^{-rA(j/p,t)} \\ &\leq 2\sum_{k=1}^{t} \binom{t}{k} 9^{-kr} = 2[(1+9^{-r})^t - 1] \\ &\leq 2[e^{t9^{-r}} - 1]. \end{aligned} \tag{1}$$

The second inequality in (1) used the easily proved bound $|j: A(\frac{j}{p}, t) = k| \leq 2\binom{t}{k}$. Now, if $n = rt$ with $r = \frac{\log t}{\log 9} + s$, the upper bound lemma gives

$$||P_n - U||^2 \leq \frac{1}{2}[e^{9^{-s}} - 1]$$

as claimed.

For general odd p, define t by $w^{t-1} < p < 2^t$. For r as chosen above, partition the initial $n = rt$ digits in the binary expansion of j/p into r blocks of length t: $B(i,j) 1 \leq i \leq r$:

$$j/p = \overbrace{\alpha_1 \ldots \alpha_t}^{B(1,j)}\ \overbrace{\alpha_{t+1} \ldots \alpha_{2t}}^{B(2,j)} \ldots\ \overbrace{\alpha_{(r-1)t+1} \ldots \alpha_{rt}}^{B(r,j)}.$$

Thus,

$$\sum_{j=1}^{p-1} \prod_{a=0}^{n-1} h(\{\frac{2^a j}{p}\}) \leq \sum_{j=1}^{p-1} 9^{-A(B(1,j))-\ldots-A(B(r,j))}. \tag{2}$$

By the choice of t, all the blocks $B(1,j)$, $1 \le j \le p-1$ in the first column are distinct and have at least one alternation. Furthermore, since $(2,p)=1$, the set of blocks in the ith column does not depend on i. This information will be used together with the following interchange lemma: If $0 < \alpha < 1$, and $a \le a'$, $b \le b'$, then

$$\alpha^{a+b'} + \alpha^{a'+b} \le \alpha^{a+b} + \alpha^{a'+b'}.$$

To prove this, simply expand $(\alpha^a - \alpha^{a'})(\alpha^b - \alpha^{b'}) \ge 0$. The lemma implies that collecting together terms with the same blocks in the exponent only increases the upper bound. Thus, the right side of (2) is no bigger than

$$\sum_{j=1}^{p-1} 9^{-r\ A(j/2^t-1,t)},$$

the sum that appears in equation (1) above! Using the bound there completes the proof. □

Remark 1. A more careful version of the argument implies that for any odd p, the cutoff is at $c^* \log_2 p \operatorname{loglog}_2 p$ with $c^* = 1/\log_2 \pi_1$ where

$$\pi_1 = \prod_{a=1}^{\infty} (\frac{1}{3} + \frac{2}{3}\cos(2\pi/2^a))^2.$$

Chung, Diaconis and Graham (1987) show that for p of form $2^t - 1$, $c^* t \log t$ steps are required. The lower bound technique again uses the "slow" terms in the upper bound lemma to define a random variable $Z(j) = \Sigma_{|k|=1} e^{2\pi i jk/p}$ where the sum is over k's with a single 1 in their binary expression. Under the uniform distribution Z has mean 0 and "variance" $(= E(Z\overline{Z})) = t$. Under P_n, Z has mean approximately $t\pi_1^{\frac{1}{2}}$ and variance of order $\sqrt{t}$.

Chung, Diaconis and Graham also prove that for most odd p, $1.02 \log_2 p$ steps are enough. A curious feature of the proof is that we do not know single explicit sequence of p's such that $2 \log p$ steps suffice to make the variation distance smaller than $\frac{1}{10}$.

Remark 2. There is a natural generalization of this problem which may lead to further interesting results. Let G be an Abelian group. Let $A: G \to G$ be an automorphism (so A is 1-1, onto and $A(s+t) = A(s) + A(t)$). Consider the process

$$X_n = A(X_{n-1}) + \epsilon_n$$

where $X_0 = \text{id}$ and ϵ_i are iid. This can be represented as a convolution of independent random variables

$$X_n = A^{n-1}(\epsilon_1) + A^{n-2}(\epsilon_2) + \ldots + \epsilon_n.$$

If $A^k = \text{id}$, these can be further grouped in blocks of k (when k divides n) to give a sum of iid variables. Then, methods similar to those used in the present example may provide rates.

It is not necessary to use an automorphism; $f(s) = A(s) + t$, with $t \varepsilon G$ fixed and A an automorphism works in a similar way. It is not necessary that G be Abelian. If the Law of ϵ_i is constant on conjugacy classes so is the law of $A(\epsilon_i)$ *and* the random variables commute in distribution (see exercise 2.7).

One natural example to try has $G = \mathbb{Z}_n^2$, A *a* 2×2 matrix, and ϵ_i the nearest neighbor random variable taking values $(00), (01), (0-1), (10), (-10)$ each with probability $\frac{1}{5}$.

Remark 3. Fourier techniques can be used to bound other distances. This remark gives a result for the maximum over all "intervals" of Z_p. The next remark discusses arbitrary compact groups. The techniques are close to work of Joop Kemperman (1975).

Let P and Q be probabilities on Z_p. Define $D(P,Q) = \sup_J P(J) - Q(J)$ where the sup is over all "intervals" in Z_p.

LEMMA. $D(P,Q) \leq \frac{1}{\sqrt{2}} \sum_{j=1}^{p-1} |\hat{P}(j) - \hat{Q}(j)|/j^*$ *where* $j^* = \min(j, p-j)$.

Proof. For J an interval on the circle, $|P(J) - Q(J)| = |P(J^c) - Q(J^c)|$, where of course J^c is an interval too. It follows that only intervals not containing zero need be considered. Let $[\ell_1, \ell_2]$ be such an interval, with $\ell_1 < \ell_2$ (clockwise). Then

$$P([\ell_1, \ell_2]) - Q([\ell_1, \ell_2]) = P([0, \ell_2]) - P([0, \ell_1)) - Q([0, \ell_2]) + Q([0, \ell_1)).$$

Now

$$P([0,\ell]) = \sum_{a=0}^{\ell} P(a) = \frac{1}{p} \sum_{a=0}^{\ell} \sum_{j=0}^{p-1} \hat{P}(j) e^{-\frac{2\pi i j a}{p}}$$

$$= \frac{1}{p} \sum_{j=0}^{p-1} \hat{P}(j)(1 - e^{-2\pi i(\ell+1)j/p})/(1 - e^{-2\pi i j/p}).$$

This implies that $P - Q$ equals

$$\frac{1}{p} \sum_{j=1}^{p-1} [\hat{P}(j) - \hat{Q}(j)][e^{-2\pi i \ell_1 j/p} - e^{-2\pi i(\ell_2+1)j/p}]/(1 - e^{-2\pi i j/p}).$$

Thus $D(P,Q)$ is bounded above by

$$\frac{\sqrt{2}}{p} \sum_{j=1}^{p-1} |\hat{P}(j) - \hat{Q}(j)|/\sqrt{1 - \cos(2\pi j/p)}.$$

Now for $0 \leq x \leq \frac{\pi}{2}$, $1 - \cos x \geq \frac{x^2}{3}$, so for $1 \leq j \leq p/4$, $\frac{\sqrt{2}}{p}(1 - \cos(2\pi j/p))^{-\frac{1}{2}} \leq \sqrt{6}/2\pi j \leq \frac{1}{j\sqrt{2}}$. For $\frac{\pi}{2} \leq x \leq \pi, 1 - \cos x \geq 1$, so for $p/4 \leq j \leq p/2$, $\frac{1}{p}(1 - \cos(1\pi j/p))^{-\frac{1}{2}} \leq \frac{1}{p} \leq \frac{1}{2j}$. For the rest, $\cos(2\pi j/p) = \cos(2\pi(p-j)/p)$. □

EXERCISE 11. Using this lemma, with P_n as defined in Example 3, show there are constants a and b such that for every odd p,

$$D(P_n, U) \leq ae^{-bn/\log p}.$$

Remark 4. There must be similar bounds for any compact group. To see the use of such results let T be the unit circle: $T = \{z \in C: |z| = 1\}$. Fix an irrational $\alpha \in T$ and consider simple random walk with step size α, thus $X_0 = 0$, and $X_n = X_{n-1} \pm \alpha$. Since X_n is concentrated on a discrete set, the variation distance to uniform is always 1. Nonetheless, the distribution of X_n converges to the uniform distribution in the weak star topology. To discuss rates, a metric must be chosen. A convenient one is

$$D(P, Q) = \sup_I |P(I) - Q(I)|$$

for I ranging over intervals of T. This metrizes weak star convergence.

Kemperman (1975) proves two useful inequalities that give bounds on D involving the Fourier transform for P a probability on T, and $m \in Z$,

$$\hat{P}(m) = \int_0^1 e^{2\pi i m x} P(dx).$$

(1) $D(P, U) = \sup_I P(I) - U(I)| \leq \{12 \sum_{m=1}^{\infty} |\hat{P}(m)|^2/(2\pi m)^2\}^{\frac{1}{3}}$.

(2) $D(P, U) \leq \frac{2}{\pi} \sum_{m=1}^{\infty} |\hat{P}(m)|/m$.

Niederreiter and Philipp (1973) discuss multivariate versions.

EXERCISE 12. Consider simple random walk on the unit circle, as in remark 3 above, with α a quadratic irrational. Use bounds (1) and (2) above to estimate rates of convergence. A direct combinatorial argument can be used to show that $D(P^{*n}, U) \leq c(\log n)/\sqrt{n}$.

It seems quite possible to carry over bounds like (2) in Remark 4 to any compact group G. Choose a bi-invariant metric $d(x, y)$ on G and consider $D(P, Q) = \sup|P(I) - Q(I)|$ where I ranges over all translates of balls centered at the identity. Then $D(P, Q)$ can be bounded as in remark 2; Lubotzky, Phillips, and Sarnak (1986) give results for the sphere. Their paper makes fascinating use of deep number theory which must be useful for other problems. Chapter 6 below discusses bi-invariant metrics.

Example 4. Random walks on the affine group A_p. (An elaborate exercise). Let p be a prime. Random numbers are often generated by the recursive scheme $X_n = aX_{n-1} + b (\text{mod } p)$. This sequence of exercises allows estimates of the rate of convergence when a and b are random. The transformation $x \to ax + b$ with a non-zero (mod p) will be written $T_{ab}(x)$. The set of such transformations form

a finite group A_p. We write (a, b) for the typical group element. The product is $(a, b)(c, d) = (ac, ad + b)$, the identity is $(1, 0)$ and $(a, b)^{-1} = (a^{-1}, -ba^{-1})$. This group has $p(p-1)$ elements. The subgroups $\{(1, b)\} \cong Z_p$ and $\{(a, 0)\} \cong Z_p^*$ are useful.

(1) Identify the p distinct conjugacy classes. Explain why measures constant on conjugacy classes are not very interesting.

(2) From part (1) there are p distinct irreducible representations; $p-1$ of these are the 1-dimensional representations given by choosing a character χ_i of Z_p^* and defining $\rho_i(a, b) = \chi_i(a)$. Show that these are distinct irreducible representations. Show that there is one other irreducible representation ρ of degree $p-1$. Use Serre's exercise 2-6 to construct this representation by considering the action of A_p on Z_p. By explicitly choosing a basis, show

$$\chi_\rho(1, 0) = p - 1,$$
$$\chi_\rho(1, b) = -1, \quad b \neq 0$$
$$\chi_\rho(a, b) = 0, \quad a \neq 1.$$

(3) Let ρ^+ and ρ^* be the restriction of ρ in Part 2 to Z_p and Z_p^* respectively. Using the character of ρ and the inner product machinery, show that ρ^* is the regular representation of Z_p^* and ρ^+ contains each non-trivial irreducible representation of Z_p once.

(4) Let P^+ be a probability on Z_p and P^* a probability on Z_p^*. Let χ_i^+ and χ_i^* be characters of Z_p and Z_p^*. Let $P(a, b) = P^*(a) \cdot P^+(b)$. Show

(a) $\hat{P}(\rho) = \hat{P}^+(\rho^+) \cdot \hat{P}^*(\rho^*)$.

(b) The eigenvalues of the matrix $\hat{P}^*(\rho^*)$ are the p-1 numbers $\hat{P}^*(\chi_i^*)$; the eigenvalues of $\hat{P}^+(\rho^+)$ are the p-1 numbers $\hat{P}^+(\chi_i^+)$, χ_i^+ non-trivial.

(5) Let p be an odd prime such that 2 is a primitive root mod p. Consider the random walk on A_p which starts at 0 and is based on P^*, P^+, with $P^*(1) = P^*(2) = P^*((p+1)/2) = \frac{1}{3}$ and $P^+(0) = P^+(1) = P^+(-1) = \frac{1}{3}$. Show that $k = c(p)p^2$ log steps are enough to get arbitrarily close to random if $c(p) \to \infty$ as p does. Use this to argue that the random point $T_{X_n}(0)$ is close to uniformly distributed on Z_p after this many steps.

One way through the computations uses the following fact. Let $\tau(A)$ be the spectral radius (= max|eigenvalue|) of the matrix A. If A and B are diagonalizable matrices then $\tau(AB) \leq \tau(A)\tau(B)$.

(6) Show by considering the first coordinate of (a, b) that $k = cp^2$ steps are not enough if c is fixed.

Remark. The argument sketched above gives $c(p)p^2 \log p$. I presume that $c(p)p^2$ is the correct answer. Actually, numerical computation strongly suggests that the random walk $X_n = aX_{n-1} + b$, where (a, b) has the distribution described in part 5, becomes uniform in order $(\log p)^A$ steps for $A = 1$ or 2.

When p is composite there are more conjugacy classes. It is an interesting exercise to determine these. I have not succeeded in finding a natural "small" measure constant on conjugacy classes which permits analysis.

D. Random transpositions: an introduction to the representation theory of the symmetric group.

As described in Section A, repeated random transpositions of n cards in a row can be modeled as repeatedly convolving the following measure:

$$(1) \qquad P(\mathrm{id}) = \frac{1}{n},\ P(\tau) = \frac{2}{n^2} \text{ for } \tau \text{ a transposition}, P(\pi) = 0 \text{ otherwise}.$$

This section presents a proof of the following theorem

Theorem 5. *Let $k = \frac{1}{2}n \log n + cn$. For $c > 0$,*

$$||P^{*k} - U|| \leq ae^{-2c}$$

for a universal constant a. Conversely, for $c < 0$, as n tends to infinity

$$||P^{*k} - U|| \geq (\frac{1}{e} - e^{-e^{-2c}}) + o(1).$$

The proof introduces some basic results about the representation theory of the symmetric groups. Most all of these will be treated in greater detail in Chapter 7. This problem was first treated by Diaconis and Shahshahani (1981). The present argument is based on simplifications suggested by Leo Flatto, Andrew Odlyzko, and Hansmartin Zeuner. After the proof, several further problems, *to which the same analysis applies*, are described.

Discussion The measure P is constant on conjugacy classes: $P(\eta\pi\eta^{-1}) = P(\pi)$. Thus Schur's lemma implies, for any irreducible representation ρ, $\hat{P}(\rho) = \text{constant} \cdot I$. Taking traces, the constant equals $(\frac{1}{n} + \frac{n-1}{n}r(\rho))$ with $r(\rho) = \chi_\rho(\tau)/d_\rho$. Here $\chi_\rho(\tau)$ denotes the character of ρ at any transposition τ and d_ρ denotes the dimension of ρ (see proposition 6 of Chapter 2). Now, the upper bound lemma yields

$$||P^{*k} - U||^2 \leq \frac{1}{4}\sum_{\rho}^{*} d_\rho^2(\frac{1}{n} + \frac{n-1}{n}r(\rho))^{2k}.$$

The following heuristic discussion may help understanding. Table 1 gives d_ρ and $\chi_\rho(\tau)$ for $n = 10$. There are 42 irreducible representations of S_{10}. For example, the first entry is $d_\rho = 1$, $\chi_\rho(\tau) = 1$ for the trivial representation. The second entry is $d_\rho = 9$, $\chi_\rho(\tau) = 7$ for the 9-dimensional permutation representation. Except for a few representations at the ends, the ratio $|\chi_\rho(\tau)/d_\rho|$ is small. Suppose it could be shown that $\chi_\rho(\tau)/d_\rho \leq \frac{1}{2}$ for most ρ, then, approximately for n large, $|\frac{1}{n} + \frac{n-1}{n}r(\rho)| \leq \frac{1}{2}$ and the upper bound above would be smaller than

$$\frac{1}{4}(\frac{1}{2})^{2k}\Sigma d_\rho^2 = \frac{1}{4}(\frac{1}{2})^{2k}n! \text{ (using proposition 5 of Chapter 2).}$$

Table 1
Characters of S_{10} (from James and Kerber (1981, pg. 354))

Partition	dim	$\chi_\rho(\tau)$
[10]	1	1
[9,1]	9	7
[8,2]	35	21
[8,1,1]	36	20
[7,3]	75	35
[7,2,1]	160	64
[7,1,1,1]	84	28
[6,4]	90	34
[6,3,1]	315	91
[6,2,2]	225	55
[6,2,1,1]	350	70
[6,1,1,1,1]	126	14
[5,5]	42	14
[5,4,1]	288	64
[5,3,2]	450	70
[5,3,1,1]	567	63
[5,2,2,1]	525	35
[5,2,1,1,1]	448	0
[5,1,1,1,1,1]	126	-14
[4,4,2]	252	28
[4,4,1,1]	300	20
[4,3,3]	210	14
[4,3,2,1]	768	0
[4,3,1,1,1]	525	-35
[4,2,2,2]	300	-20
[4,2,2,1,1]	567	-63
[4,2,1,1,1,1]	350	-70
[4,1,1,1,1,1,1]	84	-28
[3,3,3,1]	210	-14
[3,3,2,2]	252	-28
[3,3,2,1,1]	450	-70
[3,3,1,1,1,1]	225	-55
[3,2,2,2,1]	288	-64
[3,2,2,1,1,1]	315	-91
[3,2,1,1,1,1,1]	160	-64
[3,1,1,1,1,1,1,1]	36	-20
[2,2,2,2,2]	42	-14
[2,2,2,2,1,1]	90	-34
[2,2,2,1,1,1,1]	75	-35
[2,2,1,1,1,1,1,1]	35	-21
[2,1,1,1,1,1,1,1,1]	9	-7
[1,1,1,1,1,1,1,1,1,1]	1	-1

Now using Stirling's formula,

$$(\frac{1}{2})^{2k} n! \doteq e^{-2k \log 2 + n \log n - n + \cdots}.$$

It follows that if k is $n \log n$, the upper bound will tend to zero. To complete this heuristic discussion, consider the term arising from the $n-1$ dimensional representation: $d_\rho = n-1$, and $\chi_\rho(\tau) = n-3$. This is easy to see: the trace of the permutation representation for a transposition is $n-2$. The permutation representation is the direct sum of the trivial representation and the $n-1$ dimensional representation so $n-2 = \chi_\rho(\tau) + 1$. Here $(\frac{1}{n} + \frac{n-1}{n} r(\rho))^{2k} = (1 - \frac{2}{n})^{2k}$. This is a far cry from $(\frac{1}{2})^{2k}$. Persevering, in the upper bound lemma k has to be chosen large enough to kill

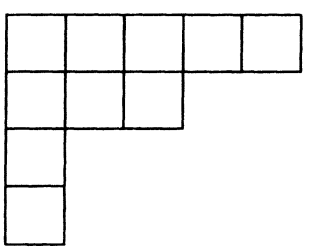
$$(n-1)^2 (1 - \frac{2}{n})^{2k} = e^{2k \log(1-\frac{2}{n}) + 2 \log(n-1)} = e^{-\frac{4k}{n} + 2 \log n + 0(\frac{k}{n^2})}.$$

For $k = \frac{1}{2} n \log n + cn$ this is asymptotic to e^{-4c}. Taking square roots gives the e^{-2c} of the theorem.

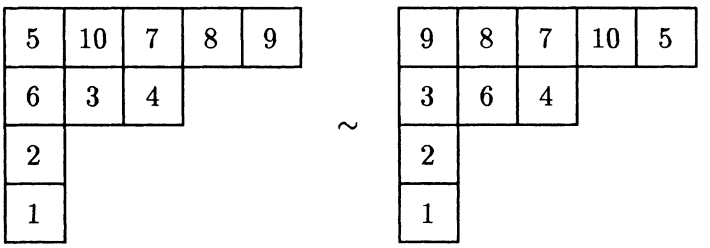
It will turn out that this is the slowest term, the other terms being geometrically smaller, and most terms being smaller than $(\frac{1}{2})^{2k}$.

This argument is somewhat similar to the bounds for simple random walk on Z_p: terms near the trivial representation needed to be summed up carefully, terms far away were geometrically small and easily dealt with. Putting in the details for Z_p required properties of cosine. For the symmetric group, the representations are usefully thought of as 2-dimensional shapes. Properties of d_ρ and $\chi_\rho(\tau)$ will have to be developed.

To begin a more detailed discussion, consider a partition of n, say $\lambda = (\lambda_1, \ldots \lambda_m)$ with $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_m$ positive integers with $\lambda_1 + \ldots + \lambda_m = n$. There is a one to one correspondence between irreducible representations of S_n and partitions of n. This is carefully described in Chapter 7. For present purposes, the notion of the *diagram* associated to a partition is helpful. An example says things best: the diagram corresponding to (5,3,1,1) is



The first row of the diagram contains λ_1 squares, etc. A diagram containing numbers 1, 2,..., n is called a *tableaux*. Two tableaux are considered equivalent if they have the same row sets:



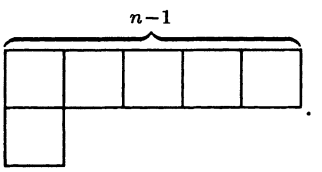
An equivalence class of tableaux is called a called a *tabloid.* There are $n!/\lambda_1!\ldots\lambda_m!$ distinct tabloids of a given shape. These are used to build a representation called M^λ as follows. Consider a vector space with basis vectors $\{e_t\}$ where t runs over all tabloids of shape λ. For π a permutation, define $\rho(\pi)$ by defining on basis vectors:

$$\rho(\pi)(e_t) = e_{\pi t}$$

where for example, π applied to the tabloid

5 10 7 8 9		$\pi(5)\ \pi(10)\ \pi(7)\ \pi(8)\ \pi(9)$
6 3 4	is the	$\pi(6)\ \pi(3)\ \pi(4)$
2	tabloid	$\pi(2)$
1		$\pi(1)$

Here are some examples: the partition $(n-1,1)$ has $n!/(n-1)! = n$ distinct tabloids, all of shape



These are evidently completely determined by the one number in the second row. Hence the vector space $M^{n-1,1}$ is just the n-dimensional space spanned by the usual basis $e_1,\ldots e_n$ with $\rho(\pi)e_i = e_{\pi(i)}$. The partition $n-2,1,1$ gives rise to a vector space $M^{n-2,1,1}$ with basis $\{e_{(i,j)}\}$ and $\rho(\pi)e_{(i,j)} = e_{(\pi(i),\pi(j))}$. The partition $n-2,2$ gives rise to $M^{n-2,2}$ with basis $\{e_{\{i,j\}}\}$ where $\{i,j\}$ runs through all unordered pairs.

The representations M^λ are all reducible except of course for $\lambda = n$. It will be argued that each M^λ contains a well-defined irreducible subspace S^λ, and as λ varies the S^λ range over all the irreducible representations of S_n. The following two facts are all that is needed to prove Theorem 5.

FACT 1. The dimension of the irreducible representation corresponding to partition λ is the number of ways of placing the numbers $1,2,\ldots,n$ into the diagram of λ such that the entries in each row and column are increasing.

This fact is Theorem (8.4) in James (1978) who discusses other formulas for the dimension. These are also described in Chapter 7 below. A useful computational corollary is

$(D-1)$ The dimension d_λ of the irreducible representation corresponding to the partition λ satisfies $d_\lambda \leq \binom{n}{\lambda_1} d_{\lambda^*}$ with $\lambda^* = (\lambda_2, \lambda_3, \ldots, \lambda_m)$ a partition of $n - \lambda_1$.

Proof. The first row may be chosen in $\binom{n}{\lambda_1}$ ways. For each choice of first row, the number of ways of placing the $n - \lambda_1$ remaining numbers is d_{λ^*}. Of course not all of these combine with the choice of first row to give a monotone total placement. This gives the inequality. □

FACT 2. Let $r(\lambda) = \chi_\lambda(\tau)/d_\lambda$ where $\chi_\lambda(\tau)$ is the character at a transposition and d_λ is the dimension of the irreducible representation corresponding to the partition λ of n. Then

$$(D-2) \qquad r(\lambda) = \frac{1}{n(n-1)} \Sigma[\lambda_j^2 - (2j-1)\lambda_j] = \frac{1}{\binom{n}{2}} \sum_j \binom{\lambda_j}{2} - \binom{\lambda'_j}{2}$$

with λ' the transpose of λ.

This is a special case of a result due to Frobenius who essentially determined formulas for all of the characters. These become unwieldly for complex conjugacy classes. An accessible proof of (D-2) is given in Ingram (1950).

Using Frobenius' formula, Shahshahani and I proved a simple monotonicity result: Call partition λ^1 larger than partition λ^2 if it is possible to get from the diagram of λ^2 to the diagram of λ^1 by moving boxes up to the right. This is a partial order. For example $(5,1) \geq (4,2) \geq (3,3)$, but $(3,3)$ and $(4,1,1)$ are not comparable, though both are larger than $(3,2,1)$. This order is widely used in statistics under the name of majorization (see e.g. Marshall and Olkin (1979)). James (1978, pg. 8) contains further examples.

LEMMA 1. *If $\lambda \geq \lambda'$, then $r(\lambda) \geq r(\lambda')$ where $r(\lambda) = \chi_\lambda(\tau)/d_\lambda$ is given by (D-2).*

Proof. It suffices to consider the case where one box is switched from row b to $a (b > a)$, i.e. $\lambda_a = \lambda'_a + 1$, $\lambda_b = \lambda'_b - 1$, $\lambda_c = \lambda'_c$ for $c \neq a, b$. Formula (D-2) shows that

$$\begin{aligned} r(\lambda) - r(\lambda') &= \frac{1}{n(n-1)} \{\lambda_a^2 - (2a-1)\lambda_a - \lambda_a'^2 + (2a-1)\lambda'_a + \\ &\qquad \lambda_b^2 - (2b-1)\lambda_b - \lambda_b'^2 + (2b-1)\lambda'_b\} \\ &= \frac{1}{n(n-1)} \{2\lambda'_a + 1 - (2a-1) - (2\lambda'_b - 1) + (2b-1)\} \\ &= \frac{1}{n(n-1)} \{\lambda'_a - \lambda'_b + b - a + 1\} \geq \frac{4}{n(n-1)} > 0 \end{aligned}$$

since $\lambda'_a \geq \lambda'_b$ and $b - a \geq 1$. This argument is correct even if $\lambda_b = 0$. □

LEMMA 2. *Let λ be a partition of n. Then*

(a) $$r(\lambda) \leq 1 - \frac{2(n-\lambda_1)(\lambda_1+1)}{n(n-1)} \text{ if } \lambda_1 \geq n/2,$$

(b) $$r(\lambda) \leq \frac{\lambda_1 - 1}{n-1}.$$

Proof.

(a) By assumption $\lambda \leq (\lambda_1, n-\lambda_1)$, so it follows from Lemma 1 and (D-2) that

$$\begin{aligned} r(\lambda) &\leq \frac{1}{n(n-1)}\{\lambda_1^2 - \lambda_1 + n^2 - 2n\lambda_1 + \lambda_1^2 - 3n + 3\lambda_1\} \\ &= 1 + \frac{2(\lambda_1^2 + \lambda_1 - n\lambda_1 - n)}{n(n-1)} \\ &= 1 - \frac{2(\lambda_1+1)(n-\lambda_1)}{n(n-1)} \end{aligned}$$

(b) $r(\lambda) = \frac{1}{n(n-1)} \sum_{j=1}^{m} (\lambda_j^2 - (2j-1)\lambda_j) \leq \frac{1}{n(n-1)} \sum_{j=1}^{m} \lambda_j(\lambda_j - 1) \leq \frac{\lambda_1 - 1}{n(n-1)} \sum_{j=1}^{m} \lambda_j =$
$\frac{\lambda_1 - 1}{n-1}$. □

COROLLARY. *Let λ be such that $r(\lambda) \geq 0$. Then*

$$\left|\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right| \leq \begin{cases} 1 - \frac{2(\lambda_1+1)(n-\lambda_1)}{n^2} & \text{if } \lambda_1 \geq n/2 \\ \frac{\lambda_1}{n} & \text{for all } \lambda. \end{cases}$$

Proof of Theorem 5: If λ^t denotes the transpose of λ we certainly have either $r(\lambda) \geq 0$ or $r(\lambda^t) > 0$, because $\chi_\lambda = -\chi_\lambda^t$ (see James 1978, p. 25). Hence

$$\begin{aligned} \Sigma_\lambda^* d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right)^{2k} &\leq \sum_{\lambda: r(\lambda) \geq 0}^{*} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right)^{2k} \\ &\quad + \sum_{\lambda: r(\lambda) > 0}^{*} d_\lambda^2 \left(\frac{1}{n} - \frac{n-1}{n} r(\lambda^t)\right)^{2k} \\ &\leq 2 \sum_{\lambda: r(\lambda) \geq 0}^{*} d_\lambda^2 \left|\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right|^{2k}. \end{aligned}$$

For $\lambda = (n)$, which is contained in the next to last, but not the last sum, this used

$$\begin{aligned} & d_n^2 \left|\frac{1}{n} - \frac{n-1}{n} r(n)\right|^{2k} + d_{(n-1,1)}^2 \left|\frac{1}{n} - \frac{n-1}{n} r(n-1,1)\right|^{2k} \\ &\quad = (1 - \frac{2}{n})^{2k} + (n-1)^2 (1 - \frac{4}{n})^{2k} \leq (n-1)^2 (1 - \frac{2}{n})^{2k} \\ &\quad = d_{n-1}^2 \left|1 + \frac{n-1}{n} r(n-1,1)\right|^{2k}. \end{aligned}$$

In order to bound this sum we split it into two parts according as $\lambda_1 \gtrless (1-\alpha)n$ (where $\alpha \in (0, \frac{1}{4})$ will be chosen below)

$$\sum_{\lambda:r(\lambda)\geq 0}^{*} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right)^{2k} = \sum_{j=1}^{n} \sum_{\substack{\lambda:r(\lambda)\geq 0 \\ \lambda_1 = n-j}} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right)^{2k}$$

$(*)$

$$\leq \sum_{j=1}^{\alpha n} \binom{n}{j} \frac{n!}{(n-j)!} \left(1 - \frac{2j(n-j+1)}{n^2}\right)^{2k} + \sum_{j>\alpha n}^{n-1} \binom{n}{j} \frac{n!}{(n-j)!} (1 - \frac{j}{n})^{2k}.$$

To obtain this we used the corollary to (D-2) above and

$$\sum_{\lambda:\lambda_1=\ell} d_\lambda^2 \leq \binom{n}{\ell}^2 \sum_{\lambda::\lambda_1=\ell} d^2_{(\lambda_2,\lambda_3,\ldots,\lambda_m)} = \binom{n}{\ell}^2 \sum_{\lambda'} d^2_{\lambda'} = \binom{n}{\ell} \frac{n!}{\ell!}$$

(where the λ' are the irreducible representations of $S_{n-\ell}$).

In order to give a bound for the first sum shown in $(*)$ above recall that $k = \frac{n}{2}\log n + cn$;

$$\sum_{j=1}^{\alpha n} \binom{n}{j} \frac{n!}{(n-j)!} e^{-4k(\frac{j}{n} - \frac{j^2-j}{n^2})}$$

$$\geq n^2 \cdot e^{-4k/n} \cdot \sum_{j=1}^{\alpha n} \frac{(n-1)!^2}{(n-j)!^2 j!} \cdot e^{-2\log n \cdot (1-\frac{j}{n})(j-1)},$$

we observe that the factor in front of the sum is exactly e^{-4c} and so all we have to do is to bound $\sum_{j=1}^{\alpha n} \frac{n^{2(j-1)}}{j!} \cdot n^{-2(1-\frac{j}{n})(j-1)} = \sum_{j=1}^{\alpha n} \frac{1}{j!} \cdot n^{\frac{2j(j-1)}{n}}$ for large values of n. The ratio between two consecutive terms in this sum is $\frac{1}{j+1} \cdot n^{4j/n}$, which, as a function of j, is decreasing if $j < \frac{n}{4\log n}$ and increasing if $j > \frac{n}{4\log n}$. So if both the first and the last ratio are less than $q < 1$ we may bound the sum by $\frac{1}{1-q}$. But the first ratio is < 1 if $n \geq 17$ and the last one is < 1 if $\frac{1}{\alpha n} n^{4\alpha} < 1$, i.e. $n > (\frac{1}{\alpha})^{1/1-4\alpha}$. This works well if $\alpha < 1/4$.

Now let's consider the second sum

$$\sum_{j>\alpha n} \binom{n}{j} \frac{n!}{(n-j)!} (1 - \frac{j}{n})^{2k} \leq (1-\alpha)^{2cn} \cdot \sum_{j>\alpha n}^{n-1} \binom{n}{j} \frac{n!}{(n-j)!} (1 - \frac{j}{n})^{n\log n}.$$

The factor in front of the sum is $\leq e^{-4c}$ if $n \geq 7$ and α is close enough to $\frac{1}{4}$. Hence it is sufficient to bound the sum for large values of n. The ratio between two consecutive terms is

$$\frac{(n-j)^2}{j+1} (1 - \frac{1}{n-j})^{n\log n},$$

which is decreasing in j. So, if the first of these ratios,

$$\frac{(n-\alpha n)^2}{\alpha n+1}(1-\frac{1}{n-\alpha n})^{n\log n} \leq \frac{(1-\alpha)^2}{\alpha} n^{1-\frac{1}{1-\alpha}}$$

is less than one (this happens if $n \geq (\frac{(1-\alpha)^2}{\alpha})^{(\frac{1-\alpha}{\alpha})}$) we may bound the sum by n times the first term, that is, by

$$n \cdot \binom{n}{\alpha n} \frac{n!}{(n-\alpha n)!} \cdot (1-\alpha)^{n\log n}.$$

Using Stirling's formula one can show that this tends to 0 (very slowly) if n tends to ∞ and so it must be bounded. This completes the proof of the upper bound part of Theorem 5.

The following argument for the lower bound produces an explicit set A where the variation distance is large. Intuitively, if not enough transpositions have been made, there will be too many cards that occupy their original positions. Let A be the set of all permutations with one or more fixed points. Under U, the chance of one or more fixed points is well known under the name of the matching problem. Feller (1968, Sec. IV.4) implies

$$U(A) = 1 - \frac{1}{e} + 0(\frac{1}{n!}).$$

To bound $P^{*k}(A)$, consider the process for generating P^{*k} using random transpositions $(L_1, R_1), \ldots, (L_k, R_k)$. Let B be the event that the set of labels $\{L_i, R_i\}_{i=1}^k$ is strictly smaller than $\{1, \ldots, n\}$. Clearly $A \supset B$. The probability of B is the same as the probability that when $2k$ balls are dropped into n boxes, one or more of the boxes will be empty. Arguments in Feller (1968, Sec. IV.2) imply that the probability of B equals

$$1 - e^{-ne^{-2k/n}} + o(1) \text{ uniformly in } k, \text{ as } n \to \infty.$$

With $k = \frac{1}{2}n\log n + cn$, $P^{*k}(A) \geq 1 - e^{-e^{-2c}+o(1)}$. Thus

$$\begin{aligned} ||P^{*k} - U|| &\geq |P^{*k}(A) - U(A)| \geq (P^{*k}(A) - U(A)) \\ &\geq (\frac{1}{e} - e^{-e^{-2c}}) + o(1). \end{aligned}$$

□

Remarks.

1) *On Lower Bounds.* The argument for the lower bound is satisfying in that it produces a simple set that "explains" why $\frac{1}{2}n\log n$ steps are needed. On the other hand, the computation involved two classical results which would not generally be available for other random walk problems. It is therefore of

some interest to see how the general approach to lower bounds works out in this case.

The general approach begins with the upper bound argument and chooses the difficult or "slowest" representations to construct a random variable to work with. In the present case, the difficult representation is $S^{n-1,1}$. A reasonable candidate for a random variable is thus the character χ of this representation. Observe that this is exactly the number of fixed points minus one. Under the uniform distribution

$$E_U(\chi(\pi)) = 0, \ \mathrm{Var}_U(\chi) = \frac{1}{|G|}\Sigma_\pi \chi^2(\pi) = (\chi|\chi) = 1.$$

Under the convolution measure, with ρ the $n-1$ dimensional representation,

$$E_k(\chi) = \Sigma P^{*k}(\pi)\ \mathrm{Tr}\rho(\pi) = \ \mathrm{Tr}\ \Sigma P^{*k}(\pi)\rho(\pi) = \ \mathrm{Tr}\ \hat{P}^{*k}(\rho) = (n-1)(1-\frac{2}{n})^k.$$

Observe that in order to drive this to its correct value zero, k must be $\frac{1}{2}n \log n + cn$ for c large. However, this is not enough to lower bound the variation distance since there are random variables with large means which are small with high probability. A second moment is needed. To compute this $E_k(\chi^2)$ is needed. Now χ^2 is the character of the tensor product of χ with itself. It is not difficult to argue that

$$\begin{array}{rccccccc} & S^{n-1,1} \otimes S^{n-1,1} & = & S^n & \oplus & S^{n-1,1} & \oplus & S^{n-2,2} \quad \oplus \quad S^{n-2,1,1} \\ \dim & (n-1)^2 & & 1 & & n-1 & & \frac{n(n-3)}{2} \qquad \frac{(n-1)(n-2)}{2}. \end{array}$$

An explicit proof of this result can be found on page 97 of James and Kerber (1981).

EXERCISE 13. Using the data above, compute $\mathrm{Var}_k(\chi)$ and use Chebychev's inequality to show that $\frac{1}{2}n \log n - cn$ steps are not enough.

2) While limited, the approach developed above gives precise results for some other problems. To begin with, consider random transpositions. The identity is chosen much more frequently than any specific transposition. It is straightforward to carry out the analysis for the probability

$$P_n(\mathrm{id}) = p_n, \ P_n(\tau) = \frac{1-p_n}{\binom{n}{2}}, \ P_n(\pi) = 0 \text{ otherwise}.$$

If $p_n = 1/(1+\binom{n}{2})$, all possible permutations are equally likely. In this case the argument shows that $k = c(n)n^2$ transpositions are needed where $c(n) \to \infty$ with n. This is somewhat surprising; usually, for a given support set, the probability that approaches the uniform most rapidly is uniform on the support set.

Similarly, any simple probability on S_n which is constant on conjugacy classes can be worked with. A key tool is a uniform bound on the characters developed by Vershik and Kerov (1981). A readable account of this is given by Flatto, Odlyzko and Wales (1985). They work out details for probabilities uniform on a fixed conjugacy class c (e.g., all 3 cycles). Their results imply that $\frac{1}{2}n \log n$ steps are

always sufficient. This is not surprising — choosing a random 3-cycle mixes more cards each time and should result in faster convergence.

A simple problem not covered by available techniques is the rate of convergence for a random involution ($\pi^2 = \mathrm{id}$). There are $\Sigma_\rho d_\rho$ of these, which is asymptotically $(\frac{n}{e})^{n/2} \cdot e^{\sqrt{n}}/\sqrt{2}e^{\frac{1}{4}}$. For this and other properties of involutions see Stanley (1971, pg. 267). Such a measure is constant on conjugacy classes, but the asymptotics haven't been worked out. It is not hard to show that any non trivial conjugacy class generates S_n. See Arad and Herzog (1985). Thus there are many open problems.

Finally, it is straightforward to handle random walks based on measures constant on conjugacy classes of the alternating group A_n. The characters of A_n are available as simple functions of the characters of S_n. James and Kerber (1981) Chapter 2 give these results.

EXERCISE 14. Let n be odd. Let Q be uniform on the set of n cycles in A_n. Show that Q^{*2} is close to uniform for large n. (Hint: See formula 2.3.17 in James and Kerber (1981) or Stanley (1983).)

3) *Connections with Radon Transforms.* The analysis developed in this section has been applied to the study of uniqueness and inversion of the Radon transform by Diaconis and Graham (1985a). Here is a brief description: let G be a group with $d(s,t)$ a bi-invariant metric: $d(rs,rt) = d(sr,tr) = d(s,t)$. Let $f: G \to \mathbb{R}$ be a function. Suppose we are told not $f(s)$ but

$$\overline{f}(s) = \sum_{d(s,t)\le a} f(t) \text{ for all } s \text{ and fixed } a.$$

When do these averages determine f? If $S = \{s: d(\mathrm{id}, s) \le a\}$ the Radon transform is $\overline{f}(s) = I_S * f(s)$. Taking Fourier transforms, the Radon transform is unique if and only if $\hat{I}_S(\rho)$ is invertible for every irreducible representation ρ.

The study of this problem leads to interesting questions of probability and computational complexity even for groups as simple as Z_2^n. In this case, with d as Hamming distance, when $a = 1, f \to \overline{f}$ is $1-1$ iff n is even; when $a = 2$, iff n is not a perfect square; for $a \ge 4$ iff n is not in a finite set of numbers.

John Morrison (1986) derived exact results for this problem using Gelfand pair tools (Section F below). Jim Fill (1987) gives comprehensive results for Z_n. For applications to data analysis, see Diaconis (1983). For general background, see Bolker (1987).

For $G = S_n$, choose $d(\pi, \eta)$ as the minimum number of transpositions needed to bring π to η. This metric is further discussed in Chapter 6-B. For any bi-invariant metric, I_S is constant on conjugacy classes, so $\hat{I}_S(\rho) = cI$. For $a = 1, c = (1 + \binom{n}{2} r(\rho))$. Diaconis and Graham use this result and Frobenius' formula for $r(\rho)$ to argue that $f \to \overline{f}$ is invertible iff $n \in \{1,3,4,5,6,8,10,12\}$.

4) *Perfect codes.* Very similar computations occur in a seemingly different problem treated by Rothaus and Thompson (1966). Let G be a group and T be a subset of G. Say that T *divides* G if there is a set S in G such that each $g \in G$

has a unique representation $st = g$ with s in S and t in T. For example, if T is a subgroup, then T divides G. If $G = S_3$ and $T = \{\mathrm{id}, (12), (13), (23)\}$, then T does not divide S_3.

The construction of codes leads naturally to questions of divisibility: Let G be a group and $d(s,t)$ a G invariant metric on G (i.e., $d(s,t) = d(gs, gt)$). For example, G might be Z_2^n and d might be the Hamming distance, or G might be S_n and $d(s,t)$ might be Cayley's distance: the minimum number of transpositions required to bring s to t (see Chapter 6-B).

A subset $S \subset G$ is called a *code*; S *corrects* k *errors* if any two code words are at distance more than $2k$ apart; S *is perfect* if G is the disjoint union of balls of fixed radius centered at the elements of S.

Perfect codes are elegant efficient ways of coding data with minimum waste. On Z_2^n the perfect codes have been classified; see MacWilliams and Sloane (1977). The search for codes in other groups is an active area of research.

To see the connection with group divisibility, consider S_n with Cayley's distance. Take T to be a (solid) ball of radius k about the identity. Observe that T divides S_n if and only if there is a perfect code S of this radius — indeed, balls centered at points of S would be disjoint if $TS = S_n$ uniquely.

Rothaus and Thompson considered $k = 1$, i.e. T as the identity together with the set of all transpositions in S_n. To explain their result, observe that a necessary condition for divisibility is $(1 + \binom{n}{2}) | n!$ (after all, disjoint balls of radius 1 have to cover). This rules out $n = 3, 4, 5$ but not 6 for example. They proved that if $(1 + \binom{n}{2})$ is divisible by a prime exceeding $\sqrt{n} + 2$, then T does not divide S_n.

Their argument is *very* similar to the argument for analyzing repeated random transpositions. Interpret the equation $ST = G$ as an equation about the convolution of the indicator functions of the sets S and T ($f_S * f_T = 1$ say). Taking Fourier transforms at an irreducible representation leads to $c(\rho)\hat{f}_S(\rho) = 0$, where $c(\rho) = 1 + \binom{n}{2}\chi_\rho(\tau)/d_\rho$. Now one must study when $c(\rho)$ vanishes (see the previous remark). One really new thing in the Rothaus-Thompson paper is the skillful use of transforms at non-irreducible representations to give checkable divisibility constraints on n. The argument is fairly detailed and will not be given here. Sloane (1982) connects this work with the modern coding literature and gives many further applications. Chihara (1987) extends the results to Chevalley groups.

EXERCISE 15. Rothaus and Thompson report 1, 2, 3, 6, 91, 137, 733, and 907 as the only integers less than 1,000 which fail to satisfy the theorem. The naive criterion does 3, (and S_2 is divisible). Show that S_6 is not divisible.

5) *Varying the measure.* The ideas developed above can be used for some related problem like transpose a random card with the top card, or switch the top k cards with k randomly chosen cards. Here we have a measure on S_n invariant under conjugation by S_k and bi-invariant under S_{nk}. The Fourier transform can be shown to be diagonal with explicitly computable elements. See Diaconis (1986) or Greenhalgh (1988) for further details.
6) *Random reflections.* Similar analyses are possible for other random walks

constant on conjugacy classes. For example, let $G = 0_p$ — the p-dimensional orthogonal group. One practical problem involves algorithms for choosing a random element of G when p is large (e.g. $p = 256$). The usual algorithm begins with p^2 standard normal random variables X_{ij}, forms a matrix $M = \{X_{ij}\}$ and makes M orthogonal using the Gram-Schmidt algorithm. It is easy to show that this results in a random orthogonal matrix uniformly distributed on G. Diaconis and Shahshahani (1987a) discuss this and other algorithms. In carrying out the Gram-Schmidt algorithm, the ith row of M must be modified by subtracting out the inner product of all rows above it. This entails computation of $i - 1$ inner products. Each inner product involves p multiplications and additions. The whole procedure takes order $p \sum_{i=1}^{p} i = 0(p^3)$ operations. This is often too large for practical use.

Sloane (1983) contains a fascinating application to encrypting telephone conversations. Sloane suggested generating a matrix by using random reflections. Geometrically this involves choosing a random point U in the p-sphere and reflecting in the hyperplane orthogonal to U. Algebraically the matrix is $\Gamma = (I - 2UU')$. Observe that the distribution of Γ is constant on conjugacy classes because $\Gamma_1(I - 2UU')\Gamma_1' = (I - 2\Gamma_1 U(\Gamma_1 U)')$. If U is uniform on the p-sphere, $\Gamma_1 U$ is uniform as well. There is a straightforward extension of the upper bound lemma to compact groups. The analysis can be carried out to show that $\frac{1}{2}p \log p + cp$ steps are enough (while $\frac{1}{2}p \log p - cp$ steps are too few). Some details can be found in Diaconis and Shahshahani (1986a).

In this problem, P is singular with respect to the uniform distribution, but P^{*k} has a density for $k \geq p$. Thus variation distance bounds make sense. For random walks on continuous compact groups involving a discrete measure, the distribution is always singular and only bounds in a metic for the weak star topology can be hoped for.

7) *Random walks on linear groups over finite fields.* The problem described above can be carried out over other fields such as C (to generate a random unitary matrix) or F_q - a finite field with $q = p^d$ elements. Here is another problem which should lead to interesting mathematics. Let V be a vector space of dimension d over F_q. Let $SL_d(V)$ be the $d \times d$ invertible matrices with determinant 1. This is a finite group of order $q^{\binom{d}{2}} \Pi_{i=2}^{d}(q^i - 1)$. A *transvection* is a linear transformation in $SL_d(V)$ which is not the identity but fixes all elements of a hyperplane. Suzuki (1982, Sec. 9) shows that if $d \geq 3$, the transvections form a single conjugacy class that generates $SL_d(V)$. Thus, the question "how many random transvections are required to get close to the uniform distribution on $SL_d(V)$?" can be attacked by the method of this section.

E. The Markov chain connection.

1. Introduction.

There is another approach to random walks on groups: treat them as Markov chains with state space G and $|G| \times |G|$ transition matrix $Q(s,t) = Q(ts^{-1})$. In early attempts to understand the problem of random transpositions Joseph Deken did *exact* computations of the second largest eigenvalue for decks of size $n = 2, 3, \ldots, 10$. He found it to be $(1 - 2/n)$. This is precisely the constant in the Fourier transform at the "slow" representation (see Theorem 5 of Section D). This striking numerical coincidence suggested that (a) the $(1 - 2/n)$ result must hold for all n, and (b) there is a close connection between the Markov chain and group representation approach. Some of this was worked out by Diaconis and Shahshahani (1981), who showed that the eigenvalues of the transition matrix are precisely the eigenvalues of $\hat{Q}(\rho)$, each appearing with multiplicity d_ρ.

The following discussion uses work of Matthews (1985). It results in a sort of diagonalization of the transition matrix and an exact determination of eigenvalues and eigenvectors where these are available. This allows us to use results from classical Markov chain theory.

2. A spectral decomposition of the transition matrix.

Let G be a finite group with elements $\{s_1, \ldots, s_N\}$, $N = |G|$. For a probability Q on G, construct $Q(i,j) = Q(s_j s_i^{-1})$ — the chance that the walk goes from s_i to s_j in one step. Suppose that the irreducible representations are numbered $\rho_1, \ldots, \rho_K$. Define

$$M_k = \begin{pmatrix} \hat{Q}(\rho_k) & & 0 \\ & \ddots & \\ 0 & & \hat{Q}(\rho_k) \end{pmatrix} \tag{1}$$

a $d_k^2 \times d_k^2$ block matrix with $\hat{Q}(\rho_k)$ the Fourier transform of Q at ρ_k.

(2) Let M be the $N \times N$ block diagonal matrix $\begin{pmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_K \end{pmatrix}$.

Suppose that a basis has been chosen so that each irreducible representation is given by a unitary matrix. Define

$$\psi_k(s) = \sqrt{\frac{d_k}{N}}(\rho_k(s)_{11}, \rho_k(s)_{21}, \ldots, \rho_k(s)_{d_k 1}, \rho_k(s)_{12}, \ldots, \rho_k(s)_{d_k d_k})^T, \tag{3}$$

a column vector of length d_k^2. Let $\phi(s) = (\psi_1(s)^T, \psi_2(s)^T, \ldots, \psi_K(s)^T)^T$ be a column vector of length N obtained by concatenating the $\psi_k(s)$ vectors.

(4) Let ϕ be the $N \times N$ matrix $(\phi(s_1), \ldots, \phi(s_N))$ and ϕ^* its conjugate transpose .

Theorem 6. *The transition matrix $Q(i,j)$ can be written*

$$Q = \phi^* M^* \phi \tag{5}$$

Remarks. The Schur orthogonality relations show that ϕ is a unitary matrix. So (5) is a decomposition similar to the traditional eigenvalue, eigenvector decomposition. It implies that each eigenvalue of $\hat{Q}(\rho)$ is an eigenvalue of $Q(i,j)$ with multiplicity d_ρ. Together these are all the eigenvalues of $Q(i,j)$. If M is diagonal (e.g. Q constant on conjugacy classes or bi-invariant on a Gelfand pair (Section F below)), then (5) is the spectral decomposition of Q with respect to an orthonormal basis of eigenvectors.

Proof of Theorem 6: The Fourier inversion theorem gives

$$Q(i,j) = \frac{1}{N}\sum_{k=1}^{K} d_k \mathrm{Tr}[\hat{Q}(\rho_k)\rho_k(s_i)\rho_k(s_j^{-1})] = \frac{1}{N}\sum_{k=1}^{K} d_k \mathrm{Tr}[\rho_k(s_j^{-1})\hat{Q}(\rho_k)\rho_k(s_i)].$$

Expanding the trace, this equals

$$\sum_{k=1}^{K} \psi_k(s_j)^* M_k \psi_k(s_i).$$

□

3. The first hit distribution.

Let G be a finite group and Q a probability on G. For $s,t \in G$, define F_{st}^n = the probability that t is first hit at time n starting at s. For $|z| < 1$ let $F_{st}(z) = \sum_{n=1}^{\infty} F_{st}^n z^n$.

Theorem 7. *For $|z| < 1, (I - z\hat{Q}(\rho))$ is invertible and*

$$F_{st}(z) = \frac{\Sigma_\rho d_\rho \mathrm{Tr}[\rho(st^{-1})(I - z\hat{Q}(\rho))^{-1}]}{\Sigma_\rho d_\rho \mathrm{Tr}[I - z\hat{Q}(\rho)]^{-1}}.$$

Proof. Using the notation of Section 2, $Q(z) = \sum_{n=1}^{\infty} z^n Q^n = \{\phi^*(I - zM^*)^{-1}\phi\}$. Kemperman (1961, pg. 18–19) gives the standard result

$$F_{st}(z) = Q_{st}(z)/Q_t(z).$$

The result follows from this and (5) above. It is given a direct independent proof in Section H. It is mentioned here to underscore the availability of the Markov chain machine in situations where all the eigenvalues and eigenvectors of the transition matrix are known.

4. On generalized circulants.

The technique we have developed for analyzing random walks gives rise to a class of "patterned matrices" for which we can explicitly determine all the eigenvalues and eigenvectors. Let G be a finite group of order g. Let $s_1, \ldots, s_g$ be an enumeration of the elements of G. Let P be a probability measure on G. *The transition matrix associated with* P is the $g \times g$ matrix with i, j entry $P(s_j s_i^{-1})$. If a random walk on G is thought of as a Markov chain with G as state space, the i, j entry is the probability of a transition from state s_i to state s_j. We have been working with measures which are constant on conjugacy classes. Generalizing this somewhat define a *G-circulant* as a $g \times g$ matrix with i, j entry $f(s_j s_i^{-1})$ with f constant on conjugacy classes.

Examples. If G is Abelian, then the equivalence classes consist of single elements. If G is cyclic, then a G circulant is an ordinary circulant: a $g \times g$ matrix in which each row is a cyclic shift of the first row. For $G = S_3$ the equivalence classes are $\{\mathrm{id}\}, \{(1\ 2), (1\ 3), (2\ 3)\}\{(1\ 2\ 3), (1\ 3\ 2)\}$. If $f(\mathrm{id}) = a, f(1\ 2) = b, f(1\ 2\ 3) = c$ and the group is labelled in order (using $\begin{pmatrix} 1 & 2 & 3 \\ \alpha & \beta & \gamma \end{pmatrix}$ notation) (1 2 3)(1 3 2)(2 1 3)(2 3 1)(3 1 2)(3 2 1), we get

$$\begin{pmatrix} a & b & b & c & c & b \\ b & a & c & b & b & c \\ b & c & a & b & b & c \\ c & b & b & a & c & b \\ c & b & b & c & a & b \\ b & c & c & b & b & a \end{pmatrix}$$

value	typical vector	dim
$a + 3b + 2c$	(1 1 1 1 1 1)	1
$a - 3b + 2c$	(1 −1 −1 1 1 −1)	1
$a - c$	(2 0 0 −1 −1 0)	4

Let G be the 8 element quarternion group $G = \{\pm 1, \pm i, \pm j, \pm k\}$ with multiplication given by $\overset{i}{\nearrow}\!\!\searrow$ $k \longleftarrow j$. Thus $ij = k, kj = -i$, etc. There are five conjugacy classes: $\{+1\}, \{-1\}, \{\pm i\}, \{\pm k\}$. Let them have weight a, b, c, d, e. Label the group 1, -1, i, -i, j, -j, k, -k. We get

$$\begin{pmatrix} a & b & c & c & d & d & e & e \\ b & a & c & c & d & d & e & e \\ c & c & a & b & e & e & d & d \\ c & c & b & a & e & e & d & d \\ d & d & e & e & a & b & c & c \\ d & d & e & e & b & a & c & c \\ e & e & d & d & c & c & a & b \\ e & e & d & d & c & c & b & a \end{pmatrix}$$

value	typical vector	dim
$a + b + 2c + 2d + 2e$	(1 1 1 1 1 1 1 1)	1
$a + b + 2c - 2d - 2e$	(1 1 1 1 −1 −1 −1 −1)	1
$a + b + 2d - 2c - 2e$	(1 1 −1 −1 1 1 −1 −1)	1
$a + b + 2e - 2c - 2d$	(1 1 −1 −1 −1 −1 1 1)	1
$a - b$	(1 −1 0 0 0 0 0 0)	2

Theorem 8. *Let M be a G-circulant. Then M has an eigenvalue λ_ρ for each irreducible representation ρ of G,*

$$\lambda_\rho = \frac{1}{d_\rho} \Sigma f(g) \chi_\rho(g),$$

the eigenvalue λ_ρ occurs with multiplicity d_ρ^2.

Proof. The spectral decomposition of Section 2 above proves a stronger result: it gives the eigenvectors as an explicit arrangement of the matrix entries of the irreducible representations.

Remarks.

1. There is a lovely book called *Circulant Matrices* by Phillip Davis (1979). It seems like a nice project to go through the book and generalize all the results to G-circulants.
2. Note that the character vector $(\chi_\rho(s_1)\dots\chi_\rho(s_g))$ is always an eigenvector for λ_ρ.
3. The argument generalizes easily to a transitive action of G on a finite set X. If P is a probability on G, then P induces a Markov chain on X. The transition matrix of this chain has the same eigenvalues as the matrices $\hat{P}(\rho)$, where ρ runs over the irreducible representations of G that appear in the permutation representation of G on X. This is developed in Section F which follows.
4. Example 3 of Section C suggests some further extensions. This begins with the Markov chain $X_n = 2X_{n-1} + \epsilon_n \pmod p$ with ϵ_i i.i.d. taking values $0, \pm 1$ with probability $\frac{1}{3}$. The transition matrix M of this chain is not a circulant, but the argument shows that its ath power is a circulant, where a is the order of 2 (mod p). Thus one knows, up to an ath root of unity, all the eigenvalues of M. Remark 2 of the example suggests many further situations where a similar analysis is possible.

F. Random walks on homogeneous spaces and Gelfand pairs.

1. Homogeneous spaces.

There is an extension of the basic set up which is useful. It involves the Markov chain induced by a random walk under the action of a group. This arises in some of the introductory examples: for instance, in considering the recurrence $X_n = a_n X_{n-1} + b_n X_{n-2} \pmod p$, a random walk on 2×2 matrices was considered. The matrices act on pairs (X_n, X_{n-1}). To understand the behavior of X_n it is not necessary to bound the rate of convergence on the group of matrices, but only on the set of non-zero pairs. Similarly, the grand tour example in section A4 only involved the action of the orthogonal group on lines or planes.

Definition. Let G be a finite group and X be a finite set. An *action* of G on X is a mapping from $G \times X \to X$ which we will denote $(s, x) \to s \cdot x$ or simply sx. It must satisfy: id $\cdot x = x$ and $s \cdot (t \cdot x) = (st) \cdot x$. Define an equivalence on X by $x \sim y$ if for some $s \in G, sx = y$. The equivalence classes are called *orbits*. G *operates transitively on* X if there is only one orbit. A set with a group acting transitively is called a *homogeneous space.*

When G operates transitively, the following canonical representation of X is useful. Fix $x_0 \in X$. Let N — the *isotropy subgroup* of x_0 - be the set of $s \in G$ with $sx_0 = x_0$. The group G acts on the coset space G/N. There is an isomorphisim between X and G/N respecting the action of G. We will identify

X with $x_0, x_1, \ldots, x_n$, a set of coset representatives for N in G. It will always be assumed that $x_0 = \mathrm{id}$.

Example 1. The symmetric group S_n acts on $\{1, 2, \ldots n\}$ transitively. The isotropy subgroup is isomorphic to S_{n-1} — as all permutations fixing 1. Coset representatives can be chosen as the identity and the transpositions $(12), \ldots, (1n)$.

A probability P on G induces a probability $\tilde{P}$ on $X = G/N$ by $\tilde{P}(x_i) = P(x_i N)$. Similarly, if P^{*k} denotes the convolution of P with itself k times, P^{*k} induces a probability on X. We can think of a process with values in G, say id, s_1, $s_2 s_1$, $s_3 s_2 s_1, \ldots$. The induced process in X is x_0, $s_1 x_0$, $s_2 s_1 x_0, \ldots$.

EXERCISE 16. Let the finite group G act on the finite set X, partitioning X into orbits θ_i. If P and Q are probabilities on X which are G-invariant, then

$$||P - Q|| = \frac{1}{2}\Sigma|P(\theta_i) - Q(\theta_i)|.$$

Thus, the variation distance between P and Q equals the distance between their restrictions to the set of orbits. This is a special case of the following result: if P and Q are probabilities on a σ-algebra $\mathcal{F}$ and if a sub-σ-algebra $B \subset \mathcal{F}$ is sufficient for P and Q, then $||P - Q||_{\mathcal{F}} = ||P - Q||_{\mathcal{B}}$. See Diaconis and Zabell (1982) for details and applications.

LEMMA 3. *Let G act transitively on X. Let P be a probability on G. The induced process is a doubly stochastic Markov chain on X with transition matrix $P_x(y) = P(yNx^{-1})$.*

Proof. For the induced processes, the chance of going from x to y in one step is $P_x(y)$ defined as $P\{s: sx = y\} = P\{yNx^{-1}\}$. For a Markov chain, the chance of going from x to y in two steps is of course

$$P_x^2(y) = \sum_z P_x(z) P_z(y).$$

The chance that the chain in question is at y in two steps is

$$P * P(yN) = \sum_s P(yNs^{-1}) P(s).$$

Let $s = xn$, we get

$$= \sum_{x,n} P(yNn^{-1}x^{-1}) \cdot P(xn) = \sum_x P(yNx^{-1}) \cdot P(xN) = P_{x_0}^2(y).$$

The last computation is essentially the inductive step of a proof that the measure induced by P^{*k} on X equals $P_{x_0}^k(y)$. □

To state the next result, introduce $L(X)$ — the set of all functions from X into the complex numbers. The action of G on X induces an action of G on $L(X)$

by $sf(x) = f(s^{-1}x)$. This is a $1-1$ linear mapping of $L(X)$, and so yields a representation of G. The representation splits into a direct sum of irreducible representations ρ.

LEMMA 4. *(Upper bound lemma). Let G operate transitively on the finite set X. Let N be the isotropy subgroup. Let P be a right N invariant probability on G, $\tilde{P}$ the induced probability on X, and U the uniform distribution on X. Then*

$$||P - U||^2 \leq \frac{1}{4}\Sigma^* d_\rho Tr\{\hat{\tilde{P}}(\rho)\hat{\tilde{P}}(\rho)^*\}$$

where the sum is over all nontrivial irreducible representations that occur in $L(X)$.

Proof. Let $\tilde{U}$ be the uniform probability on G.

$$(\Sigma_x|P(x) - U(x)|)^2 \leq |X|\Sigma_x|P(x) - U(x)|^2 = |X|\ |N|\Sigma_s|\tilde{P}(s) - \tilde{U}(s)|^2$$
$$= \Sigma_\rho^* d_\rho \mathrm{Tr}(\hat{\tilde{P}}(\rho)\hat{\tilde{P}}(\rho)^*).$$

In the last step, the Plancherel theorem was used together with the facts that a right N invariant function has zero Fourier transform if ρ does not occur in $L(X)$. This follows from the following lemma and remark. □

LEMMA 5. *Let ρ, V be an irreducible representation of the finite group G. Let $N \subset G$ be a subgroup and $X = G/N$ the associated homogeneous space. The number of times that ρ appears in $L(X)$ equals the dimension of the space of N fixed vectors in $\rho, V(= \dim\{v \in V: \rho(n)v = v \text{ for all } n \in N\})$.*

Proof. Let $\{\delta_x(\cdot)\}$ be a basis for $L(X)$. The character χ for the representation of G on $L(X)$ is

$$\chi(s) = |\{x: \delta_x(s^{-1}y) = \delta_x(y)\}| = |x: sx = x|$$
$$= |x: x^{-1}sx \in N|.$$

Now, the number of times ρ appears in this representation is

$$(\chi_\rho|\chi) = \frac{1}{|G|}\sum_s \chi_\rho(s)\chi(s) = \frac{1}{|G|}\sum_s \chi_\rho(s) \sum_{\substack{x,n \\ x^{-1}sx=n}} 1$$
$$= \frac{1}{|G|}\sum_n \chi_\rho(n) \sum_{\substack{s,x \\ x^{-1}sx=n}}$$

But, for any fixed n,

$$\sum_{\substack{s,x \\ x^{-1}sx=n}} 1 = |X|.$$

To see this observe that for fixed $n, x, s = xnx^{-1}$ is determined. Further, if $x^{-1}sx = n$, then $(tx)^{-1}tst^{-1}(tx) = n$ for all $t \in G$. Since G operates transitively on X, for every $y \in X$ there is a unique s^* such that $y^{-1}s^*y = n$.

Since $|G|/|X| = |N|$,

$$(\chi_\rho|\chi) = \frac{1}{N}\Sigma_n \chi_\rho(n) = (\chi_\rho|1)_N.$$

The right side is the number of times the trivial representation appears in χ_ρ restricted to N. This is just the dimension of the space of N fixed-vectors. □

Remarks. Lemma 5 is a special case of Frobenius' reciprocity formula. The representation $L(X)$ is the trivial representation of N induced up to G. Frobenius' formula says the number of times a representation ρ of G appears in the induction of λ (a representation of N) to G equals the multiplicity of λ in ρ restricted to N. Chapters 3 and 7 of Serre (1977) give further development. The general result is proved by essentially the same combinatorial argument. For present purposes, Lemma 5 is all that is needed.

Using Lemma 5, if ρ does not occur in $L(X)$, the trivial representation does not occur in ρ restricted to N. Now, the orthogonality relations (Corollary 2 of Schurs lemma in Chapter 2) yield $\Sigma_n \rho(n) = 0$. For a right N invariant function f on G,

$$\hat{f}(\rho) = \Sigma_x\ f(x)\ \rho(x)\Sigma_n\ \rho(n) = 0.$$

This completes the proof of the upper bound Lemma 4.

The next section discusses a collection of examples where a huge simplification occurs.

2. *Gelfand pairs*

This is a class of examples where the Fourier analysis becomes simple. Consider, as above, a group G acting transitively on a finite set X with isotropy subgroup N. A function $f: G \to C$ is called *N-bi-invariant* if $f(n_1 s n_2) = f(s)$ for all $s \in G, n_1, n_2 \in N$.

Definition. G, N is called a *Gelfand pair* if the convolution of N bi-invariant functions is commutative.

One value of this definition comes from a long list of examples. Some of these are discussed later in this section. Letac (1981, 1982) or Bougerol (1983) present very readable surveys of available results. The following theorem is basic:

Theorem 9. *The following three conditions are equivalent*

(1) *G, N is a Gelfand pair.*

(2) *The decomposition of $L(X)$ is multiplicity free.*

(3) *For every irreducible representation (ρ, V) there is a basis of V such that $\hat{f}(\rho) = \begin{pmatrix} * & 0 \\ 0 & 0 \end{pmatrix}$ (a matrix with zero entries except perhaps in the $(1,1)$ position) for all N-bi-invariant functions f.*

Proof. Assume (2), so $L(X) = V_1 \oplus V_2 \oplus \ldots \oplus V_m$ say. This is multiplicity free, so by Lemma 5 above, each V_i has a non-trivial one-dimensional space of N invariant functions. Choose so called spherical functions $s_i \in V_i$ to be left N-invariant and

normalized so $s_i(\mathrm{id}) = 1$. Complete s_i to a basis for V_i chosen so $\rho_i(n) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & * \end{smallmatrix}\right)$ for all $n \in N$ (the top block is 1×1, the bottom block is $(d_i - 1) \times (d_i - 1)$).

For f an N-bi-invariant function,

$$\hat{f}(\rho_i) = \sum_s f(t)\rho_i(t) = \sum_{x,n} f(xn)\rho_i(xn) = \sum_x \rho_i(x)f(x) \sum_n \rho_i(n).$$

But ρ_i restricted to N has a one-dimensional space of fixed vectors. By the orthogonality relations for the matrix entries, the (a, b) entry satisfies

$$\sum_n \rho_i^{ab}(n) = \begin{cases} |N| & \text{if } a = b = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus $\hat{f}(\rho_i)$ has the form $\Sigma f(x)\left(\begin{smallmatrix} * & \\ \vdots & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} * & \\ \vdots & 0 \end{smallmatrix}\right)$. This argument works for any right invariant function f. For left invariant f, a similar argument shows that $\hat{f}(\rho_i)$ has form $\left(\begin{smallmatrix} * & \cdots & * \\ 0 & & \end{smallmatrix}\right)$. From Lemma 5, if ρ does not appear in $L(X)$, $\hat{f}(\rho) = 0$. This shows (2) implies (3).

Clearly (3) implies (1) by taking Fourier transforms. To finish off, suppose (1) but some ρ_i has multiplicity $j > 1$ in $L(X)$. Pick a basis of V_i with first j coordinates spanning the N-invariant space. Take M_1, M_2 any two non-commuting $j \times j$ matrices. Define f_1, f_2 on G by

$$\hat{f}_1(\rho) = \hat{f}_2(\rho) = 0 \text{ if } \rho \neq \rho_i$$
$$\hat{f}_1(\rho_i) = \begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{f}_2(\rho_i) = \begin{pmatrix} M_2 & 0 \\ 0 & 0 \end{pmatrix}.$$

By Fourier inversion, these are non-zero, N-bi-invariant functions and $f_1 * f_2 \neq f_2 * f_1$. □

COROLLARY. *Let (G, N) be a Gelfand pair with $L(X) = V_1 \oplus \ldots \oplus V_m$. Each V_i contains a unique N-invariant function s_i with $s_i(id) = 1$. If the Fourier transform of an N invariant probability P on X is defined by*

$$\hat{P}(i) = \Sigma_x s_i(x)P(x),$$

then, for U the uniform distribution on X

$$||P_{x_0}^k - U||^2 \leq \frac{1}{4} \sum_{i=1}^m d_i |\hat{P}(i)|^{2k}.$$

Remarks. The corollary follows from the theorem above and the upper bound lemma of the last section. The s_i are called *spherical functions.* They have been explicitly computed for many groups. From part (3) of the theorem $s_i(x) = \rho_i(x)_{11}$. This sometimes serves as a definition, or as a way of computing spherical functions: take (ρ_i, V_i), a unitary representation that appers in $L(X)$. By the

theorem, V_i contains a one-dimensional space of N fixed vectors. Let u be a unit N-fixed vector. Then $s_i(x) =< \rho_i(x)u, u >$. The s_i are left N invariant functions on X. They are also N-bi-invariant functions on G. Turning things around, if the spherical functions are known, the * in Theorem 9-3 can be computed as $\sum_{t \in G} f(t)s(t)$.

EXERCISE 17. Let χ_i be the character of a representation ρ_i that appears in $L(X)$. Show

$$s_i(x) = \frac{1}{|N|} \sum_{n \in N} \chi_i(xn).$$

Thus the spherical functions are determined by characters.

3. *Example: The Bernoulli-Laplace model of diffusion.*

As a specific example of the techniques discussed above consider the following model of diffusion suggested originally by Daniel Bernoulli and Laplace. Feller (1968, p. 378) contains the history. There are two urns, the first containing n red balls, the second containing n white balls. At each stage, a ball is picked at random from each urn and the two are switched. Evidently, many switches mix things up and it is not difficult to show that once things reach equilibrium they evolve (approximately) as an Ornstein-Uhlenbeck process (at least for large n). The problem is, how many switches are required to reach equilibrium? In what follows, we show that $\frac{n}{4} \log n + cn$ switches suffice.

It is just as simple to solve the same problem with r red balls in the first urn and b black balls in the second urn. Let $n = r + b$. A convenient mathematical model for this has $X = S_n/S_r \times S_b$; thus X can be thought of as the set of r element subsets of a set with n elements. For $x, y \in X$ define the distance $d(x, y) = r - |x \cap y|$. This is a metric (see Chapter 6-D), and the random walk problem becomes the following: start at $x_0 = \{1, 2, \ldots, r\}$. Choose an element inside x_0 and an element outside x_0 and switch them. This chooses a set x at distance one from x_0 at random. The following result is proved by Diaconis and Shahshahani (1987b).

Theorem 10. *For nearest neighbor random walk on the r sets of an n set, if $k = \frac{r}{2}(1 - \frac{r}{n}) \log n + cr$ then*

$$||P^k_{x_0} - U||^2 \leq ae^{-dc}$$

for positive universal constants a and d.

Proof. Without loss, take $r \leq n/2$. The space decomposes as $L(X) = V_0 \oplus V_1 \oplus \ldots \oplus V_r$ where V_i is the irreducible representation of the symmetric group S_n corresponding to the partition $(n - i, i)$. James (1978) gives this result as well as $\dim(V_i) = \binom{n}{i} - \binom{n}{i-1}$. In particular, the pair $(S_n, S_r \times S_b)$ is a Gelfand pair.

The spherical functions have been determined by Karlin and McGregor (1961) in studying an equivalent formulation in a genetics example (Moran's model). Stanton (1984) contains this result in modern language. The spherical functions

turn out to be classically studied orthogonal functions called the Dual Hahn or Eberlein polynomials. The function $s_i(x)$ only depends on the distance $d(x, x_0)$ and is a polynomial in d given by

$$s_i(d) = \sum_{m=0}^{i} \frac{(-i)_m(i-n-1)_m(-d)_m}{(r-n)_m(-r)_m m!} \quad 0 \le i \le r,$$

where $(j)_m = j(j+1)\dots(j+m-1)$. Thus,

$$s_0(d) = 1, \; s_1(d) = 1 - \frac{nd}{r(n-r)},$$
$$s_2(d) = 1 - \frac{2(n-1)d}{r(n-r)} + \frac{(n-1)(n-2)d(d-1)}{(n-r)(n-r-1)r(r-1)}.$$

The basic probability P for this problem is supported on the $r(n-r)$ sets of distance one from the set $\{1, \dots, r\}$. Thus the Fourier transform of P at the ith spherical function is

$$\hat{P}(i) = s_i(1) = 1 - \frac{i(n-i+1)}{r(n-r)} \quad 0 \le i \le r. \tag{3}$$

Now the corollary to Theorem 9

$$||P^k - U||^2 \le \frac{1}{4}\sum_{i=1}^{r}\{\binom{n}{i} - \binom{n}{i-1}\}(1 - \frac{i(n-i+1)}{r(n-r)})^{2k}.$$

To bound this sum, consider first the term for $i = 1$,

$$(n-1)(1 - \frac{n}{r(n-r)})^{2k}.$$

This is essentially

$$e^{-\frac{2kn}{r(n-r)} + \log n}.$$

Thus k must be $\frac{r}{2}(1 - \frac{r}{n}) \log n$ at least to kill this term. If $r = n/2$, this becomes $\frac{r}{4} \log r$. If $r = o(n)$, this becomes $\frac{r}{2} \log n$.

Next consider the final term

$$\left(\binom{n}{r} - \binom{n}{r-1}\right)(\frac{1}{n-r})^{2k}.$$

This is certainly bounded above by

$$\frac{n^r}{r!}\frac{1}{(\frac{n}{2})^{2k}} = e^{-2k \log \frac{n}{2} + r \log n - \log r!}.$$

In any case, if k is of order $\frac{r}{2}(1 - \frac{r}{n}) \log n$, this tends to zero exponentially fast. The intermediate terms are always geometrically smaller than the extreme

terms, just as with the argument for random transpositions. Further details are in Diaconis and Shahshahani (1987b). □

Remark 1. As described in Section E, the analysis gives a precise formula for the eigenvectors and eigenvalues of the transition matrix of this problem treated as a Markov chain. Karlin and McGregor (1961) essentially derived this result without using group theory. Their application was to a similar problem arising as a genetics model due to Moran. A clear discussion of Moran's model can be found in Ewens (1979, Sec. 3.3). Diaconis and Shahshahani give applications to a learning problem discussed by Piaget.

Remark 2. As usual with approximation, some precision has been lost to get a clean statement. The basic result is the bound of the corollary to Theorem 9. When $r = 1$ for example there is only one term: $(n-1)(\frac{1}{n-1})^{2k}$. For $k = 1$ taking square roots gives $\frac{1}{2}\frac{1}{\sqrt{n-1}}$ as an upper bound for the variation distance. Elementary considerations show that the exact distance in this case is $1/n$. Here, when n is large, use of the upper bound lemma gives the right answer for the number of steps required (namely 1) to make the distance small but an overestimate for the distance itself.

EXERCISE 18. Consider two urns, the left containing n red balls, the right containing n black balls. At each stage "a" balls are chosen at random from each urn and the two sets are switched. Show that this is bi-invariant. Show that for fixed a, as $n \to \infty$, this speeds things up by a factor of a (so $\frac{1}{4a}n \log n$ moves suffice).

Remark 3. A reasonable number of other problems allow very similar analysis. Stanton (1984) contains a list of finite homogeneous spaces arising from Chevalley groups where (a) the associated $L(X)$ is a Gelfand pair, and (b) the spherical functions are explicitly known orthogonal polynomials. One case of particular interest is a walk in the set of r-dimensional subspaces of an s dimensional vector space over a finite field. See Greenhalgh (1988) for details. In all cases, there is a natural metric so that nearest neighbor walks on X allow a 1-dimensional analysis. For the example of r-dimensional subspaces the distance is $d(x, y) = r - \dim(x \cap y)$.

A special case of this analysis is nearest neighbor walk on the cube $X = Z_2^n$. Here G can be represented as the semi-direct product of Z_2^n with S_n. This is the group of pairs (x, π) for $x \in Z_2^n$, $\pi \in S^n$. It acts on $y \in Z_2^n$ by $(x, \pi)(y) = \pi y + x$. Multiplication in G is composition of repeated transformations. Choosing $x_0 = 0$, the isotropy subgroup is $N = \{(0, \pi) : \pi \in S_n\} \cong S_n$. It is easy to verify that $L(X) = V_0 \oplus V_1 \oplus \ldots \oplus V_n$ where V_j is the subspace spanned by the functions $\{f_y(x)\}_{|y|=j}$ and $f_y(x) = (-1)^{x \cdot y}$. Thus G, N is a Gelfand pair and $\dim V_j = \binom{n}{j}$. The spherical functions $s_j(x)$ again only depend on $d(x, 0)$ (with $d(x, y) = \#\text{places}(x_i \neq y_i)$) and are polynomials in d called Krawtchouk polynomials:

$$s_j(d) = \frac{1}{\binom{n}{j}} \sum_{m=0}^{j} (-1)^m \binom{d}{m}\binom{n-d}{j-m}.$$

The upper bound found by treating this problem as a Gelfand pair is the same as the upper bound by treating the problem on the group Z_2^n (Example 2 of Section C).

Remark 4. The theory of Gelfand pairs can be developed without using group theory. One advantage of the present program is that it offers a route to follow for problems where the representation is not multiplicity free. For example, consider the Bernoulli-Laplace urn model with 3 urns; the first containing n red, the second containing n white, the third containing n blue balls. At each stage, a pair of urns is chosen at random, then a randomly picked pair of balls is switched. Analysis of the contents of even the first urn is complicated by the fact that the associated representation of S_{3n} on $L(X)$, with $X = S_{3n}/S_n \times S_n \times S_n$, has multiplicity. (See Young's rule in Chapter 7.) This is an open problem.

There is a useful sufficient condition for showing that (G, N) is Gelfand without explicitly determining the decomposition of $L(X)$.

LEMMA 6. *(Gelfand's lemma). Let τ be $1-1$ homomorphism $\tau: G \to G$ with the property $s^{-1} \in N\tau(s)N$ for all $s \in G$. Then (G, N) is a Gelfand pair.*

Proof. Note first that for bi-invariant functions $f(s^{-1}) = f(\tau(s))$ and $\tau(N) \subset N$. If f is bi-invariant, define $\check{f}(s) = f(s^{-1})$, $f^\tau(s) = f(\tau(s))$. Thus $\check{f} = f^\tau$. Now $f * g(t) = \Sigma_s f(ts^{-1})g(s)$, so

$$f \check{*} g(t) = \Sigma_s f(t^{-1}s^{-1})g(s) = \Sigma_z f(z^{-1})g(zt^{-1}) = \Sigma_z \check{f}(z)\check{g}(tz^{-1}) = \check{g} * \check{f}(t),$$
$$(f * g)^\tau(t) = \Sigma_s f(\tau(t)s^{-1})g(s) = \Sigma f(\tau(t)\tau(s)^{-1})g(\tau(s)) = f^\tau * g^\tau(t).$$

It thus follows that for all bi-invariant f, g

$$\check{f} * \check{g} = g \check{*} f = (g * f)^\tau = g^\tau * f^\tau = \check{g} * \check{f},$$

so $f * g = g * f$. □

Example 1. Let N be any group, A an Abelian group and suppose N acts on A. Form the semi-direct product $G = N \times_s A$ as the set of pairs (n, a) with $(n_2, a_2)(n_1, a_1) = (n_2 n_1, n_2 a_1 + a_2); (n, a)^{-1} = (n^{-1}, -n^{-1}a)$. These are all Gelfand pairs as one sees by considering the $1-1$ homomorphisms $\tau(n, a) = (n, -a)$. This satisfies $(n, a)^{-1} = (n^{-1}, 0)(n, -a)(n^{-1}, 0)$.

As examples we have the dihedral groups, the group of the cube $(S_n \times_s Z_2^n)$, the affine group $Z_m^* \times_s Z_m$. The Euclidean group $SO^d \times_s \mathbb{R}^d$ is also a Gelfand pair.

Example 2. (Groups of isometries). Let (X, d) be a finite metric space on which G acts. Suppose d is G invariant. Say G acts 2 *point homogeniously* if for all (x_1, y_1), (x_2, y_2) with $d(x_1, y_1) = d(x_2, y_2)$ there is an s such that $sx_1 = x_2$, $sy_1 = y_2$. Observe that G operates transitively (take $x_1 = y_1$, $x_2 = y_2$, then $sx_1 = x_2$ for some s). Pick $x_0 \in X$, let $N = \{s \in G: sx_0 = x_0\}$. Then (G, N) is Gelfand with $\tau(s) = s$. To see this observe $d(x_0, sx_0) = d(x_0, s^{-1}x_0)$. Thus there is an n so $nsx_0 = s^{-1}x_0$. This implies $sns \in N$, so $s^{-1} \in NsN$.

There are *many* special cases for this construction – most notably graphs whose automorphism groups act 2 point homogeniously. Biggs (1984) gives an extensive list, and a detailed recipe for determining the associated spherical functions. As a special case, consider X as the k sets of an n set with distance $d(x,y) = k - |x \cap y|$. The symmetric group S_n operates 2 point homogeniously. The isotropy subgroup is $S_k \times S_{n-k}$, and we have recaptured the Bernoulli-Laplace model. A continuous example has $X = S^n$ (the n-sphere), $G = SO(n)$.

It is interesting to know when (G,N) can be shown Gelfand by the existence of a homomorphism τ. Diaconis and Garsia (1988) show that $\tau(s) = s$ works if and only if the representation of G in the space of real functions on X is multiplicity free. They also present counter examples (a Gelfand pair that doesn't admit an automorphisim) and discussion.

We have seen that Fourier analysis of bi-invariant functions on a Gelfand pair offers a rich theory and collection of examples. The commutativity, which makes life so easy here, is also present in the analysis of functions that are constant on conjugacy classes. It is not surprising that one can be regarded as a special case of the other.

EXERCISE 19. Let G be a finite group. Let $G \times G$ act on G by $(s,t)x = s^{-1}xt$. The isotropy subgroup N in $G \times G$ is isomorphic to G as is the quotient space X. Show that $(G \times G, N)$ is a Gelfand pair. The decomposition of $L(X)$ is into $\rho \oplus \tilde{\rho}$ with $\tilde{\rho}(s) = \rho^*(s^{-1})$ as ρ varies over irreducible representations of G. These are all distinct irreducible representations of $G \times G$. Find the spherical functions in terms of the characters and show how Fourier analysis of N invariant functions on X via Gelfand pair techniques is the same as Fourier analysis of functions constant on conjugacy classes as developed in section D.

There are two generalizations of Gelfand pairs worth mentioning here: association schemes and Hypergroups.

An *association scheme* is a finite set X with a collection of relations $R_0, R_1, \ldots, R_d$. Take R_i as a zero-one matrix indexed by X with a 1 in position (x,y) if x and y are related in the ith relation. The R_i's satisfy axioms: (1) $R_0 = \mathrm{Id}$, (2) $\Sigma R_i = J$ (matrix of all ones), (3) for every i there is an i' such that $R_i^t = R_{i'}$, (4) $R_i R_j = \Sigma p_{ij}^k R_k$ for non-negative integers p_{ij}^k. If $R_i R_j = R_j R_i$, the association scheme is called commutative.

Commutative association schemes have an interesting elementary theory. MacWilliams and Sloane (1981) give an efficient development. Bannai and Ito (1986, 1987) give a very well done encyclopedic treatment.

Because of (4) the set of all linear combinations of the R_i form a_n algebra. For commutative association schemes the R_i can be simultaneously diagonalized. For many examples, this diagonalization is very explicit.

As one example, take G a group, H a subgroup with $X = G/H$ and (G,H) a Gelfand pair. Then G acts on $X \times X$ by $g(x,y) = (gx, gy)$. Take the orbits of this action as relations on X. These relations form a commutative association scheme with algebra isomorphic to the convolution algebra $L(X)$.

In the other direction, consider a commutative association scheme X. The axioms imply that every row and column of R_i have the same number k_i of ones

in each row and column. Thus R_i/k_i is a doubly stochastic matrix. If $w_i \geq 0$ sum to 1,

$$M = \sum_{i=0}^{d} \frac{w_i}{k_i} R_i$$

is doubly stochastic and so defines a Markov chain on X. The point is, for hundreds of examples, this Markov chain is explicitly diagonalizable using available information. Classical Markov chain techniques can then be used to derive answers to the usual questions. Diaconis and Smith (1987) derive an appropriate upper bound lemma and carry through some examples that don't arise from groups.

Association schemes were originally developed by statisticians for analysis of variance problems. Speed (1987) shows how they have come to life recently for new statistical applications. Coding theorists, combinatorialists, and finite group theorists have been the principal developers of assocation schemes in recent years. Bannai and Ito (1986, 1987) survey these developments and examples.

A *Hypergroup* begins with a set X and introduces a product on probabilities on X — so the product of two pointmasses is a probability (which is not usually a point mass). For example, a product can be introduced on the conjugacy classes of a group: e.g. in the symmetric group, the product of two transpositions can be the identity, a 3-cycle on the product of two 2-cycles. These occur with certain mass. As a second example, the set of irreducible representations on a compact group can be made into a Hypergroup using the Tensor product and its associated weights.

A reasonable amount of theory and examples have been developed. There has started to be a payoff to more classical areas. For example, Bochner's theorem for Gelfand pairs or class functions follows from Hypergroup Theorems of Johanson (1981). It is still unknown in general cases. Gallardo (1987) presents a nice example of Fourier analysis for a class of birth and death processes that is available by interpreting the decomposition of Tensor products on $SU(2)$ as rules for births and deaths. Zeuner (1987) gives a unified treatment of central limit problmes on Hypergroups and pointers to related literature.

Hypergroups offer a continuous generalization of association schemes. They appear to offer an extension worth keeping track of.

There are many further topics to discuss relating to Gelfand pairs. The interested reader is referred to the annotated bibliography in Section G.

G. Some references on Gelfand pairs.

The literature on Gelfand pairs is already sizeable. I hope the following annotated bibliography will help. The articles by Bougerol and Stanton are very clear and give details. The articles by Sloane and Heyer have extensive bibliographies.

Bailey, R. and Rowley, C. A. (1987). General balance and treatment permutations. Techical Report, Statistics Department, Rothamsted Experimental Station. Harpenden, Herts, AL5 2JQ, United Kingdom.

This paper is important in offering a bridge betwen the mathematics of Gelfand pairs and an important component of designed experiments — gener-

alized balance. Many experimental designs are constructed using group theory. The paper shows that many such designs automatically have nice statistical properties.

Biggs, N. (1974). *Algebraic Graph Theory.* Cambridge University Press, London.

Chapters 20, 21 discuss "distance transitive graphs." These are what we called two-point homogeneous. Graph theorists have derived lots of facts about the eigenvalue, eigenvectors of these groups redeveloping the tools of Gelfand pairs.

Bougerol, P. (1983). *Un Mini-Cours Sur Les Couples de Guelfand.* Pub. du Laboratoire de Statistique et Probabilities, Universite Paul Sabatier, Toulouse.

A terrific set of lectures with complete proofs and no "baloney," many examples.

Bougerol, P. (1981). Theoreme Central Limite Local Sur Certains Groupes de Lie. *Ann. Scient. Ec. Norm. Sup. 4th Ser. 1*, **14**, 403–432.

A serious application in probability, showing how general results (*not* restricted to bi-invariant functions) can be derived using the machinery of Gelfand pairs.

Cartier, P. (1972). Fonctions Harmoniques Sur Un Arbe. *Symposia Math.* **9**, 203–270.

An elegant combinatorial derivation of all properties of this Gelfand pair. See Sawyer (1978) for an application.

Diaconis, P. and Graham, R. L. (1985). The Radon Transform Z_2^k. *Pacific Jour.* **118**, 323–345.

This can all be carried over to bi-invariant neighborhoods on Gelfand pairs.

Diaconis, P. and Shahshahani, M. (1987). Time to reach stationarity in the Bernoulli-Laplace diffusion model. *SIAM Jour. Math. Anal.* **18**, 208–218.

A longer version of Section F-3 above.

Dieudonne, J. (1978). *Treatise on Analysis VI.* Academic Press, New York.

A reasonably self-contained single source. Weighted toward the analyst, but it's possible to read.

Farrell, R. (1976). *Techniques of Multivariate Calculation.* Lecture Notes in Math, No. 520. Springer-Verlag, Berlin.

The only attempt at a beginning to end treatment of the mathematics of multivariate analysis that really does zonal polynomials.

Gangolli, R. (1972). Spherical functions on semi-simple Lie groups. In *Symmetric Spaces*, W. Boothby and G. Weiss. Marcel Dekker, New York.

Gangolli's article is a well written introduction to computations on continuous groups involving the Laplacian and its generalizations. The whole book consists of survey articles, roughly on the same topic.

Guillemin, V. and Sternberg, S. (1984). Multiplicity free spaces. *J. Differential Geometry* **19**, 31–56.

This is included to show this area is still under active development.

Helgason, S. (1978). *Differential Geometry Lie Groups and Symmetric Spaces.* Academic Press, New York.

Helgason, S. (1984). *Groups and Geometric Analysis: Integral Geometry Invariant Differential Operators, and Spherical Functions.* Academic Press, New York.

These two books give a comprehensive modern treatment of continuous Gelfand Pairs.

Helgason, S. (1973). Functions on symmetric spaces, pp. 101–146 in Harmonic Analysis on Homogeneous Spaces. *Proc. Symposia Math.* **24**, American Mathematical Society. Providence.

This entire volume shows how "grown-ups" use Gelfand pairs to do general representation theory.

Heyer, H. (1983). Convolution semigroups of probability measures on Gelfand pairs. *Expo. Math.* **1**, 3–45.

Contains a 62 item bibliography (mainly analytic, but useful).

James, A. T. (1975). Special functions of matrix and single argument in statistics. *Theory and Application of Special Functions*, R. Askey ed.

This is a summary of years of work on the example GL_n/O_n. This is a central example in the piece of math statistics known as multivariate analysis. The spherical functions, known as zonal polynomials, are used to derive distributions of things like the largest eigenvector in the covariance matrix of a normal sample.

Karlin, S. and McGregor, J. (1961). The Hahn polynomials, formulas and an application. *Scripta Math.* **23**, 33–46.

One of the earliest derivations of the special functions of $S_n/S_k \times S_{n-k}$. The applications are to a genetics model for random mating in a population with two alleles due to Moran. Many useful properties of the spherical functions are derived without mention of group theory.

Kramer, M. (1979). Sphärische Untergruppen in Kompacten Zusammenhangenden Lie Gruppen. *Composito Math.* **38**, 129–153.

He classifies, for G compact, simple, connected, Lie, all subgroups K such that (G, K) is Gelfand.

Letac, G. (1981). Problemes classiques de probabilite sur un couple de Gelfand. In *Lecture Notes in Math.* **861** (Springer-Verlag).

A very clear, elementary survey explaining a dozen applications in probability. Highly recommended.

Letac, G. (1982). Les fonctions spheriques d'un couple de Gelfand symetrique et les chaines de Markov. *Advances Appl. Prob.* **14**, 272–294.

A very clear, useful survey, explaining in particular a method of computing the spherical functions in "small" cases, so one can hope to guess at the answer.

Saxl, J. (1981). On multiplicity — free permutation representations. In *Finite Geometries and Designs*, London Math. Soc. Lecture notes, Series 48, Cambridge University Press, 337–353.

This classifies all subgroups of S_n which yield a Gelfand pair. Aside from $S_k \times S_{n-k}$ and small twists like $A_k \times A_{n-k}$ (A_k the alternating groups) the only "interesting" example is $S_2 Wr S_n$ which gives the Zonal Polynomials. See Diaconis (1987).

Saw, J. G (1977). Zonal polynomials: an alternative approach. *Jour. Multivariate Analysis* **7**, 461–467.

Derives properties of the spherical functions of GL_n/O_n without any group theory (but lots of "standard" properties of the Wishart distribution).

Sawyer, S. (1978). Isotropic random walks in a tree. *Zeit. Wahr.* **42**, 279–292.

A fascinating application of Gelfand pairs and p-adic numbers to salmon fishing!

Sloane, N. J. A. (1975). An introduction to association schemes and coding theory. *Theory and Applications of Special Functions*, R. Askey, ed.

Long, friendly introduction to the use of the tools of interest to coding theory.

Sloane, N. J. A. (1982). Recent bounds for codes, sphere packings and related problems obtained by linear programming and other methods. *Contemp. Math* **9**, 153–185.

Great, friendly article on the use of Gelfand pairs. Bibliography of 163 items.

Soto-Andrade, J. (1985). En torna a las funciones esfericas (caso finito). *Notas de la Sociedad de Matematica de Chile* IV, 71–94.

There is an active group working in Chile on Gelfand pairs. There are several other papers in this volume on this subject. A valuable thesis: *Caracteres de Espacios de Gelfand Finitos* by S. Garcia Zambrano (1984), also contains much of interest, in particular a careful discussion of spherical functions for the "orthogonal group" over a finite field.

Stanton, D. (1984). Orthogonal polynomials and Chevalley groups. In R. Askey et al (eds.) *Special Functions: Group Theoretical Aspects and Applications*, 87–92.

An important, clear, friendly survey of a dozen explicit calculations of spherical functions for *finite* spaces. Highly recommended.

Takemura, A. (1984). *Zonal Polynomials.* Institute of Mathematical Statistics. Hayward.

The best introduction to zonal polynomials for statisticians. No group theory, but lots of Wishart distributions.

H. First hitting times

Fourier analysis has been used to bound rates of convergence through the upper bound lemma. In this section a different application is presented. This permits sharp approximation of first passage probabilities and first return times for random walk. As an application, the classical gambler's ruin is given a new presentation. The arguments lean heavily on Good (1951).

Let Q be a probability on a finite group G. The random walk determined by Q starting at x is denoted Q_x^{*k}. Thus $Q_x^{*0}(y) = \delta_x(y)$, $Q_x^{*1}(y) = Q(yx^{-1})$, $Q_x^{*2}(y) = \Sigma_z Q(y(zx)^{-1})Q(z)$. In general, $Q_x^{*k} = Q_{id}^{*k} * \delta_x$.

Let $S \subset G$ be a set of elements called "sinks." We consider the random walk, starting at x and absorbed the first time it hits S. To rule out trivialities, assume $x \notin S$.

Let $a_k(t)$ be defined as the probabililty of *arriving* at the group element t at time k. If $t \notin S$, this is the chance of the walk being at t at time k, without having hit any sites in S. If $t \in S$, this is the chance of first being absorbed at t at time k.

Let $b_k(t)$ be defined as the probability of arriving at t, at time k, in the unrestricted random walk ($S = \phi$). The a's and b's are related via

LEMMA 7.

$$b_k(t) = a_k(t)\delta_{S^c}(t) + \sum_{s \in S} \sum_{j=0}^{k} a_j(s) b_{k-j}(ts^{-1}x).$$

Proof. Divide the set of paths of length k from x to t into $1 + (k+1)|S|$ classes. The first consists of paths that avoid all sinks. A typical path in one of the other classes hits a sink s for the first time at j (Probability $a_j(s)$) and then goes from s to t in the next $k - j$ steps (Probability $b_{k-j}(ts^{-1}x)$). By finite additivity, $b_k(t)$ is the sum of the probabilities of the classes. □

The convolution suggests generating functions (Fourier analysis on Z). Let $A(t,z), B(t,z)$ be the generating functions

$$\sum_{k=0}^{\infty} a_k(t) z^k, \quad \sum_{k=0}^{\infty} b_k(t) z^k.$$

COROLLARY.

$$B(t,z) = \Sigma b_k(t) z^k = \Sigma\Big(a_k(t)\delta_{S^c}(t) + \sum_{s \in S} \sum_{j=0}^{k} a_j(s) b_{k-j}(ts^{-1}x)\Big) z^k$$

$$= \delta_{S^c}(t) a(t,z) + \sum_{s \in S} A(s,z) B(ts^{-1}x, z).$$

Remark. Here is the way this formulation works: usually, we have a closed form expression for $B(t,z)$ for all t. Letting t run through S, the corollary gives $|S|$ equations in the $|S|$ unknowns $A(s,z)$. These can be solved (if $|S|$ is not too large, or is "symmetric") and then the corollary gives an expression for $A(t,z)$ for all t. The group theory enters as follows:

LEMMA 8. *With notation as above, $B(t,z)$ equals*

$$\frac{1}{|G|} \Sigma_\rho d_\rho \, Tr(\rho(xt^{-1}) \cdot (I - z\hat{Q}(\rho))^{-1}).$$

The inverse exists, at least for $|z| < 1$. *The sum is over all irreducible representations of* G.

Proof. This is just the Fourier inversion theorem applied to the definition of $B(t, z)$. □

Classical Gambler's Ruin: Peter and Paul flip a fair coin, Peter wins \$1 if the coin comes up heads; Paul wins \$1 if the coin comes up tails. Peter starts with $\$\ell$, Paul starts with $\$m$. The game stops at time T when one of them has no money left.

This can be analyzed as simple random walk on Z_n, where $n = \ell + m$. The particle starts at ℓ, and the game ends the first time zero is hit. For example, suppose Peter has 1 and Paul has 4

$$\begin{array}{ccc} & 0 & \\ 4 & & \textcircled{1} \\ 3 & & 2 \end{array}$$

A walk starting at 1 stops after 1 step to the left (Peter wiped out) or 4 steps to the right (Paul wiped out), etc.

Here there is one sink, namely, zero. From Lemmas 7, 8

$$A(0,z) = \frac{B(0,z)}{B(\ell,z)} = \frac{\sum_{j=0}^{n-1} \frac{e^{2\pi ij\ell/n}}{1 - z\cos(\frac{2\pi j}{n})}}{\sum_{j=0}^{n-1} \frac{1}{1 - z\cos(\frac{2\pi j}{n})}}.$$

Note that the numerator and denominator both have simple poles at $z = 1$. It follows that the left side is analytic in $|z| \leq 1$. Writing

$$\frac{1}{1-z} + \sum_{j=1}^{n-1} \frac{e^{2\pi ij\ell/n}}{1 - z\cos(\frac{2\pi j}{n})} = \left\{\frac{1}{1-z} + \sum_{j=1}^{n-1} \frac{1}{1 - z\cos(\frac{2\pi j}{n})}\right\}\{A(0,1) - (1-z)A'(0,1) + \ldots\}$$

Here A' denotes differentiation with respect to the second argument. Comparing coefficient as $z \to 1$ (set $1 - z = \varepsilon$) gives

Result 1. $A(0,1) = 1$ *so absorption is certain.*

Result 2.
$E(T) = A'(0,1) = \sum_{j=1}^{n-1} \frac{1-e^{2\pi ij\ell/n}}{1-\cos(\frac{2\pi j}{n})} = \sum_{j=1}^{n-1} (1 - \cos(\frac{2\pi j\ell}{n}))\ (1 - \cos(\frac{2\pi j}{n}))^{-1}$. Here is a curious consequence. By an elementary argument (Feller (1968, Chapter 14) $E(T) = \ell(n - \ell)$. This gives a curious trigonometric identity. Pass to the limit, with ℓ fixed, $n \to \infty$, the following emerges:

$$\int_0^1 \frac{1 - \cos(2\pi\ell t)}{1 - \cos(2\pi t)} = \ell.$$

It is also straightforward to pass to the limit in the original generating function:

Result 3.
Case 1. Let ℓ be fixed and let $n \to \infty$:

$$A(0,z) \to \frac{\int_0^1 \frac{\cos(2\pi t\ell)}{1-z\cos(2\pi t)}}{\int_0^1 \frac{1}{1-z\cos(2\pi t)}} = \left(\frac{1-(1-z^2)^{\frac{1}{2}}}{z}\right)^{\ell}.$$

The second identity is derived as follows: by expanding both sides, verify $\int_0^{2\pi} \frac{1}{1-z\cos t} = 2\pi(1-z^2)^{-\frac{1}{2}}$. Then, for $\ell = 1$, $A(0,z) = \frac{N(z)}{D(z)}$ with $D - zN = 1$. This gives the right side when $\ell = 1$. The general case follows from the convolution interpretation of the left side.

Case 2. Let $\ell = \theta n$ for $0 < \theta < 1$ fixed. Make the change of variables $z = e^{-\lambda/n^2}$. Then as n tends to ∞, $E\{e^{\lambda T/n^2}\}$ tends to

$$\frac{\sum_{j=0}^{\infty} \frac{\cos(2\pi\theta j)}{\lambda+(2\pi j)^2/2}}{\sum_{j=0}^{\infty} \frac{1}{\lambda+(2\pi j)^2/2}}.$$

This last function is the Laplace transform of a probability measure on $\mathbb{R}^+$.

EXERCISE 20. Consider nearest neighbor walk on the cube Z_2^n as described in Example 2 of Section C. Let T be the first return time to zero. Prove that $E(T) = 2^n$, and show that $T/2^n$ has a limiting exponential distribution.

Remarks. Flatto, Odlyzko, and Wales (1985) carried out similar computations for random transpositions. All of these computations use only 1 sink. Using 2 sinks, one can derive the chance that Peter wins in gambler's ruin, or the law of the maximum of random walk, given that its first return is at time $2k$; see Smith and Diaconis (1988) for references to the literature. Similar results on the cube would give results for fluctuations and excursions of the Ornstein-Uhlenbeck process, or any of the birth and death chains described in Section F.

An elegant application of first hitting distributions for simple random walk on the n-cube to the analysis of algorithms appears in Aldous (1983b). Consider finding the minimum of a function $f: Z_2^n \to \mathbb{R}$. Clearly general functions take order 2^n steps for any algorithm on average. People in operations research hoped that "local-global" functions with the property that if $f(x)$ is not a minimum, then $f(y) < f(x)$ for some neighbor y of x, would be a useful special class.

The obvious algorithm is: start someplace, and take your smallest neighbor as your next place, etc. Craig Tovey showed there were some exponentially bad examples, but naturally created functions seemed to work in order n^2 steps.

Aldous treated the problem as a two-person game: nature picks a function, we pick an algorithm. We pay nature the number of steps it takes our algorithm

to find the minimum. Because both sets of possibilities are finite (things only depend on the relative values) the game has a value v. Aldous showed that the value was approximately $2^{n/2}$.

The two strategies are easy to describe: your (randomized) strategy is to pick vertices $x_1, x_2 \ldots, x_J$ at random ($J \doteq 2^{n/2}$) and then use the obvious algorithm starting at $\min(f(x_i))$.

Nature's strategy involves choosing a random local-global function as follows. Start simple random walk at a random point x_0. Let $f(x)$ be the number of steps until the walk first hits x. Thus $f(x_0) = 0$, and for any other x there is a neighbor y of x which was visited first. Thus f is local-global. By careful analysis of random walk, Aldous is able to show it takes order $2^{n/2}$ steps to find the minimum with any algorithm.

Chapter 4. Probabilistic Arguments

A. Introduction — Strong uniform times.

There are a number of other arguments available for bounding the rate of convergence to the uniform distribution. This chapter discusses the method of strong uniform times and coupling. Let's begin with a simple example, drawn from Aldous and Diaconis (1986).

Example 1. Top in at random. Consider mixing a deck of n cards by repeatedly removing the top card and inserting it at a random position. This corresponds to choosing a random cycle:

$$(1) \qquad P(\mathrm{id}) = P(21) = P(321) = P(4321) = \ldots = P(nn-1\ldots1) = \frac{1}{n}.$$

The following argument will be used to show that $n \log n$ shuffles suffice to mix up the cards. Consider the bottom card of the deck. This card stays at the bottom until the first time a card is inserted below it. This is a geometric waiting time with mean n. As the shuffles continue, eventually a second card is inserted below the original bottom card (this takes about $n/2$ further shuffles). The two cards under the original bottom card are equally likely to be in relative order low-high or high-low.

Similarly, the first time a third card is inserted below the original bottom card, each of the six possible orders of the three bottom cards is equally likely. Now consider the first time T that the original bottom card comes to the top. By an inductive argument, all $(n-1)!$ arrangements of the lower cards are equally likely. When the original bottom card is inserted at random, all $n!$ possible arrangements of the deck are equally likely.

When the original bottom card is at position k from the bottom, the waiting time for a new card to be inserted is geometric with mean n/k. Thus the waiting time T has mean $n + \frac{n}{2} + \frac{n}{3} + \ldots + \frac{n}{n} \doteq n \log n$.

To make this argument rigorous, introduce strong uniform times. Let G be a finite group. Intuitively, a stopping time is a rule which looks at a sequence of elements in G and says "stop at the jth one." The rule is allowed to depend on what appears up to time j, but not to look in the future. Formally, *a stopping time* is a function $T: G^\infty \to \{1, 2, \ldots, \infty\}$ such that if $T(\underset{\sim}{s}) = j$ then $T(\underset{\sim}{s}') = j$ for all $\underset{\sim}{s}'$ with $s_i' = s_i$ for $1 \le i \le j$. Let Q be a probability on G, X_k the associated random walk, P the associated probability on G^∞. A *strong uniform time* T is a stopping time T such that for each $k < \infty$,

$$(2) \qquad P\{T = k,\ X_k = s\} \quad \text{is constant in } s.$$

Note that (2) is equivalent to independence of the stopping time and the stopped process:

$$P\{X_k = s|T = k\} = 1/|G| \tag{3}$$

or to

$$P\{X_k = s|T \leq k\} = 1/|G|. \tag{4}$$

In Example 1, the time T that the first card takes to reach the top and has been inserted into the deck is certainly a stopping time. The inductive argument given shows that, given $T = k$, all arrangements of the deck are equally likely, so T is a strong uniform time. Many other examples will be given in the remainder of this chapter. The following lemma relates strong uniform times to the distance between Q^{*k} and the uniform distribution U.

LEMMA 1. *Let Q be a probability on the finite group G. Let T be a strong uniform time for Q. Then for all $k \geq 0$*

$$||Q^{*k} - U|| \leq P\{T > k\}.$$

Proof. For any $A \subset G$,

$$\begin{aligned} Q^{*k}(A) &= P\{X_k \in A\} \\ &= \sum_{j \leq k} P\{X_k \in A, T = j\} + P\{X_k \in A, T > k\} \\ &= \sum_{j \leq k} U(A)P(T = j) + P\{X_k \in A|T > k\}\ P\{T > k\} \\ &= U(A) + [P\{X_k \in A|T > k\} - U(A)]\ P\{T > k\}. \end{aligned}$$

Thus,

$$|Q^{*k}(A) - U(A)| \leq P\{T > k\}.$$

□

Using this result we can deduce a sharp bound for the first example: $n \log n$ steps are both necessary and sufficient to drive the variation distance to zero.

Theorem 1. *For the top in at random shuffle defined in (1), let $k = n \log n + cn$. Then,*

$$||P^{*k} - U|| \leq e^{-c} \text{ for } c \geq 0,\ n \geq 2, \tag{5}$$

$$||P^{*k} - U|| \to 1 \text{ as } n \to \infty,\ \textit{for } c = c(n) \to -\infty. \tag{6}$$

Proof. As argued above, $T = T_1+(T_2-T_1)+\ldots+(T_{n-1}-T_{n-2})+(T-T_{n-1})$ where T_1 is the time until the 1st card is placed under the bottom card and $T_{i+1} - T_i$ has a geometric distribution $P\{T_{i+1} - T_i = j\} = \frac{i+1}{n}(1 - \frac{i+1}{n})^{j-1}$; $j \geq 1$. Further, these differences are independent.

The time T has the same distribution as the waiting time in the coupon collector's problem; to define this, consider a random sample with replacement from an urn with n balls. Let V be the number of balls required until each ball has been drawn at least once. Let $m = n \log n + cn$. For each ball b, let A_b be the event "ball b is not drawn in the first m draws." Then,

$$(7) \quad P\{V > m\} = P\{\cup_b A_b\} \leq \Sigma_b\, P\{A_b\} = n(1 - \frac{1}{n})^m \leq n\, e^{-m/n} = e^{-c}.$$

Now V can be written

$$V = (V - V_{n-1}) + (V_{n-1} - V_{n-2}) + \ldots + (V_2 - V_1) + V_1$$

where V_i is the number of draws required until i distinct balls have been drawn at least once. After i distinct balls have been drawn, the chance that a draw produces a new ball is $\frac{n-i}{n}$, so $V_{i+1} - V_i$ is geometric,

$$P\{V_{i+1} - V_i = j\} = \frac{n-i}{n}(1 - \frac{n-i}{n})^{j-1},\ j \geq 1.$$

It follows that the laws of T and V are the same. So (7) and lemma 1 (the upper bound lemma) combine to give a proof of (5).

To prove (6), fix j and let A_j be the set of configurations of the deck such that the bottom j original cards remain in their original relative order. Plainly $U(A_j) = 1/j!$. For $k = n \log n + c_n n$, $c_n \to -\infty$, we argue that

$$(8) \qquad P^{*k}(A_j) \to 1 \text{ as } n \to \infty,\ j \text{ fixed.}$$

Then $||P^{*k} - U|| \geq \max_j\{P^{*k}(A_j) - U(A_j)\} \to 1$ as $n \to \infty$, establishing (6).

To prove (8), observe that $P^{*k}(A_j) \geq P(T - T_{j-1} > k)$. For $T - T_{j-1}$ is distributed as the time for the card initially jth from bottom to come to the top and be inserted; and if this has not occurred by time k, then the original bottom j cards must still be in their original relative order at time k. Thus it suffices to show

$$(9) \qquad P(T - T_{j-1} \leq k) \to 0 \text{ as } n \to \infty;\ j \text{ fixed}\ .$$

We shall prove this using Chebyshev's inequality:

$$P(|Z - EZ| \geq a) \leq \frac{\text{var}(Z)}{a^2}, \text{ where } a \geq 0, \text{ and } Z \text{ is any random variable}\ .$$

For a geometric variable

$$E(T_{i+1} - T_i) = \frac{n}{i+1}, \quad \text{var}(T_{i+1} - T_i) = (\frac{n}{i+1})^2(1 - \frac{i+1}{n}),$$

and so

$$E(T - T_j) = \sum_{i=j}^{n-1} \frac{n}{i+1} = n \log n + 0(n),$$

$$\text{var}(T - T_j) = \sum_{i=j}^{n-1} (\frac{n}{i+1})^2(1 - \frac{i+1}{n}) = 0(n^2),$$

and Chebyshev's inequality applied to $Z = T - T_{j-1}$ readily yields (9). □

B. Examples of strong uniform times.

Example 2. Simple random walk on Z_2^d. For simplicity we work with the following probability

$$Q(0\ldots0) = \frac{1}{2}, \; Q(10\ldots0) = Q(01\ldots0) = \ldots = Q(0\ldots1) = \frac{1}{2d},$$

(10) $$Q = 0 \quad \text{otherwise}.$$

The following simple stopping time has been developed by Andre Broder. It involves "checking off coordinates" according to the following scheme: at each time, pick one of the d coordinates at random and check it off. Then flip a fair coin. If the coin comes up heads, take a step in the direction of the chosen coordinate. If the coin comes up tails, the random walk stays where it is. Stop at time T when all coordinates have been checked.

Clearly the particle evolves according to the probability (10). To see that T is a strong uniform time, observe that because of the randomized coin toss, the particle is equally likely to have a zero or one in each checked coordinate.

Theorem 2. *For simple random walk on Z_2^d (10), and $k = n \log n + cn$,*

$$||P^{*k} - U|| \leq e^{-c}.$$

Proof. This follows from the upper bound lemma and the bound from the coupon collector's waiting time (7). □

Remark 1. Fourier analysis and the lower bound arguments of Chapter 3 show that $\frac{1}{2}n \log n + cn$ steps is the right answer for this version of random walk. The discrepancy is explained in Section C (exercise 4) below.

Remark 2. In Example 2, the uniform time depends on added, external, randomization. It was not constructed just by looking at the past of the process. The

upper bound, lemma 1, holds for such randomized strong uniform times without change.

EXERCISE 1. Give a strong uniform time for random walk on Z_2^d determined by $Q(00\ldots0) = Q(10\ldots0) = \ldots = Q(0\ldots01) = \frac{1}{d+1}$.

Example 3. General random walk on a finite group.

Theorem 3. *Let G be a finite group and Q a probability on G such that for some $c(0 < c < 1)$ and k_0,*

$$Q^{*k}\{A\} \geq cU(A) \tag{11}$$

for all $A \subset G$, $k \geq k_0$. Then

$$||Q^{*k} - U|| \leq (1-c)^{\lfloor k/k_0 \rfloor}.$$

Proof. Suppose first that $k_0 = 1$. Define a probability Q_1 on G by

$$Q_1(s) = \frac{Q(s) - cU(s)}{1-c}.$$

Thus

$$Q(s) = cU(s) + (1-c)Q_1(s).$$

This gives the following recipe for choosing steps according to Q: flip a coin with probability of heads equal to c. If the coin comes up heads, step according to U, if tails, step according to Q_1. Let T be the first time a head occurs. This T is clearly a strong uniform time and

$$P\{T > k\} = (1-c)^k.$$

For general k_0, apply the argument to Q^{*k_0}. □

Remarks. The argument above extends easily to compact groups with condition (11) required to hold for all open sets. In this generality, the theorem appears in Kloss (1959) whose proof is a Fourier version of the same argument Athreya and Ney (1978) apply this idea to prove convergence to stationarity for general state space Markov chains.

The simplicity of the proof, coupled with the generality of the argument, should make the reader suspicious. While the result seems quantitative, all depends on estimating c and k_0. I do not know how to use this theorem to get the right rate of convergence in a single example.

Example 4. Random transpositions. This problem was discussed at some length in Chapter 3. Here are two constructions of strong uniform times. Both involve the notion of "checking" the backs of certain cards as they are chosen successively

in pairs. This argument is capable of use in a variety of other random walks. It is due to Andre Broder.

Construction A (Broder). The basic mixing procedure involves switching pairs of cards (L_i, R_i). If either

a) both hands touch the same unchecked card; or
b) the card touched by the left hand is unchecked but the card touched by the right hand is checked,

then check the card touched by the left hand. Stop at time T when all cards are checked.

Construction B (Matthews). If both hands touch unchecked cards, then check the card touched by the left hand.

In each construction stop at the time T that only one card remains unchecked.

Proof. Construction A. First consider the situation informally. The procedure starts when both hands hit the same card (say card 1). This is checked. Nothing happens until either both hands hit a different card (say 2) or the left hand hits an unchecked card (say 2) and the right hand hits card 1 whereupon these cards are switched. At this stage, conditional on the positions of the two checked cards A_1, A_2 say, and the labels 1, 2, the positions are equally likely to correspond $(A_1, 1)(A_2, 2)$ or $(A_1, 2)(A_2, 1)$. This is because the chance of choosing card 2 is $\frac{1}{n^2}$ for both possibilities.

In general, the position may be described as follows:

$$\{L, \{A_1 \dots A_L\}, \{C_1 \dots C_L\}, \Pi_L\}.$$

Where

$$\begin{aligned} L &= \text{number of checked cards} \\ \{A_1 \dots A_L\} &= \text{set of positions of the checked cards} \\ \{C_1 \dots C_L\} &= \text{labels (or names) of the checked cards.} \\ \Pi_L\colon \{A_1 \dots A_L\} &\to \{C_1 \dots C_L\} \text{ records the card at each position.} \end{aligned}$$

□

Claim. *At each time, conditional on* L, $\{A_1 \dots A_L\}$, $\{C_1 \dots C_L\}$, *the permutation* Π_L *is uniform.*

The claim is proved by induction. It is clearly true for $L = 0$ and 1. Assume it for $L = \rho$. The claim remains true until a new card c is checked. This can occur by both hands hitting the same new card or by the left hand hitting c and the right hand hitting one of the ρ checked cards. For any new card c, each of these $\rho + 1$ possibilities has the same chance $\frac{1}{n^2}$. It follows that Π_{L+1} is uniform.

The proof for Construction B is similar. The state at any time can again be taken as above. This time the inductive step is that, given L,

a) $\{A_1 \dots A_L\}$ and $\{C_1 \dots C_L\}$ are independent and uniformly distributed
b) Given L, $\{A_1 \dots A_L\}$ and $\{C_1 \dots C_L\}$, the permutation Π_L is uniform.

It can be verified that both a and b hold at each time for Construction B. Here, (a) is needed to check (b) in the argument. Note that (a) is not valid (or needed) for Construction A. ($\{A_1 \dots A_L\}$ and $\{C_1 \dots C_L\}$ are marginally uniform but not independent.)

The analysis of T in Construction A is similar to that in Example 2. Write $T = \sum_{i=1}^{n}(T_i - T_{i-1})$ where T_i is the number of transpositions until i cards are checked. The random variables $(T_i - T_i)$ are independent with geometric distributions of mean $n^2/[(i+1)(n-i)]$. Thus

$$E(T) = \sum_{i=0}^{n-1} n^2/[(i+1)(n-i)] = (2 + 0(\frac{1}{n}))n \log n$$
$$\mathrm{Var}(T) = 0(n^2).$$

Now the central limit theorem implies for $k = 2n \log n + c(n)n$, with $c(n) \to \infty$.

$$||P^{*k} - U|| \to 0 \text{ as } n \to \infty.$$

The T given by Construction B turns out to give $k = 0(n^2)$ as required. However, Construction B starts out by checking cards rapidly. Peter Matthews (1986b) observes that the two constructions can be combined: use Construction B until m cards have been checked (for fixed m, say $m = \frac{n}{2}$). Then use Construction A. Because (a) and (b) are valid throughout the time involved for Construction B, when A takes over, (b) remains valid until the time T that all cards are checked. This time gives $k = n \log n$ sufficient to drive the variance distance to zero. Matthews has suggested variants which give the correct number of steps $\frac{1}{2}n \log n$.

EXERCISE 2. To emphasize the need for careful proof, show that checking each card as it is touched or checking each card the left hand touches do not yield strong uniform times in the random transposition problem. (Hint: consider a three-card deck, and see what the distribution is given $T = 3$.)

Further examples (simple random walk on Z_p or $X_{k+1} = a_k X_k + b_k$) are given in Aldous and Diaconis (1986, 1987a, 1987b) and Matthews (1986a,b).

C. A CLOSER LOOK AT STRONG UNIFORM TIMES.

The success of strong uniform times in the examples above and a variety of other examples given below prompts obvious questions: can one always find a useful strong uniform time? Are there strong uniform times that achieve the variation distance? To answer these questions it is useful to introduce a different notion of distance from uniformity.

Definition. Let Q be a probability on the finite group G. Define the n step *separation* by

$$s(n) = |G| \max_s \{\frac{1}{|G|} - Q^{*n}(s)\}.$$

Clearly, $0 \le s(n) \le 1$ with $s(n) = 0$ iff $Q^{*n} = U$, $s(n) = 1$ iff $Q^{*n}(s) = 0$ for some s. The separation is an upper bound for the variation distance:

$$||Q^{*n} - U|| \le s(n)$$

because

$$||Q^{*n} - U|| = \sum_{s:Q^{*n}(s)<1/|G|} \left\{ \frac{1}{|G|} - Q^{*n}(s) \right\}.$$

Note that the two distances can be very different: If Q is uniform on $G - \{\mathrm{id}\}$ then $||Q - U|| = 1/|G|$ but $s(1) = 1$. The following theorem improves the upper bound of lemma 1.

Theorem 4. *If T is a strong uniform time for the random walk generated by Q on G, then for all k*

$$s(k) \le P\{T > k\}. \tag{12}$$

Conversely, for every random walk there is a strong uniform time such that (12) holds with equality.

Proof. Let k_0 be the smallest value of k such that $P\{T \le k_0\} > 0$. The result holds vacuously if $k_0 = \infty$ and for $k < k_0$. For $k \ge k_0$, $s \in G$

$$\begin{aligned} |G|\{\frac{1}{|G|} - Q^{*k}(s)\} &= 1 - |G|Q^{*k}(s) \le 1 - |G|P\{X_k = s \text{ and } T \le k\} \\ &= 1 - |G|P\{X_k = s | T \le k\} \cdot P\{T \le k\} \\ &= 1 - P\{T \le k\} = P\{T > k\}. \end{aligned}$$

This proves (12).

For the converse, the random time T will be defined as follows: at time k, given that the random walk is at t, flip a coin with probability of heads

$$p_k(t) = \frac{\alpha_k - \alpha_{k-1}}{Q^{*k}(t) - \alpha_{k-1}}$$

where $\alpha_k = \min_s Q^{*k}(s)$. If heads comes up, stop. If tails comes up, take another step and flip again with probability p_{k+1}. Observe that $p_k(t) \ge 0$. Let k_0 be the smallest integer such that $\alpha_{k_0} > 0$. Clearly $p_k = 0$ for $k < k_0$, and

$$p_{k_0}(t) = P\{T = k_0 | X_{k_0} = t\} = \frac{\alpha_{k_0}}{P\{X_{k_0} = t\}}.$$

Thus, $P\{T = k_0\} = \Sigma_t\, P\{T = k_0 | X_{k_0} = t\} \cdot P\{X_{k_0} = t\} = |G|\alpha_{k_0}$. Further,

$$P\{X_{k_0} = s | T = k_0\} = P\{T = k_0 | X_{k_0} = s\} \cdot \frac{P\{X_{k_0} = s\}}{P\{T = k_0\}} = \frac{1}{|G|}.$$

This is the first step in an inductive argument to show that T is a strong uniform time. For general k,

$$P\{X_k = s, T = k\} = \alpha_k - \alpha_{k-1}. \tag{13}$$

This follows because

$$\begin{aligned} P\{X_k = s, T = k\} &= P\{T = k | X_k = s, T \geq k\} \cdot P\{X_k = s, T \geq k\} \\ &= \frac{\alpha_k - \alpha_{k-1}}{P\{X_k = s\} - \alpha_{k-1}} \cdot [P\{X_k = s\} - P\{X_k = s; \\ &\qquad T \leq k-1\}] \end{aligned}$$

If (13) holds for all integers smaller than k, then

$$P\{X_k = s, T \leq k-1\} = \alpha_{k-1}.$$

This shows T is strong uniform. □

EXERCISE 3. Prove that the strong uniform time T^* constructed in the course of proving Theorem 4 is the stochastically fastest strong uniform: $P\{T^* > k\} \leq P\{T > k\}$ for all k. Now consider Example 1 (top in at random). The stopping time defined there can be improved: consider T^* — the first time that the card originally *second* from the bottom comes up to the top. Show that T^* is a fastest strong uniform time.

EXERCISE 4. As an example of Theorem 4, consider the model for random walk on the d-cube treated in Section B. The cutoff point for variation distance is $\frac{1}{2}\, d \log d$, and the stopping time argument gives d log d. Show that this is sharp: it takes $d \log d + cd$ steps to have a reasonable probability of reaching the vertex opposite $\underline{0}$, namely $(1 \ldots 1)$. Hint: try Fourier analysis.

The following result, proved in Aldous and Diaconis (1987) shows that the factor of 2 found above is no accident. Roughly, if the variation distance becomes small after k steps, the separation becomes small after at most $2k$ steps. To make this precise, let $\phi(\varepsilon) = 1 - (1 - 2\varepsilon^{\frac{1}{2}})(1 - \varepsilon^{\frac{1}{2}})^2$. Observe that $\phi(\varepsilon)$ decreases with ε and $\phi(\varepsilon) \sim 4\varepsilon^{\frac{1}{2}}$ as $\varepsilon \to 0$.

Theorem 5. *For any probability Q on any finite group G, and all $k \geq 1$,*

$$s(2k) \leq \phi(2||Q^{*k} - U||) \text{ provided } ||Q^{*k} - U|| < \frac{1}{8}.$$

Further discussion of separation can be found in Aldous and Diaconis (1986, 1987) or Diaconis and Fill (1988).

D. AN ANALYSIS OF REAL RIFFLE SHUFFLES.

How many ordinary riffle shuffles are required to bring a deck of cards close to random? We will show that the answer is 7. The discussion proceeds in two sections: (1) practical discussion and data analysis, (2) a model for riffle shuffling.

(1) *Practical shuffling.* Of course, people shuffle cards all the time for card games. We begin by asking "Does it matter?" That is, even if people don't shuffle really well, will it make any practical difference? One answer to this question is in Berger (1973). Berger uses the fact that tournament bridge went from hand shuffling to computer shuffling in the late 1960's. Berger obtained records of the suit distribution of the south hand in 2000 deals, one thousand before the computer, one thousand after the computer. A summary is in Table 1.

Inspection of the table shows that hands with an even suit distribution occur with higher than the expected frequency in hand shuffling. A chi-squared test rejects uniformity of the suit distribution in hand shuffling. Uniformity is accepted for computer shuffling. Something is going on that *does* make a practical, observable difference. Here is a first explanation: The way bridge tends to be played, cards are collected in groups of 4, by suit. If the riffle shuffling was clumpy and clustery, cards of the same suit would tend to clump together and then, when the deck was dealt into 4 hands, tend to be separated.

Table 1
Frequency of Computer-dealt Hands Versus Theoretical
Expected Frequencies from Berger (1973)

Distribution of the 4 suits	Expected Frequencies	Actual Frequencies of Computer-dealt Hands	Actual Frequencies of Man-dealt Hands
4,4,3,2*	216	198	241
5,3,3,2	155	160	172
5,4,3,1	129	116	124
5,4,2,2	106	92	105
4,3,3,3	105	103	129
6,3,2,2	56	64	46
6,4,2,1	47	53	36
6,3,3,1	34	40	41
5,5,2,1	32	40	19
4,4,4,1	30	35	25
7,3,2,1 and others	90	99	62
	1,000	1,000	1,000

* by "4,4,3,2" we mean that the thirteen cards contained 4 cards in one suit, 4 cards in another suit, 3 cards in another suit, and 2 cards in the remaining suit.

This would make "even splits" like 4 3 3 3 occur more often than they should. One objection to this is that the cards in duplicate bridge are not usually collected in groups of 4 (they are in non-duplicate games). In duplicate, the cards are collected into 4 piles of 13, each pile being roughly in the same suit order. If these piles were placed on top of one another and riffle shuffled twice, the cards would

tend to clump in suit groups of 4 and we are back to the previous case.

Ely Culbertson (1934) discusses ways of taking advantage of poor shuffling in Bridge. Thorp (1973) discusses other card games.

A second practical view comes from considering the results of a single shuffle. An example is presented in Table 2 below. This records a shuffle for which the deck was cut 1 through 29, and 30 through 52. Card 29 was dropped first, then 28, then 52, then, ..., then 1, finally 30.

Table 2
A Single Riffle Shuffle

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
$\pi(i)$	2	3	5	7	9	11	13	14	16	18	20	22	24	26	27	29	31	33	35	36	38	39	41	42	45	46
	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
	48	51	52	1	4	6	8	10	12	15	17	19	21	23	25	28	30	32	34	37	40	43	44	47	49	50

A single riffle shuffle can have at most 2 rising sequences. There are $2^n - n$ possible arrangements of n cards after a single riffle shuffle. Similarly, there are at most 4 rising sequences after 2 riffle shuffles. This generalizes:

Theorem 6 (Shannon).
(1) Let π be a permutation with R rising sequences. π is the outcome of k riffle shuffles if and only if $R \leq 2^k$.
(2) Each π with exactly 2^k rising sequences can be obtained by k riffle shuffles in only one way.

This theorem appears in E. Gilbert (1955), "Theory of Shuffling," Bell Laboratories Technical Memorandum. Part (1) has been used as the basis of a card trick for many years. In this trick, a deck of cards is mailed to a spectator who is instructed to riffle shuffle the deck 3 times, giving the deck any number of straight cuts during the shuffling. Then the top card is removed, noted, and placed into the center of the pack. This is followed by more cuts. The pack is mailed back to the magician who unerringly finds the card. The secret is that there will be eight rising sequences and 1 card in its own rising sequence. It is not hard to show that a random permutation has about $\frac{n}{2}$ rising sequences so a few shuffles (on order $\log_2 \frac{n}{2}$) will not be enough to randomize n cards.

These arguments yield a lower bound. In a bit more generaltiy, if P is a probability on a finite group G supported and uniform on the set $A \subset G$, then

$$||P - U|| \geq P(A) - U(A) = 1 - \frac{|A|}{|G|}.$$

This can be combined with the observations on rising sequences to give a lower bound that works for a few shuffles. Let $F_n(R)$ be the number of permutations of n items with exactly R rising sequences. Thus $F_n(1) = 1$ and $F_n(2) = 2^n - (n+1)$. A formula for $F_n(R)$ is derived in Sade (1949); see also Riordan (1950):

$$F_n(R) = \sum_{j=0}^{R} (-1)^j \binom{n+1}{j} (R-j)^n.$$

In k shuffles, the total number of permutations that can be achieved is $T_n(k) = \sum_{R=1}^{2^k} F_n(R)$. Thus $1 - T_n(k)/n!$ is a lower bound for the variation distance. For $n = 52$, the lower bound is larger than .99 for $1 \leq k \leq 4$. For $k = 5$ it is .38, for $k = 6$ it is zero.

Observe that this approach makes *no* assumptions about the stochastic mechanism for shuffling, but the argument breaks down at 6 shuffles.

(2) *A probability model.* The following model for riffle shuffling was suggested by Shannon and Gilbert, and Reeds.

1st description: Cut the n card deck according to a binomial distribution with parameters $\frac{1}{2}, n$. Suppose k cards are cut off. Pick one of the $\binom{n}{k}$ possible riffle shuffles uniformly.

2nd description: Cut the n card deck according to a binomial distribution with parameters $\frac{1}{2}, n$. Suppose k cards are cut off and held in the left hand and $n - k$ held in the right hand. Drop cards with probability proportional to packet size. Thus the chance that a card is dropped first from the left hand is $\frac{k}{n}$. If this happens, the chance that the left hand drops a second card is $\frac{k-1}{n-1}$; and so on.

3rd description: To generate the inverse shuffle, label the back of each card with the result of an independent fair coin flip: $\{0, 1\}$. Remove all cards labeled 0 and place them on top of the deck, keeping the cards otherwise in the same relative order.

LEMMA 2. *The three descriptions yield the same probability distribution.*

Proof. The 1st and 3rd descriptions are equivalent: indeed, the binary labeling chooses a binomial number of zeros and conditional on this choice, all possible placements of the zeros are equally likely. The lst and 2nd descriptions are equivalent: Suppose k cards have been cut off. Under the 2nd description, the chance of a shuffle is the chance of the sequence of drops $D_1, D_2, \ldots, D_n$, where each D_i can be L or R and k D_i's must be L and $n - k$ D_i's must be R. The chance of any such sequence is $k!(n-k)!/n!$. □

Remarks. This shuffling mechanism has some claim to being the "most random" subject to the binomial cutting. It has the largest entropy, for example. As a model for shuffling, it yields shuffles a bit "clumpier" than either the shuffles of Diaconis or Reeds discussed in remark (e) below. Only half the packets are expected to be of size 1, a quarter of size 2, etc. Of course, extremely neat shuffles are not necessarily good for randomization. A perfect shuffle is completely non-random for example, eight perfect shuffles bring the deck back to order. See Diaconis, Graham, and Kantor (1983). Mellish (1973) discusses these issues.

To proceed further, we construct a strong uniform time for this model of shuffling. To begin, observe that the variation distance is invariant under 1-1 transformations, so it is the same problem to bound the number of inverse shuffles required to get close to random.

The results of repeated inverse shuffles of n cards can be recorded by forming

a binary matrix with n rows. The first column records the zeros and ones that determine the first shuffle, and so on. The ith row of the matrix is associated to the ith card in the original ordering of the deck, recording in coordinate j the behavior of this card on the jth shuffle.

LEMMA 3 *(Reeds). Let T be the first time that the binary matrix formed from inverse shuffling has distinct rows. Then T is a strong uniform time.*

Proof. The matrix can be considered as formed by flipping a fair coin to fill out the i, j entry. At every stage, the rows are independent binary vectors. The joint distribution of the rows, conditional on being all distinct, is invariant under permutations.

After the first inverse shuffle, all cards associated to binary vectors starting with 0 are above cards with binary vectors starting with 1. After two shuffles, cards associated with binary vectors starting (0,0) are on top followed by cards associated to vectors beginning (1,0), followed by (0,1), followed by (1,1) at the bottom of the deck.

Inductively, the inverse shuffles sort the binary vectors starting with 0 are above cards with binary vectors starting with 1. After two shuffles, cards associated with binary vectors starting (0,0) are on top followed by cards associated to vectors beginning (1,0), followed by (0,1), followed by (1,1) at the bottom of the deck.

Inductively, the inverse shuffles sort the binary vectors (from right to left) in lexographic order. At time T the vectors are all distinct, and all sorted. By permutation invariance, any of the n cards is equally likely to have been associated with the smallest row of the matrix (and so be on top). Similarly, at time T, all $n!$ orders are equally likely. □

To complete this analysis, the chance that $T > k$ must be computed. This is simply the probability that if n balls are dropped into 2^k boxes there are not two or more balls in a box. If the balls are thought of as people, and the boxes as birthdays, we have the familiar question of the birthday problem and its well known answer. This yields:

Theorem 7. *For Q the Gilbert-Shannon-Reeds distribution defined in Lemma 2,*

$$||Q^{*k} - U|| \leq P\{T > k\} = 1 - \prod_{i=1}^{n-1}(1 - \frac{i}{2^k}). \tag{14}$$

Standard calculus shows that if $k = 2\log_2(n/c)$,

$$P\{T > k\} \underset{\infty}{\overset{n}{\approx}} 1 - e^{-\frac{c^2}{2}} \underset{0}{\overset{c}{\approx}} \frac{c^2}{2}.$$

In this sense, 2 log n is the cut off point for this bound. Exact computation of the right side of (14) when $n = 52$ gives the bounds

k	upper bound
10	.73
11	.48
12	.28
13	.15
14	.08

Remark (a). The lovely new idea here is to consider shuffling as inverse sorting. The argument works for any symmetric method of labelling the cards. For example, biased cuts can be modeled by flipping an unfair coin. To model cutting off exactly j cards each time, fill the columns of the matrix with the results of n draws without replacement from an urn containing j balls labelled zero and $n-j$ balls labelled one. The first time all vectors are different is a strong uniform time. These lead to slightly unorthodox birthday problems which turn out to be easy to work with.

Observe that the shuffle in which only 1 card is cut off and randomly riffled into the deck is the "top in at random" shuffle of example 1. The two stopping times are the same!

Remark (b). The argument can be refined. Suppose shuffling is stopped slightly before all rows of the matrix are distinct — e.g., stop after 2 log n shuffles. Cards associated to identical binary rows correspond to cards in their original relative positions. It is possible to bound how far such permutations are from uniform and get bounds on $||Q^{*k} - U||$. Reeds (1981) has used such arguments to show that 9 or fewer shuffles make the variation distance small for 52 cards.

Remark (c). A variety of ad hoc techniques have been used to get lower bounds. One simple method that works well is simply to follow the top card after repeated shuffles. This executes a Markov chain on n states with a simple transition matrix. For n in the range of real deck sizes, $n \times n$ matrices can be numerically multiplied and then the variation distance to uniform computed. Reeds (1981) has carried this out for decks of size 52 and shown that $||Q^{*6} - U|| \geq .1$. Techniques which allow asymptotic verification that $k = 3/2 \log_2 n$ is the right cutoff for large n are described in Aldous (1983). These analyses and the results quoted above suggest that seven riffle shuffles are needed to get close to random.

Remark (d). Other mathematical models for riffle shuffling are suggested in Donner and Uppuluri (1970), Epstein (1977), and Thorp (1973). Borel and Cheron (1955) and Kosambi and Rao (1958) discuss the problem in a less formal way. Where conclusions are drawn, 6 to 7 shuffles are recommended to randomize 52 cards.

Remark (e) some data analysis. Of course, our ability to shuffle cards depends on practice and agility. The model produces shuffles with single cards being dropped about 1/2 of the time, pairs of cards being dropped about 1/4 of the time, and i card blocks being dropped about $1/2^i$ of the time.

To get a feeling for the difference between shufflers, the following experiment was performed: Diaconis and Reeds each shuffled a deck of 52 cards about 100

times; every permutation was recorded. The following summary statistics are relevant.

Diaconis — 103 Shuffles

# cut off top	23	24	25	26	27	28	29
	2	4	22	32	33	9	1

In shuffling, the left hand dropped first 44 times. In all there were 4,376 "packets" dropped. The counts and proportions were

	1	2	3	4	5
	3501	793	63	15	4
	.80	.18	.01	.00	.00

The packet size distribution of first dropped packets was

	1	2	3	4	5
	.37	.37	.17	.1	.01

Reeds — 100 Shuffles

# cut off top	23	24	25	26	27	28	29	30	31
	2	2	8	16	23	26	16	5	2

In shuffling, the left hand dropped first 18 times. In all there were 3,375 "packets" dropped. The counts and proportions were

	1	2	3	4	5	6	7	8	9	10	11	12	13
	2102	931	228	68	24	12	3	3	2	0	0	1	1
	.62	.28	.07	.02	.01	.00	.00	.00	.00	.00	.00	.00	.00

Diaconis does very neat shuffles and can be compared to a Las Vegas dealer. Reeds shuffles like an "ordinary person." Observe that the first drops for Diaconis are quite different from the average drop. Even though the two types of shufflers are fairly different, to a first approximation they are quite similar, both dropping 1, 2, or 3 cards most of the time.

Remark (f). There is another equivalent way to describe repeated riffle shuffles under the Gilbert-Shannon-Reeds model that suggests much further research. The following evolved in conversations with Izzy Katznelson and Jim Reeds. Begin by dropping n points at random into the unit interval and labeling them, left to right, as $1, 2, \ldots, n$. The transformation $T(x) = 2x (\mathrm{mod}\ 1)$ (sometimes called the Baker's transformation) maps the unit interval into itself and so permutes the points. T takes each of the two half intervals and stretches it out to cover $[0, 1]$. There are a binomial number of points in each half and T shuffles them together. Arguing as in Lemma 1 above, it is easy to see that the induced permutation is precisely a basic riffle shuffle. Further, successive shuffles are independent (they depend on successive bits of the underlying uniform variables).

To complete the argument, consider k chosen so large that n points dropped at random into $[0, 1]$ fall into disjoint pieces of a partition with pieces of length $1/2^k$, with high probability. Picture a point in a piece of the partition. After k shuffles, the piece is stretched out to cover $[0, 1]$. The point is randomly distributed in the piece of the partition. After k shuffles it's randomly distributed in $[0, 1]$. Further, points in disjoint pieces are independent. After k shuffles, the n points are in random relative arrangement (given that they fall into disjoint pieces).

This argument generalizes to some extent ($x \to k_j x (\text{mod } 1)$ on shuffle j for integer k_j). It should be possible to take other measure preserving transformations (toral endomorphisms of the unit square (see Walters (1982)) and convert them to other shuffles.

E. Coupling.

There is a more widely known purely probabilistic argument called coupling. Again, perhaps it is best to begin with an example, this one is due to David Aldous.

Example 1-Borel's shuffle. Borel and Cheron (1955) discuss several methods of mixing up cards. They give the following as an open problem: begin with n cards. Take the top card off and put it into the deck at a random position. Take the bottom card off and put it into the deck at a random position. Continue alternately placing top in at random, bottom in at random.

Observe that there is no longer an obvious stopping time. The following elegant coupling argument has been suggested by David Aldous. It is better to work with the inverse shuffle that removes a random card and places it alternately on top or bottom. Because of the invariance of variation distance under 1-1 maps ($||Q - U|| = ||Qh^{-1} - Uh^{-1}||$ for any 1-1 function $G \to G$) the two shuffles have the same rates of convergence (see exercise 3 of Chapter 3).

To describe a "coupling", consider a second deck of cards. The first deck starts in order $\{1, 2, 3, \ldots, n\}$. The second deck starts in a random order. A card is determined at each stage by shuffling a *third* deck and choosing a card at random. Say the first card chosen is the six of hearts. Remove the six from deck 1 and place it on top. Remove the six from deck two and place it on top. Note that from each deck's marginal vantage point, a card was removed at random and placed on top.

The second step is to reshuffle the 3rd deck and choose a second card, say the Ace of spades. This is removed and placed at the bottom of each deck. Continue in this way, each time choosing a card at random from the third deck, removing the card from decks one and two, and placing the card alternately on top and bottom.

As this process continues, decks one and two match up. The same cards being in the same order at top and bottom. If the same card is chosen again, the procedure keeps the same number of matches. A new match is created for each new card touched. Let T be the first time that each card has been touched. Clearly, the two decks are in the same order, but deck two started at random, and so remains random. It follows that deck one is random at time T. The bound on the coupon collector's problem yields

Theorem 8. *For Borel's shuffle, if* $k = n \log n + cn$ *for* $c > 0$,

$$||P^{*k} - U|| \leq e^{-c}.$$

EXERCISE 5. Find a strong uniform time to get a bound in Borel's problem.

To discuss coupling more carefully, we need the following fact about variation distance:

LEMMA 4. *Let S be a finite set. Let P_1 and P_2 be probability measures on S. Let Q be a probability on $S \times S$ with margins P_1 and P_2. Let Δ be the diagonal: $\Delta = \{(s,s): s \in S\}$ then,*

$$||P_1 - P_2|| \leq Q(\Delta^c).$$

Proof.

$$\begin{aligned} |P_1(A) - P_2(A)| &= |Q(A \times S) - Q(S \times A)| \\ &= |Q(A \times S \cap \Delta) + Q(A \times S \cap \Delta^c) - Q(S \times A \cap \Delta) \\ &\qquad - Q(S \times A \cap \Delta^c)|. \end{aligned}$$

The first and third numbers in the absolute value sign are equal. The second and fourth give a difference between two numbers, both non-negative and smaller than $Q(\Delta^c)$. □

Remarks. The inequality is sharp in the sense that there is a Q which achieves equality. A proof and discussion may be found in V. Strassen (1965). This Q allows the following interpretation of variation distance: $||P_1 - P_2|| = \varepsilon$ if and only if there are two random variables, X_1 distributed as P_1 and X_2 distributed as P_2, such that $X_1 = X_2$ with probability $1 - \varepsilon$; X_1 and X_2 may be arbitrarily different with probability ε. Another interpretation: the optimal Q is most concentrated about the diagonal with these fixed margins.

Let us define a coupling for a Markov chain on state space I, with transition probability $P_i(j)$, and stationary distribution π. We will work with Markovian couplings. These are processes on $I \times I$ with transition probability Q satisfying

$$\sum_t Q_{i,j}(s,t) = P_i(s) \text{ for all } j$$

$$\sum_s Q_{i,j}(s,t) = P_j(t) \text{ for all } i.$$

These conditions just say that the transition mechanism of each component of the vector process is $P_i(j)$. Call the vector process (X_k^1, X_k^2). Suppose that X^1 starts in i and X^2 starts according to π. Let $T = \min\{k: X_k^1 = X_k^2\}$. This T is a stopping time. Suppose that T is finite with probability 1. Let

$$X_k^3 = \begin{cases} X_k^2 & k \leq T \\ X_k^1 & k > T. \end{cases}$$

The process $(X_k^1,\ X_k^3)$ is called a coupling, the interpretation being that the two processes evolve until they are equal, at which time they couple, and thereafter

remain equal. The usefulness of coupling depends on being able to get our hands on T: Let P_i^k be the law of the process after k steps started from i.

LEMMA 5. *(Coupling inequality).* $||P_i^k - \pi|| \leq P(T > k)$.

Proof. Take Q to be the distribution of (X_k^1, X_k^3). This Q has marginal distributions $P_i^k(\ )$ and π. Lemma 4 implies that

$$||P_i^k - \pi|| \leq Q(\Delta^c) = P(T > k).$$

□

Remarks. It is instructive to note that while the distribution of X_T is stationary (and so uniform in our examples) the time T is *not* a strong uniform time as we have defined it; for this requires $P(X_k \in A | T = k) = U(A)$ for all k.

Remarks. Example 1 gives an actual construction of $Q_{ij}(k\ell)$. It might be instructive to write down what $Q_{ij}(k\ell)$ is for this example. Griffeath (1975), Pitman (1976) or Goldstein (1979) show that the argument is tight in the sense that there is a coupling that achieves the total variation distance. This coupling cannot be taken as Markovian in general (that is, the bivariate process needn't be Markov).

Example 2. Random walk on the d-cube. Here $G = Z_2^d$. Take

$$P(\underline{0}) = p,\ P(1\ 0 \ldots 0) = P(0\ 1\ 00 \ldots 0) \ldots = P(0 \ldots 1) = (1-p)/d.$$

Here is a coupling argument, due to David Aldous, for bounding convergence to uniform. Consider two cubes. The process X_0^1 starts at zero, X_0^2 starts in a uniformly distributed position. The pair (X_i^1, X_i^2) evolves as follows: if X_i^1 and X_i^2 differ in an odd number of places, the two processes take independent steps according to P. If X_i^1 and X_i^2 differ in an even number of places, then with probability p each remains unchanged. If they don't stay the same, then pick an index j at random in $\{1, 2, \ldots, d\}$. If the jth component of X_i^1 and X_i^2 agree, change that component to its opposite (mod 2) in both processes. If the jth component of X_i^1 and X_i^2 do not agree, complement the jth component of X_i^1 and the *next* non-agreeing component of X_i^2 from j counting cyclically. This forces X_i^1 and X_i^2 to agree in two more coordinates. Once the number of disagreeing places is even, it stays even, so the coupled process "gets together" very rapidly. Of course, once $X_i^1 = X_i^2$, they stay coupled.

EXERCISE 6. Analyze the coupling time T and get a bound on the rate of convergence for Example 2 . Compare this with the right rate derived from Fourier analysis.

Matthews (1986b) has constructed non-Markovian couplings that give the right rate of convergence for the cube.

Coupling is a very widely used tool which has many success stories to its credit. Aldous (1983a) gives a number of card shuffling examples. Robin Pemantle

(1989) has given a marvelous coupling analysis of the familiar over-hand shuffle. For a range of reasonable models he shows that order n^2 shuffles are required to mix up n cards. Thus about 2,500 shuffles are required for 52 cards. This should be compared with 7 or 8 riffle shuffles, and the computation that a single riffle and single over-hand shuffle produce the same number of distinct permutations.

Aldous and Diaconis (1987a) and Thorisson (1987) study the relation between coupling and strong uniform times. Briefly, for any strong uniform time there is a coupling with the same time. Thus couplings can occur faster in principle.

Theorem 5 of Section C shows that couplings can only speed things up by a factor of at most 2. The example of simple random walk on the cube shows that this actually happens: it takes $\frac{1}{4}$ n log n $+cn$ steps to make the variation distance small; $\frac{1}{2}$ n log n $+cn$ steps are needed to make the separation small.

Despite the similarities, the connection is fairly formal. The way of thinking, and basic examples, can be very different. There is no known direct coupling argument to get anything better than n^2 for random transpositions, while strong uniform times or Fourier analysis show the right rate is order n log n. Similarly, there is no strong uniform time for the over-hand shuffle, or the shuffle that picks a card at random and switches it with a neighbor. Coupling can handle these problems.

F. First hits and first time to cover all.

(1) *Introduction.* Most of the work in this and the previous chapter has been devoted to estimating rates of convergence to uniformity. There are many other natural questions connectd to random walk. One may ask

- How long does it take to hit a fixed state (or set of states) from a given (or random) start?
- How long does it take to hit every state?
- How long until the first return to the starting state? How far away is the walk likely to get before first returning? How many states does the walk hit before first returning? What is the maximum number of times any state has been visited at first return?
- How long does a walk take before it hits a point previously hit (the birthday problem for random walk)?

David Aldous has introduced an important heuristic which suggests and explains answers to such questions, and sometimes allows a proof using only bounds on convergence to stationarity.

The idea is as follows. Suppose a specific random walk on a group G is rapidly mixing in the sense that the variation distance is less than $\frac{1}{2}$ after k steps with log k of order a polynomial in log$|G|$. Then, the random walk forgets where it is rapidly, and successive steps may be thought about as the position of balls, dropped at random, into $|G|$ boxes.

Questions about balls in boxes are well understood. For example, the mean waiting time T until a ball is dropped into a fixed box is $|G|$ and

$$P\{\frac{T}{|G|} > t\} \to e^{-t} \text{ as } G \to \infty.$$

This suggests that a rapidly mixing random walk takes about $|G|$ steps to hit a fixed point and the waiting time is approximately exponential. A precise version is given in (2) below.

As a second example, the waiting time V for all boxes to have at least one ball is well studied as the coupon collector's problem. For balls dropped at random into $|G|$ boxes, it takes about $|G| \log|G|$ balls to have a good chance of filling all boxes. Results in Feller (1968, pg. 106) yield

$$P\{\frac{V - |G| \log|G|}{|G|} \leq x\} \to e^{-e^{-x}} \text{ as } |G| \to \infty.$$

This suggests that a rapidly mixing random walk takes about $|G| \log|G|$ steps to cover all points. (3) below gives some precise results due to Aldous and Matthews.

Section 4 points to what little is known about other problems on the list above.

(2) *First hit distributions.* The heuristics above are right "up to constants." One remarkable finding of Aldous (1982, 1983b) is that only one other feature of the walk enters. This is a measure of the amount of time the walk spends in its starting state in a short time period. Consider throughout a random walk on a finite group G. The transition mechanism is assumed to be aperiodic, and the uniform distribution on G is the stationary distribution.

Standard renewal theory implies that $R(s;t)$, the amount of time the walk spends in a fixed state s up to time t, is asymptotically $t/|G|$. Moreover

$$R = \lim_{t\to\infty} E\{R(s;t)\} - t/|G|$$

exists and is finite. By homogeneity R doesn't depend on s. Aldous (1983b) argues that for rapidly mixing walks, R can be interpreted as the mean number of visits the random walk spends in its initial state in a short time. For most of the examples in this and the previous chapter, $R = 1$.

With this notation, some careful results can be stated.

Theorem 9. *(Aldous). Let T_s be the first time a random walk starting in a uniformly chosen position hits state s. Then*

$$E(T_s) = R|G| \text{ for all } s \in G. \tag{1}$$

Let $\tau = \inf_k\{||P^{*k} - U|| \leq 1/2e\}$. Then

$$\sup_{t\geq 0} |P\{T_s > t\} - e^{-t/R|G|}| \leq \psi(\tau/R|G|), \tag{2}$$

with $\psi(x)$ tending monotonically to zero as x tends to 0.

Remarks. Part (2) makes precise the heuristics of the previous section. Consider a process like simple random walk on the d-cube. Then $|G| = 2^d$, and $\tau \doteq \frac{1}{4}d \log d$, $R = 1$, and (1) and (2) recapture the limiting results derived in Chapter

3H. Aldous (1983b) gives similar results for the first hitting time to arbitrary sets with any starting distribution.

Most random walks considered above have $\tau/|G| \to 0$. An exception is simple random walk on Z_n, where τ is of order n^2. The wait to first hit a point has a rather complicated distribution (see Chapter 3H). Flatto, Odlyzko, and Wales (1985) use Fourier analytic methods to get higher order correction terms.

(3) *Time to cover all.* Let G be a finite group and P a probability with convolutions that converge to uniform aperiodically. Let V be the first time that a random walk hits every point in G. Note that the distribution of V doesn't depend on the starting state. Let τ and R be as defined in Section 2. Aldous (1983a) proves

Theorem 10.

$$E\left|\frac{V}{R|G|\ log|G|} - 1\right| \leq \psi\left(\frac{log(1+\tau)}{log|G|}\right)$$

with $\psi(x)$ tending monotonically to zero as x tends to 0.

Remark. In the case of the cube, $\log(1+\tau) \sim \log$ d, $\log|G| \sim$ d log 2, so the ratio tends to zero. Analogs of the extreme value limits for the coupon collector's problem are not established in this generality. However, Matthews (1985) has established limit theorems for many of the examples where Fourier analysis can be successfully applied.

Usually the results follow the heuristic. For the cube there is an extra factor of 2. Matthews shows

$$P\left\{\frac{V - 2^n \log 2^{n+1}}{2^n} \leq x\right\} \to e^{-e^{-x}}$$

for all fixed x as n tends to infinity. Here $R \sim 1 + \frac{1}{n}$ which explains the 2.

Matthews' argument works by getting upper and lower bounds on the required probability. These apply to problems like first time for a Markov chain to hit every point in a finite state space or first time for Brownian motion to come within ε of every point on a high-dimensional sphere. The bounds merge as $|G| \to \infty$ for random walk problems.

(4) *Other problems.* There has been some work on special cases of the problems listed in (1) above. Aldous (1985) started to classify the kind of limiting behavior that can occur in the birthday problem for random walk. Diaconis and Smith (1988) have begun to develop a fluctuation theory (as in Chapter 3 of Feller (1968)). Some neat results emerge for nearest neighbor random walk on a 2-point homogeneous space. For example, on the n-cube, the probability that random walk starting at $(00\ldots0)$ hits a given point at any specified distance less than n before returning to zero tends to 1 as n tends to ∞. The probability tends to 1/2 for $(1,\ldots,1)$.

This seems like a rich collection of reasonably tractable problems. Passing to the limit should give results for the approximating diffusions (e.g., Ornstein-Uhlenbeck process for the cube) in much the same way as results about simple random walk lead to results for Brownian motion.

G. Some open problems on random walk and strong uniform times.

Here is a small list of problems that seem worth careful work.

(1) *The slowest shuffle.* Arunas Rudvalis has suggested the following candidate for the slowest shuffle: At each time, the top card is placed either at the bottom, or second from the bottom, each with chance $\frac{1}{2}$. How long does it take to get random? Is this the slowest shuffle equally supported at 2 generating permutations?

(2) Let $G = Z_n$. Pick k points in G, and repeatedly choose one of them at random. This determines a random walk. What are the slowest k points (given no paritiy problems) — a "arc" near zero? (i.e. the set of points j with $|j| < k/2$.) What are the fastest k points? Andy Greenhalgh has shown how to get rate $n^{1/k}$ by an appropriate choice. What's the rate for "most" sets of k points? These questions are already non-trivial for $k = 3$. They are also worth studying when k grows with n.

(3) Moving on to other groups, Aldous and Diaconis showed that for most measures P on a finite group G, $||P * P - U|| \leq \frac{1}{|G|}$, so for G large, most measures are random after two steps. To get an interesting theory, constraints must be put on the support. Andre Broder asked the following: pick a pair of elements in S_n. Consider the walk generated by choosing both of these elements at random. It can be shown that such a pair generates S_n with probability 3/4 asymptotically. Is the walk random after a polynomial number of steps? Similar problems are worth investigating for any of the classical infinite families of finite simple groups (I'd try $PGL_n(q)$). Back on S_n; it seems that any "reasonable" shuffle gets random in at most a polynomial number of steps.

(4) *The 15 puzzle.* This familiar puzzle has 15 blocks arranged in a 4×4 grid. At each state, any of the blocks can be slid into the blank. Suppose uniform choices are made among the current possibilities.

Here is a simplified version: Consider the blank as a 16th block, and consider the puzzle on a "torus." An allowable move now involves picking one of the 16 squares at random, and then a direction (North, South, East, West) and "cycling" that square in the chosen direction. For example, the bottom row might change from 13, 14, 15, 16 to 16, 13, 14, 15 or to 14, 15, 16, 13. It is not hard to show that it takes order n^3 steps to randomize a single square (on an $n \times n$ grid). I presume that order n^3 log n steps suffice to randomize everything. For a 4×4, this gives about 90 "moves" to randomize. I presume this simplified version converges to uniform faster than the original 15 puzzle. Similar questions can be asked for other puzzles such as Rubic's cube.

(5) *The affine group.* Consider random walks of form $X_n = a_n X_{n-1} + b_n (\text{mod } p)$. Here p is a fixed number (perhaps a prime) and (a_n, b_n) are chosen at random: e.g., $a_n = 2$ or $\frac{1}{2} (\text{mod } p), b_n = \pm 1$. It seems that the right answer for these is $(\log p)^a$ for $a = 1$ or 2. The best that has been proved at present is order p^2 (see Diaconis and Shahshahani (1986a).

(6) *Thorp's shuffle.* A simple model for a random riffle shuffle has been described by Thorp (1973). Cut the deck exactly in half. Start to drop the cards from left or right hand as with an ordinary shuffle. At each time, choose left or

right with chance $\frac{1}{2}$. Whatever is chosen, drop that card, and then a card from the opposite half. Continue inductively. I think use of the mathematics of shuffle nets (or work on sorting in parallel) will allow an elegant solution to this problem.

(7) *Continuous groups.* We have no examples of a strong uniform time argument being used to get rates of convergence for a random walk on a compact, infinite group. It may be necessary to change the distance to the Prohorov metric. For problems like random reflections (see Diaconis and Shahshahani (1968a)) or random walk on the circle determined by repeatedly choosing a point in a small arc uniformly, there is convergence in total variation metric.

(8) *The cutoff phenomenon.* The most striking finding is the existence of sharp phase transition, $||P^{*k} - U||$ cutting down from 1 to zero in a relatively short time. It would be great to understand if this usually happens. As explained in problem (3) above, restrictions will have to be put on the support.

(9) *Relations between various approaches.* A curious feature of the examples is that *usually* if one method of attack works (e. g., Fourier analysis, or coupling, or strong uniform times), then all the methods work. There must be a reason. The greatest mystery is to understand the connections between the analytic and probabilistic methods. One place to start is "top in at random," the first example of Chapter 4. This can be done by strong uniform times and coupling. There *must* be a way to do it Fourier analytically.

Chapter 5. Examples of Data on Permutations and Homogeneous Spaces

To fix ideas, as well as to make contact with reality, it is useful to have a collection of real data sets on hand.

A. PERMUTATION DATA.

(1) Large sets of rankings are sometimes generated in psychophysical experiments (rank these sounds for loudness), taste testing experiments (rank these 5 types of coffee ice cream), or surveys. To give an example, in 1972, the National Opinion Research Center included the following question in one of their surveys: Where do you want to live? Rank the following 3 options: in a big city; near a big city ($\leq$ 50 miles); far from a big city ($>$ 50 miles). The data from 1439 respondents was

city	suburbs	country	#
1	2	3	242
1	3	2	28
2	1	3	170
3	1	2	628
2	3	1	12
3	2	1	359

Let us briefly discuss this data. The modal rank is $\left(\begin{smallmatrix}1&2&3\\3&1&2\end{smallmatrix}\right)$ — people prefer the suburbs, then country, then city. This is born out by simple averages: 270 people ranked city first, 798 ranked suburb first, 371 ranked country first.

The 2 small counts lead to an interesting interpretation. Both violate the unfolding hypothesis of Coombs (1964). To spell this out a bit, suppose people's rankings are chosen in accordance with the ideal distance from the city, different people having different preferences. Thus, one chooses the rank one location and then "unfolds" around it. In this model $\left(\begin{smallmatrix}1&2&3\\1&3&2\end{smallmatrix}\right)$ is impossible since if one most prefers being in the city, one must prefer being close to the city to being far away. The number of permutations of the set $1, 2, \ldots, n$ consistent with unfolding is about 2^{n-1}, so many arrangements are ruled out. Unfolding is a nice idea, but distance to the city might not determine things for someone who works in the suburbs and doesn't want to live where they work. If you ask people to rank order temperature for tea (hot, medium, cold), you don't expect the unfolding restriction to hold, but if you ask people to rank order sugar teaspoons (0, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2) you do expect the data to be consistent with unfolding.

Further analysis of the distance to cities data is in Chapter 8. Duncan and Brody (1982) discuss these data in some detail.

Ranked data often comes with other variables — rankings for men and women, or by income being examples. In the data on distance to cities, the actual dwelling place of the respondent is available. Methods for dealing with covariates are developed in Chapter 9.

It is worth pointing to a common problem not represented in the cities data. Because n! grows so rapidly, one can have a fairly large data set of rankings and still only have a small proportion of the possible orders represented. For example, I am considering a data set in which 129 black students and 98 white students were asked to rank "score, instrument, solo, benediction, suite" from the least related to "song" to the most strongly related to "song." Here, there cannot be very many repeats in each ranking. In another data set, quoted in Feigin and Cohen (1978), 148 people ranked 10 occupations for desirability. Clearly, the ratio of the sample size to n! has a limiting effect on what kind of models can be fit to the data.

(2) Pairs of permutations often arise as in "rank order the class on the midterm and final." Similarly, small sets of rankings arise as in a panel of judges ranking a set of contestants. A large collection of examples appears in Chapter 7A.

(3) *The Draft Lottery.* In 1970, a single "random" permutation in S_{365} was chosen. This permutation was used to fix the order of induction into the army. The actual permutation is shown in Table 1. For discussion of this data set, see the article by S. E. Fienberg (1971).

As Fienberg reports, it was widely claimed that the permutation tended to have lower order months Jan., Feb., ... having higher numbers. The Spearman rank correlation coefficient is -.226, significant at the .001 level. Figure 2, based on Figure 1, shows the average lottery number by month. The evidence seems strong until we reflect on the problems of pattern finding in a single data source after agressive data analysis.

Further analysis of this data is given in example 1 of Chapter 7A.

B. Partially ranked data.

There are numerous examples in which people rank a long list only partially. For example, people might be asked to rank their favorite 10 out of 40 movies, a typical ranking yielding $(a_1, a_2, \ldots, a_{10})$ with a_1 the name of the movie ranked first, etc. Alternatively people might be asked to choose a committee of 10 out of 40, not ranking within. Then a typical selection yields the set $\{a_1, a_2, \ldots, a_{10}\}$.

In each case the symmetric group S_{40} acts transitively on the partial rankings which may thus be represented as homogeneous spaces for S_{40} (see Chapter 3-F for definitions). For ranked 10 out of 40 the homogeneous space is S_{40}/S_{30}. For unranked 10 out of 40, the homogeneous space is $S_{40}/S_{10} \times S_{30}$.

Here are some real examples of such data.

Example 1. American Psychological Association data. The American Psychological Association is a large professional group (about 50,000 members). To vote for a president, members rank order five candidates. A winner is chosen by the *Hare system*: Look at the first place votes for all five candidates. If there is no majority candidate ($\geq 50\%$) delete the candidate with the fewest first place votes. Ballots

Figure 1
The 1970 Random Selection Sequence by Month and Day

Day	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1	305	086	108	032	330	249	093	111	225	359	019	129
2	159	144	029	271	298	228	350	045	161	125	034	328
3	251	297	267	083	040	301	115	261	049	244	348	157
4	215	210	225	081	276	020	279	145	232	202	266	165
5	101	214	293	269	364	028	188	054	082	024	310	056
6	224	347	139	253	155	110	327	114	006	087	076	010
7	306	091	122	147	035	085	050	168	008	234	051	012
8	199	181	213	312	321	366	013	048	184	283	097	105
9	194	338	317	219	197	335	277	106	263	342	080	043
10	325	216	323	218	065	206	284	021	071	220	282	041
11	329	150	136	014	037	134	248	324	158	237	046	039
12	221	068	300	346	133	272	015	142	242	072	066	314
13	318	152	259	124	295	069	042	307	175	138	126	163
14	238	004	254	231	178	356	331	198	001	294	127	026
15	017	039	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	055	274	120	044	207	254	107	096
17	235	189	033	260	112	073	058	154	255	288	143	304
18	140	292	332	090	278	341	190	141	246	005	146	128
19	058	025	200	236	075	104	227	311	177	241	203	240
20	280	302	239	346	123	360	187	344	063	192	185	135
21	186	363	334	062	250	060	027	291	204	243	156	070
22	337	290	265	316	326	247	153	339	160	117	009	053
23	118	057	256	252	319	109	172	116	119	201	182	162
24	059	236	258	002	031	358	023	036	195	196	230	095
25	052	179	343	351	361	137	067	286	149	176	132	084
26	092	365	170	340	357	022	303	245	018	007	309	173
27	355	205	268	074	296	064	289	352	233	264	047	078
28	077	299	223	262	308	222	088	167	257	094	281	123
29	349	285	362	191	226	353	270	061	151	229	099	016
30	164		217	208	108	209	287	333	315	038	174	003
31	211		030		313		193	011		079		100

Figure 2

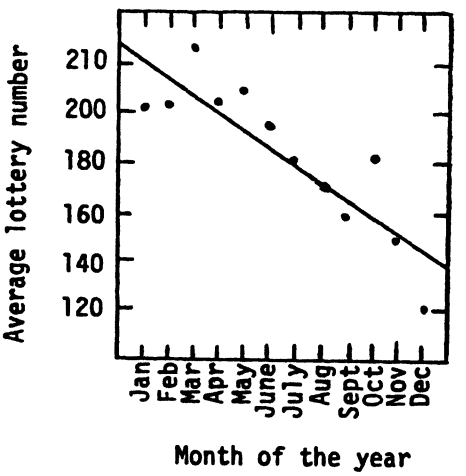


Fig. 2. Average lottery numbers by month. The line is the least squares regression line, treating the months as being equally spaced.

with this candidate are relabelled to have the remaining candidates in the same relative order. The procedure is now continued with the four remaining candidates. Fishburn (1973), Doran (1979), or Brams and Fishburn (1983) discuss the system and relevant literature.

A considerable number of voters do not rank all five candidates. For example, in the year being considered the number of voters ranking q of the candidates was

q	#	
1	5141	} 9711
2	2462	
3	2108	
5	5738	
	15,449	

Thus there were 5,738 complete rankings, but 5,141 only voted for their first choice. In all, more than half of the ballots were incomplete. It is assumed that people who rank 4 candidates meant to rank the 5th candidate last.

It is natural to inquire whether the partially ranked ballots are different from the restriction of the complete ballots (or vary with q). Such considerations should play a role in deciding on a final voting rule, and on deciding on ballot design and election publicity in following years.

Table 1 gives the complete data. The data are arranged as (rank, #) where rank is a five-digit number, whose ith digit represents the rank given to candidate i (a zero or blank means that this is a partial ranking, in which candidate i has not been ranked). For example, the first entry (1, 1022) indicates that candidate 5 was ranked first by 1022 people who didn't rank anyone else. The second entry (10, 1145) indicates that candidate 4 was ranked first by 1145 people (who didn't rank anyone else). The first 5 entries give the totals for singly ranked items. The next 20 entries give totals for people ranking 2 of the 5 candidates. For example 143 people ranked candidate 5 first and candidate 4 second (and didn't rank anyone else). These data are analyzed by Diaconis (1989).

Example 2. k sets of an n set. If people are asked to choose their favorite k of n, *without* ranking within (as in choosing a committee or set of invitees to a meeting), then the relevant homogeneous space is $S_n/S_k \times S_{n-k}$, where $S_k \times S_{n-k}$ is the subgroup of S_n allowing arbitrary permutations among $\{1, \ldots, k\}$ and among $\{k+1, \ldots, n\}$. Approval voting, recommended by Brams and Fishburn (1983) yields such data.

Here is an example where large amounts of such data occur. The State of California has a state lottery game called 6/49 or Lotto. To play, you select a 6 set from $\{1, 2, \ldots, 49\}$. Then, 6 of 49 numbered balls are chosen at random. The grand prize is divided between the people choosing this subset.

There are about 14 million subsets, and 11 million players per week in this game at present. Of course, people do not choose subsets at random — they play favorate combinations. One can get a distinct advantage in this game by avoiding popular numbers and subsets. After all, if you are the only person on the subset you don't have to split with anyone. This can actually overcome the "house take"

Table 1
American Psychological Association Election Data

Partial Ranking	# of Votes Cast of This Type	Partial Ranking	# of Votes Cast of This Type	Partial Ranking	# of Votes Cast of This Type	Partial Ranking	# of Votes Cast of This Type
1	1022	23100	83	45213	24	24135	96
10	1145	20103	74	45132	38	23541	45
100	1198	132	19	45123	30	23514	52
1000	881	123	15	43521	91	23451	53
10000	895	2103	16	43512	84	23415	52
21	143	1302	15	43251	30	23154	186
12	196	1032	45	43215	35	23145	172
201	64	1320	17	43152	38	21543	36
210	48	1203	8	43125	35	21534	42
102	93	31002	38	42531	58	21453	24
120	56	31020	45	42513	66	21435	26
2001	70	31200	32	42351	24	21354	30
2010	114	21003	17	42315	51	21345	40
2100	89	21030	31	42153	52	15432	40
1002	80	1023	55	42135	40	15423	35
1020	87	1230	9	41532	50	15342	36
1200	51	21300	31	41523	45	15324	17
20001	117	10032	35	41352	31	15243	70
20010	104	10203	49	41325	23	15234	50
20100	547	10302	41	41253	22	14532	52
21000	72	10320	21	41235	16	14523	48
10002	72	13002	31	35421	71	14352	51
10020	74	13020	22	35412	61	14325	24
10200	302	13200	79	35241	41	14253	70
12000	83	10023	44	35214	27	14235	45
30021	75	10230	30	35142	45	13542	35
30201	32	12003	26	35124	36	13524	28
32001	41	12030	19	34521	107	13452	37
20031	62	12300	27	34512	133	13425	35
20301	37	54321	29	34251	62	13254	95
23001	35	54312	67	34215	28	13245	102
3201	15	54231	37	34152	87	12543	34
2301	14	54213	24	34125	35	12534	35
3021	59	54132	43	32541	41	12453	29
2031	50	54123	28	32514	64	12435	27
321	20	53421	57	32451	34	12354	28
231	17	53412	49	32415	75	12345	30
30012	90	53241	22	32154	82		
30210	13	53214	22	32145	74		
32010	51	53142	34	31542	30		
20013	46	53124	26	31524	34		
20310	15	52431	54	31452	40		
23010	28	52413	44	31425	42		
3012	62	52341	26	31254	30		
3210	18	52314	24	31245	34		
2310	21	52143	35	25431	35		
2013	54	52134	50	25413	34		
312	46	51432	50	25341	40		
213	16	51423	46	25314	21		
2130	17	51342	25	25143	106		
3120	26	51324	19	25134	79		
3102	16	51243	11	24531	63		
30102	47	51234	29	24513	53		
32100	57	45321	31	24351	44		
30120	15	45312	54	24315	28		
20130	39	45231	34	24153	162		

and yield a favorable game. Chernoff (1981) gives details for the Massachusetts lottery. In any case, the data must be analyzed.

While it is not possible to present such data here, the following smaller example shows that interesting analyses are possible.

There are various gadgets sold to generate a six-element subset of $\{1, 2, \ldots, 49\}$. These are used to help players pick combinations for the California state Lotto game.

One such gadget is pictured in Figure 1. There are 49 numbered holes and six balls enclosed by a plastic cover. One shakes the balls around and uses the six set determined by their final resting place.

Figure 1.

PICK 6			LOTTO				& WIN	
1	2	3	4	5	6	7	8	9
o	o	o	o	o	o	o	o	o
10	11	12	13		14	15	16	17
o	o	o	o		o	o	o	o
18	19	20	21		22	23	24	25
o	o	o	o		o	o	o	o
26	27	28	29		30	31	32	33
o	o	o	o		o	o	o	o
34	35	36	37		38	39	40	41
o	o	o	o		o	o	o	o
42	43	44	45		46	47	48	49
o	o	o	o		o	o	o	o

This gadget seems at first like other classical devices to generate random outcomes: if vigorously shaken, it should lead to random results. Further thought suggests that the outer, or border numbers might be favored over the inner numbers.

To test this, 100 trials were performed. The gadget was vigorously shaken and set down on a flat surface. The results are given in Table 2.

Following each six set is X — the number of balls falling on the outer perimeter in that 6-set. For example, the first 6-set $\{10, 11, 13, 25, 36, 42\}$ had 3 outside numbers — 10, 25, 42 — so $X = 3$. There are 25 outside numbers out of 49.

Table 2

100 6–sets of $\{1, 2, \ldots, 49\}$

10,11,13,25,36,42/3	5,10,21,26,42,46/5	4,17,18,22,32,41/4	1,17,22,25,29,31/3
25,27,34,39,45,46/4	16,23,37,41,43,45/3	6, 9,10,12,16,32/3	2,13,23,24,26,30/2
3, 5,18,20,33,39/4	8,10,13,34,43,49/5	2, 5,17,19,36,40/3	2, 6,15,18,32,37/3
3,10,23,26,45,49/5	2,10,11,12,13,15/2	2, 6,10,25,33,38/5	2,14,15,17,18,35/3
3, 7,15,19,26,34/4	10,13,15,22,26,43/3	3,17,29,40,41,45/4	4,10,20,31,32,37/2
4,15,32,33,36,49/3	15,19,22,30,32,39/0	4, 7,11,23,35,36/2	7,13,17,27,31,44/2
10,11,23,33,43,46/4	2,15,22,25,29,48/3	1,18,31,33,34,46/5	19,22,28,32,42,44/2
1, 6, 7,18,26,34/6	6, 7,10,11,17,31/3	11,13,15,28,34,39/1	7,13,19,33,47,48/4
6,11,15,19,26,46/3	6,17,24,29,42,43/4	4, 7,15,18,31,33/4	1, 2, 4,15,19,40/3
1,11,15,18,26,29/3	2, 9,21,36,43,45/4	1, 3,12,15,20,41/3	2, 5,25,26,30,39/4
10,11,19,31,36,42/2	7,12,18,35,42,44/4	4, 9,12,22,39,41/3	
6,15,25,27,42,47/4	5,16,18,33,36,39/3	3,10,12,28,34,39/3	
17,18,33,36,43,46/5	2, 6, 7,11,31,47/4	2, 7,12,27,34,35/3	
1, 2,32,36,43,48/4	18,22,28,36,42,47/3	1, 4, 7,12,20,43/4	
16,20,30,35,45,46/2	4,18,29,35,39,46/3	5, 7,14,16,18,31/3	
16,20,26,37,42,49/3	3, 6,16,25,29,42/4	6,23,28,34,36,40/2	
3,18,27,30,42,43/4	1,28,31,37,42,43/3	2, 5, 9,15,23,27/3	
9,10,27,42,43,45/5	1,18,23,27,42,43/4	2, 3,19,34,39,44/4	
6,23,32,39,42,46/3	4, 5, 7, 8,40,42/5	6,12,14,16,23,39/1	
5,19,36,39,42,44/3	6, 7, 9,12,39,49/4	2,12,15,26,38,43/3	
7,18,20,29,35,43/3	12,13,18,19,22,36/1	5, 7,12,17,29,35/3	
12,14,23,29,41,48/2	4, 7, 8,10,33,49/6	4, 9,16,23,27,42/3	
4, 6,17,20,33,48/5	7, 9,31,32,41,46/4	2,13,15,20,21,48/2	
4,18,27,30,43,49/4	9,12,14,37,46,48/3	1, 5,34,42,44,46/6	
6,10,18,30,35,45/4	7, 9,16,29,41,46/4	15,16,17,24,27,30/1	
26,27,38,42,43,44/4	14,19,21,28,33,42/2	8,15,18,21,30,39/2	
6, 8,19,38,43,49/4	6,14,15,17,31,49/3	6,15,21,23,32,47/2	
1,20,25,42,43,49/5	8,33,35,41,45,47/5	5, 7, 8,19,23,49/4	
4,34,27,39,43,46/4	2, 8,25,29,42,47/5	7, 8, 9,14,20,22/3	
3, 4,11,33,46,49/5	8,11,24,25,37,48/3	10,15,29,34,46,49/4	

If the six sets were chosen at random, X would have a hypergeometric distribution $H\{X = j\} = \frac{\binom{25}{j}\binom{24}{6-j}}{\binom{49}{6}}$. These numbers are given in Table 3 which also shows the empirical counts from Table 2.

Table 3

Hypergeometric and Empirical Probabilities for X.

j	0	1	2	3	4	5	6
$H\{X = j\}$	.013	.091	.250	.333	.228	.016	.010
Empirical	.01	.04	.14	.35	.30	.13	.03

The differences are *not* overwhelming visually. They do show up in two straightforward tests.

A first test was based on $p = H\{X = 4, 5, 6\} = .353$ versus the empirical ratio $\hat{p} = .46$. Then $(\hat{p} - p)/\sqrt{p(1-p)/100} = 2.23$. This difference, more than two standard deviations, is convincing evidence against uniformity.

Colin Mallows suggested using the average, $\overline{X}$, as a statistic. Under the null distribution, $E(\overline{X}) \doteq 3.06$, $SD(\overline{X}) \doteq 0.116$. The observed $\overline{X}$ is 3.40. This yields a standardized (z value) of 2.92.

Remarks.

1) As is well known, the omnibus chi-square test is to be avoided for these kinds of problems. Because it tries to test for all possible departures from uniformity, chi-square only works well for large deviations or sample sizes. Interestingly, here it *fails* to reject the null (10.23 on six degrees of freedom

with all 7 categories or 9.01 on five degrees of freedom with the first and last categories combined).

2) Other questions can be asked of these data. To begin with, the central numbers

$$20,\ 21,\ 22,\ 23$$
$$28,\ 29,\ 30,\ 31$$

presumably occur less often. More generally, a test that looks at all numbers, but takes into account the distance from the edge, could be constructed. A preliminary graphical analysis was *not* instructive.

Interesting questions arise about the corners and about individual numbers. With more data, some second order questions can be entertained.

3) It seems clear that this style of randomization mechanism is badly flawed. Possible physical explanations can be entertained to explain these flaws. The balls lose most of their energy on impact with the sides, and then "trickle back" to the edge. A slight tilt draws the balls toward an edge.

4) One practical application of this kind of testing problem comes in the actual lottery. A quick test to detect marked departures is needed for a pre-game screening (someone might have switched for loaded balls during the night).

Example 3. Q sort data. The General Social Survey lists thirteen qualities a child could possess. From this list, respondents are asked to choose the most desirable quality, the two next most desirable qualities, the least desirable quality and the next two least desirable qualities. In an obvious way, this is data on $S_{13}/S_1 \times S_2 \times S_7 \times S_2 \times S_1$. More generally, if λ is a partition of n, so $\lambda = (\lambda_1, \ldots, \lambda_m)$ with $\lambda_1 + \ldots + \lambda_m = n$, one can consider data of the form: choose the first λ_1 objects (but do not order between), choose the next λ_2 objects, etc., finishing with λ_m objects ranked last. Such a scheme is called Q sort data in psychology experiments. It is not unusual to ask for a list of 100 items to be ranked for its degree of concordance or similarity with a fixed object. For example, the object might be a person (spouse, national leader) and the items might be descriptive levels of aggression. Suppose 9 categories of similarity are used, ranging from 1 - "most uncharacteristic," through 5 "neither characteristic nor uncharacteristic," up to 9 - "most characteristic." To aid in different rates, a forced distribution is often imposed. For $n = 100$, the numbers permitted in each category are often chosen from binomial considerations as 5, 8, 12, 16, 18, 16, 12, 8, 5. A novel application and references to the older literature may be found in L. E. Moses et al (1967). For more recent discussion see Heavlin (1980).

Example 4. Other actions of S_n. The symmetric group acts on many other combinatorial objects, such as the set of partitions or labelled binary trees. It follows that there is a wide variety of objects to which the analysis of this and succeeding chapters may be applied.

C. The d-sphere S^d.

Sometimes data are collected on the circle – which way do birds leave their nests. Data are also collected on the sphere – for example, in investigating the

theory of continental drift, geologists looked at magnetization direction of rock samples on two "sides" of a purported boundary. Roughly, small pieces of certain kinds of rocks have a given magnetic orientation giving points on the sphere in $\mathbb{R}^3$. This leads to two-sample and other data analytic problems. Such considerations led Fisher (1953) to invent his famous family of distributions on the sphere.

Here is an example of data on higher dimensional spheres: consider testing whether measurement errors are normal. Samples of size p are available from a variety of different sources. Say sample i is normal with parameters μ_i, σ_i^2:

$$(X_{11}, \ldots, X_{1p}) \text{ i.i.d. } n(\mu_1, \sigma_1^2)$$
$$(X_{21}, \ldots, X_{2p}) \text{ i.i.d. } n(\mu_2, \sigma_2^2)$$
$$(X_{n1}, \ldots, X_{np}) \text{ i.i.d. } n(\mu_p, \sigma_p^2).$$

Think of p small (say 10) and n large (say 50). All samples are assumed independent. Let $\overline{X}_i$ and S_i be the ith sample mean and standard deviation.

$$Y_i = \Big(\frac{X_{i1} - \overline{X}_i}{S_i}, \ldots, \frac{X_{ip} - \overline{X}_i}{S_i}\Big).$$

The spherical symmetry of the normal distribution implies that Y_i are randomly distributed over a $p-2$ dimensional sphere. Standard tests for uniformity thus provide tests for normality.

The group of $n \times n$ orthogonal matrices $O(n)$ acts transitively on the n sphere. The subgroup fixing a point (say the north pole (1,0,...,0)) is clearly $O(n-1)$. Thus the sphere can be thought of as $O(n)/O(n-1)$ and the rich tools of harmonic analysis become available.

Further introductory discussion is in Chapter 9B. Mardia (1972) and Watson (1983) give motivated, extensive treatments of data on the sphere.

D. Other groups.

Many other groups occur. For example binary test results (e.g. right/wrong on the ith question $1 \le i \le k$) lead to data on Z_2^k. Here, for $x \in Z_2^k$, $f(x)$ is the number of people answering with pattern x. In *panel studies* a subject is followed over time. For example, 5,000 people may be followed for a year, each month a one or zero is recorded as the person is employed or not. This leads to data on Z_2^{12}.

There is a curious data set for $Z_{365} \times Z_{365}$ connected to the birthday-deathday question. Some researchers claim famous people tend to die close to the date of their birth. See Diaconis (1985) for a review of this literature.

Data on yet other groups arises in testing Monte Carlo algorithms for generating from the uniform distribution. Such group valued random variables are useful in doing integrals over groups. Testing a generator leads to a sample on the group in question. I have looked at data for the orthogonal and unitary groups in this regard.

It seems inevitable that data on other groups and homogeneous spaces will arise naturally in applications. One final example: with many scatterplots, one

has many covariance matrices. The set of positive definite 2×2 matrices is usefully represented as GL_2/O_2. Several other examples are given in the following chapters.

E. STATISTICS ON GROUPS.

The examples described above suggest a wealth of statistical problems. In classical language, there is

- Testing for uniformity (is the sample really random?)
- Two sample tests (is there a difference between men and women's rankings?)
- Assessing association (is husband's ranking close to wife's?)
- Model building (can this huge list of data be summarized by a few parameters?)
- Model testing

More inclusively, there is the general problem of data analysis: how to make sense of this type of data; how to discover structure and find patterns.

The next four chapters offer three different approaches to these problems. Chapter 6 develops measures of distance on groups and homogeneous spaces. These are used to carry all sorts of familiar procedures into group valued examples.

Chapter 8 develops an analog of the spectral analysis of time series for group valued data. This is explored in the examples of partially ranked data. These examples make full use of the representation theory of the symmetric group. Chapter 7 is devoted to a self-contained development of this theory.

Chapter 9 uses representation theory to develop a natural family of models. In familiar cases, these reduce to models introduced by applied workers. The theory shows how to go further, and gives a unified development for all groups at once.

Of course, there is no substitute for trying things out in real examples, where special knowledge and insight can be brought to bear. There has not been much Bayesian work on these problems that I know of. The problems of developing natural prior distributions with respect to invariance seem fascinating. Consonni and Dawid (1985) or Fligner and Verducci (1988) offer steps in this direction.

Chapter 6. Metrics on Groups, and Their Statistical Uses

In working with data, it is often useful to have a convenient notion of distance. Statisticians have used a number of different measures of closeness for permutations. This chapter begins by analyzing some applications. Then a host of natural metrics (and basic properties) is provided. Next, some abstract principles for constructing metrics on any group are shown to yield the known examples. Finally, the ideas are carried from groups to homogeneous spaces.

A. Applications of metrics.

Example 1. Association. Let ρ be any metric on the permutations in S_n. Thus, $\rho(\pi,\pi) = 0, \rho(\pi,\sigma) = \rho(\sigma,\pi)$ and $\rho(\pi,\eta) \leq \rho(\pi,\sigma) + \rho(\sigma,\eta)$. Many possible metrics will be described in Section B. To fix ideas, one might think of ρ as Spearman's footrule: $\rho(\pi,\sigma) = \sum_i |\pi(i) - \sigma(i)|$. One frequent use is calculation of a measure of nonparametric association between two permutations. A standard reference is the book by Kendall (1970).

As an example, consider the draft lottery example in Figure 2 of Chapter 5. The data consists of 12 pairs of numbers, (i, Y_i), and Y_i being the rank of the average lottery number in month i. It is hard to get the value of Y_i out of the figure, but easy to get the rank of Y_i (i.e., biggest, next biggest, etc.). I get

π Month	J	F	M	A	M	J	J	A	S	O	N	D
σ Rank Y_i	5	4	1	3	2	6	8	9	10	7	11	12

The two rows can be thought of as two permutations in S_{12}. Are they close together? Taking ρ as the footrule, $\rho(\pi, \sigma) = 18$. Is this small? The largest value ρ can take is 72. This doesn't help much. One idea is to ask how large $\rho(\pi, \sigma)$ would be if σ were chosen at random, uniformly. Diaconis and Graham showed the following result (proved in Section B below).

Theorem 1. *Let* $\rho(\pi, \sigma) = \sum |\pi(i) - \sigma(i)|$. *If* σ *is chosen uniformly in* S_n *then*

$$AV(\rho) = \frac{1}{3}(n^2 - 1)$$

$$Var(\rho) = \frac{1}{45}(n+1)(2n^2+7)$$

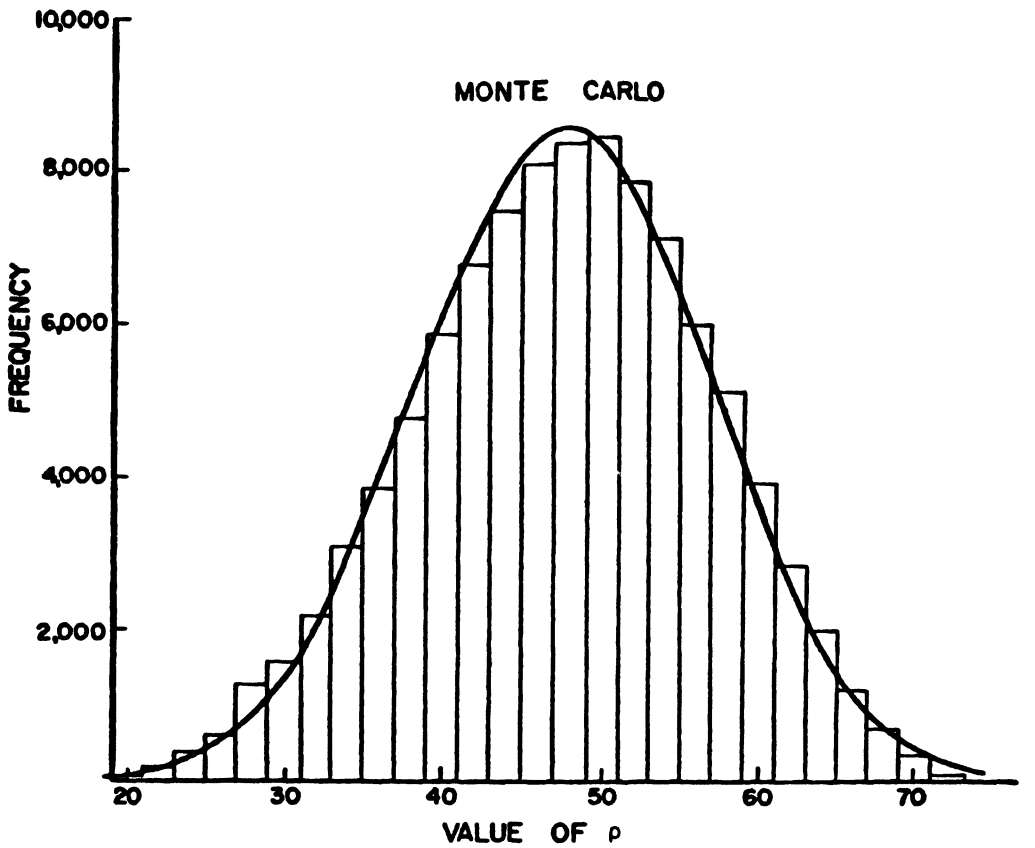
$$P\{\frac{\rho - AV}{SD} \leq t\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx + o(1).$$

In the example, $AV \doteq 47.7, SD \doteq 9.23$. The value 18 is more than 3 standard

deviations from the mean. Thus 18 is small in that it (or a smaller value) is quite unlikely to have occurred under a simple chance model.

The approximate normality is valid as n tends to infinity and one might worry about $n = 12$. Figure 4 below shows the result of a Monte Carlo experiment based on 100,000 choices of σ from a uniform distribution. The normal approximation seems fine. The graph was supplied by Hans Ury who also published tables of the footrule for $n \leq 15$, in Ury and Kleinecke (1979). From their tables, the p value for the draft lottery data is $P\{\rho \leq 18\} = .001$.

Figure 1
Normal approximation for $n = 12$



Many further tests of randomness for the Draft Lottery data are described by Fienberg (1971). This test is natural starting from Figure 2.

Statisticians often normalize metrics to lie in [-1,1] like correlation coefficients. If $\rho(s,t)$ is a metric with maximum value m, then $R(s,t) = 1 - 2\rho/m$ lies in [-1,1].

I find it interesting that the standard "non-parametric measures of associ-

ation" arise from metrics. I've never been able to get much mileage out of the triangle inequality, which translates to

$$R(s,u) \geq R(s,t) + R(t,u) - 1.$$

Example 2. Scaling A second use of metrics for permutation data adapts such data for a run through a standard multidimensional scaling or clustering program. Multidimensional scaling takes points in any metric space and finds points in the plane such that the distances between the points in the plane are close to the distances between the points in the metric space. Imagine a collection of several hundred rankings of 10 items. It can be hard to get a preliminary "feel" for such data. Scaling finds representative points or sometimes "nonlinear mapping" which can be visualized. Obviously, a metric is necessary, since the input to a multidimensional scaling program is the set of distances between the original points. A nice discussion of scaling is in Chapter 14 of Mardia, Kent, and Bibby (1978). Critchlow (1985, pg. 116–121) gives an example with permutation data. Cohen and Mallows (1980) use the biplot in a similar way. See Figure 2 below.

Example 3. Mallows' model. A third use of metrics is as a means of model building. Following Mallows (1957), let's use a metric to put a probability measure on S_n. This measure will have a location parameter $\pi_0 \in S_n$ and a scale parameter $\lambda \in R^+$. Set

$$P(\pi) = ce^{-\lambda\rho(\pi,\pi_0)}; \quad c^{-1} = \sum_{\pi} e^{-\lambda\rho(\pi,\pi_0)}.$$

The largest probability is assigned to π_0 and probability decreases geometrically as the distance from π_0. Increasing λ makes the distribution more and more peaked about π_0. Of course, $\lambda = 0$ gives the uniform distribution.

A nice application of this approach to analyzing agreement between several judges in a contest is in Feigin and Cohen (1978). Critchlow (1985) gives other examples where Mallows' model provides a good fit to ranking data.

Mallows' original derivation of this model is less ad hoc. He considers generating a ranking of n items by making paired comparisons. Suppose π_0 is the true ranking, but a subject errs in comparing i and j with probability p. Mallows shows that conditional on the comparisons yielding a ranking, the ranking is distributed as above, with ρ given by Kendall's measure of association τ and λ a function of p. This is discussed in Section B below. Fligner and Verducci (1986, 1988b) develop and extend this justification for Mallows model.

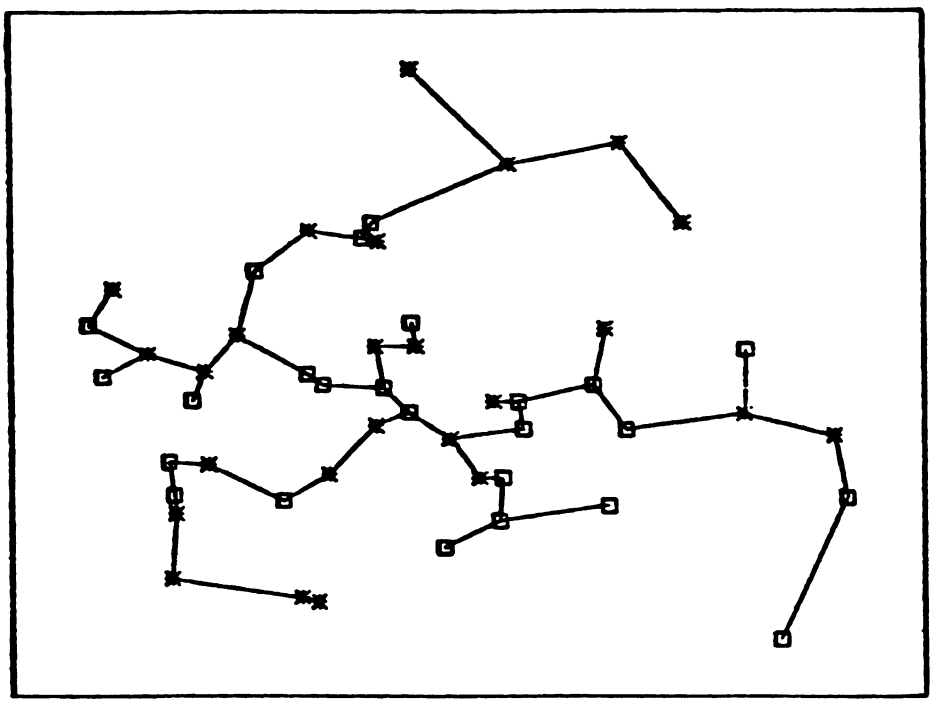
Example 4. Two-sample problems. Here is a fourth use of metrics: as a means of looking at 2 sample problems. In such problems we consider two sets of permutations $\pi_1, \ldots, \pi_n$ and $\sigma_1, \ldots, \sigma_m$ and ask about their similarities and differences. One classical question: "Can these two sets be regarded as samples from a single population of permutations?". If the $\pi's$ and σ's were permutations of a small number of items and n and m were large, there would be no problem. The question could be treated by well-known techniques for the multinomial distribution. Consider though, the problem of distinguishing between the distribution of riffle shuffles generated by Reeds and Diaconis in Chapter 5. Here $n = 100$, $m = 103$

and the permutations are in S_{52}. Here is an idea, borrowed from J. Friedman and L. Rafsky (1979).

Choose a metric ρ. Regard the 2 sets of permutations as points in a metric space. Form the minimal spanning tree for the combined data – that is, the shortest connected graph having a unique path between every pair of points. "Color" the points of one set (say the set $\{\pi_i\}$) red. Count T, the number of edges in the tree that join two nodes of different colors. The idea is that if the distributions of π and σ differ, the 2 types of points will tend to be separated, and only a few edges in the tree will cross over. If the distributions of π and σ are the same, there will be many cross-overs.

Figure 2

A 'scaling' picture of the minimal spanning tree in a metric space. The squares are sample 1, the stars are sample 2.



The distribution of T can be simulated by fixing the tree and randomly relabelling the vertices, drawing the m values without replacement from an urn containing $n + m$ balls. Friedman and Rafsky give a normal approximation. See also Stein (1986).

The discussion above used the minimal spanning tree. Any graph that connects points together if they are close can be used. Friedman and Rafsky also obtained good results for the graph connecting each point to its k-nearest neigh-

bors. Critchlow (1985, Chapter 6) used the union of all minimal spanning trees – for discrete metrics, the tree need not be unique.

Feigin and Alvo (1986) give another approach to assessing variability between groups using metrics on permutations. Fligner and Verducci (1988a) develop these ideas into a new approach for judging athletic competitions.

Example 5. Generalized association. Friedman and Rafsky (1983) have developed a method of testing association for data of the form (x_1, y_1), $(x_2, y_2), \ldots,$ (x_n, y_n). Here x takes values in a metric space X, and y takes values in a metric space Y. In an epidemiology application it might be that x_i are times of occurrence, and y_i are spatial locations of cases of a rare disease. One suspects trouble if points that are close in time are close in space.

In a more mundane setting, X and Y may both be symmetric groups, the data representing rankings of items on two occasions.

To test "association" they suggest forming a nearest neighbor graph for the x_i, and a separate nearest neighbor graph for the y_i. These graphs might both be minimal spanning trees. This gives two graphs on the vertex set 1, 2, $\ldots, n$. Now take T to be the number of edges that occur in both graphs. T is large if points close in X are close in Y.

One can get a null hypothesis distribution for T by comparing it with repeated values from the samples $(x_1, y_{\pi(1)}), \ldots, (x_n, y_{\pi(n)})$ where π is a random permutation. After all, if x_i and y_i have no connection, the value of T should be about the same as for $(x_i, y_{\pi(i)})$. Friedman and Rafsky give a normal approximation for this statistic. See also Stein (1986).

One final idea: this test of association includes the 2 sample test described in Example 4! To see this, consider the $m + n$ values as points in a space Y, and let x_i be one or zero as y_i is from the first or second sample. Use the discrete metric on X. The association statistic T_a counts the number of edges that appear in both graphs. This is the number of edges in the graph in Y space that have the same colored edges in the two sample setting. Thus $T = n + m - 1 - T_a$, so the two tests are equivalent; distributions are judged different if there is association with sample labels.

Jupp and Spurr (1985) give a different approach to testing for independence on groups using metrics.

Example 6. Goodness of fit tests. Given a model for data in a metric space X, one can carry out standard chi-squared goodness of fit tests by splitting X into pieces based on a metric and comparing observed and expected.

Example 7. Robust regression. Here is an approach to non-linear regression using a metric on S_n. Consider a family of real valued functions from a space X;

$$f(x, \theta): X \to \mathbb{R},\ \theta \in \Theta$$

e.g., $f(x) = a + bx$, or $f(x) = a + b \cos(cx + d)$. Suppose we observe $(y_1, x_1), \ldots,$ (y_n, x_n) and desire a value of θ such that y_i is close to $f(x_i, \theta)$. The classical approach to this is to fit by least squares: find a value of θ minimizing $\sum(y_i -$

$f(x_i, \theta))^2$. In recent years, people have noted that this approach is very sensitive to a few "wild values". If 1 or 2 of the x or y values are far away from the rest, those values have a huge effect on the minimizing θ. Here is a simple idea: choose a value of θ so that the rank of $f(x_i, \theta)$ is as close as possible to the rank of y_i. In simple linear cases, this gives the line with correlation replaced by the nonparametric measure of correlation induced by ρ. Sen (1968) develops properties of the estimator. Bhattacharya, Chernoff, and Yang (1983) apply it to a fascinating cosmology problem involving truncated regression.

Example 8. Social choice functions. A common problem in social choice theory is the choice of the "best alternative" based on a committee's ranking of the available alternatives. Classical examples include

Plurality:	Choose the alternative with the most first place votes
Borda's rule:	Assign a weight of 0 to the least preferred alternative. 1 to the next least preferred, and so on. The total score of each alternative is computed and the alternative(s) with the highest score is chosen as winner.
Condorcet's rule:	If there is some alternative that defeats every other in pairwise comparison, then that alternative should be chosen as the winner.

Even when applicable, the different rules need not lead to the same choice. Consider 19 rankers choosing between three alternatives a, b, c. If the rankings are

a	b	c	#
1	2	3	3
1	3	2	4
2	1	3	2
3	1	2	4
3	2	1	6
			19

then a is chosen by plurality but b is chosen by Borda's rule (it gets score 21 versus 16 for a and 20 for c) and c is chosen by Condorcet's rule (it defeats each of a and b in 10 votes). A famous theorem of Arrow says that there is no "reasonable" social choice function. A review of this literature may be found in Fishburn (1973). Grofman and Owen (1986) contains several further review articles.

For some tasks it may be desirable to choose a winner and a runner up. Other tasks require a committee of the top three choices or a complete permutation, representing the group's ranking. These may all be subsumed under the problem of choosing a partial ranking of shape λ, where λ is a partition of n, the number of alternatives (see Section B of Chapter 5). We will focus on the choice of a complete ranking given a probability P on rankings. Usually, $P(\pi)$ is the proportion of rankers choosing π.

One usable route through this problem uses metrics on groups as a way of

defining a "mean" or "median". Let P be a probability on a finite group G. Let ρ be a metric on G. Define

$$f(s) = \sum_t P(t)\rho(s,t).$$

The group element η is a *ρ-median of P* if η minimizes $f(s)$. The number $f(\eta)$ is called the *ρ-spread* of P. Substitution of ρ^2 for ρ in the formula for $f(s)$ yields a *ρ-mean.*

John Kemeny has proposed choosing a group ranking by using the metric induced by Kendall's tau on S_n. In Young and Levenglick (1978), a list of properties of Kemeny's procedure are shown to characterize it. Here is a version of their result:

A *preference function* μ assigns a set of permutations to each probability P on S_n. For example $\mu(P)$ could be the set of ρ-medians of P. A preference function is *neutral* if it transforms correctly under relabeling. In symbols, let $P_\eta(\pi) = P(\eta^{-1}\pi)$, then μ is neutral if

$$\mu(P_\eta) = \eta\mu(P) \text{ for all } \eta \text{ and } P.$$

A preference function is *consistent* if for any a in (0,1),

$$\mu(P_1) \cap \mu(P_2) \neq \phi \Rightarrow \mu(aP_1 + (1-a)P_2) = \mu(P_1) \cap \mu(P_2).$$

If P_1 and P_2 represent the rankings of n and m judges respectively, then the pooled panel is represented by $P_1 n/(n+m) + P_2 m/(n+m)$. Consistency says that if P_1 and P_2 lead to common preferences then the combined judges choose these preferences.

Given a probability P, let $n(P, ij)$ be the difference between the probabilities of all π preferring i to j and all π preferring j to i. Condorcet's proposal was that alternative i was preferred if $n(P, ij) > 0$ for all $j \neq i$ (thus i would beat any j in a pairwise popularity contest).

If a complete ranking is desired, a natural extension of Condorcet's idea is this: if i beats j in a pairwise popularity contest, then i should be ranked above j in any consensus ranking. Formally, it suffices to deal only with adjacent rankings. A preference function μ is called *Condorcet* if $n(P, ij > 0)$ (for fixed i and j) implies no π with $\pi(i) = \pi(j)+1$ is in $\mu(P)$. (For this, the condition becomes $n(P, ij) = 0$ implies $\pi^{-1}(k) = i, \pi^{-1}(k+1) = j \varepsilon \mu(P)$ iff $\pi^{-1}(k) = j, \pi^{-1}(k+1) = i \varepsilon \mu(P)$). Thus, no π ranking j as the immediate predecessor of i is in the consensus ranking.

Young and Levenglick show that medians based on Kendall's τ are neutral, consistent, and Condorcet. They further show that these three properties characterize τ-medians among preference functions.

These ideas can be carried over to choosing a final ranking of shape λ. Each $\pi \in S_n$ can be naturally assigned to such a partial ranking. The image of P under this map gives a probability on partial ranks and a choice of distance on partial rankings leads to a mean.

In Section 8.7 of Grenander (1981), a notion of a centroid set is introduced. This is very similar to a ρ-median, based on a distance defined using characters.

Example 9. Moments of probabilities on groups.

It is not clear how the ρ-medians and ρ-spreads relate to group operations like convolution. There is a little theory for moments of probabilities on groups that share, with the mean and variance, the property of being homomorphisms from probabilities under convolution into G (so the mean of a convolution is the sum of the means) or R^+ (so the variance of a convolution is the sum of the variances). This is elegantly surveyed in Heyer (1981).

Here is an example due to Levy (1939). Consider a random variable X taking values on the circle $T = \{z \varepsilon C: |z| = 1\}$. Levy defined variance as

$$V(X) = \inf_{a \in T} \int_T [\arg(z\bar{a})]^2 P_X(dz)$$

(where arg z is the unique $\phi \in (-\pi, \pi]$ such that $e^{i\phi} = z$). Every $a \in T$ which achieved the infimum he called a mean. He used these notions to prove the following version of the Kolmogorov three series theorem: Let $X_1, X_2, \ldots,$ be T valued random variables. A necessary and sufficient condition for convergence of $\sum_{j=1}^{\infty} X_j$ a.s. is

$$\text{(a) } \Sigma V(X_j) < \infty \quad \text{(b) } \Sigma E(X_j) < \infty$$

where (b) is interpreted as holding for any choice of expectations. This has been somewhat improved by Bartfai (1966).

Note that Levy's mean is the mean of example 8, with the usual metric.

Example 10. Tests for uniformity. Let X be a homogeneous space on which G acts transitively. We have data $x_1, x_2, \ldots, x_n$ and want to test if it is reasonable to suppose that these are independent and uniform.

As an example, X might equal G and the x_i might be the output of a computer routine to generate random elements of G — one wants to test such things. See Diaconis and Shahshahani (1987a) for examples.

The amount of data will play a role in choosing a test. If n is small, one can only hope to pick up fairly strong departure from uniformity.

One simple example is the following variant of the empty cell test: Let $\rho(x, y)$ be a G invariant metric. Look at $m = \min \rho(x_i, x_j)$, and compare with its null distribution. The null distribution can be approximated using Poisson limit theorems for U-statistics.

To fix ideas, take $X = G = Z_2^d$ with $\rho(x, y)$ the Hamming distance — the number of coordinates where x and y disagree. If x and y are chosen at random, $P\{\rho(x, y) \leq a\} = P\{B(a)\}$, with B(a) the ball of radius a. This has $\sum_{j=0}^{a} \binom{d}{j}$ points, and so $P(B(a)) = \frac{1}{2^d} \sum_{j=0}^{a} \binom{d}{j}$.

The expected number of pairs (x_i, x_j) within distance a is thus $\lambda = \binom{n}{2} P(B(a))$. For d large, and a chosen so that e.g. $1 \leq \lambda \leq 10$, the number

of close pairs is approximately Poisson(λ). The chance that no two are within distance a is thus approximately $e^{-\lambda}$.

For example, if $d = 10, n = 50, a = 0$, then $\lambda \doteq 1.2, e^{-\lambda} \doteq .3$.

The argument can be made rigorous by checking the conditions in Sevastyanov (1972), Silverman and Brown (1978), or Stein (1986). Note that theorems giving the null distributions of metrics (see Example 1 above) now are useful to compute volumes of spheres $B(a)$.

A collection of tests for uniformity on groups is suggested by Beran (1968), developed by Giné (1973), with a practical implementation by Wellner (1979). These all use distances and are mainly specialized to continuous examples such as the circle or sphere. Jupp and Spurr (1985) apply similar ideas.

Example 11. Loss functions. Investigating statistical aspects of the examples presented here leads to estimating parameters in a group. Metrics can be used as loss functions. For a classical example, consider n observations from a multinomial distribution with k categories and unknown probability vector $p_1, p_2, \ldots, p_k$. It may be desired to rank the p_i, deciding on the largest, next largest, and so on. Thus the parameter and estimate are permutations, and a decision theoretic formulation will involve a distance.

Estimation of Gaussian covariance matrices could stand some work from this viewpoint using the observation that GL_n/O_n is identified with the space of positive definite matrices; now the techniques of Section D below can be used.

The location parameter in Mallows' model (Example 3 above) is an element of S_n, and evaluation of estimators again necessitates a metric.

Andrew Rukhin (1970, 1977) began a systematic development of statistical estimation on groups that is well worth consulting.

Example 12. Random walk again. In investigating the rate of convergence of random walk on groups to the uniform distribution we used the total variation distance. It is natural to try other distances between probabilities. Several of these may be defined starting from a metric on G. Let G be a compact group, P and Q probabilities on G, and d a metric on G. We assume d is compatible with the topology on G (so $d(s,t)$ is jointly continuous). Usually d is invariant or bi-invariant. Also assume $d \leq 1$.

The Wasserstein or dual bounded Lipschitz metric is defined by $d_W(P,Q) = \sup |P(f) - Q(f)|$; the sup being over all f satisfying the Lipschitz condition $|f(x) - f(y)| \leq d(x,y)$.

It can be shown that the following statements are equivalent:

(a) $d_W(P,Q) \leq \varepsilon$.

(b) There are random variables taking values in G with $X \sim P, Y \sim Q$ and

$$E(d(X,Y)) \leq \varepsilon.$$

Dudley (1968) and Huber (1981) contain proofs of this result. Rachev (1986) contains an extensive survey. These papers also describe the Prohorov distance between P and Q — this also depends on the underlying metric. It seems extremely hard to get our hands on these metrics.

Inequality (b) above suggests that strong uniform times and coupling techniques can be used to bound these distances. I do not know of any examples.

Example 13. Rank tests. Doug Critchlow (1986) has recently found a remarkable connection between metrics and nonparametric rank tests. It is easy to describe a special case: consider two groups of people — m in the first, n in the second. We measure something from each person which yields a number, say $x_1, x_2, \ldots, x_m; y_1, \ldots, y_n$. We want to test if the two sets of numbers are "about the same."

This is the classical two-sample problem and uncountably many procedures have been proposed. The following common sense scenario leads to some of the most widely used nonparametric solutions.

Rank all $n + m$ numbers, color the first sample red and the second sample blue, now count how many moves it takes to unscramble the two populations. If it takes very few moves, because things were pretty well sorted, we have grounds for believing the numbers were drawn from different populations. If the numbers were drawn from the same population, they should be well intermingled and require many moves to unscramble.

To actually have a test, we have to say what we mean by "moves" and "unscramble." If moves are taken as "pairwise adjacent transpositions," and unscramble is taken as "bring all the reds to the left," we have a test which is equivalent to the popular Mann-Whitney statistic. If $m = n$, and moves are taken as the basic insertion deletion operations of Ulam's metric (see Section B below) we get the Kolmogorov-Smirnov statistic.

Critchlow begins by abstracting slightly: consider the positions of sample 1 as an m set out of $m + n$. The procedures above measure the distance to the set $\{1, 2, \ldots, m\}$. A two-sided procedure measures the smaller of the distances to $\{1, 2, \ldots, m\}$ or $\{n + 1, n + 2, \ldots, n + m\}$.

Every metric on $S_{n+m}/S_n \times S_m$ gives a naturally associated test. This is just the beginning. With k sample problems, having sample size λ_i from the ith population, we get testing problems on $S_N/S_{\lambda_1} \times S_{\lambda_2} \times \ldots \times S_{\lambda_k}$. Metrics on these spaces give rise to natural test statistics. Critchlow shows how essentially all of the basic testing problems in nonparametric statistics can be put into this framework.

This leads to a unified approach – there is a straightforward extension of the Mann-Whitney statistic for k sample problems, two-way layouts, two-sample spread problems, and others. Further, some procedures popular in two-sample problems have not been previously generalized, so many new tests are possible.

To those of us who have marveled at the cleverness of nonparametricians in cooking up new tests, this new unified view comes as a breath of fresh air. It offers hope for a lot more.

We all realize that normal theory testing is essentially testing with respect to the orthogonal group. Consider the ordinary t test for mean 0 versus mean $\mu > 0$. One normalizes the data vector $x_1, x_2, \ldots, x_n$ to lie on the unit sphere in $\mathbb{R}^n$, and calculates the distance to $(1, 1, \ldots, 1)/\sqrt{n}$. If $\mu = 0$, the point on the sphere is random. If $\mu > 0$, the point should be close to the vector with all equal

coordinates. The t-test amounts to the cosine of the angle between the vectors of interest. See Efron (1969) for discussion and pictures.

The F test in classical ANOVA has a similar interpretation as the distance between the observed vector and a subspace where some coordinates are equal. If in the robust regression of example 7, one uses the orthogonal group, ordinary least squares results. Many other normal theory procedures can be similarly interpreted.

Of course, the permutation group sits inside the orthogonal group. One may try to interpolate between nonparametrics and normal theory by considering intermediate groups. The sign change group is a natural starting place.

More examples will be discussed as we go along. Most of the applications can be carried over to other groups and homogeneous spaces. It is time to get to some metrics and their properties.

B. Some metrics on permutations.

Let π and σ be permutations in S_n, with the interpretation that $\pi(i)$ is the rank assigned by π to item i.

The following metrics have been used in various statistical problems.

$D(\pi, \sigma) = \Sigma|\pi(i) - \sigma(i)|$(Footrule)

$S^2(\pi, \sigma) = \Sigma\{\pi(i) - \sigma(i)\}^2$ (Spearman's rank correlation)

$H(\pi, \sigma) = \#\{i: \pi(i) \neq \sigma(i)\}$ (Hamming distance)

$I(\pi, \sigma) =$ minimum number of pairwise adjacent transpositions taking π^{-1} to σ^{-1}(Kendall's tau)

$T(\pi, \sigma) =$ minimum number of transpositions taking π to σ (Cayley distance)

$L(\pi, \sigma) = n -$ length of longest increasing subsequence in $\sigma\pi^{-1}$ (Ulam's distance)

This seems like a lot of metrics although it is only the tip of the iceberg. Table 2 gives the distance to the identity for all 6 metrics on S_4. The metrics have all been defined to be right-invariant in a way which will now be explained.

Invariance. In the most general situation, permutations are presented as $1-1$ maps between 2 different sets of the same cardinality:

$$\pi_i: A \to B, |A| = |B| = n.$$

The way we wind up labeling A or B may be fairly arbitrary and it is reasonable to consider distances that are invariant in some way. Here, if η is a $1-1$ map $\eta: A \to A$, right invariance means

$$\rho(\pi_1, \pi_2) = \rho(\pi_1\eta, \pi_2\eta).$$

Example. Consider 3 students ranked on the midterm and final:

Table 1
Values of the six metrics when $n = 4$

π	Cycles	$T(\pi)$	$I(\pi)$	$D(\pi)$	$S^2(\pi)$	$H(\pi)$	$L(\pi)$
1 2 3 4	(1)(2)(3)(4)	0	0	0	0	0	0
1 2 4 3	(1)(2)(3 4)	1	1	2	2	2	1
1 3 2 4	(1)(2 3)(4)	1	1	2	2	2	1
1 3 4 2	(1)(2 3 4)	2	2	4	6	3	1
1 4 2 3	(1)(2 4 3)	2	2	4	6	3	1
1 4 3 2	(1)(2 4)(3)	1	3	4	8	2	2
2 1 3 4	(1 2)(3)(4)	1	1	2	2	2	1
2 1 4 3	(1 2)(3 4)	2	2	4	4	4	2
2 3 1 4	(1 2 3)(4)	2	2	4	6	3	1
2 3 4 1	(1 2 3 4)	3	3	6	12	4	1
2 4 1 3	(1 2 4 3)	3	3	6	10	4	2
2 4 3 1	(1 2 4)(3)	2	4	6	14	3	2
3 1 2 4	(1 3 2)(4)	2	2	4	6	3	1
3 1 4 2	(1 3 4 2)	3	3	6	10	4	2
3 2 1 4	(1 3)(2)(4)	1	3	4	8	2	2
3 2 4 1	(1 3 4)(2)	2	4	6	14	3	2
3 4 1 2	(1 3)(2 4)	2	4	8	16	4	2
3 4 2 1	(1 3 2 4)	3	5	8	18	4	2
4 1 2 3	(1 4 3 2)	3	3	6	12	4	1
4 1 3 2	(1 4 2)(3)	2	4	6	14	3	2
4 2 1 3	(1 4 3)(2)	2	4	6	14	3	2
4 2 3 1	(1 4)(2)(3)	1	5	6	18	2	2
4 3 1 2	(1 4 2 3)	3	5	8	18	4	2
4 3 2 1	(1 4)(2 3)	2	6	8	20	4	3

		Bill	Bob	Jane
midterm	π_1	2	1	3
final	π_2	3	1	2

So the set $A = \{\text{Bill, Bob, Jane}\}$ and $B = \{1, 2, 3\}$. Suppose the data had been recorded as

	Bob	Bill	Jane
midterm	1	2	3
final	1	3	2

This is the same situation: Bob finished first in both exams, etc. It seems reasonable to insist that whatever measure of distance is used not change under this type of relabeling. If one naively uses the minimum number of pairwise adjacent transpositions it takes to bring the second row to the first, then the original way of writing things down takes 3 transpositions and the second way of writing things down takes 1 transposition.

Obviously, data can be presented in a form where left invariance is the natural requirement:

rank	1	2	3
midterm	Bob	Bill	Jane
final	Bob	Jane	Bill

Finally, here is an example in which two-sided invariance is a natural requirement. Imagine 5 people and 5 "descriptions" e.g., a psychological profile like MMPI or a psychic's description. A judge matches people with descriptions giving a $1-1$ map $\{\text{descriptions}\} \leftrightarrow \{\text{people}\}$. With 2 or more judges, the question of how close the judges' rankings are to one another arises. A two sided invariant distance seems appropriate.

Of the six distances in Section B, only H and T are invariant on both sides. Of course, any metric can be made invariant by averaging it.

EXERICSE 1. Show that T is bi-invariant. Show that Spearman's footrule, averaged to also make it left invariant, is the same as Hamming distance up to a constant multiple.

There are examples in which invariance, on either side, is not compelling. Consider a psycho-physical experiment in which a subject is asked to rank seven musical tones from high to low. If the tones are not uniformly distributed on some natural scale it might be natural to give different weights to differences in different parts of the scale. A measure like $\Sigma w_i|\pi(i) - \sigma(i)i|$ is not invariant on either side.

All of the six metrics are invariant under reversing order — changing i to $n + 1 - i$ — i.e. interchanging high and low.

Invariance considerations are natural in other problems as well. Consider an empirical set of data $g_1, \ldots, g_n$ taking values in the finite group G. In testing whether the data is uniform it is sometimes natural to require that a test statistic $T(g_1, \ldots, g_n)$ is invariant under translation: $T(g_1, \ldots, g_n) = T(g_1\eta, \ldots, g_n\eta)$. An example of a non invariant test takes $T(g_1, \ldots, g_n)$ equal to the number of $g_i \varepsilon A$ (e.g., the number of even permutations). Two easy ways to make statistics invariant are averaging and maximizing. Averaging replaces T by $T_1 =$

$\frac{1}{|G|}\sum_{\eta} T(g_1\eta, \ldots, g_n\eta)$. Maximizing replaces T by $T_2 = \max_{\eta} T(g_1\eta, \ldots, g_n\eta)$. Again, there are problems in which invariance is not compelling: In testing a shuffling mechanism for uniformity it is perfectly reasonable to pay special attention to the top and bottom cards.

We next turn to a case-by-case discussion of the six metrics and their properties.

1. *Spearman's footrule* $D(\pi, \sigma) = \Sigma|\pi(i) - \sigma(i)|$. Clearly this is a right invariant metric. Thus $D(\pi, \sigma) = D(\mathrm{id}, \sigma\pi^{-1})$. If either π or σ is uniformly chosen from S_n, the distribution of $D(\pi, \sigma)$ is the same as the distribution of $D(\mathrm{id}, \eta)$ with η chosen uniformly in S_n. The mean of D is computed as

$$E\{D\} = \frac{1}{n!}\sum_{\pi} D(\mathrm{id}, \pi) = \frac{1}{n!}\sum_{\pi}\sum_{i=1}^{n} |i - \pi(i)|$$
$$= \frac{1}{3}(n^2 - 1).$$

EXERCISE 2. Prove this last assertion.

A more tedious computation (see Diaconis and Graham (1977)) gives

$$\mathrm{Var}\{D\} = \frac{1}{45}(n + 1)(2n^2 + 7).$$

Finally, we indicate how the asymptotic normality of D can be shown (see the Theorem in example 1 of Section A for a careful statement). One approach uses Hoeffding's (1951) combinatorial central limit theorem: Consider $\{a_{ij}^n\}$, $i, j = 1, \ldots, n$ a sequence of arrays. Define

$$W_n = \sum_{i=1}^{n} a_{i\pi(i)}^n$$

where π is a random permutation in S_n. Then, subject to growth conditions on a_{ij}^n, W_n is asymptotically normal. The expression for the variance given above allows verification of the sufficient condition (12) in Hoeffding (1951) for the array $a_{ij}^n = |i - j|, i, j = 1, \ldots, n$. Bolthausen (1984) gives a version of the combinatorial limit theorem with a Berry-Esseen like error bound.

Ury and Kleinecke (1979) gave tables for the footrule when $n \leq 15$. The asymptotics seem quite accurate for n larger than 10. See Example 1 in Section A above.

Diaconis and Graham (1977) give some relations between the footrule and other measures of association that appear in the list of metrics. In particular

$$I + T \leq D \leq 2I.$$

So the more widely used metric I underlying Kendall's tau is close to the footrule D in the sense $I \leq D \leq 2I$.

Ian Abramson has pointed out a sampling theory interpretation for the footrule. Consider using the footrule to measure association. We are given n pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Assume that $P\{X_1 < s, Y_1 < t\} = H(s,t)$ and that the pairs are iid (of course, X_1 and Y_1 may well be dependent). To transform things to permutations, let the rank $R_i = \#\{j: X_j \leq X_i\}$. Similarly, let S_i denote the rank of Y_i. Assuming no tied values, Spearman's footrule defines a measure of association between the two samples by

$$(*) \qquad D = \sum_{i=1}^{n} |R_i - S_i|.$$

LEMMA 1. *Let $\{X_i, Y_i\}$ be iid from joint distribution function H, with margins $H_1(s), H_2(t)$. Then, Spearman's footrule D satisfies $\frac{1}{n^2}D = E\{|H_1(X) - H_2(Y)|\} + 0_p(\frac{1}{\sqrt{n}})$.*

Proof. From the Kolmogorov-Smirnov limit theorem

$$H_1(X_i) = \frac{R_i}{n} + 0_p(\frac{1}{\sqrt{n}}) \text{ uniformly in } i.$$

Thus

$$\frac{1}{n^2}\Sigma|R_i - S_i| = \frac{1}{n}\Sigma|H_1(X_i) - H_2(Y_i)| + 0_p(\frac{1}{\sqrt{n}}).$$

The sum converges to its mean as a sum of iid random variables. □

Remarks. Of course $H_1(X)$ and $H_2(Y)$ are uniform random variables. If $H(s,t) = H_1(s)H_2(t)$, then $E\{|H_1(X) - H_2(Y)|\} = \frac{1}{3}$, so the lemma agrees with the mean of D derived above. If X and Y are perfectly correlated (so $H(s,t) = (H_1(s) \wedge H_2(t))$ and have equal margins, the parameter $E|H_1 - H_2| = 0$. If X and Y are perfectly negatively correlated (so $H(s,t) = (H_1(s) + H_2(t) - 1)_+$), then $E|H_1 - H_2| = \frac{1}{2}$.

The test based on D is clearly not consistent (there are marginally uniform variables on the unit square which are dependent but for which $E|X - Y| = \frac{1}{3}$). Lehmann (1966) discusses consistent tests under these assumptions.

2. *Spearman's rank correlation* $S^2(\pi, \sigma) = \Sigma(\pi(i) - \sigma(i))^2$. This metric is the L^2 distance between two permutations. It is right invariant. When transformed to lie in $[-1, 1]$ as in example 1 of Section A, it arises naturally as the correlation R between the ranks of two samples. It is widely used in applied work.

S^2 has mean $(n^3 - n)/6$ and variance $\frac{n^2(n-1)(n+1)^2}{36}$. Normalized by its mean and variance, S^2 has a limiting normal distribution. These results can all be found in Kendall (1970). Normality can be proved using Hoeffding's theorem as above.

EXERCISE 3. Compute S^2 for the draft lottery example in Section A above and test for randomness.

The correlation version R has an interpretation as an estimate of a population parameter. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed pairs drawn from the joint distribution function $H(x, y)$. Then as in the lemma for Spearman's footrule,

$$\frac{S^2}{n^3} = \frac{1}{n^3}\Sigma|R_i - S_i|^2 = E|H_1(X) - H_2(Y)|^2 + 0_p(\frac{1}{\sqrt{n}}).$$

If X and Y are marginally uniform, $E(H_1 - H_2)^2 = 2(\frac{1}{12} - \text{cov}(X, Y))$.

There is a different population interpretation: $R = 1 - \frac{S^2(\pi,\sigma)}{(n^3-n)/3}$ is the sample correlation between the ranks. The expected value of R can be shown to the three times the covariance of $X = \text{sgn}(X_2 - X_1)$ and $Y = \text{sgn}(Y_2 - Y_1)$. This and further interpretations are carefully discussed by Kruskal (1958, Sec. 5.6) and Hoeffding (1948, Sec. 9). Lehmann (1966, Sec. 3) gives some further properties of R.

3. *Hamming distance* $H(\pi, \sigma) = n - \#\{i: \pi(i) = \sigma(i)\}$. Hamming's distance is widely used in coding theory for binary strings. It is a bi-invariant metric on permutations. Following Exercise 1 in Chapter 7, under the uniform distribution $E\{H\} = n - 1$, $\text{Var}\{H\} = 1$, and $n - H$ has a limiting Poisson (1) distribution. These results are all familiar from the probability theory of the matching problem (Feller (1968, pg. 107)). I have shown that the total variation distance between $n - H$ and Poisson (1) is smaller than $2^n/n!$.

The null distribution of H is thus close to its maximum with very little variability. This doesn't mean that H is useless: for instance, in the draft lottery example (section A above) $H(\pi, \sigma) = 9$ which has a p-value of .08.

4. *Kendall's tau* $I(\pi, \sigma) =$ min # *pairwise adjacent transpositions to bring* π^{-1} *to* σ^{-1}. This metric has a long history, summarized in Kruskal (1958, Sec. 17). It was popularized by Kendall who gives a comprehensive discussion in Kendall (1970). The definitions in terms of inverses is given to make the metric right invariant. It has a simple operational form: given π, σ e.g., $\pi = \begin{smallmatrix}1&2&3&4\\3&2&4&1\end{smallmatrix}, \sigma = \begin{smallmatrix}1&2&3&4\\4&3&1&2\end{smallmatrix}$, write them on top of each other, $\begin{smallmatrix}\pi&3&2&4&1\\\sigma&4&3&1&2\end{smallmatrix}$, sort the columns by the top row, $\begin{smallmatrix}1&2&3&4\\2&3&4&1\end{smallmatrix}$, and calculate the number of inversions in the second row (= # pairs $i < j$ with ith entry >jth entry). This is 3 in the example. This number of inversions is also the minimum number of pairwise adjacent transpositions required to bring the 2nd row into order. The letter I is used to represent inversions.

$I(\pi, \sigma)$ has mean $\binom{n}{2}/2$ and variance $n(n-1)(2n+5)/72$. Standardized by its mean and variance I has a standard normal limiting distribution. Kendall (1970) gives tables for small n. An elegant argument for the mean, variance and limiting normality is given in (C-3) below. This also gives fast computational algorithms and correction terms to the normal limit. A second argument is sketched in 5. below.

Kruskal (1958) and Hoeffding (1948) show that the correlation version of I has a sampling interpretation. Using the notation introduced for Spearman's S^2, $E(1-2I/\binom{n}{2})$ is the covariance of $X = \text{sgn}\,(X_2 - X_1)$ and $Y = \text{sgn}\,(Y_2 - Y_1)$.

5. *Cayley's distance* $T(\pi, \sigma) =$ min # *transpositions required to bring* π *to* σ. This is a bi-invariant metric on S_n. It was named after Cayley because he

discovered the simple relationship

$$T(\pi, \sigma) = n - \#\text{ cycles in } (\pi\sigma^{-1}).$$

This is easy to prove. By invariance, take $\sigma =$ id. If π is a k cycle, it takes $k-1$ moves to sort, and disjoint cycles take separate sorting operations.

For the distribution theory, under the null hypothesis the mean is asymptotically $n - \log n$, the variance is asymptotically $\log n$, and T normed by its mean and standard deviation has a limiting standard normal distribution.

These results have an easy derivation. Without loss, take $\sigma =$ id . Sort π by transposing pairs, first switching 1 to place 1, then 2 to place 2, etc. The chance that 1 is already at 1 is $1/n$. Whether or not 1 is switched, after it is in place 1 the relative order of $2, \ldots, n$ is uniform. The chance that 2 does not need to be switched is $1/(n-1)$, and so on. Thus T has the same distribution as

$$X_1 + X_2 + \ldots + X_n$$

with $X_i's$ independent having $P\{X_i = 1\} = 1 - 1/i = 1 - P\{X_i = 0\}$. From here,

$$E(T) = n - (1 + \frac{1}{2} + \ldots + \frac{1}{n}),\ \mathrm{Var}(T) = 1 + \frac{1}{2} + \ldots + \frac{1}{n} - (1 + \frac{1}{2^2} + \ldots + \frac{1}{n^2}).$$

The central limit theorem for sums of independent variables gives the limiting normality. This proof appears in Feller (1968, pg. 257). Section C-3 below gives an algebraic connection.

The same argument works to give the distribution of the number of inversions for Kendall's tau. There the sum is $Y_1 + \ldots + Y_n$, with Y_i uniform on $0, 1, \ldots, i-1$.

EXERCISE 4. Compute Cayley's distance for the Draft Lottery example A-1 and show it *doesn't* reject the null hypothesis.

6. *Ulam's distance* $L(\pi, \sigma) = n -$ *length of longest increasing subsequence in* $\sigma\pi^{-1}$. If $\sigma = \begin{smallmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 2 & 1 & 4 & 5 & 9 & 6 & 7 & 8 \end{smallmatrix}$, the longest increasing subsequence is of length 6 (e.g., $\sigma(3) < \sigma(4) < \sigma(5) < \sigma(7) < \sigma(8) < \sigma(9)$). This natural metric is defined to be right invariant. To motivate it, consider n books on a shelf in order σ. We want to sort the books by deletion-insertion-operations — taking a book out and inserting it in another place. Thus 3 moves are required to sort σ above.

LEMMA 2. *The smallest number of moves to sort* π *is* $n -$ *length of longest increasing subsequence in* π.

Proof. If $\pi(i_1) < \pi(i_2) < \ldots < \pi(i_k)$ is a longest increasing subsequence, then clearly inserting and deleting other letters doesn't change the ordering of this subsequence. It follows that $n - k$ moves suffice. Since each move can increase the longest increasing subsequence by at most 1, $n - k$ moves are required. □

This metric is used by biologists and computer scientists. See Knuth (1978, 5.1.4). Gordon (1983) has suggested it for statistical tasks. If n is large, it is

not so obvious how to compute L in a reasonable amount of time. The following solitaire game gives an efficient algorithm.

Floyd's game. Consider a deck containing n cards labelled $1, 2, \ldots, n$. Shuffle, so the top card is labeled $\pi(1)$, etc. Start to play solitaire (turning cards up one at a time) subject to the following rule: you can only put a lower card on a higher card. If a card is turned up that is higher than the ones on top of piles, it starts a new pile. The object is to have as few piles as possible. Thus, if the deck starts as 6 3 1 5 2 4 7, the game goes

			1		1			1			1			1					
6		3		3		3			3	2		3	2			3	2		
		6		6		6		5	6		5	6	5		4	6	5		4	7

It seems clear that the best strategy is to place a lower card on the smallest card higher than it. We will always assume that the game is played this way.

EXERCISE 5.

(a) Show that the number of piles equals the length of the longest increasing subsequence in π.

(b) Show that the expected number of cards in the first pile is $\log n$ asymptotically, in the 2nd pile $(e-1)\log n$, in the 3rd pile $c \log n$, with $c = \sum_{j=1}^{\infty} [\binom{2j}{j} \frac{1}{j+1} - 1]/j!$. It can be shown that the expected number of cards in the kth pile is of order $\log n$ for fixed k. The remarks below show there are order $2\sqrt{n}$ piles asymptotically.

This game was invented by Bob Floyd (1964). It gives an order $n \log n$ algorithm for finding the longest increasing subsequence. Fredman (1975) shows this is best possible.

The distribution theory of $L(\pi, \sigma)$ is a hard unsolved problem. The mean is asymptotically $n - 2\sqrt{n}$, see Logan and Shepp (1977). The rest of the distribution is unknown. The analysis leads into fascinating areas of group theory; see, e.g. Kerov-Vershik (1985).

C. GENERAL CONSTRUCTIONS OF METRICS.

The preceding section discussed a variety of metrices that have been suggested and used by applied researchers. In this section we give general recipes for constructing metrics on groups. Specialized to the symmetric group, these recapture the examples, and a good deal more.

1. *Matrix norm approach.*

Let G be a finite group. Let $\rho: G \to GL(V)$ be a unitary representation of G which is *faithful* in the sense that if $s \neq t$ then $\rho(s) \neq \rho(t)$. Let $||\ ||$ be a unitarily invariant norm on $GL(V)$. Thus $||AM|| = ||M|| = ||MA||$ for A unitary. Define $d_\rho(s,t) = ||\rho(s) - \rho(t)||$. Observe that this is a bi-invariant metric on G.

Example. Let $||M||^2 = \Sigma_{i,j} M_{ij} \overline{M}_{ij} = Tr(MM^*)$, the sum of squared lengths of the rows. This is unitarily invariant and leads to interesting special cases.

Case 1. Take $G = S_n$. Take ρ as the n-dimensional permutation representation. Then, $d_\rho^2(\mathrm{id}, \pi) = Tr(I - \rho(\pi))(I - \rho(\pi)^T) = Tr(2I - \rho(\pi) - \rho(\pi)^T) = 2H(\mathrm{id}, \pi)$ where H is the Hamming metric, where on the right, d_ρ is the dimension of ρ.

Case 2. For general G and ρ, the argument above shows that characters yield metrics. Thus $d_\rho(s,t) = (d_\rho - re\ \chi_\rho(st^{-1}))^{\frac{1}{2}}$ is a metric, where on the right, d_ρ is the dimension of ρ.

Case 3. Specializing the above to the usual n-dimensional representation of the orthogonal group, $d_\rho(s,t) = (n - Tr(st^{-1}))^{\frac{1}{2}}$ is a metric on O_n. Consider the distance to the identity $d(s, \mathrm{id}) = \sqrt{n - Tr(s)}$. The (i,i) element of s is the cosine of the angle between se_i and e_i, where e_i is the ith basis vector. Thus $d(s, \mathrm{id}) = \{\Sigma 1 - (se_i, e_i)\}^{\frac{1}{2}}$. Since the metric is bi-invariant, it can be expressed in terms of eigenvalues $e^{i\theta_j}$: $d(s, \mathrm{id}) = \{\Sigma(1 - \cos\theta_j)\}^{\frac{1}{2}}$.

Despite these natural properties, and its ease of computation, this is *not* the "natural" metric on O_n. Mathematicians prefer a metric arising from the Riemannian structure on O_n as a Lie group. In terms of the eigenvalues this metric is $\{\Sigma\theta_j^2\}^{\frac{1}{2}}$. See E-5 below.

Case 4. The regular representation R of G gives the discrete metric

$$d_R(s,t) = \begin{cases} 2|G| & \text{if } s \neq t \\ 0 & \text{if } s = t. \end{cases}$$

To determine the distribution of $d_\rho(\mathrm{id}, t)$ requires knowing the distribution of characters. That is, pick t at random on G, and treat $\chi_\rho(t)$ as a random variable. This is a problem that is interesting on its own. It has not been well studied.

EXERCISE 6. Show that $E(\chi_\rho)$ and $E(\chi_\rho - E\chi_\rho)\overline{(\chi_\rho - E\chi_\rho)}$ can be expressed as follows: Let $\chi_\rho = a_1\chi_1 + \ldots + a_h\chi_h$ be a decomposition into irreducibles, with repetitions. If χ_1 is the trivial representation, then $E(\chi_\rho) = a_1$, and $E(\chi_\rho\overline{\chi}_\rho) = a_1^2 + \ldots + a_h^2$. In particular, if ρ is real irreducible, $E(\chi_\rho) = 0, Var(\chi_\rho) = 1$. Find the mean and variance of d_ρ^2 described in Case 4 above.

Remark. Exercise 6 suggests that metrics defined as $(d_\rho - re\ \chi(st^{-1}))^{\frac{1}{2}}$ will not be very "spread out." For real irreducible representations, d_ρ^2 has mean dim ρ and variance one. Nonetheless, they can have interesting distributions. For example $n - H(\mathrm{id}, \pi)$ has a limiting Poisson(1) distribution. Further, the first n moments of $n - H(\mathrm{id}, \pi)$ equal the first n moments of Poisson(1). Similarly, the first $2n + 1$ moments of the trace of a random orthogonal matrix equal the first $2n + 1$ moments of a standard normal variable. Thus, the distance defined for the orthogonal group (Case 3 above) has an approximate standard normal distribution. See Diaconis and Mallows (1985) for these results.

EXERCISE 7. Take G as S_n. Let ρ be the $\binom{n}{2}$ dimensional representation derived by the action of π on the set of unordered pairs $\{i, j\}$. Show that for large n, $\chi_\rho(\pi)$ has as limiting distribution the same distribution as $\frac{X(X-1)}{2} + Y$ where X and Y are independent, X is Poisson(1) and Y is Poisson(1/2).

EXERCISE 8. Compute distances suggested by the discussion above for $G = Z_n$, and Z_2^n. What are the limiting distributions for n large?

All of the above examples used the L^2 or Frobenius norm. There are many other unitarily invariant norms. Indeed, these have been classified by von Neumann (1937). To state his result, define a *symmetric gauge function* as a function $\phi: \mathbb{R}^n \to \mathbb{R}$ satisfying

(a) $\phi(u_1, \ldots, u_n) \geq 0$, ϕ continuous.
(b) $\phi(u_1, \ldots, u_n) = 0$ implies $u_1 = \ldots = u_n = 0$.
(c) $\phi(tu_1, \ldots, tu_n) = t\phi(u_1, \ldots, u_n)$, $t \geq 0$.
(d) $\phi(u_1 + u_1', \ldots u_n + u_n') \leq \phi(u_1, \ldots, u_n) + \phi(u_1', \ldots, u_n')$.
(e) ϕ is invariant under permuting and sign changes of coordinates.

For $M \in GL_n$, let $w_1, \ldots, w_n$ be the eigenvalues of MM^*. Define $\|M\| = \phi(|w_1|^{\frac{1}{2}}, \ldots, |w_n|^{\frac{1}{2}})$. This is a matrix norm: $\|cM\| = |c|\|M\|, \|M + N\| \leq \|M\| + \|N\|$. It is unitarily invariant and $\|M\| = \|M^*\|$.

Von Neumann showed that, conversely, every such norm arises in this way. Examples include

$$\phi = (\Sigma|w_i|^p)^{\frac{1}{p}}, \ \max|w_i|, \text{ or } \{\sum_{i_1 \leq i_2 \cdots \leq i_j} w_{i_1} w_{i_2} \ldots w_{i_j}\}^{\frac{1}{j}}.$$

The first of these, for $p = 2$, becomes the already considered matrix norm $(\Sigma|M_{ij}|^2)^{\frac{1}{2}}$. The second choice becomes the maximum length of Mu subject to $uu^t = 1$. These last two norms also satisfy $\|MN\| \leq \|M\|\|N\|$. It would be instructive to try some of these norms out on the symmetric group.

2. *The fixed vector approach.*

Let G be a group, (ρ, V) a representation. Suppose that V has an inner product, and ρ is unitary. Fix a vector $v \in V$ and define

$$d(s,t) = \|\rho(s^{-1})v - \rho(t^{-1})v\|.$$

This distance has been defined to be right invariant. It clearly satisfies the triangle inequality and symmetry. One must check that $d(\mathrm{id}, t) \neq 0$ unless $t = \mathrm{id}$. It is not even necessary that $\|\|$ come from an inner product. All that is needed is that $\rho(s)$ be norm preserving for $s \in G$.

Example. Take $G = S_n$, ρ the usual n-dimensional representation, so $\rho(\pi^{-1})(v_1, v_2, \ldots, v_n) = (v_{\pi(1)}, v_{\pi(2)}, \ldots, v_{\pi(n)})$. Take $v = (1, 2, \ldots, n)^T$. Then $d^2(\pi, \eta) = \Sigma|\pi(i) - \eta(i)|^2$. If the distance on $\mathbb{R}^n$ is chosen as the L^1 distance, Spearman's footrule results. These considerations emphasize that Spearman's rho and footrule depend on the choice of v. They make it easy to change v to emphasize differences at one end of the scale. The distribution theory of these variants follows from Hoeffding's combinatorial limit theorem. The strong points of the fixed vector approach will become apparent when it is applied to homogeneous spaces in the next section.

3. *Lengths.*

Let G be a finite group. Let S be a subset of G that generates G in the sense that any element can be written as a finite product of elements in S. Assume id $\notin S$ and $S^{-1} = S$. Define the *length* of an element t as the smallest integer $q \geq 0$ such that $t = s_1 s_2 \ldots s_q$ with each $s_i \in S$. Write $q = \ell(t)$. Thus id is the unique element of length zero, and each element of S has length 1.

Define a metric on G by $d(t, u) = \ell(tu^{-1})$.

LEMMA 3. *The length metric $d(t, u) = \ell(tu^{-1})$ is a right invariant metric. It is two-sided invariant if $tSt^{-1} = S$ for every $t \in G$.*

Proof. Clearly, lengths satisfy $\ell(tu) \leq \ell(t) + \ell(u)$, and $\ell(t) = \ell(t^{-1})$. Thus $d(t, u)$ is a right invariant metric. For the last claim, $d(\eta t, \eta u) = \ell(\eta t u^{-1} \eta^{-1}) = \ell(tu^{-1})$ because S is invariant under conjugation by η. □

Example. Take $G = S_n$. If S is chosen as the set of all transpositions one gets the Cayley metric T. Choosing S as the set of transpositions of form $(i, i+1), 1 \leq i \leq n-1$ gives the metric form I of Kendall's tau. To get Ulam's metric L, take S_1 as the set of all cycles $(a, a+1, \ldots, b), 1 \leq a < b \leq n$. Let $S = S_1 \cup S_1^{-1}$. These amount to the basic insertion deletion operations described in example 6 of Section B.

Not all metrics arise this way. For instance, the Hamming distance on S_n is not based on lengths. To see this observe that elements in S have length 1 and two permutations cannot disagree in only one place. The Hamming distance on Z_2^n *is* based on lengths.

There is a curious application of some fairly deep group theory to the distribution theory of length metrics. When specialized, it gives the neat representations of Kendall's and Cayley's distances as sums of independent random variables.

Each of the classical groups (e.g. orthogonal, unitary, symplectic) has associated a finite Weyl group W. A Weyl group is a group W with a set of generators $s_1, s_2, \ldots, s_n$ such that $s_i^2 =$ id and for some integers $n_{ij}, (s_i, s_j)^{n_{ij}} =$ id, these being the only relations. For example, S_n with generators $(i, i+1)$ is a Weyl group; n_{ij} being 2 if the generators are disjoint and 3 otherwise. The sign change group (permute coordinates and change signs arbitrarily) is another familiar Weyl group.

Modern motivation for studying these groups comes from Lie theory and combinatorics. Bourbaki (1968) and Stanley (1980) are readable surveys.

Let (W, S) be a Weyl group. Let $F(t) = \Sigma_{w \in W} t^{\ell(w)}$ be the generating function of the lengths. A basic theorem in the subject states that there exist an m and integers e_i called the exponents of W such that

$$F(t) = \prod_{i=1}^{m} (1 + t + \ldots t^{e_i}).$$

Letting $t = 1$, this shows $|W| = \Pi(e_i + 1)$. Dividing both sides by $|W|$, we have the following.

COROLLARY 1. *Let (W, S) be a Weyl group with exponents e_i. Then the length of a random $w \in W$ has the distribution of $X_1 + \ldots + X_m$, where the X_i are chosen as independent uniform variables on $\{0, 1, 2, \ldots, e_i\}$.*

The factorization can be found as Exercise 10 of Section 4 in Bourbaki (1968) or Stanley (1980). As a convolution of symmetric unimodal distributions, $P\{\ell(w) = j\}$ is a symmetric unimodal sequence as j varies.

As an example, on S_n with pairwise adjacent transpositions as generators, the exponents are $e_i = i - 1$ for $i = 1, 2, \ldots, n$, and the factorization becomes the representation of the number of inversions as a sum of uniforms discussed under Cayley's distance in Section B above.

There is a second general theorem of the same type. Let (W, S) be a Weyl group. Take $\overline{S} = \{tSt^{-1}\}$ as a new set of generators obtained by closing up the old set under conjugation. This gives a new length function, say $\overline{\ell}(w)$. It is an amazing fact that the generating function of $\overline{\ell}$ factors as

$$(*) \qquad \sum_w t^{\overline{\ell}(w)} = \prod_{i=1}^{m} (1 + e_i t).$$

COROLLARY 2. *Let (W, S) be a Weyl group with exponents e_i. Then the length $\overline{\ell}$ of a random $w \in W$ has the distribution of $X_1 + \ldots + X_m$ where X_i are independent with $P\{X_i = 0\} = 1/1 + e_i, P\{X_i = 1\} = e_i/1 + e_i$.*

The factorization $(*)$ was proven by Coxeter and Shephard Todd. See Solomon (1963) or Proposition 4.6 in Stanley (1979).

These representations make the means and variances of $d(s, t)$ easy to compute. They also make the distribution easy to work with: sums of independent uniforms have an easy limit theory, with correction terms readily available. Further, Harding (1984) shows how such factorizations lead to an easy algorithm for fast exact computation of distributions in small cases.

Of course, in the case of Cayley's distance or Kendall's tau, the representations are well known in statistics. In the next section we show how a similar factorization holds for the natural extension of these metrics to homogeneous spaces.

EXERCISE 9. Consider Hamming distance on Z_2^n. Show its length generating function factors as $(1 + t)^n$.

An Application. Here is an application of the factorizations in Corollaries 1 and 2 above. Consider Monte Carlo generation of a sample from the Mallows model (example 3 of Section A) based on the metric I of Section B:

$$* \qquad P_\lambda(\pi) = C(\lambda) e^{-\lambda I(\pi, \pi_0)}.$$

We begin by recalling a correspondence between permutations and sequences. Let $(a_1, \ldots, a_n)$ be a sequence of integers $0 \le a_i \le i - 1$. Associate a permutation by insertion; starting with $n, n-1, n-2, \ldots$ insert $n - i + 1$ so it has a_i previously

inserted numbers to its left. Thus, if $n = 7$, the sequence $(0, 0, 1, 3, 2, 3, 6)$ develops as

$$7 \to 67 \to 657 \to 6574 \to 65374 \to 653274 \to 6532741.$$

The final permutation has $a_1 + \ldots + a_n$ inversions (here 15). This gives a 1-1 correspondence between permutations and sequences, with the sum of the sequence equal to the number of inversions. The correspondence is equivalent to the Weyl group factorization of Corollary 1 above.

If the initial sequence is chosen uniformly: $0 \le a_i \le i - 1$, then a random permutation results. If $P\{a_i = j\} = e^{-\lambda j}[\frac{e^{-\lambda}-1}{e^{-\lambda i}-1}]$, $0 \le j \le i - 1$, the final permutation has probability $*$ with $\pi_0 = \text{id}$. The distribution of a_i is easy to generate by inversion (Chapter III.2 of Devroye (1986)).

It is easy to modify things to incorporate π_0, or to work for any other metric with a similar factorization.

Fligner and Verducci (1986, 1988b) have pointed out that the normalizing constant $C(\lambda)$ in $*$ is known from the factorization in Corollary 1. They apply this in doing maximum likelihood estimation and as a way of extending the models. Steele (1987) discusses some other combinatorial problems where similar factorizations arise.

D. Metrics on homogeneous spaces.

Most of the considerations of previous sections can be generalized to homogeneous spaces. Let X be a homogeneous space on which a group G operates from the right, transitively. Fix $\overline{y}_0 \in X$, let $K = \{s \in G : \overline{y}_0 s = \overline{y}_0\}$. In this section X will be identfied with *right* cosets of K in G, $X \cong \{Kx_i\}$ where $\text{id} = x_0, x_1, \ldots, x_j \in G$ are coset representatives for K in G (so $G = K \cup Kx_1 \ldots \cup Kx_j$ as a disjoint union). Here G acts (from the right) on cosets by $\overline{x}s = (Kx)s = Kxs$ for any $s \in G$ and any $\overline{x} = Kx \in X$.

We have made a slight change of notation (from left to right cosets) to agree with the notation in Critchlow (1985). Critchlow's monograph develops a host of metrics for partially ranked data. He gives numerous applications, computer programs, and tables for popular cases. It is very readable and highly recommended.

There are several ways to choose a metric on X which is right-invariant in the sense that $d(\overline{x}, \overline{y}) = d(\overline{x}s, \overline{y}s)$, i.e. $d(Kx, Ky) = d(Kxs, Kys)$.

a) *Hausdorff metrics.* Let G be a compact group, K a closed subgroup and d a metric on G. Let X be a space on which G acts transitively with isotropy subgroup K. Write $X = G/K$ to denote the representation of X by right cosets.

A metric d^* is induced on G/K by the formula

$$d^*(\overline{x}, \overline{y}) = d^*(Kx, Ky) = \max(a, b)$$

with

$$a = \max_{s \in Kx} \min_{t \in Ky} d(s, t), \quad b = \max_{s \in Ky} \min_{t \in Kx} d(s, t).$$

The metric d^* is the Hausdorff distance between the sets Kx and Ky — the smallest amount that each must be "blown up" to include the others. It is a

standard way to metrize the homogeneous space X, see e.g., Dieudonne (1970, pg. 53), Nadler (1978), or Roelcke and Dierolf (1981).

EXERCISE 10.
(a) Show that d^* is a metric.
(b) If d is right invariant then so is d^*.
(c) If d is left invariant, then $d^*(Kx, Ky) = \min_{k \in K} d(x, ky)$.

The definition of d^* seems more theoretically than practically useful — it seems hard to explicitly compute the minimum. However, Critchlow (1985) has given reasonably elegant closed form expressions for partially ranked data and d any of the classical metrics of Section B. Some of his results will be given here.

Example 1. k sets of an n set. Let $\overline{x}$ and $\overline{y}$ be k element subsets of $\{1, 2, \ldots, n\}$. Note $\overline{x}$ and $\overline{y}$ can be identified with points in the homogeneous space $S_n/(S_k \times S_{n-k})$, where $S_k \times S_{n-k}$ is the subgroup $\{\pi \in S_n : \pi(i) \leq k \ \forall i = 1, \ldots, k$ and $\pi(i) > k \ \forall i = k+1, \ldots, n\}$. Let H be the Hamming distance on the symmetric group S_n. Then the induced Hausdorff metric is

$$H^*(\overline{x}, \overline{y}) = 2(k - |\overline{x} \cap \overline{y}|).$$

To see this, realize $\overline{x}$ and $\overline{y}$ as ordered sets $x_1 < \ldots < x_k, y_1 < \ldots < y_k$. Associate permutations x and y to $\overline{x}$ and $\overline{y}$ by choosing coset representatives. Since $H(x, y) = H(x^{-1}, y^{-1})$, the permutations can be taken as

$$x = \begin{pmatrix} x_1 x_2 & \ldots & x_k & x'_1 & \ldots & x'_{n-k} \\ 1\ 2 & \ldots & k & k+1 & \ldots & n \end{pmatrix}$$

$$y = \begin{pmatrix} y_1 y_2 & \ldots & y_k & y'_1 & \ldots & y'_{n-k} \\ 1\ 2 & \ldots & k & k+1 & \ldots & n \end{pmatrix}$$

Now using part c) of the exercise above

$$H^*(\overline{x}, \overline{y}) = \min_{\pi \in S_k \times S_{n-k}} H(x, \pi y).$$

Multiplying on the left by π allows us to permute the y_i with $i \in \{1, \ldots, k\}$ among themselves and the $y_{i'} \in \{k+1, \ldots, n\}$ among themselves in the first row of y. This permits matching elements and proves the result.

The null distribution of $|\overline{x} \cap \overline{y}|$ is the well known hypergeometric distribution.

Example 2. Rank k out of n. Here people rank order their favorite k out of n, in order. Represent a ranking as $(x_1, x_2, \ldots, x_k)$ where x_1 is the item ranked first, x_2 is the item ranked second, etc. Critchlow (1985, Chapter 3) shows

$$H^*(\overline{x}, \overline{y}) = \#\{i \leq k : x_i \neq y_i\} + (k - |\overline{x} \cap \overline{y}|).$$

Again, this is a very reasonable distance, albeit, perhaps, a bit crude. Critchlow gives similar explicit, interpretable formulas for the induced Hausdorff distances derived from the other classical metrics.

Example 3. The n-sphere. Using the distance $d^2(s,t) = n - Tr(st^{-1})$ on the orthogonal group, then choosing reflections $I - 2xx^t$ as coset representatives (for x on the unit sphere), leads to

$$d^*(\overline{x},\overline{y}) = \sqrt{1 - <\overline{x}|\overline{y}>} = \frac{1}{\sqrt{2}}\|\overline{x} - \overline{y}\|.$$

Diaconis and Shahshahani (1983, Sec. 3) discuss the choice of coset representatives more carefully. If $\overline{x}$ or $\overline{y}$ is chosen at random, $\sqrt{n}(d^2 - 1)$ is approximately standard normal for large n. This last result is proved with good error bounds in Diaconis and Freedman (1987).

b) *The fixed vector approach.* Here is another large class of invariant metrics on a homogeneous space $X = G/K$. Let (ρ, V) be any unitary representation of G. Say ρ has a K *fixed-vector* $v \in V$ if $\rho(k)v = v$ for every $k \in K$. Usually it is easy to find such a ρ and v, see the examples below. It follows from Chapter 3F that ρ has a K fixed vector if and only if ρ appears in the decomposition of $L(X)$. Define a metric on X by

$$d_\rho(\overline{x},\overline{y}) = d_\rho(Kx, Ky) = \|(\rho(x^{-1}) - \rho(y^{-1}))v\|.$$

Note that this is well defined (it is independent of the choice of coset representatives). Note further that this distance is right G-invariant:

$$\begin{aligned} d_\rho(\overline{x}s, \overline{y}s) &= d_\rho(Kxs, Kys) = \|\rho(s^{-1})[\rho(x^{-1}) - \rho(y^{-1})]v\| \\ &= d_\rho(\overline{x},\overline{y}), \end{aligned}$$

because ρ is unitary. This d_ρ clearly satisfies the properties of a metric except perhaps for $d_\rho(\overline{x},\overline{y}) = 0$ implying $\overline{x} = \overline{y}$. This must be checked separately. The fixed vector approach was suggested by Andrew Rukhin as a way to choose loss functions in statistical problems on groups.

Example 1. k sets of an n set. For the $\binom{n}{k}$ k-element subsets of $\{1, 2, \ldots, n\}$, choose ρ as the usual n-dimensional representation on $\mathbb{R}^n$ with the usual inner product. Take $v = (a, \ldots, a, b, \ldots, b)$ with a run of k a's followed by $n - k$ b's. Choosing coset representatives as the reverse shuffles of example 1 above yields

$$d_\rho(x, y) = |a - b|\sqrt{2}(k - |\overline{x} \cap \overline{y}|)^{\frac{1}{2}}$$

Cf. Example 1 of the Hausdorff approach.

Again, Critchlow (1985) gives a variety of results, giving extensions of Spearman's footrule and rho to partially ranked data.

Example 2. The n-sphere. Take $X = S^n$ $G = 0_n$, $K = 0_{n-1}$. Take ρ as the usual n-dimensional representation of 0_n, and $e_1 = (10\ldots0)^t$ as a K-fixed vector. Finally take coset representatives as $I - 2vv^t$ where $v = (e_1 + x)/c, c = |e_1 + x|$, and x runs over S^n. An easy computation yields $d^2(x, y) = \|x - y\|^2$.

Constructions a, b, make it clear that there are a wealth of tractable metrics on homogeneous spaces. Critchlow gives examples and applications carrying

over much of the material of Section A to partially ranked data. He has subsequently developed many further applications to standard nonparametric problems as remarked in Example 13 of Section A.

There are a reasonable number of nice distributional problems open — the null distribution of metrics on homogeneous spaces needs to be better developed. The following special case hints at what's lurking there.

Example. A metric on partially ranked data. Consider six flavors, a, b, c, d, e, f. Suppose two rankers rank them, choosing their two favorites, and two least favorite, not distinguishing within:

$$(1) \qquad \begin{matrix} a & b & c & d & e & f \\ 1 & 2 & 1 & 2 & 3 & 3 \end{matrix} \qquad \begin{matrix} a & b & c & d & e & f \\ 1 & 1 & 3 & 3 & 2 & 2 \end{matrix}$$

How close are these ranks? It is natural to try the minimum number of pairwise adjacent transpositions it takes to bring one bottom row to the other. This is 5 in the example. Recall however that the labelling of the top row is arbitrary. The two arrays could just as easily have been presented with first and last columns switched. This yields

$$\begin{matrix} f & b & c & d & e & a \\ 3 & 2 & 1 & 2 & 3 & 1 \end{matrix} \qquad \begin{matrix} f & b & c & d & e & a \\ 2 & 1 & 3 & 3 & 2 & 1 \end{matrix}$$

These are the same rankings, but now their distance is 3.

A simple way to have invariance rearranges the two rankings in order of (say) the first, and then computes inversions. Thus (1) becomes

$$\begin{matrix} a & c & b & d & e & f \\ 1 & 1 & 2 & 2 & 3 & 3 \end{matrix} \qquad \begin{matrix} a & c & b & d & e & f \\ 1 & 3 & 1 & 3 & 2 & 2 \end{matrix} \qquad \#\text{ inversions } = 5.$$

If we had sorted by the 2nd ranking (1) becomes

$$\begin{matrix} a & b & e & f & c & d \\ 1 & 2 & 3 & 3 & 1 & 2 \end{matrix} \qquad \begin{matrix} a & b & e & f & c & d \\ 1 & 1 & 2 & 2 & 3 & 3 \end{matrix} \qquad \#\text{ inversions } = 5.$$

This example has $n = 6$, and partial rankings of shape 222. More generally,

Definition. Let λ be a partition of n. Let π and η be partial rankings of shape λ. Define $I(\pi, \eta)$ as follows: arrange the columns of π and η so that π is in order, beginning with λ_1 ones, λ_2 twos, etc. This must be done using the minimum number of pairwise adjacent transpositions. Then count the minimum number of pairwise adjacent transpositions required to bring the 2nd row of η into order.

EXERCISE 11. Show that I is a right invariant metric.

One reason for working with the metric I is the following elegant closed form expression for its null distribution. By right invariance, this only needs to be computed for $I(id, \pi) \stackrel{\mathrm{d}}{=} I(\pi)$.

Theorem 2. *Let $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_r\}$ be a partition of n. Let π range over the $n!/\Pi\lambda_i!$ partial rankings of shape λ. Then*

$$\Sigma_\pi \, q^{I(\pi)} = \left(\!\!\binom{n}{\lambda_1 \lambda_2 \ldots \lambda_r}\!\!\right) = \frac{((n))!}{((\lambda_1))! \ldots ((\lambda_r))!}$$

where $((\lambda))! = ((\lambda-1))((\lambda-2))\ldots((1))$ *with* $((j)) = 1 + q + q^2 + \ldots + q^{j-1}$.

Remarks. Theorem 2 was proved by Netto when $r = 2$, and by Carlitz in the general case. Stanley (1985, Proposition 1.3.17) presents several elementary proofs. Stanley (1980) proves that the coefficients $P\{I(\pi) = j\}$ are symmetric and unimodal. The factorization and unimodality generalize to other Weyl groups. The expressions on the right side are known as q-nomial coefficients.

Fligner and Verducci (1986, 1988b) use this factorization as a base for extending and interpreting Mallows model on partially ranked data.

Note that $((\lambda))!/\lambda!$ is the generating function of the convolution of λ independent uniform variables $U_1 + \ldots + U_\lambda$ with U_i uniform on $\{0, 1, 2, \ldots, i-1\}$. This gives an easy way to compute means, variances, and asymptotic distributions where necessary. The following neat argument evolved in work with Andy Gleason and Ali Rejali. For clarity, it is given for $\lambda = (k, n-k)$.

The null distribution can be described this way: let x be a pattern of k ones and $(n-k)$ twos. Let $I(x)$ be the number of inversions (e.g. 2121 has 3 inversions). For x chosen at random, the generating function of $I(x)$ satisfies

$$\frac{1}{\binom{n}{k}} \Sigma_x \, q^{I(x)} = \frac{((\binom{n}{k}))}{\binom{n}{k}}.$$

Rearranging, the right hand side is

$$\frac{g_n(q) g_{n-1}(q) \cdots g_{n-k+1}(q)}{g_k(q) g_{k-1}(q) \cdots g_2(q)},$$

with $g_j(q) = \frac{1}{j}(1 + q + \ldots + q^{j-1})$, the generating function of U_j - a uniform random variable on $\{0, 1, 2, \ldots j-1\}$. This has mean $= \mu_j = \frac{j-1}{2}$ and variance $\sigma_j^2 = \frac{j^2-1}{12}$.

Cross-multiplying, the identity has the probabilistic interpretation

$$I + U_2 + U_3 + \ldots + U_k \stackrel{D}{=} U_n + U_{n-1} + \ldots + U_{n-k+1},$$

where the D means equality in distribution. All of the uniform variables are independent. From this we have

PROPOSITION 1.

a) $E(I) = \mu_n + \ldots + \mu_{n-k+1} - \mu_k - \mu_{k-1} \cdots - \mu_2 = \frac{k(n-k)}{2}$

b) $Var(I) = \sigma_n^2 + \ldots + \sigma_{n-k+1}^2 - \sigma_k^2 - \sigma_{k-1}^2 \cdots - \sigma_2^2 = \frac{k(n+1)(n-k)}{12}$

c) As n and k tend to infinity in any way, provided $n-k$ also tends to infinity, the distribution of I, standardized by its mean and standard deviation, has a standard normal limit.

Proof. The mean and variance are derived in the remarks preceding the statement. For the distribution, suppose without loss that $k \geq n/2$. Write the distributional identity as $I + \overline{U}_k = \overline{U}_{n-k}$. Then standardize

$$\left\{\frac{I - \mu_I}{\sigma_I}\right\}\frac{\sigma_I}{\overline{\sigma}_{n-k}} + \left\{\frac{\overline{U}_k - \overline{\mu}_k}{\overline{\sigma}_k}\right\}\frac{\overline{\sigma}_k}{\overline{\sigma}_{n-k}} \stackrel{D}{=} \frac{\overline{U}_{n-k} - \overline{\mu}_{n-k}}{\overline{\sigma}_{n-k}}.$$

The right side converges to a standard normal distribution, as does $\{\frac{\overline{U}_k - \overline{\mu}_k}{\overline{\sigma}_k}\}$. Since this last is independent of $\{\frac{I-\mu_I}{\sigma_I}\}$, it must be that $\{\frac{I-\mu_I}{\sigma_I}\}$ converges, and by Cramer's theorem, to a standard normal. □

Remark 1. The argument above works for virtually any type of partition, in particular $1^q, (n-q)^1$ — for rankings of q out of n.

Remark 2. The proof is similar to the standard argument for the Mann-Whitney statistic given in Kendall and Stuart (1967, pg. 505).

Remark 3. The generating function is a *ratio* of generating functions. We took advantage of this by cross-multiplying. That is different from having a direct probabilistic interpretation. Indeed, I do not know how to generate random partial rankings from the associated Mallows model as suggested for full rankings at the end of the last section. Fligner and Verducci (1988b, Sec. 3.2) have made some progress here.

E. Some philosophy.

We have seen examples and applications of metrics. We pause for a moment to reflect on the big picture. What makes a natural metric; how can we compare metrics? Important issues here are

1) *Interpretability.* Is the metric easy to think about directly, easy to explain to a non-professional? Does it measure something with real world significance such as the actual number of steps required to sort, or the running time of an algorithm, or the cost of an error?

Along these lines, observe that Cayley's, Kendall's tau, and Ulam's metric have sorting interpretations. The footrule, Kendall's tau, and Spearman's rho have a statistical interpretation as estimates of population parameters.

2) *Tractability.* Is the metric easy to compute? The footrule, Hamming and rho are trivial to program, Cayley and tau require a bit of thought, and Ulam's metric can be tricky if n is large. Is the null distribution available for small samples? Are useful asymptotics available? Ulam's metric fares badly here — its asymptotic distribution is unknown. Of course, null distributions can always be simulated.

3) *Invariance.* In the application, is right or left invariance natural and available?

4) *Sensitivity.* Does the metric effectively use its range or does it just take on a few values? Among two sided invariant metrics this is a problem. Worst is the discrete metric ($d(s,t) = 0$ or 1 as $s = t$ or not). Next is Hamming distance, which effectively takes on only a few values around n under the uniform distribution. Finally, Cayley's distance takes about $\sqrt{\log n}$ values effectively. It should be possible to find bi-invariant metrics that naturally take on more values. Since variance can be changed by multiplication by a constant, perhaps the limiting coefficient of variation μ/σ should be used to measure effective range.

5) *Available theory.* Has the metric been studied and used enough so that its strengths and pitfalls are known? Does it link into other aspects of the analysis?

A nice example arises for continuous groups. Mathematicians seem to agree on a unique bi-invariant way of metrizing Lie groups such as the orthogonal group. When pushed "what makes *that* metric so natural?" they respond with theorems like "there is a unique differential (smooth except at id, like $|x|$) bi-invariant metric compatible with the Riemannian structure." See Milnor (1976, Lemma 7.6). Metrics can sometimes be derived from axioms (as in Example A-8).

6) The bottom line. There is a fairly harsh test: did somebody actually use the metric in a real application? Was it used to prove a theorem? Could this have been done without the metric just as easily? Failing this, does the metric lead to interesting theoretical questions or results?

A first pass through this list suggests Kendall's tau as the metric of choice. It's easy to interpret and explain, having both an algorithmic and statistical interpretation. It's highly tractable because of the factorization. It's been well studied, tabled for small values of n, and widely used. It's quite sensitive in the coefficient of variation scale, and links into nice mathematics. It also has a natural extension to partially ranked data. The bottom line judgement is left to the reader.

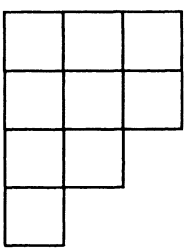
Chapter 7. Representation Theory of the Symmetric Group

We have already built three irreducible representations of the symmetric group: the trivial, alternating and $n-1$ dimensional representations in Chapter 2. In this chapter we build the remaining representations and develop some of their properties.

To motivate the general construction, consider the space X of the unordered pairs $\{i,j\}$ of cardinality $\binom{n}{2}$. The symmetric group acts on these pairs by $\pi\{i,j\} = \{\pi(i),\pi(j)\}$. The permutation representation generated by this action can be described as an $\binom{n}{2}$ dimensional vector space spanned by basis vectors $e_{\{i,j\}}$. This space splits into three irreducibles: A one-dimensional trivial representation is spanned by $\overline{v} = \Sigma\, e_{\{i,j\}}$. An $n-1$ dimensional space is spanned by $v_i = \Sigma_j\, e_{\{i,j\}} - c\overline{v}, 1 \le i \le n$, with c chosen so v_i is orthogonal to $\overline{v}$. The complement of these two spaces is also an irreducible representation. A direct argument for these assertions is given at the end of Section A. The arguments generalize. The following treatment follows the first few sections of James (1978) quite closely. Chapter 7 in James and Kerber (1981) is another presentation.

A. Construction of the irreducible representations of the symmetric group.

There are various definitions relating to diagrams, tableaux, and tabloids. Let $\lambda = (\lambda_1, \ldots, \lambda_r)$ be a partition of n. Thus, $\lambda_1 \ge \lambda_2 \ldots \ge \lambda_r$ and $\lambda_1 + \ldots + \lambda_r = n$. The *diagram* of λ is an ordered set of boxes with λ_i boxes in row i. If $\lambda = (3,3,2,1)$, the diagram is



If λ and μ are partitions of n we say λ *dominates* μ, and write $\lambda \trianglerighteq \mu$, provided that $\lambda_1 \ge \mu_1, \lambda_1 + \lambda_2 \ge \mu_1 + \mu_2, \ldots$, etc. This partial order is widely used in various areas of mathematics. It is sometimes called the order of *majorization.* There is a book length treatment of this order by Marshall and Olkin (1979). They show that $\lambda \trianglerighteq \mu$ if and only if we can move from the diagram of λ to the diagram of μ by moving blocks from the right hand edge upward, one at a time, such that at each stage the resulting configuration is the diagram of a partition. Thus, $(4,2) \triangleright (3,3)$, but $(3,3)$, and $(4,1,1)$ are not comparable. See Hazewinkel and Martin (1983) for many novel applications of the order.

A λ-*tableau* is an array of integers obtained by placing the integers from 1 through n into the diagram for λ. Clearly there are $n!$ λ-tableaux.

The following lemma is basic:

LEMMA 0. *Let λ and μ be partitions of n, suppose that t_1 is a λ-tableau and t_2 is a μ-tableau. Suppose that for each i the numbers from the ith row of t_2 belong to different columns of t_1. Then $\lambda \trianglerighteq \mu$.*

Proof. Since the numbers in the first row of t_2 are in different columns of t_1, $\lambda_1 \geq \mu_1$. The numbers in the second row of t_2 are in distinct columns of t_1, so no column of t_1 can have more than two of the numbers in the first or second row of t_2. Imagine "sliding these numbers up to the top of the columns of t_1." They fit in the first two rows, so $\lambda_1 + \lambda_2 \geq \mu_1 + \mu_2$. In general, no column of t_1 can have more than i of the numbers from the first i rows of t_2. □

If t is a tableau, its *column-stabilizer* C_t is the subgroup of S_n keeping the columns of t fixed. For example, when $t = \begin{matrix} 1 & 2 & 4 & 5 \\ 3 & 6 & 7 \\ 8 \end{matrix}$, $C_t \cong S_{\{138\}} \times S_{\{26\}} \times S_{\{47\}} \times S_{\{5\}}$. The notation $S_{\{i,j,\ldots,k\}}$ means the subgroup of S_n permuting only the integers in brackets.

Define an equivalence relation on the set of λ-tableaux by considering $t_1 \sim t_2$ if each row of t_1 contains the same numbers as the corresponding row of t_2. The *tabloid* $\{t\}$ is the equivalence class of t. Think of a tabloid as a "tableau with unordered row entries." The permutations operate on the tabloids in the obvious way. The action is transitive and the subgroup stablizing the tabloid with $1, \ldots, \lambda_1$ in the first row, $\lambda_1 + 1, \ldots, \lambda_1 + \lambda_2$, in the second row, etc., is

$$S_{\{1,2,\ldots,\lambda_1\}} \times S_{\{\lambda_1+1,\ldots,\lambda_1+\lambda_2\}} \times \cdots.$$

It follows that there are $n!/\lambda_1! \ldots \lambda_r!$ λ-tabloids.

Define the *permutation representation* associated to the action of S_n on tabloids as a vector space with basis $e_{\{t\}}$. It is denoted M^λ. This representation is reducible but contains the irreducible representation we are after. To get this, define for each tableau t a *polytabloid* $e_t \in M^\lambda$ by

$$e_t = \sum_{\pi \in C_t} \operatorname{sgn}(\pi) e_{\pi\{t\}}.$$

Check that $\pi e_t = e_{\pi t}$, so the subspace of M^λ spanned by the $\{e_t\}$ is invariant under S_n (and generated as an "S_n module" by any e_t). It is called the *Specht module* S^λ. The object of the next collection of lemmas is to prove that S^λ is irreducible and that all the irreducible representations of S_n arise this way. These lemmas are all from Section 4 of James.

LEMMA 1. *Let λ and μ be partitions of n. Suppose that t is a λ tableau and s is a μ-tableau. Suppose that*

$$\sum_{\pi \in C_t} sgn(\pi) e_{\pi\{s\}} \neq 0.$$

Then $\lambda \trianglerighteq \mu$, and if $\lambda = \mu$, the sum equals $\pm e_t$.

Proof. Suppose for some a, b that a and b are in the same row of s and in the same column of t. Then

$$(\mathrm{id} - (ab))e_{\{s\}} = e_{\{s\}} - e_{\{s\}} = 0.$$

Since a and b are in the same column of t, the group $< \mathrm{id}, (ab) >$ is a subgroup of C_t. Let $\sigma_1, \ldots, \sigma_k$ be coset representatives, so $\sum_{\pi \in C_t} \mathrm{sgn}(\pi) e_{\pi\{s\}} = \sum_{i=1}^{k} \mathrm{sgn}(\sigma_i)\sigma_i\{\mathrm{id} - (ab)\}e_{\{s\}} = 0$. This is ruled out by hypothesis, so the numbers in the ith row of s are in different columns of t. Lemma 0 implies that $\lambda \trianglerighteq \mu$.

Suppose $\lambda = \mu$, and the sum does not vanish; then, again, numbers in the ith row of s appear in different columns of t. It follows that for a unique $\pi^* \in C_t, \pi^*\{t\} = \{s\}$ and this implies that the sum equals $\pm e_t$ (replace $\{s\}$ by $\pi^*\{t\}$ in the sum). □

LEMMA 2. *Let $\mu \in M^\mu$, and let t be a μ tableau. Then for some scalar c*

$$\sum_{\pi \in C_t} sgn(\pi)\pi u = c\, e_t.$$

Proof. u is a linear combination of $e_{\{s\}}$. For $u = e_{\{s\}}$, Lemma 1 gives the result with $c = 0$ or ± 1. □

Now put an inner product on M^μ which makes $e_{\{s\}}$ orthonormal: $< e_{\{s\}}, e_{\{t\}} > = 1$ if $\{s\} = \{t\}$ and 0 otherwise. This is S_n invariant. Consider the "operator" $A_t = \sum_{\pi \in C_t} \mathrm{sgn}(\pi)\pi$. For any $u, v \in M^\mu$, $< A_t u, v > = \Sigma \mathrm{sgn}\, \pi < \pi u, v > = \Sigma \mathrm{sgn}\, \pi < u, \pi^{-1} v > = < u, A_t v >$. Using this inner product we get:

LEMMA 3. *(Submodule Theorem). Let U be an invariant subspace of M^λ. Then either $U \supset S^\lambda$ or $U \subset S^{\lambda\perp}$. In particular, S^λ is irreducible.*

Proof. Suppose $u \in U$ and t is a λ-tableau. By Lemma 2, $A_t u$ is a constant times e_t. If we can choose u and t such that this constant is non-zero, then $e_t \in U$ and, since $\pi e_t = e_{\pi t}$, $S^\lambda \subset U$. If $A_t u = 0$ for all t and u, then $0 = < A_t u, e_{\{t\}} > = < u, A_t\, e_{\{t\}} > = < u, e_t >$. So $U \subset S^{\lambda\perp}$. □

At this stage we have one irreducible representation for each partition λ of n. The number of irreducible representations is the same as the number of conjugacy classes: see Theorem 7 of Chapter 2. This number is also the number of partitions of n as explained at the beginning of Chapter 2D. Hence, if we can show that the S^λ are all inequivalent, we have finished determining all of the irreducible representations of S_n.

LEMMA 4. *Let $T: M^\lambda \to M^\mu$ be a linear map that commutes with the action of S_n. Suppose that $S^\lambda \not\subset \ker T$. Then $\lambda \trianglerighteq \mu$. If $\lambda = \mu$, then the restriction of T to S^λ is a scalar multiple of id.*

Proof. By lemma 3, $\operatorname{Ker} T \subset S^{\lambda\perp}$. Thus for any $t, 0 \neq Te_t = TA_t e_{\{t\}} = A_t T e_{\{t\}}$. But $Te_{\{t\}}$ is a linear combination of μ-tabloids $e_{\{s\}}$ and for at least one such $e_{\{s\}}$, $A_t\, e_{\{s\}} \neq 0$. By Lemma 1, $\lambda \trianglerighteq \mu$. If $\lambda = \mu$, then $Te_t = c\, e_t$ by the same argument. □

LEMMA 5. *Let $T: S^\lambda \to S^\mu$ be a linear map that commutes with the action of S_n. If $T \neq 0$, $\lambda \trianglerighteq \mu$.*

Proof. Any such T can be extended to a linear map from M^λ to M^μ by defining T to be 0 on $S^{\lambda\perp}$. The extended map commutes with the action of S_n. If $T \neq 0$, then Lemma 4 implies $\lambda \trianglerighteq \mu$. □

Theorem 1. *The S^λ are all of the irreducible representations of S_n.*

Proof. If S^λ is equivalent to S^μ, then, using Lemma 5 in both directions, $\lambda = \mu$. □

Remark. The argument for Lemma 4 shows that the irreducible representations in M^μ are S^μ (once) and some of $\{S^\lambda: \lambda \trianglerighteq \mu\}$ (possibly with repeats). In fact, S^λ occurs in M^μ if and only if $\lambda \trianglerighteq \mu$.

To complete this section, here is a direct proof of the decomposition of $M^{n-2,2}$ discussed in the introductory paragraph to this chapter. We begin with a broadly applicable result.

A USEFUL FACT.

Let G be a finite group acting on a set X. Extend the action to the product space X^k coordinatewise. The number of fixed points of the element $s \in G$ is $F(s) = |\{x: sx = x\}|$. For any positive integer k:

(1) $\frac{1}{|G|} \sum_s F(s)^k = |$ orbits of G acting on $X^k|$.

(2) Let R, V be the permutation representation associated to X. Thus V has as basis δ_x and $R_s(\delta_x) = \delta_{sx}$. The character of this representation is $\chi_R(s) = F(s)$. If R decomposes into irreducibles as $R = m_1\rho_1 \oplus \ldots \oplus m_n\rho_n$; then

$$\Sigma m_i^2 = |\text{ orbits of } G \text{ acting on } X^2|.$$

Proof. For (1) we have the action of G on X^k given by $s(x_1, \ldots, x_k) = (sx_1, \ldots, sx_k)$. Let C_i be a decomposition of X^k into G orbits. Then

$$\sum_s f(s)^k = \sum_s \sum_{x_1} \delta_{sx_1}(x_1) \ldots \sum_{x_k} \delta_{sx_k}(x_k) = \sum_{X^k} \sum_s \delta_{s\underset{\sim}{x}}(\underset{\sim}{x})$$

$$= \sum_i \sum_{\underset{\sim}{x} \in C_i} \sum_s \delta_{s\underset{\sim}{x}}(\underset{\sim}{x}).$$

The innermost sum is the cardinality of the stabilizer of G at $\underset{\sim}{x}$: $|N_{\underset{\sim}{x}}|$ with $N_{\underset{\sim}{x}} = \{s: s\underset{\sim}{x} = \underset{\sim}{x}\}$. Observe $N_{s\underset{\sim}{x}} = s\, N_{\underset{\sim}{x}} s^{-1}$. In particular, the size of $N_{\underset{\sim}{x}}$ doesn't depend on the choice of $\underset{\sim}{x}$ in a given orbit. Since $|G| = |N_{\underset{\sim}{x}}|\ |C_i|$ the inner sum equals $|G|/|C_i|$. The sum over $\underset{\sim}{x} \in C_i$ multiplies this by $|C_i|$. The final sum yields $|G| \cdot |\text{Orbits}|$ as required. To prove (2), we use the orthogonality of characters: $\chi_R = m_1\chi_1 + \ldots + m_n\chi_n$ so $< \chi_R|\chi_R >= m_1^2 + \ldots + m_n^2$. On the other hand, it is clear $\chi_R(s) = F(s)$ and $F(s^{-1}) = F(s)$, so $< \chi_R|\chi_R >= \frac{1}{|G|}\Sigma F(s)^2$. □

REMARKS AND APPLICATIONS

(a) With $k = 1$, part (1) is called Burnside's lemma. It is at the heart of Serre's exercise 2.6 which we have found so useful. It also forms the basis of the Polya-Redfield "theory of counting." See e.g., de Bruijn (1964).

(b) If G acts doubly transitively on X, then there are two orbits of G acting on $X \times X$: $\{(x,x)\}$ and $\{(x,y): y \neq x\}$. It follows that V decomposes into two irreducible components: One of these is clearly spanned by the constants. Thus its complement $\{v: \Sigma v_i = 0\}$ is irreducible.

(c) When G acts on itself we get back the decomposition of the regular representation.

(d) There is an amusing connection with probability problems. If G is considered as a probability space under the uniform distribution U, then $F(s)$ is a "random variable" corresponding to "pick an element of G at random and count how many fixed points it has." When $G = S_n$ and $X = \{1, 2, \ldots, n\}$, $F(\pi)$ is the number of fixed points of π. We know that this has an approximate Poisson distribution with mean 1. Part (1) gives a "formula" for all the moments of $F(g)$.

EXERCISE 1. Using (1), prove that the first n moments of $F(\pi)$ equal the first n moments of Poisson(1), where π is chosen at random on S_n.

(e) Let us decompose $M^{n-2,2}$. The space X is the set of unordered pairs $\{i,j\}$ with $\pi\{i,j\} = \{\pi(i), \pi(j)\}$. The permutation representation has dimension $\binom{n}{2}$. There are 3 orbits of S_n acting on $X \times X$ corresponding to pairs $\{i,j\}, \{k,\ell\}$ with 0, 1, or 2 integers in common. Thus, clearly S_n acts transitively on the set of pairs $(\{i,j\}, \{i,j\})$. Also for $(\{i,j\}, \{j,\ell\})$ $\ell \neq i,j$ and for $(\{i,j\}, \{k,\ell\})$ with $\{k,\ell\} \cap \{i,j\} = \phi$. It follows that V splits into 3 irreducible subspaces. These are, the 1-dimensional space spanned by $\overline{v} = \Sigma e_{\{ij\}}$, the $n-1$-dimensional space spanned by $\overline{v}_i = \Sigma_j e_{\{ij\}} - c\overline{v}$ $1 \leq i \leq n$, and the complement of these two spaces. Clearly, the space spanned by $\overline{v}$ gives the trivial representation and the space spanned by $\overline{v}_i$ gives the $n-1$ dimensional representation. What is left is $n(n-3)/2$ dimensional. If we regard the permutation representation as the set of all functions on X with $sf(x) = f(s^{-1}x)$, then the trivial and $n-1$ dimensional representations are the set of functions of form $f\{i,j\} = f_1(i) + f_1(j)$.

EXERCISE 2. Show that for fixed j, $0 \leq j \leq n/2$, $M^{n-j,j}$ splits into $j+1$ distinct irreducible representations, the ith having dimension $\binom{n}{i} - \binom{n}{i-1}$. Hint: use the useful fact and induction, (e) above is the case $j = 2$.

We can build some new irreducible representations directly by tensoring the representation we know about with the alternating representation. Tensoring the alternating representation with the $n-1$ dimensional representation always gives a different irreducible representation. For $n = 4$ we already have all irreducible representations: 2 of 1 dimension, 2 of 3 dimensions and 1 of dimension $n(n-3)/2 = 2$. The sum of squares adds to 24. For $n > 4$ (but not $n = 4$) the $n(n-3)/2$ dimensional representation yields a new irreducible representation of the same dimension. For $n = 5$ this gives all the irreducible representations but 1. We can build this by considering the action of S_n on ordered pairs (i, j). That is, $M^{3,1,1}$.

B. More on representations of S_n

The books by James and James-Kerber are full of interesting and useful facts. Here is a brief description of some of the most useful ones, along with pointers to other work on representations of S_n.

(1) *The Standard Basis of* S^λ. We have defined S^λ as the representation of M^λ generated by elements e_t. There are $n!$ different e_t and the dimension of S^λ can be quite small. For example, if $\lambda = (n-1, 1)$, we know S^λ is $n-1$ dimensional. It turns out that a few of the e_t generate S^λ. Define t to be a *standard tableau* if the numbers increase along the rows and down the columns. Thus $\begin{array}{|c|c|c|}\hline 1 & 3 & 5 \\ \hline 2 & 4 \\ \cline{1-2}\end{array}$ is a standard $[3, 2]$ tableau. There is only 1 standard (n) tableau. There are $n-1$ standard $(n-1, 1)$ tableaux. In Section 8, James proves that $\{e_t | t$ is a standard λ-tableau$\}$ is a basis for S^μ. This is a beautiful result, but not so helpful in "really understanding S^λ." What one wants is a set of objects on which S_n acts that are comprehensible. The graphs in Section 5 of James are potentially very useful in this regard for small n. As far as I know, a "concrete" determination of the representations of S_n is an open problem. See (6) below.

(2) *The Dimension of* S^λ. There are a number of formulas for the dimension (and other values of the character) of the representation associated to λ. The dimensions get fairly large; they are bounded by $\sqrt{n!}$ of course, but they get quite large:

n	2	3	4	5	5	7	8	9	10
max dim	1	2	3	6	16	35	90	216	768

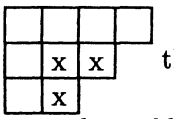
We know that $\dim(S^\lambda)$ equals the number of ways of placing the numbers $1, \ldots, n$ into the Young diagram for λ in such a way that the numbers increase along rows and down columns. From this follows bounds like the following which was so useful in Chapter 3:

$$\dim(S^\lambda) \leq \binom{n}{\lambda_1} \sqrt{(n-\lambda_1)!}\,.$$

There is a classical determinant formula (James, Corollary 19.5)

$$\dim(S^\lambda) = n!\ \mathrm{Det}|1/(\lambda_i - i + j)!|, \text{ where } 1/r! = 0 \text{ if } r < 0.$$

Finally, there is the hook formula for dimensions. Let λ be a partition of n. The (i, j) *hook* is that part of the Young diagram that starts at (i, j) and goes as far

as it can either down or to the right. Thus, if $\lambda = (4,3,2)$,  the (2,2) hook is indicated by $x's$. The length of the (i,j) hook is the number of boxes in the hook. Using this terminology, the hook length formula says

$$\dim(S^\lambda) = n!/\text{product of hook lengths in } \lambda.$$

For example, when $\lambda = (4,3,2)$, the hook lengths are

6	5	3	1
4	3	1	
2	1		

The dimension is $9!/6! \cdot 3 = 168$.

The dimension of $S^{(n-1,1)}$ is $n-1$. The dimension of $S^{11\cdots1}$ is 1.

Greene, Nijenhuis, and Wilf (1979, 1984) give an elegant, elementary proof of the hook length formula involving a random "hook walk" on a board of shape λ.

Hooks come into several other parts of representation theory – in particular, the Murnaghan-Nakayama rule for calculating the value of a character (section 21 of James).

(3) *Characters of the Symmetric Group.* To begin, we acknowledge a sad fact: there is no reasonable formula for the character $\chi_\lambda(\mu)$ where λ is a partition of n, χ_λ the associated irreducible character of S_n, and μ stands for a conjugacy class of S_n. This is countered by several facts.

(a) For small n (≤ 15) the characters have been explicitly tabulated. James-Kerber (1981) give tables for $n \leq 10$ and references for larger n.

(b) For large n there are useful asymptotic results for $\chi_\lambda(\mu)$. These are clearly explained in Flatto, Odlyzko, and Wales (1985).

(c) For any specific λ and μ there is an efficient algorithm for calculating the character called the Murnaghan-Nakayama rule. Section 21 of James (1978) or Theorem 2.4.7 of James-Kerber (1981) give details.

EXERCISE 3. Define a probability Q on S_n as follows: with probability p_n choose the identity with probability $1-p_n$ choose a random n-cycle. Determine the rate of convergence. How should p_n be chosen to make this as fast as possible? Hint: See (d) below.

(d) For "small" conjugacy classes μ, and arbitrary n and λ, there are formulas like Frobenius' formula used in Chapter 3D. Ingram (1950) gives useful references. See also Formula 2.3.17 in James-Kerber (1981).

(e) For some special shapes of λ, such as hooks $\lambda = (k,1,1,\ldots,1)$, closed form formulas are known, see e.g. Stanley (1983) or Macdonald (1979).

(f) There are also available rather intractable generating functions for the characters due to Frobenius. This analytic machinery is nicely presented in Chapter 1.7 of Macdonald (1979).

(4) *The Branching Theorem* (Section 9 of James). Consider ρ the $n-1$ dimensional representation of S_n. Let S_{n-1} be considered as a subgroup of S_n (all permutations that fix 1). Then ρ is a representation of S_{n-1}, "by restriction" James writes $S^{(n-1,1)} \downarrow S_{n-1}$. Observe that ρ restricted to S_{n-1} is reducible. If we choose the basis $e_1 - e_2, e_1 - e_3, \ldots, e_1 - e_n$; then the sum of the basis elements generates a one-dimensional invariant subspace. Since S_{n-1} operates doubly transitively on the basis elements, we have $\rho \downarrow S_{n-1}$ splitting into two irreducible subspaces; one of dimension 1 and one of dimension $n-2$.

The branching theorem gives the general result on how $S^\mu \downarrow S_{n-1}$ decomposes: there is one irreducible representation for each way of removing a "box" from the right hand side of the Young diagram for μ in such a way that the resulting configuration is a diagram for a partition. Thus, the diagram for $[n-1,1]$ can be reduced to $(n-1)$ or $(n-2,1)$ and these are the two irreducible components. The branching theorem is used to give a fast Fourier transform for computing all $\hat{f}(\rho)$ in Diaconis and Rockmore (1988).

EXERCISE 4. (Flatto, Odlyzko, Wales). Let ρ be an irreducible representation of S_n. Show that ρ restricted to S_{n-1} splits in a multiplicity free way. Using this, show that if P is a probability on S_n that is invariant under conjugation by S_{n-1} (so $P(\pi) = P(\sigma\pi\sigma^{-1})$ for $\sigma \in S_{n-1}$), then $\hat{P}(\rho)$ is diagonal for an appropriate basis which does not depend on P.

(5) *Young's Rule.* This gives a way to determine which irreducible subspaces occur in the decomposition of M^λ. It will be extremely useful in Chapter 8 in dealing with partially ordered data "in configuration λ." For example, data of the form "pick the best m of n" can be regarded as a vector in $M^{(n-m,m)}$, the components being the number of people who picked the subset corresponding to the second row of the associated tabloid. The decomposition of $M^{(n-m,m)}$ into irreducibles gives us a spectral decomposition of the frequencies and a nested sequence of models. See Chapter 8B and 9A.

Young's rule depends on the notion of semi-standard tableaux. This allows repeated numbers to be placed in a diagram. Let λ and μ be partitions of n. A *semi-standard tableau* of *shape* λ and *type* μ is a placement of integers $\leq n$ into a Young tableau of shape λ, with numbers nondecreasing in rows and strictly increasing down columns, such that the number i occurs μ_i times. Thus, if $\lambda = (4,1)$ and $\mu = (2,2,1)$, there are two tableaux of shape λ and type μ:

$$\begin{array}{cccccccc} 1 & 1 & 2 & 2 & \quad & 1 & 1 & 2 & 3 \\ 3 & & & & & 2 & & & \end{array}$$

Young's Rule: The multiplicity of S^λ in M^μ equals the number of semi-standard λ tableaux of type μ. As an example, consider, for $m \leq n/2$, $\mu = (n-m, m)$. We are decomposing M^μ. The possible shapes λ are

$$\begin{array}{c} \overbrace{1\,1\ldots1}^{n-m}\overbrace{2\ldots2}^{m},\overbrace{1\,1\ldots1}^{n-m}\overbrace{2\ldots2}^{m-1},\ldots \\ \hline 2 \end{array} \qquad \begin{array}{c} 1\ 1\ \ldots\ 1\ 1\ 1\ \ldots\ 1 \\ 2\ 2\ \ldots\ 2\ 2 \end{array}$$

Each occurs once only. Thus $M^{(n-m,m)} = S^{(n)} \oplus S^{(n-1,1)} \oplus S^{(n-2,2)} \oplus \ldots \oplus S^{(n-m,m)}$. By induction dim $S^{(n-m,m)} = \binom{n}{m} - \binom{n}{m-1}$. When we translate this decomposition into an interpretation for "best m out of n" data, the subspaces $S^{(n-m,m)}$ have interpretations:

S^n – The grand mean or # of people in sample.

$S^{n-1,1}$ – The effect of item i, $1 \leq i \leq n$.

$S^{n-2,2}$ – The effect of items $\{i, j\}$ adjusted for the effect of i and j.

$\vdots$

$S^{n-k,k}$ – The effect of a subset of k items adjusted for lower order effects.

Remarks. Many further examples of Young's rule appear in Chapter 8. Young's rule does not give an algorithm for decomposing M^μ or interpreting the S^μ. It just says what pieces appear. Section 17 of James (1978) solves both of these problems in a computationally useful way. This remark is applied in Chapter 8C below.

Young's rule is a special case of the Littlewood-Richardson rule which describes how a given representation of S_n restricts to the subgroup $S_k \times S_{n-k}$. See James and Kerber (1981, Sec. 2.8).

(6) *Kazhdan-Lusztig Representations.* The construction of the irreducible representations given in Section A constructs S^λ as a rather complicated subspace of the highly interpretable M^λ. Even using the standard basis ((1) above), S^λ is spanned by the mysterious Young symmetrizers e_t. It is desirable to have a more concrete combinatorial object on which the symmetric group acts, with associated permutation representation isomorphic to S^λ. An exciting step in this direction appears in Kazhdan and Lusztig (1979). They construct graphs on which S_n acts to give S^λ. For $n \leq 6$, these graphs are available in useful form. Kazhdan and Lusztig construct these representations as part of a unified study of Coxeter groups. The details involve an excursion into very high-powered homology. Garsia and McLarnan (1988) gives as close to an "in English" discussion as is currently available, showing the connections between Kazdahn and Lusztig's representations and Young's natural representation as developed in Chapter 3 of James-Kerber.

(7) *The Robinson-Schensted-Knuth (RSK) Correspondence.* There is a fascinating connection between the representation theory of S_n and a host of problems of interest to probabilists, statisticians, and combinatorialists centered about the R-S-K correspondence. The connected problems include sweeping generalizations of the ballot problem: if one puts λ_1-ones, λ_2-twos, ..., $\lambda_k - k$'s into an urn and draws without replacement, where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_k$ is a partition of n, then the chance that # ones $\geq$ # two $\geq \ldots \geq$ # k's at each stage of the drawing equals $f(\lambda)/n!$ where $f(\lambda) = \dim(S^\lambda)$ discussed in (2) above. This links into formulas for the coverage of Kolmogorov-Smirnov tests, the distribution of the longest increasing subsequence in a permutation, and much else.

The connection centers around a 1-1 onto map $\pi \to (P, Q)$ between S_n and pairs of standard Young tableaux of the same shape. Since there are $f(\lambda)$ of

these tableaux of shape λ, we have an explicit interpretation of the formula $n! = \Sigma_\lambda f(\lambda)^2$.

One route to accessing this material starts with Section 5.1.4 in Knuth (1975). Then try Stanley (1971), then some of the papers in Kung (1982). Narayana (1979) gives pointers to some statistical applications. Kerov and Virshik (1985) give applications to statistical analysis of other aspects of random permutations. White (1983) discusses the connection between the R-S-K correspondence and the Littlewood-Richardson rule.

Chapter 8. Spectral Analysis

Often data are presented as a function $f(x)$ defined on some index set X. If X is connected to a group, the function f can be "Fourier expanded" and one may try to interpret its coefficients. This is a rich idea which includes the usual spectral analysis of time series and the analysis of variance.

This chapter develops the idea in stages. First, for data on groups (time series, binary vectors, and permutation data). Then the idea is developed for data on homogeneous spaces (the sphere and partially ranked data). Next some theory needed for practical computation is derived. All of this is illustrated for some classical designs in the analysis of variance. Finally, some research projects are spelled out.

A. Data on groups

1. *Time series analysis.* Fourier analysis of time series or other signals is a familiar scientific activity. For example, data on the number of babies born daily in New York City over a five year period are studied by Izenman and Zabell (1978). Here $X = \{1, 2, \ldots, n\}$ with $n = 5 \times 365 + 1$. The data are represented as a function $f(x) = \#$ born on day x. Inspection of this data shows strong periodic phenomena: about 450 babies are born on each week day and about 350 on each day of the weekend. (Physicians don't like to work on weekends.) There might also be monthly or quarterly effects.

To examine (and discover) such phenomena, scientists pass from the original data $f(x)$ to its Fourier transform

$$\hat{f}(y) = \Sigma_x f(x)\, e^{2\pi i x y/n}$$

where the sum runs over $x = 0, 1, \ldots, n-1$. Fourier inversion gives

$$f(x) = \frac{1}{n}\Sigma_y \hat{f}(y)\, e^{-2\pi i x y/n}.$$

It sometimes happens that a few values of $\hat{f}(y)$ are much larger than the rest and determine f in the sense that f is closely approximated by the function defined by using only the large Fourier coefficients in the inversion formula. When this happens, we have f approximated by a few simple periodic functions of x, e.g. $e^{-2\pi i x y/n}$, and may feel we understand the situation.

The hunting and interpretation of periodicities is one use of spectral analysis. A splendid introduction to this subject is given by Bloomfield (1976). A more advanced treatment from the same point of view is given by Brillinger (1975).

There are other interpretations of spectral analysis. The discussion papers by Jenkins (1961) and Parzen (1961) present a useful survey of different views. Some further discussion is given in the last section of this chapter. For now, we will stick to the data analytic view outlined above.

2. *Permutation data.* Spectral analysis can be carried out for any group using the matrix entries of the irreducible representations as a basis. Before developing this in general, here is an example.

In Chapter 5-A we discussed rankings of three items. People were asked to rank where they wanted to live: in a city, suburbs, or country. The rankings were

π	city	suburbs	country	#
id	1	2	3	242
(23)	1	3	2	28
(12)	2	1	3	170
(132)	3	1	2	628
(123)	2	3	1	12
(13)	3	2	1	359

Here $X = S_3$, and $f(\pi)$ is the number choosing π. There are three irreducible representations of S_3: the trivial, sgn, and two-dimensional representation ρ. The Fourier inversion theorem gives

$$f(\pi) = \frac{1}{6}\{\hat{f}(\text{triv}) + \text{sgn}(\pi) \cdot \hat{f}(\text{sgn}) + 2Tr(\rho(\pi^{-1})\hat{f}(\rho)).$$

Expanding the trace gives a spectral analysis of f as a sum of orthogonal functions.

To facilitate comparison between functions in this basis, let us choose an orthogonal version of ρ. Thus, using cycle notation on S_3

π	id	(1 2)	(2 3)	(1 3)	(1 2 3)	(1 3 2)
$\rho(\pi)$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} 1 & \sqrt{3} \\ \sqrt{3} & -1 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} 1 & -\sqrt{3} \\ -\sqrt{3} & -1 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} -1 & -\sqrt{3} \\ \sqrt{3} & -1 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} -1 & \sqrt{3} \\ -\sqrt{3} & -1 \end{pmatrix}$

These are arrived at by choosing $w_1 = \frac{1}{\sqrt{2}}(e_1 - e_2)$, $w_2 = \frac{1}{\sqrt{6}}(e_1 + e_2 - 2e_3)$ as an orthonormal basis for $\{v \in \mathbb{R}^3 : v_1 + v_2 + v_3 = 0\}$. The matrices $\rho(\pi)$ give the action of π in this basis. Now

$$\hat{f}(\text{triv}) = 1439,\ \hat{f}(\text{sgn}) = 242 - 28 - 170 + 628 + 12 - 359 = 325,$$
$$\hat{f}(\rho) = \begin{pmatrix} -54.5 & 285\sqrt{3}/2 \\ -947\sqrt{3}/2 & -101.5 \end{pmatrix}.$$

Define four functions on S_3 by

$$\sqrt{2}\ \rho(\pi^{-1}) = \begin{pmatrix} a(\pi) & b(\pi) \\ c(\pi) & d(\pi) \end{pmatrix}.$$

With this definition, the functions 1, sgn π, $a(\pi)$, $b(\pi)$, $c(\pi)$, $d(\pi)$ are orthogonal and have the same length.

Expanding the trace in the Fourier inversion theorem gives

$$f(\pi) = \frac{1}{6}\{1439 + 325\ \mathrm{sgn}(\pi) - 54.5\sqrt{2}\ a(\pi) - 947\sqrt{3/2}\ b(\pi) + 285\sqrt{3/2}\ c(\pi) - 101.5\sqrt{2}\ d(\pi)\}.$$

$$\doteq \frac{1}{6}\{1439 + 325\ \mathrm{sgn}(\pi) - 77\ a(\pi) - 1160\ b(\pi) + 349\ c(\pi) - 144\ d(\pi)\}.$$

As a check, when $\pi = \mathrm{id}$, this becomes

$$242 = \frac{1}{6}\{1439 + 325 - 109 - 203\}.$$

The largest non-constant coefficient, 1160, multiplies $b(\pi)$. This is the function

π	id	(1 2)	(2 3)	(1 3)	(1 2 3)	(1 3 2)
$b(\pi)$	0	0	$\sqrt{3/2}$	$-\sqrt{3/2}$	$\sqrt{3/2}$	$-\sqrt{3/2}$

or

$$b(\pi) = \begin{cases} -\sqrt{3/2} & \text{if cities are ranked 3rd } (\pi(1) = 3) \\ 0 & \text{if country is ranked 3rd } (\pi(3) = 3) \\ \sqrt{3/2} & \text{if suburbs are ranked 3rd } (\pi(2) = 3). \end{cases}$$

Spectral analysis gives fresh insight into this little data set: After the constant, the best single predictor of f is what people rank last. Now $b(\pi)$ enters with a *negative* coefficient. It "follows" that people "hate" the city most, the suburbs least and the country in between. Going back to the data $\#\{\pi(1) = 3\} = 981$, $\#\{\pi(2) = 3\} = 40$, $\#\{\pi(3) = 3\} = 412$, so the effect is real. It seems at variance with the unfolding hypothesis, but is in fact necessary. If people make individual rankings by unfolding, the proportions connected to extreme ranks will be monotone with the unfolding point in the center.

The kind of data analytic interpretation given in the last paragraph seems mandatory – we must seek subject matter interpretation of our findings. As a word of warning, almost any set of summary statistics can have a story woven about them – we are good at making up stories.

EXERCISE 1. Carry out the analysis of this section for the non-orthogonal representation of S_3 given in Chapter 2-A. Show that the main conclusions do not change.

3. *A general case, with inference.* For G a finite group, and f a function on G, the Fourier inversion theorem gives

$$f(s) = \frac{1}{|G|}\Sigma_\rho\ d_\rho\ Tr(\rho(s^{-1})\hat{f}(\rho)).$$

Now let ρ be a unitary representation, so $\rho(t)^* = \rho(t^{-1})$. Corollaries 2 and 3 to Schurs lemma of Chapter 2 yield orthogonality relations for the matrix entries of ρ. Define the usual inner product on all functions on G: $< \phi|\psi > = \frac{1}{|G|}\Sigma\, \phi(t)\psi(t)^*$. Then

$< \rho_{ij}|\eta_{k\ell} > = 0$ if ρ and η are inequivalent unitary representations, for any $ij, k\ell$.

$$< \rho_{ij}|\rho_{k\ell} > = \begin{cases} 0 & \text{unless } i = k \text{ and } j = \ell, \\ \frac{1}{d_\rho} & \text{if } i = k,\ j = \ell. \end{cases}$$

It follows that the functions $\tilde{\rho}_{ij}(s) = \sqrt{d_\rho}\ \rho_{ij}(s^{-1})$ are orthonormal on G with respect to $< \cdot|\cdot >$.

To numerically compute the spectral representation, compute $\hat{f}(\rho)$ and expand the trace giving

$$f(s) = \frac{1}{|G|}\sum_{\rho,i,j} \sqrt{d_\rho}\ \hat{f}(\rho)_{ji}\ \tilde{\rho}_{ij}(s).$$

The squared length of the projection indexed by functions associated to ρ is $d_\rho\ Tr(\hat{f}(\rho)\hat{f}(\rho)^*) = d_\rho\|\hat{f}(\rho)\|^2$.

Elementary inference. In carrying out spectral analysis, it is natural to wonder "if the data had come out a bit different, would the inferences change?" This is susceptible to a wealth of interpretations – from an elementary sensitivity analysis, through a bootstrap analysis, through a frequentist model through a Bayesian analysis. At present, very little is available in off the shelf tools. We here develop the obvious normal theory. Some more speculative suggestions are contained in the final section of this chapter.

Let G be a finite group. Suppose that an observed function $f(s)$ can be regarded as a true function $\mu(s)$ plus an error or perturbation function $\varepsilon(s)$

$$f(s) = \mu(s) + \varepsilon(s).$$

The strongest possible assumptions that can be made are $\mu(s)$ fixed (or known) and $\varepsilon(s) \sim N(0, \sigma^2)$ (normal, with mean 0 and variance σ^2 independent for each s). Then for an orthogonal representation ρ, the coefficients $\hat{f}(\rho)_{ij}\sqrt{d_\rho}$ are all independent normals, with mean $\sqrt{d_\rho}\hat{\mu}(\rho)_{ij}$ and variance $\sigma^2|G|$. (I have assumed that all of the representations are real. For unitary representations, complex normal distributions occur). Further, for ρ and η inequivalent representations, $\hat{f}(\rho)_{ij}$ and $\hat{f}(\eta)_{k\ell}$ are independent.

EXERCISE 2. Prove the results in the last paragraph.

If σ^2 is assumed known, or $\hat{\mu}(\rho) = 0$ is assumed for some irreducible so σ^2 can be estimated, all of the usual inferential tools associated to the general linear model are available. In particular, $d_\rho\|\hat{f}(\rho)\|^2$ is distributed as σ^2 times

a chi-squared variable on d_ρ^2 degrees of freedom (if $\hat{\mu}(\rho) = 0$). Fisher's test for the significance of the largest Fourier coefficient (see e.g. Bloomfield (1976) or Anderson (1971)) is easily adapted to this setting.

If σ^2 is assumed to depend on s, a rich array of tools from variance components and classical multivariate analysis are available. Work of the Danish school is particularly relevant – see Perlman (1988) for entry into this literature.

The normality of the Fourier transforms is approximately true under much less restrictive assumptions than normality of $f(s)$. After all, $\hat{f}(\rho)_{ij}$ is an average of a lot of things, and will be approximately normal and independent of other coefficients under fairly mild assumptions. This deserves to be worked out more carefully. Anderson (1971, Sec. 2.6), Freedman and Lane (1980), or Diaconis and Freedman (1984) are useful places to start.

This seems like a good place to point out

Elementary fact. Let ρ be an irreducible unitary representation of G. Let $L(G)$ be all functions on G. Consider the subspace L_j spanned by the matrix entries of the jth column of ρ. Then L_j is an invariant subspace isomorphic to ρ.

Proof. Let $s\ \rho_{ij}(t) = \rho_{ij}(s^{-1}t)$. The matrix equation $\rho(s^{-1})\rho(t) = \rho(s^{-1}t)$ shows that $\rho_{ij}(s^{-1}t) = \Sigma_k\ a_k\rho_{kj}(t)$ as a function of t. Thus L_j is an invariant subspace. Choosing the functions ρ_{ij}, $1 \leq i \leq d_\rho$ as a basis gives $\rho(s)$ being the associated matrix representation. □

Remarks. As j varies, the representations L_j give the d_ρ copies of ρ that appear in the decomposition of the regular representation. In applications, the columns of ρ sometimes have a natural meaning. For example, on S_n the jth column of the $n - 1$-dimensional representation codes which i is mapped to position j in one choice of basis. Expanding the trace preserves this interpretation. Further discussion of interpretability of the basis functions appears in the last section of this chapter.

4. *Bahadur's item analysis.* Bahadur (1961) introduced a spectral analysis when $X = Z_2^k$ consists of data on binary k-tuples. For example, fix k at 5 and consider a population of families with five children. Take $f(x)$ as the proportion of families whose children were born in birth order x – so if $x = 01010$, the birth order was girl, boy, girl, boy, girl. Klotz (1970) gives an analysis of such data.

Economics has massive amounts of "panel study data." Here k might be 12 and x might represent the pattern of employed/unemployed behavior for a person in the study. It is not uncommon to have samples larger than 5,000 people. Hsiao (1986) is a recent book on this subject.

Bahadur's original motivation was test score analysis, where x records the pattern of correct/incorrect guesses, and $f(x)$ records the proportion of students answering in a given pattern.

Let x_i, $1 \leq i \leq k$ be the coordinate projection from Z_2^k. Let $f(x)$ be a probability on Z_2^k. Define $\alpha_i = E_f(x_i)$ and $z_i = (x_i - \alpha_i)/\sqrt{\alpha_i(1-\alpha_i)}$. Define

$$r_{ij} = E_f(z_i z_j),\ i < j$$
$$r_{ij\ell} = E_f(z_i z_j z_\ell),\ i < j < \ell$$
$$\vdots$$
$$r_{1\ldots k} = E_f(z_1 z_2 \ldots z_k).$$

The $\binom{k}{2}$ parameters r_{ij} are correlations. The $r_{ij\ell}$ can be thought of higher order correlations, etc. Bahadur wrote the spectral decomposition of f as follows.

PROPOSITION 1. *(Bahadur). For f a probability on Z_2^k,*

$$f = f_1 \cdot f_2$$

where

$$f_1(x) = \Pi\ \alpha_i^{x_i}(1-\alpha_i)^{1-x_i} \text{ is the product distribution with margins matching } f.$$

$$f_2(x) = 1 + \sum_{i<j} r_{ij} z_i z_j + \sum_{i<j<\ell} r_{ij\ell} z_i z_j z_\ell + \ldots + r_{1\ldots k} z_1 \ldots z_k.$$

Proof. Consider the vector space of all functions on Z_2^k with inner product $< g, h >= E_{f_1}(g \cdot h)$. The set

$$S = \{1; z_1, z_2, \ldots, z_k;\ z_1 z_2, \ldots, z_{k-1} z_k;\ z_1 z_2 z_3, \ldots; \ldots z_1 \ldots z_k\}$$

consists of orthonormal functions: $\|g\| = 1$ for $g \in S$ and $< g, h >= 0$ for $g, h \in S$ but $g \neq h$. There are 2^k functions in S, so they form a basis. Now, set $f_2 = f/f_1$. Then $< f_2, g >= \Sigma f_2 g f_1 = E_f(g)$. Also $E_f(1) = 1, E_f(z_i) = 0$, so the proposition follows. □

The function f_2 measures departure from independence. It is natural to look at the individual coefficients to try to understand the nature of the dependence. Let $\delta_{(k)}^2 = \|f_2 - 1\|^2$. Thus define

$$\delta_{(k)}^2 = \sum_{i<j} r_{ij}^2 + \sum_{i<j<\ell} r_{ij\ell}^2 + \ldots + r_{1\ldots k}^2$$
$$\stackrel{\mathrm{d}}{=} \delta_2^2 + \ldots + \delta_k^2.$$

The ratios $\delta_j^2/\delta_{(k)}^2$ provide an index of the relative importance of the jth order terms. Similarly, if f_m is the approximation to f in which all terms in f_2 involving a product of more than m z_i's are omitted, then

$$\|f_m/f_1\|^2/\|f/f_1\|^2 = \frac{1 + \delta_1^2 + \ldots + \delta_m^2}{1 + \delta_1^2 + \ldots + \delta_k}$$

provides an index of quality for the approximation of f by f_m.

The analysis above is just the standard spectral analysis shifted to a more natural base measure (matched marginals versus uniform). As Bahadur remarks, it is also natural to expand and analyze $\log f$. This gets away from the problems of interpreting negative coefficients and links into the usual log-linear analysis of such data as a $2 \times 2 \times \ldots \times 2$ contingency table.

B. Data on homogeneous space

A natural and useful extension of spectral analysis occurs for homogeneous spaces. These were defined and illustrated in Chapter 3F.

1. *Definitions.* Let X be a finite set. Let G act transitively on X with N the isotropy subgroup: $N = \{s \in G: sx_0 = x_0\}$ where x_0 is some fixed point in X. Let $L(X)$ be all functions from X into the complex numbers C. Then $L(X)$ decomposes into irreducibles as

$$V_0 \oplus V_1 \oplus \ldots \oplus V_k.$$

Suppose we are given $f(x) \in L(X)$, regarded as data – $f(x)$ is the number (or proportion) of people having property x.

Spectral analysis is the decomposition of $f(x)$ into its projections on the irreducible invariant subspaces of $L(X)$, and the approximation of $f(x)$ by as small a number of projections as give a reasonable fit.

Here $L(X)$ is regarded as an inner product space using $< f|g > = \frac{1}{|X|}\Sigma f(x)g(x)^*$, and projections are orthogonal. Of course we usually want even more: the coefficients of f projected into V_i in some natural or interpretable basis help connect the analysis to the original subject matter.

As shown in Corollary 1 of Section 3 below the decomposition of the regular representation falls into this domain. The following are less standard.

Example 2. Partially ranked data. Let λ be a partition of n. Consider data consisting of partial rankings of n items of shape λ: thus people rank their favorite λ_1 items (but not within) and their next λ_2 items (but not within) and their final λ_k items (but not within). Here $n = \lambda_1 + \lambda_2 + \ldots + \lambda_k$ and we do not assume λ_i are ordered. Chapter 5B gives examples.

If $S_{\lambda_1} \times S_{\lambda_2} \times \ldots \times S_{\lambda_k}$ denotes the subgroup of S_n which allows permutations among the first λ_1 coordinates, the next λ_2 coordinates, and so on, then $X = S_n/S_{\lambda_1} \times \ldots \times S_{\lambda_k}$. The space $L(X)$ can be taken as all real valued functions on X because all irreducible representations are real. Thus $L(X) = M^\lambda$ of Chapter 7. The decomposition of $L(X)$ is given by Young's rule (see Chapter 7-B). Here are some special cases.

Case 1. $\lambda = (n-1, 1)$. This is simple choice data, people choosing one out of n items. The set X may be regarded as $\{1, 2, \ldots, n\}$ and $f(x)$ is the number of people choosing x. The decomposition is

$$L(X) = S^n \oplus S^{n-1,1}$$

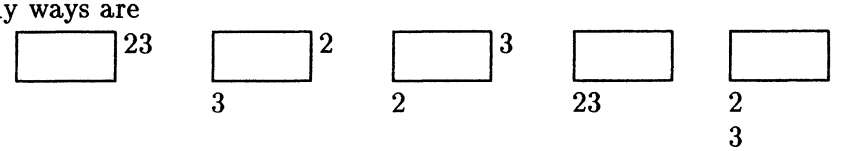
with S^n the trivial representation and $S^{n-1,1}$ the $n-1$ dimensional representation. This amounts to $f(i) = \overline{f} + (f(i) - \overline{f})$ with $\overline{f} = \frac{1}{n}\Sigma f(i)$.

Case 2. $\lambda = (n-2,1,1)$. Here people pick their favorite and second favorite items from a list of n (order matters). The space X may be regarded as $\{(i,j),\ 1 \le i \ne j \le n\}$ so $|X| = n(n-1)$. The decomposition of $L(X)$ is

$$L(X) = S^n \oplus 2S^{n-1,1} \oplus S^{n-2,2} \oplus S^{n-2,1,1}$$

$$\dim\ n(n-1)\ \ 1\quad 2(n-1)\ \ n(n-3)/2\ (n-1)(n-2)/2.$$

To derive this, start with $\lambda = (n-2,1,1)$. Consider $n-2$ ones, one two, and one three. Young's rule asks for the number of ways of arranging these symbols into arrays which are increasing along rows, and strictly increasing down columns. The only ways are



where the block indicates $n-2$ ones. The number of times each shape appears gives the multiplicity of the corresponding irreducible. The dimensions are computed from the hook length formula of Chapter 7.

Here there is a natural interpretation of the decomposition. The projection into S^n represents the best constant approximation to f, i.e. $\overline{f} = \frac{1}{n(n-1)}\Sigma f(i,j)$. The two $S^{n-1,1}$ spaces give the effect of the first coordinate and the second coordinate respectively. The projection of f into $S^n \oplus 2S^{n-1,1}$ gives the best approximation to f of the form

$$a + f_1(i) + f_2(j) \text{ with } \Sigma f_1(i) = \Sigma f_2(j) = 0.$$

The projection into $S^{n-2,2}$ can be thought of as the best approximation of f, orthogonal to $S^n \oplus 2S^{n-1,1}$, of form

$$f_3\{i,j\}$$

the brackets indicating unordered pairs. The projection onto $S^{n-2,1,1}$ can be thought of as a residual, or what's left when the first four terms are subtracted off.

EXERCISE 3. Verify the decomposition

$$\begin{array}{lllllllll} M^{n-3,1,1,1} & = S^n \oplus 3S^{n-1,1} \oplus 3S^{n-2,2} \oplus S^{n-3,3} & \oplus 2S^{n-3,2,1} & \oplus 3S^{n-2,1,1} \oplus S^{n-3,1,1,1} \\ n(n-1)(n-2) & 1 \quad 3(n-1) & \frac{3n(n-3)}{2} & \frac{n(n-1)(n-5)}{6} & \frac{2n(n-2)(n-4)}{3} & 3\binom{n-1}{2} & \binom{n-1}{3} \end{array}$$

Can you give an interpretation to the subspaces involved?

3. *Computing projections.* We turn now to the problem of actually computing the projections, choosing bases, and so on.

Method 1 – *Character Theory.* Let G be a group, $\rho: G \to GL(V)$ a representation. The decomposition of V into irreducibles usually requires choosing a basis. There is a coarser direct sum decomposition that is basis free. Let $\chi_1, \chi_2, \ldots, \chi_h$ be the distinct characters of the irreducible representations $W_1, \ldots, W_h$ of G. Let

$d_1, d_2, \ldots, d_h$ be their degrees. Let $V = U_1 \oplus U_2 \oplus \ldots \oplus U_m$ be the decomposition of V into a direct sum of irreducible representations. For $i = 1, \ldots, h$, denote by V_i the direct sum of the spaces U_j which are isomorphic to W_i. Then

$$V = V_1 \oplus \ldots \oplus V_h.$$

This decomposition is canonical in the following sense:

Theorem 1.

(1) The decomposition $V = V_1 \oplus \ldots \oplus V_h$ does not depend on the initially chosen decomposition of V.

(2) The projection Π_i of V onto V_i associated with this decomposition is given by

$$\Pi_i = \frac{d_i}{|G|} \sum_{t \in G} \chi_i(t)^* \rho(t).$$

(3) If the original representation is unitary, then Π_i is an orthogonal projection.

Proof. We first prove (2). Restrict attention to an irreducible subspace U_j. Then ρ restricted to U_j is an irreducible representation with character χ say. Let Π_i be defined by the formula in (2). Since U_j is invariant, Π_i maps U_j into itself. Since Π_i and ρ commute, Π_i restricted to U_j is a constant times the identity. This constant is

$$< \chi_i | \chi >^* = \begin{cases} 0 & \text{if } \chi \neq \chi_i \\ 1 & \text{if } \chi \neq \chi_i. \end{cases}$$

This proves that Π_i is a projection onto V_i. Since Π_i does not depend on the originally chosen decomposition, (1) follows.

For (3), clearly Π_i is a projection. To show it is orthogonal, we must show that $\Pi_i^* = \Pi_i$. This is obvious if ρ is unitary. □

EXERCISE 4. Consider the voting data of Chapter 5-B. Compute the projections into the irreducible subspaces for the people voting for only two candidates (see Case 2 in Section 2 above. The characters of S_5 are in James and Kerber (1981)) compare with people voting for only one candidate.

We will find use for this theorem as a practical tool for computing projections. Here is a theoretical application which shows that the spectral analysis on groups developed in Section A is the same as the spectral analysis developed in this section.

COROLLARY 1. *Let R, $L(G)$ be the regular representation of the finite group G. Let ρ be an irreducible representation and V_ρ the direct sum of all irreducibles in $L(G)$ isomorphic to ρ. For $f \in L(G)$, the orthogonal projection of f into the space V_ρ is given by*

$$\Pi_\rho f(s) = \frac{d_\rho}{|G|} Tr(\rho(s^{-1}) \hat{f}(\rho)).$$

Proof. Using Theorem 1, we must show

$$\Pi_\rho f(s) = \frac{d_\rho}{|G|}\sum_\rho \chi_\rho(t^{-1})R_t\, f(s) = \frac{d_\rho}{|G|}Tr\{\rho(s^{-1})\hat{f}(\rho)\}.$$

Both sides are linear in f, so take $f = \delta_u$. Cancelling common factors, the left side equals

$$\Sigma_t\, \chi_\rho(t^{-1})\delta_u(t^{-1}s) = \chi_\rho(us^{-1}) = Tr\{\rho(s^{-1})\rho(u)\} = Tr\{\rho(s^{-1})\hat{\delta}_u(\rho)\}.$$

The projection is orthogonal because the regular representation is unitary. Note that the right side of the formula doesn't depend on the basis chosen for ρ. □

Example. Take $G = S_n$, and X as the k sets of $\{1, 2, \ldots, n\}$ as discussed in example 2 of Chapter 5B. Here $X = S_n/S_k \times S_{n-k}$ and the representation decomposes without multiplicity (see Exercise 2 of Chapter 7) as

$L(X)$	$=$	S^n	$\oplus$	$S^{n-1,1}$	$\oplus$	$S^{n-2,2}$	$\oplus \ldots \oplus$	$S^{n-k,k}$
dim $\binom{n}{k}$		1		$n-1$		$\binom{n}{2}-\binom{n}{1}$	$\ldots$	$\binom{n}{k}-\binom{n}{k-1}$.

Here the projection given by Theorem 1 is simply the projection into irreducibles. This holds for data on any homogeneous space which is a Gelfand pair (see Chapter 3-F).

Example. Take $G = S_n$ and $X = S_n/S_{n-2} \times S_1 \times S_1$. Here there is multiplicity:

$$L(X) = S^n \oplus 2S^{n-1,1} \oplus S^{n-2,2} \oplus S^{n-2,1,1}.$$

A further decomposition of the projection into $2S^{n-1,1}$ is required. One way to do this is described in Exercises 2.8, 2.9 and 2.10 of Serre (1977, Sections 2.6 and 2.7). A second way to do it is outlined next.

Method 2 – Following known vectors. One problem with Theorem 1 is that it involves a sum over the group. For homogeneous spaces like $S_{49}/S_6 \times S_{43}$, this is simply not feasible.

Let G be a group and (ρ, V) a representation. Suppose V is equipped with an invariant inner product. We are often in a position where we know, or can guess at, vectors $w_1, w_2, \ldots, w_J \in V$ which generate an invariant subspace $W \subset V$. It is then straightforward to orthonormalize w_i using the Gram-Schmidt procedure, forming $w_1^*, w_2^*, \ldots, w_J^*$. Then the projection of v on W is

$$\Pi_W v = \Sigma < v, w_j^* > w_j^*.$$

Computations using this approach usually only involve a small fraction of work required for the character theory approach.

Example. Consider the decomposition arising from $X = S_n/S_{n-2} \times S_1 \times S_1$-ordered pairs out of n. The pieces are

$$L(X) = S^n \oplus 2S^{n-1,1} \oplus S^{n-2,2} \oplus S^{n-2,1,1}.$$

Here $L(X)$ is considered as the space of all real functions on ordered pairs with $< f, g > = \Sigma_{i,j}\ f(i,j)g(i,j)$. The $n-1$ functions $f_k(i,j) = \delta_k(i) - \frac{1}{n}$, $1 \leq k < n$ are linearly independent and span an $n-1$ dimensional subspace of $2S^{n-1,1}$. The $n-1$ functions $g_k(i,j) = \delta_k(j) - \frac{1}{n}, 1 \leq k < n$ are linearly independent of the f_k and each other. They span the rest of $2S^{n-1,1}$.

This gives a natural way of decomposing the remaining subspace in the example of the last section.

Remark 1. This approach is available for decomposing *any* of the subspaces M^λ. The relevant details are given in Section 17 of James (1978). In deriving a version of Young's rule that works for finite fields, James introduces a hierarchy of invariant subspaces $S^{\mu\#,\mu}$ which split M^λ into progressively more refined pieces (ending with irreducibles). He gives an explicit basis for each of these subspaces involving sums over the column stabilizer subgroups. These sums, and attendant computations, seem computationally manageable.

As an example, we know that each M^λ contains the irreducible S^λ once. Recall from Chapter 7, a standard Young tableau is a tableau which is decreasing across rows and down columns. Recall that for a tableau t the vector (or function or polytabloid) e_t is defined by

$$e_t = \sum_{\pi \in C_t} \operatorname{sgn}(\pi) e_{\pi\{t\}},$$

where $\{t\}$ is the tabloid associated to t, and C_t is the column stabilizer of t. James shows that $\{e_t : t \text{ is a standard Young tableau of shape } \lambda\}$ is a basis for S^λ in M^λ. For λ's without many parts, $|C_t|$ is manageable (e.g., it has size 2^k for $\lambda = (n-k, k)$). The number of standard Young tableaux is given by the hook length formula of Chapter 7. The e_t can be orthogonalized to e_t^*. Then the projection of $v \in M^\lambda$ can be computed through $v \cdot e_t^*$.

Remark 2. Often a subspace $W \subset V$ has a simple data analytic interpretation, but for computational convenience, the projections are computed with respect to a basis which scrambles things up. A second problem: the dimension of W can be smaller than the number of natural projections. For instance, the decomposition arising in studying unordered three sets out of six is

$$M^{3,3} = S^6 \oplus S^{5,1} \oplus S^{4,2} \oplus S^{3,3}$$

the dimension of $S^{4,2}$ is 9 $(= \binom{6}{2} - \binom{6}{1})$. Now $S^{4,2}$ is the space of unordered pair effects. There are $\binom{6}{2} = 15$ such effects it is natural to look at.

Colin Mallows has suggested a remedy for these problems. Given $v \in V$, project it to $v^* \in W$. Then take the natural vectors of interest, say $v_1, v_2, \ldots, v_\ell$, project them into $v_i^* \in W$, and plot $< v^*, v_i^* >$ versus i. This allows us to try to interpret the projections on a natural scale. It assumes v^*, v_i^* have been normalized to be unit vectors. See Diaconis (1989) for applications.

Method 3 – Radon transforms. There is a special technique available for decomposing the representations M^λ associated to partially ranked data of shape λ.

For clarity, this will be explained for the multiplicity free case $\lambda = (n-k, k)$ with $k \leq n/2$.

A vector $f \in M^{n-k,k}$ can be regarded as a function on k sets of an n set. For $1 \leq j \leq k$, define a mapping

$$R^+ : M^{n-j,j} \to M^{n-k,k} \text{ by } R^+ f(s) = \sum_{s \supset r} f(r)$$

where $|r| = j$, $|s| = k$. This Radon transform R^+ has an inverse $R^- : M^{n-k,k} \to M^{n-j,j}$ satisfying $R^- R^+ f = f$. An explicit form of the inverse is given by Graham, Li and Li (1980). If $M^{n-k,k}$ and $M^{n-j,j}$ are given bases consisting of delta functions on the k sets and j sets respectively, then the r, s element of R^- is

$$\frac{(-1)^{k-j}(k-j)}{(-1)^{|s-r|}|s-r|} \frac{1}{\binom{n-j}{|s-r|}}.$$

Note that both R^+ and R^- commute with the action of S_n. Composing maps in the other way, R^+R^- gives a map $M^{n-k,k} \to M^{n-k,k}$. This map is an orthogonal projection onto the single copy of $M^{n-j,j}$ contained in $M^{n-k,k}$:

LEMMA 1. *The map R^+R^- is an orthogonal projection on $M^{n-k,k}$ with range isomorphic to $M^{n-j,j}$.*

Proof. Since $R^+R^-R^+R^- = R^+(R^-R^+)R^- = R^+R^-$, the map is a projection. To show that it is orthogonal we must show $(R^+R^-)^t = R^+R^-$. Let s_1 and s_2 be k sets. The s_1, s_2 entry of R^+R^- is proportional to

$$\sum_{|r|=j} \delta_{rs_1} \frac{1}{(-1)^{|s_2-r|}|s_2-r|} \frac{1}{\binom{n-j}{|s_2-r|}}.$$

With $\delta_{rs} = 1$ or 0 as $r \subset s$ or not. This sum is a sum of terms of form

$$\frac{1}{(-1)^{k-\ell}(k-\ell)} \frac{1}{\binom{n-j}{k-\ell}},$$

the multiplicity of this term being $\#\{r \subset s_1 : r \cap s_2 = \ell\}$. This is a patently symmetric in s_1 and s_2, so orthogonality follows.

Finally, both R^+ and R^- commute with the action of S_n. The map R^- is onto $M^{n-j,j}$, so R^+ is an isomorphism of $M^{n-j,j}$ with the range of R^+R^- i.e. with range R^+. □

To use the lemma, define $R^+R^- = \pi_j$. Then $I - \pi_j$ is also an orthogonal projection. We know $M^{n-k,k} = S^n \oplus S^{n-1,1} \oplus S^{n-2-2} \oplus \ldots \oplus S^{n-k,k} = M^{n-j,j} \oplus S^{n-j-1,j+1} \oplus \ldots \oplus S^{n-k,k}$. One may thus proceed by taking $j = k-1, k-2, \ldots, 1$ inductively. This procedure is computationally feasiable for n large and k small.

A similar procedure is available for any partition λ using the results in Section 17 of James (1978). Diaconis (1987) gives details.

Final remarks on choice of basis. Return to data f on a group G. The Fourier transform $\hat{f}(\rho)$ can be a completely arbitrary matrix for general f. To see this, just define f by the inversion theorem to have a prescribed $\hat{f}(\rho)$. There *is* a rather complex restriction, akin to Bochner's theorem but more complex, when f is positive. For practical purposes $\hat{f}(\rho)$ is an essentially arbitrary matrix.

We have available the possibility of changing bases, changing from $\hat{f}(\rho)$ to $A\,\hat{f}(\rho)A^{-1}$. As A varies, the invariants of $\hat{f}(\rho)$ are its eigenvalues, or its "canonical form."

To compare coefficients within $\hat{f}(\rho)$ and between various ρ's, it seems natural to consider only unitary (or orthogonal) base changes. Then, there is not much that can be done to bring $\hat{f}(\rho)$ into a simple form. One possibility is to rotate to make the first few rows of $\hat{f}$ as large as possible.

Other possibilities are to change bases to bring the matrix into a simple form. For example, any matrix can be conjugated by an orthogonal matrix to the sum of a diagonal and skew symmetric matrix: $A = \frac{1}{2}(A + A^t) + \frac{1}{2}(A - A^t)$, and $A + A^t = \Gamma D \Gamma^t$, so $\Gamma A \Gamma^t = \frac{1}{2}D + S$, with D diagonal and S skew symmetric. Any matrix can be orthogonally conjugated to lower triangular form (essentially) see Golub and van Loane (1983).

The choice of bases must balance off

- computational convenience
- ease of interpretation
- convenience of orthogonality
- maximal informativeness of a few projections
- invariance considerations.

Nobody ever said statistics was easy.

At present, we can have the first three properties, for partially ranked data, by using methods 2 or 3 and Mallows' idea as outlined under method 2. There is plenty to think about.

C. Analysis of variance.

Analysis of variance (ANOVA) is a set of techniques for analyzing cross tabulated data such as two-way arrays. Data analytically, ANOVA is a special case of spectral analysis as described in Section B. This section develops the connection, and explains a technique introduced by Peter Fortini for defining the group for the array from a naturally given set of factors.

There is more to ANOVA than the data analytic aspects presented here. Section D-4 describes some of the other ideas. Hopefully they can be developed for other types of spectral analysis.

1. *Classical examples.*

Example 1. Two-way ANOVA with 1 *repetition per cell.* Here data are collected in a two way layout

x_{11}	x_{12}	...	x_{1J}
x_{21}			
$\vdots$			
x_{I1}			x_{IJ}

The rows represent the I different levels of one factor, the columns represent the J different levels of a second factor. The classical examples involve agriculture: the x_{ij} might represent the crop yield, the row factors might be different types of seed, the column factors might be different types of fertilizer. In a psychology example the data might be reaction time, the row factor might be wording of question, the column factor might be type of recording device.

There are two basic steps in analyzing such data: forming estimates of "main effects" and the testing ritual involving F statistics that is often called ANOVA. The first is simple and quite sensible: calculate the average of all numbers in the table $x_{..}$ – the grand mean. Subtract $x_{..}$ from x_{ij}. Calculate "row effects" $x_{i.} = \frac{1}{J}\Sigma_j\{x_{ij} - x_{..}\}$. Subtract the row effects and calculate column effects $x_{.j}$. Subtract the column effects forming residuals: $r_{ij} = x_{ij} - x_{..} - x_{i.} - x_{.j}$. This is attempting an additive decomposition of the data:

$$x_{ij} \doteq x_{..} + x_{i.} + x_{.j}.$$

If the residuals are small, the approximation makes a useful, simple description of the data. One then looks at the main effects – e.g. if the column effects are ordered, they might be plotted against j to see their shape, etc. The residuals can be plotted (does there seem to be a relation with $x_{i.}$ or $x_{.j}$?). If an additive fit fails, one can try different scales – an additive fit to log x_{ij}. Tukey (1977) vigorously illustrates this approach to ANOVA.

The second piece of analysis is a way of performing a statistical test to see if the main effects are zero. There are several different tests – all main effects zero, row effects zero, column effects zero, and even grand mean zero. These tests all have a geometric interpretation: they are ratios of the squared lengths of the "effect" vectors to the squared length of the residual vector. The tests can be justified in language involving normal distributions or in language involving permutations of the residuals.

I cannot hope to adequately review the statistical aspects of ANOVA here. The classical text by Scheffé (1959) is still most readable and useful. Two fine survey articles are those by Tjur (1984) and Speed (1987). Further references and discussion are given in Section D-4.

We now rephrase this as a spectral analysis problem. The data x_{ij} can be thought of as a function defined on the set of ordered pairs $X = \{(i,j)\ 1 \le i \le I,\ 1 \le j \le J\}$. The group $G = S_I \times S_J$ (where S_n is the symmetric group on n letters) acts transitively on the ordered pairs. This gives a representation of $S_I \times S_J$ on $L(X)$ which is just the product of the usual I dimensional representation we called $M^{I-1,1}$ of S_I with $M^{J-1,1}$ of S_1. Now $M^{I-1,1}$ splits into the constants, and an $I-1$ dimensional irreducible. We write $M^{I-1,1} = S^I \oplus S^{I-1,1}$. So

$$\begin{array}{lcl} & M^{I-1} \otimes M^{J-1} = (S^I \oplus S^{I-1,1}) \otimes (S^J \oplus S^{J-1,1}) = Triv \oplus S^{I-1,1} \otimes Triv \oplus S^{J-1,1} \otimes Triv \oplus S^{I-1,1} \otimes S^{J-1,1} \\ dim & I \times J & 1 \quad I-1 \quad J-1 \quad (I-1)(J-1) \end{array}$$

The observed array x_{ij} is just a vector in $L(X)$. The projection of x into the invariant irreducible subspaces gives the usual ANOVA decompositon: the projection onto the one dimensional space of constants is the array with constant entry $x_{..}$. The projection onto $S^{I-1,1} \otimes$ triv is onto all arrays which are constant in each row, with the constants summing to zero. These constants are $x_{i.}$. The final space is spanned by the residuals r_{ij}.

The lengths of these projections give the usual sums of squares required for the classical ANOVA table. The dimensions of the various irreducible subspaces give the degrees of freedom for the classical F tests.

Example 2. Two way ANOVA with repeated entries – wreath products. In this example, data are collected in a two way array, but k observations are collected at each level of the two factors. It is intuitively clear that the appropriate group of symmetries allows permutations of rows and columns and, within each row and column, an arbitrary permutation of the k elements in that cell among each other. To write this down neatly, it is useful to introduce the notion of wreath product.

Let G be any group and H a subgroup of S_n. Define a group G wr H as follows. The group consists of the set of elements in $G^n \times H = \{(g_1, g_2, \ldots, g_n;\ h)\}$. The product of two elements is

$$(g_1, \ldots, g_n;\ h)(g'_1, \ldots, g'_n;\ h') = (g_1 g'_{h^{-1}(1)}, \ldots, g_n g'_{h^{-1}(n)};\ hh').$$

The identity is $(i_G, \ldots, i_G;\ i_H)$, $(g_1, \ldots, g_n;\ h)^{-1} = (g^{-1}_{h^{-1}(1)}, \ldots, g^{-1}_{h^{-1}(n)};\ h^{-1})$. The subgroup G^n sits in G wr H as a normal subgroup. H sits in G wr H as $\{(i_G, \ldots, i_G;\ h)\}$. G wr H is the semi-direct product of G^n and H.

For an $I \times J$ table with k replications per cell, the natural symmetry group is S_k wr$(S_I \times S_J)$. The representation theory of wreath products is completely known in terms of the irreducible representations of G and H. A reasonable treatment is Chapter 4 of James and Kerber (1981). When this theory is applied to ANOVA there are no surprises; the decomposition is the one derived in all standard texts; see Example 4 below for a special case.

Example 3. The Diallel Cross Design. This design is used by animal or plant breeders to find new hybrid species of known species. Consider n known types of plant. Form all $\binom{n}{2}$ crosses of distinct strains. The data are some measurable characteristic of each offspring (yield, size, etc.). Here the data are indexed by unordered pairs $\{i, j\}$. The natural symmetry group is S_n; the standard ANOVA decomposition corresponds to the decomposition of what we have called $M^{n-2,2} \cong S^n \oplus S^{n-1,1} \oplus S^{n-2,2}$.

We will return to examples after a general definition of the symmetry group of a designed experiment.

2. *A systematic approach.* In the examples above we just wrote down a group and showed how spectral analysis gave "the usual" ANOVA.

Alan James (1957, 1982) and Ted Hannan (1965) introduced ideas to relate an algebraic object to a designed experiment. Gleason and McGilchrist (1978) give

a more detailed account of these ideas. Lederman (1967) independently pointed out the connection between ANOVA and representation theory.

Peter Fortini (1977) extended these ideas to give a reasonably systematic theory. His work associates a group in a natural way to an experiment given in terms of classical factors. Then spectral analysis can be used as before. Here is a brief account of Fortini's ideas.

Let X be a finite set. We will work with observations indexed by X.

Definition 1. A *factor* f of the design indexed by X is a set valued map from X to the elements of a finite set F. If $|X| = N$, $|F| = k$, a factor can be described by an $N \times k$ matrix f where $f_{x\ell} = 1$ if $\ell \in f(x)$, $f_{x\ell} = 0$ otherwise. A *multifactorial design* is a set X and a collection of factors $(f_1, F_1), \ldots, (f_k, F_k)$.

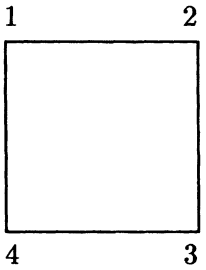
Definition 2. The *automorphism group* G of a multifactorial design is the group of all permutations g of X with the property that for each factor f_i there is a permutation g_i of F_i such that $f_i(gx) = g_i(f_i(x))$.

Thus, if x is associated with levels $f_i(x)$, gx is associated to levels $g_i f_i(x)$. The *outcome set* V of a numerical experiment indexed by X is the set of all real valued functions on X. This has basis e_x, and we have the permutation representation of G acting on V. An *analysis of variance* is a decomposition of V into a direct sum of irreducible invariant subspaces V_i. The projections of $v \in V$ onto V_i are called main effects. The squared lengths of these projections form the ith line of an ANOVA table: the number of degrees of freedom of the ith line is dim V_i.

Example 4. Two treatments with two objects per treatment. To understand the definitions, consider comparing two treatments: A or B with two objects for each treatment. Take X to be the set of four objects, labeled $\{1, 2, 3, 4\}$. There is one factor – did the object get treatment A or B? Suppose that objects 1 and 3 get treatment A and objects 2 and 4 get treatment B. The factor matrix is

	A	B
1	1	0
2	0	1
3	1	0
4	0	1

The permutations of X that are in G are id, (1 3), (2 4), (1 3)(2 4), each associated with the identity permutation of factors A and B, and (1 2)(3 4), (1 4)(2 3), (1 2 3 4), (4 3 2 1) each associated with transposing A and B. We thus get an eight element automorphism group. Observe that the group is the symmetry group of the square



It is often called D_4 – the dihedral group on four letters. Observe too that if we think about the problem directly, as we did before, the "obvious" symmetry group is S_2 wr S_2. This is an eight element group which is isomorphic to D_4.

How does the four-dimensional space V decompose? Here, the decomposition is obvious, but as an exercise, we follow Fortini and derive the result using character theory. The character table for D_4 is in Chapter 5 of Serre (1977):

Class	χ_1	χ_2	χ_3	χ_4	χ_5	Perm rep.
1	1	1	2	1	1	4
(1 3)(2 4)	1	1	-2	1	1	0
(1 2)(3 4), (1 4)(2 3)	1	-1	0	-1	1	0
(1 3), (2 4)	1	-1	0	1	-1	2
(1 2 3 4), (1 4 2 3)	1	1	0	-1	-1	0

Looking across the first row, there are 5 distinct irreducible representations, four of dimension 1 and one of dimension 2. Which of these appear in the permutation representation? The character of the permutation representation is the number of fixed points of G acting on X. This is given in the last column above. The multiplicity of χ_i in the permutation representation is given by

$$m_i = < \chi_i | \chi_{\text{perm}} > = \frac{1}{8} \Sigma \, \chi_i(g) \chi_{\text{perm}}(g).$$

We get $m_1 = 1$, $m_2 = 0$, $m_3 = 1$, $m_4 = 1, m_5 = 0$. Thus, $V = V_1 \oplus V_2 \oplus V_3$. V_1 is the 1 dimensional grand mean space, V_2 is the space: $\{y \in V: (y_1+y_3)-(y_2+y_4) = 0\}$. It represents the difference between the average of the group means. The space V_3 is a 2 dimensional space of "what's left over."

Example 3. revisited. The need for factors taking more than one value on a given observation is well illustrated by the diallel cross. Here $X = \{\{i, j\};\ i \neq j\}$. $F = \{1, 2, \ldots, n\}$, and $f\{i, j\} = \{i, j\}$. Thus the matrix has two ones in each row. The automorphism group is a subgroup of $S_{\binom{n}{2}}$. A bit of reflection shows that it is S_n, and that V is what we have been calling $M^{n-2,2}$.

Example 5. Balanced incomplete blocks. Let us begin with an example taken from Cochran and Cox (1957, pg. 443) on the effects of aging on the tenderness of beef. Six periods of storage (0, 1, 2, 4, 9, and 18 days) were tested. These treatments are denoted 1, 2, 3, 4, 5, 6 respectively.

To facilitate comparison, roasts from symmetric locations (left/right sides) were paired into blocks of size 2. There are $15 = \binom{6}{2}$ ways to treat a pair. Scoring was done by 4 judges, each marking on a scale from 0 to 10. The data shows their totals out of 40, a high score indicating very tender beef.

Block	1	2	3	4	5	6	7	8
Pair	{1,2}	{3,4}	{5,6}	{1,3}	{2,5}	{4,6}	{1,4}	{2,6}
Scores	7 17	26 25	33 29	17 27	23 27	29 30	10 25	26 37
Block Sums	24	51	62	44	50	59	35	63

Block	9	10	11	12	13	14	15
Pair	{3,5}	{1,5}	{2,4}	{3,6}	{1,6}	{2,3}	{4,5}
Scores	24 26	25 40	25 34	34 32	11 27	24 21	26 32
Block Sums	50	65	59	66	38	45	58

Thus block 1 was given treatments 1 and 2. The roast receiving treatment 1 was rated at 7. The roast receiving treatment 2 was rated 17.

The first thoughts for such an analysis run as follows: It is natural to compute the sum for all roasts receiving treatment i:

Treatment	1	2	3	4	5	6
Sum	70	115	132	139	158	155

This suggests longer aging improves tenderness, the effect perhaps peaking at treatment 5 (9 days).

Now it is natural to try to see if there is a block effect: roasts from different location vary in tenderness, and if one of the treatments was tried on more tender blocks, this would favor the treatment.

The natural adjustment subtracts from the ith treatment total the sum of the block averages (here block sum/2) for the blocks containing the ith treatment. Chapter 5 of Scheffé (1959) gives a careful, clear description. In the example, the adjusted sums are

Treatment	1	2	3	4	5	6
Adjusted Sum	-33	-5.5	4	8	15.5	11

These now sum to zero.

Here, the adjustment doesn't affect the conclusions drawn above. We leave further details to Cochran and Cox who carry out the usual analysis and conclude that storage up to about a week increases tenderness.

In general, there are t treatments to be compared in blocks of size $k < t$. An *unreduced balanced incomplete block design* involves all possible $\binom{t}{k}$ blocks and so $k\binom{t}{k}$ basic units (in the example, $t = 6$, $k = 2$ and 30 roasts were involved).

One natural way to index such data is as a pair (i,s) with $1 \leq i \leq t$ denoting the treatment applied to that unit, and s of cardinality $|s| = k - 1$ denoting the subset in the same block as the given unit. Thus $X = \{(i, s)\}$ and $|X| = t\binom{t-1}{k-1} = k\binom{t}{k}$.

The group S_t that permutes treatment labels operates transitively on X, and we see that $L(X)$ can be identified with $M^{t-k,k-1,1}$ of Chapter 7. Another way to arrive at this result begins with the idea that this experiment has two factors: "treatments" $f_1(i, s) = i$, and "blocks" $f_2\{i, s\} = \{i \cup s\}$.

What automorphisms are possible? A little reflection shows that only permutations that permute treatments among themselves are allowed; treatments that permute things within a block are ruled out and allowable treatments that move blocks around can be achieved by permuting treatments. Thus the automorphism group is isomorphic to S_t.

The decomposition now follows from Young's rule. Before stating the result, let us examine the introductory example on aging of meat. Here $t = 6, k = 2$, and we have

$$M^{4,1,1} = S^6 \oplus 2S^{5,1} \oplus S^{4,2} \oplus S^{4,1,1}$$
$$\dim\ 30 \quad 1 \quad 2 \times 5 \quad 9 \quad 10$$

(see Case 2 of Section B-2 above). The projection onto the one dimensional space S^6 is the grand mean. The treatment effects can be identified as one of the $S^{5,1}$ spaces. The block effects space, orthogonal to $S^6 \oplus S^{5,1}$, is $S^{4,2}$ (the full 15-dimensional block effect space $M^{4,2} = S^6 \oplus S^{5,1} \oplus S^{4,2}$). These projections constitute the classical analysis for this experiment, projecting into treatment and blocks adjusted for treatment.

The remaining $S^{5,1}$ gives a new piece of analysis due to Fortini. To give it a clear interpretation let us change the imagery. Suppose there are 6 amounts of vanilla added to ice cream (from none to a fair amount). Fifteen people each taste two servings each in a balanced incomplete block involving 30 servings in all. The treatment effects are interpreted as before. The block effects become subject effects – some people are systematically higher than others.

The classical analysis assumes that treatment and block effects are additive. However, this would fail if tasters give ratings partially by comparison. The 2nd copy of $S^{5,1}$ can be interpreted as the additive effect on the rating of treatment i due to being in the same block with treatment j.

In tasting examples, this (and its higher order analogs described below) makes perfect sense. In the aging/tenderness experiment it seems perfectly sensible to set this effect to zero as is done classically.

To decompose $M^{t-k,k-1,1}$ in general, Young's rule begins with $t - k$ ones, $k - 1$ twos, and one 3. These are placed into any arrangement as a tableaux which is non-decreasing in rows and strictly increasing in columns. Thus

$$M^{t-k,k-1,1} = S^t \bigoplus_{j=1}^{k-1} 2S^{t-j,j} \oplus S^{t-k,k} \bigoplus_{j=1}^{k-1} S^{t-j-1,j,1}.$$

It being understood that any improper partition above doesn't contribute.

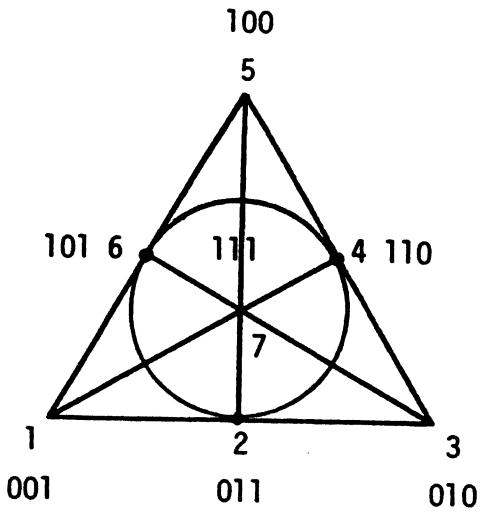
Fortini (1977) used this decomposition to build a class of standard linear models for which the decomposition gives the appropriate ANOVA. Calvin (1954) seems to be the first paper to extend the classical model to try to cope with this kind of non-additivity.

Example 6. The projective plane. As a final topic for this section we recall that there are many further types of block designs. For example, to compare 7 treatments, with block size 3, the following classical design can be used:

$$\{123\}\ \{345\}\ \{156\}\ \{246\}\ \{147\}\ \{257\}\ \{367\}.$$

Here, each pair appears together in exactly one block, so the same type of block/treatment analysis is available.

This design is constructed as the set of lines through the origan in the vector space Z_2^3. Each plane contains 3 points which gives rise to the blocks. These are edges in the figure below.



The group of this design is $GL_3(Z_2)$, a well studied simple group of order 168. For someone wishing to learn a little group theory, decomposing $L(X)$ makes a nice exercise. $L(X)$ is 21-dimensional. It decomposes into 4 irreducible pieces: the trivial representation, a 6-dimensional space of treatment effects, a 6-dimensional space of block effects, and an 8 dimensional irreducible space of residuals.

EXERCISE 5. Prove the assertions about $L(X)$. Hints: argue that $GL_3(Z_2)$ acts doubly transitively on lines and planes in Z_2^3. Using the "useful fact" in Chapter 7A, this gives two irreducible 6-dimensional representations of GL_3. Now consider $X = \{(i,p)\colon i \text{ a point}, p \text{ a plane}, i \in p\}$. Show that GL_3 acting on $X \times X$ has 4 orbits. Now use the "useful fact" of Chapter 7 to argue that $L(X)$ decomposes into 4 irreducibles.

This example is the tip of the iceberg called combinatorial design. Lander (1982) is a nice reference for the construction of block designs. Beth, Jungnichel, and Lenz (1986) is a recent encyclopedic reference. It seems like a worthwhile project to go through the classical designs, compute automorphisim groups, and compare the spectral decomposition with classical ANOVA.

D. Thoughts about spectral analysis.

This chapter builds on two well established traditions: the analysis of variance and the spectral analysis of time series. The new applications and variations of classical areas such as block designs should be regarded speculatively.

There is much of value in the parent theories as they appear in modern practice which I have not thought through in sufficient detail to carry over. This section outlines some of these techniques.

1. *Why projections onto invariant subspaces?* It is an empirical fact that the decomposition of data indexed by X into irreducible pieces is sometimes scientifically useful. Let me make an attempt to explain this. There is no one right way to analyze data. The spectral decomposition presents certain averages of the data which often seem sensible. The set of all averages is a 1-1 function of the data, so nothing is lost. Often the specific labeling of X is fairly arbitrary, and we would rather not have our conclusions depend on this labeling. In other circumstances, the order is worth taking into account, but it is useful to separate the part of the analysis that depends on the order from the part that is invariant under G.

Data indexed by X are represented by a function on X. Sometimes this function takes values in the integers, as in the case of counted data: "how many people in the population chose X?" Even here we want to be able to talk about averages and differences of averages, so we need to consider functions from X into the rationals Q. For the permutation group, the splitting of a representation over Q is the same as its splitting over the real or complex numbers.

Our function thus naturally sits in $L(X)$: the set of all complex valued functions. A subset of $L(X)$ can be thought of as a partial description. In practice these will be things like the constants or "first order functions;" sets of functions that are simple to describe or interpret. If the actual f can be well approximated by a simple f we regard this as useful. If f is "first order" and g is "first order" it seems natural that $\frac{1}{m}f$ or $f+g$ be the "first order" since, for example, $f+g$ has the interpretation of combining the two data sets (for counted data). This suggests that a basic partial description be a subspace of $L(X)$.

Finally, if G acts on X we want to consider descriptions that don't depend on the labelling: it seems unnatural that $f(x)$ could be "first order" but not $f(gx)$. Thus a natural descriptive unit is an invariant subspace of $L(X)$. Then *spectral analysis* is the decomposition of $f(x)$ into its projections on the irreducible invariant subspaces of $L(X)$, and the approximation of $f(x)$ by as small a number of its projections as give a reasonable fit.

Of course, this kind of linear analysis is just a start. Non-linear analyses may also be most useful (see Section (4) below). These are easiest to interpret if they have a simple connection to the linear theory (log f may be well approximated

by a simple linear fit).

2. *On the choice of group.* Spectral analysis begins with a group G acting on a set X. In some areas, like ANOVA, there is some theory to guide the choice of G. In other areas, like time series, physics dictates the choice. In new applications, there may be several possible groups among which to choose.

Clearly if G acts on X and H is a subgroup of G, the splitting of $L(X)$ into H irreducibles is finer than the splitting under G. Choosing the smallest group that preserves essential structure gives the largest number of projections. Preliminary analysis may start with larger groups.

Consider a two way $I \times J$ array. The group that gives classical ANOVA is $S_I \times S_J$. Another group that operates transitively is $Z_I \times Z_J$, operating by cyclically shifting each coordinate. A third transitive action is given by $S_I \times Z_J$. Any of these groups might be appropriate. Z_J preserves adjacency (or time order) while S_I invariance says adjacency isn't relevant. Thus $S_I \times Z_J$ might be deemed natural for a problem in which the rows index types of bird, the columns index months of the year, and the entries are number of birds sighted in that month by a local bird watching society.

EXERCISE 6. Derive the appropriate spectral analysis for $S_I \times Z_J$.

As a second example, consider $X = Z_2^k$. Four groups act naturally on this space: Z_2^k, $GL_k(Z_2)$, Z_2 wr S_k, and Z_2 wr Z_k. We consider these in turn.

a) Under Z_2^k, $L(X) = \bigoplus_{y \in X} V_y$ where V_y is the one dimensional space spanned by $x \to (-1)^{x \cdot y}$. Here spectral analysis amounts to the Fourier inversion theorem

$$f(x) = \frac{1}{2^k} \sum_y (-1)^{x \cdot y} \hat{f}(y).$$

b) $GL_k(Z_2)$ acts doubly transitively on X, so by Serre's exercise (2.6),

$$L(X) = V_0 \oplus V_1$$

with V_0 the constants and $V_1 = \{f : \Sigma f(x) = 0\}$.

c) Z_2 wr S_k (see Section C for notation) is the group of pairs (y, π) with $\pi \in S_k$, $y \in Z_2^k$ acting on X by $(y, \pi)x = (x_{\pi^{-1}(1)} + y_1, \ldots, x_{\pi^{-1}(k)} + y_k)$: you permute the coordinates by π and add y. It is straightforward to show that

$$L(X) = V_0 \oplus V_1 \oplus \ldots \oplus V_k$$

with V_j the linear span of $(-1)^{x \cdot y}$ for $|y| = j$. Thus spectral analysis under Z_2 wr S_k lumps together pieces of the spectral analysis under $G = Z_2^k$.

EXERCISE 7. Show that $L(X)$ splits as shown for Z_2 wr S_k. Find the decomposition with respect to Z_2 wr Z_k.

3. *On probability.* The data analytic approach to spectral analysis presented in this chapter is not based on probability. Spectral analysis is a reasonable, useful activity if f is not a sample, rather a complete enumeration of a population. Thus,

nothing is unknown, but there may still be plenty to do in condensing the data so that simple descriptions can be given. This point of view follows Tukey (1977).

Of course, there are many opportunities for probability to play a role. When we informally assess goodness of fit in a linear approximation such as $f(i,j) \doteq a + b_i + c_j$, we compare the residuals $r_{ij} = f(i,j) - a - b_i - c_j$, with the main effects a, b_i, c_j. If residuals are small, we regard the approximation as useful. The standard tests and confidence interval procedures are formalizations of this idea.

There is much more to do in constructing a believable probabilistic theory for spectral analysis. For example, the data $f(x)$ might be a sample from a larger population $F(x)$. If the sample size is small, there is no reasonable hope of estimating F, but one can hope to estimate some simple functionals of F such as its projections onto a few subspaces of interest. How the dimensionality of the subspaces should relate to the size of X and the sample size seems like a rich interesting question. Finch (1969) makes a start on these questions. See also the discussion in Section 3 of Diaconis (1985).

In some instances the bootstrap offers a reasonable way to wiggle data a bit. There are two obvious ways to bootstrap – sample iid from $f(x)$ to get $f^1(x)$, $f^2(x), \ldots, f^b(x)$, or fit a model and bootstrap residuals. Freedman (1981) describes these alternatives in regression problems. Of course, all of this is tied to some sampling like story. Without a believable sampling justification (as in complete enumeration problems) I find the bootstrap far too violent a perturbation of the data. Diaconis (1983) suggests less drastic perturbations.

I have been impressed with the usefulness of the basic normal perturbation model outlined in Section A-3. This serves as an accurate approximation for all sorts of schemes for quantifying "if the data had come out different, how would my conclusions differ?" or "if I had more data, how close is my best guess likely to be?"

Some further discussion of the need for probability is given later. It is worth emphasizing that spectral analysis can be useful without underlying probabilistic structure.

4. *Lessons from ANOVA.* A basic component of ANOVA is a test to see if projection of a given function onto a given subspace can be assumed to be zero. The standard test involves comparison of the observed projection with an estimate of "ambient noise" – usually the normalized length of the projection of the data onto the space of residuals. If the ratio of lengths is "large", then the projection cannot be asserted to be zero. Usually, "large" is quantified under the normal perturbation model of Section A-3.

Terry Speed (1987) gives an elegant survey of this material delineating a natural class of perturbation models involving patterned covariance matrices, for which the orthogonal projections in fact give the maximum likelihood estimates.

An important idea from classical ANOVA deals with fixed versus random effects. As data analytic motivation, consider a two way array with one observation per cell. Suppose the rows are different brands of typewriters and the columns are different typists. The x_{ij} might be average speed. If we are trying to evaluate typewriters, then the row averages are basic. It may be that the typists are thought of as drawn from a pool of typists. Then, the column averages are not of

particular interest – they are thought of as random effects. Their mean, variance, or distribution may be of interest as information about the population of typists. Tukey (1961) contains an interesting discussion of the distinction between fixed and random effects and its impact on ANOVA and time series data.

Modern ANOVA employs a wealth of non-linear techniques. These begin with transformations of x_{ij} to $T(x_{ij})$ aiming for linearity. Box and Cox (1964) or Hoaglin, Mosteller, and Tukey (1983, Chapter 8) give the basics. This is often supplemented by fitting more complex models such as

$$f(i,j) \doteq a + b_i + c_j + db_ic_j$$

as in Tukey (1949) or Mandel (1961). Stein (1966) gives a decision theoretic version.

More aggressive transformation techniques involving splines or more complex smoothers appear in approaches like projection pursuit (see, e.g. Huber (1985) or Henry (1983)) or Ace (see Breiman and Friedman (1985), Stone (1985), or Hastie and Tibshirani (1986)). None of these techniques are well tried out in ANOVA settings, but all seem worth thinking about and extending to other instances of spectral analysis.

Another important aspect of modern statistics is a concern for robustness – it may be that observed data is well approximated by a linear fit except for a few "wild values." For methods based on linear projections, even one wild value can foul up everything, making it seem as if no linear fit works.

One approach to these problems is to replace the usual averaging operators by robust versions such as medians. Hoaglin, Mosteller and Tukey (1985, Chapters 2-5) contains a good review. It is not at all clear if these approaches can be adapted to more complex designs which really lean on the additivity. Another approach is to try to remove or down-weight the outliers. Other approaches involve using perturbation distributions that are longer tailed than normal. Pasta (1987) contains a review of the problems and available remedies. It is fair to say that even for ANOVA this is a difficult problem, on the cutting edge of current research.

Here is another contribution of modern statistical theory to ANOVA. We now realize that while the projections have many appealing properties, they are not the most accurate estimates of population projections under the usual normal model. Non-linear shrinkage estimators can do better than the classical linear estimators. Stein (1966) shows how to modify the usual estimators in orthogonal designs to get improvement. I do not know of a systematic account for more general designs.

The Bayesian analysis of ANOVA data is important on its own – we often know quite a bit and want a way to combine this prior knowledge with the observations in a sensible way. Box and Tiao (1973) cover the basics. Consonni and Dawid (1985) give group related theory.

The Bayesian techniques give a reasonable way to get shrinkage estimates, possibly using conjugate priors and empirical Bayes ideas as surveyed by Morris (1983). Berger (1985, Chapter 4) surveys the recent literature on estimating a

normal mean. Again, most of these ideas are applicable in ANOVA and other spectral analysis problems but the details need to be worked out on a case by case basis.

An important problem that must be dealt with is missing data. The algebraic theory of ANOVA works neatly if things are neatly balanced. With a two-way layout and unequal numbers of observations per cell, there are no longer simple formulas for many quantities of interest. It is not unusual to have symmetry broken by having some subjects get sick or be lost for other reasons.

A rich collection of techniques has evolved for dealing with these problems. A most useful tool for computing "projections" has evolved as the EM algorithm of Dempster, Laird and Rubin (1977). They survey the literature. Dodge (1985) gives some special techniques for ANOVA problems.

There is some very recent work which relates the approach taken in this chapter to some of the mathematics of Chapter 3-F. In working with stochastic models for designed experiments, many workers have emphasized the importance and utility of "general balance." This is a condition introduced by Nelder (1965) which links the "design" part of the experiment to the structure of the assumed model. When the condition holds, all sorts of elegant formulas for computing best linear unbiased estimates are available. A recent survey of this material is given by Houtman and Speed (1983).

Recently, Bailey and Rowley (1986) have given a useful sufficient condition on the group of symmetries of the design (the definitions are a bit different than those in Section C) which forces generalized balance. There are too many details needed to state their results carefully. Very roughly, the group of symmetries acts on a space T of treatments, and the representation $L(T)$ must be multiplicity free (see Chapter 3F). The Bailey and Rowley paper is very clearly written and connects wonderfully with the material in this monograph.

5. *Lessons from time series.* Spectral analysis of time series and other signals is widely applied and has an enormous literature. Robinson (1982) surveys its history. It differs from ANOVA in having numerous hard science manifestations. Indeed, in some speech, geophysics, and signal processing applications the spectrum is naturally observed. Brillinger (1988) contains a list of such applications.

It is natural to try to emulate the success stories of spectral analysis for other groups. It seems fair to admit that at present there is no underlying "physical justification" for the spectral decomposition in other cases. None the less, it seems promising to try to plug into the experience and imagery of the established theory.

For example, it is well known that the periodogram (essentially the squared discrete Fourier transform) is not an accurate estimate of the amplitude of the Fourier transform at given frequency. It is asymptotically an exponential variable, and so not consistent as time goes on. Under smoothness assumptions on the underlying spectrum, smoothing the periodogram becomes sensible and leads to reasonable estimators. Olshen (1967) gives details.

The same problem is present in other spectral analyses. Under the normal perturbation model the Fourier transform has independent coordinates with fixed variance (see Example A-3). This continues to hold under broad alternative perturbation models. Without repeated observations, there is a limit to the accuracy

of estimates. Thought of in the ANOVA context this is not surprising – we can only hope for an accurate indicator of variability. We can think of formulating natural smoothness assumptions to allow increased accuracy.

When applicable, the sampling model is a natural replacement for signal plus noise – the Fourier transform applied to a population is simply a 1-1 transform and so gives a natural parameter to estimate.

Time series has goals in addition to estimating the spectrum at a point. In continuous time problems, there can be a continuous spectrum, and many other functionals are relevant. Tukey (1986) emphasizes this aspect and points to properties of ensembles of time series that are of scientific interest. For ranked data, there is an obvious prediction problem: how will a complete enumeration (or larger sample), turn out? More in line with time series, prediction might be this: people partially rank several quantities; later their true preferences are revealed. How good a guess can be made based on current knowledge?

Time series analysis has a time domain side which is based on models for the observed process. There is a healthy interplay between time and frequency domains. This is presented by Anderson (1971) or Priestly (1981). As with ANOVA, it's nice to see natural projections have an interpretation as reasonable estimates of parameters in a model. Models for data on homogeneous spaces are introduced in the next chapter. There hasn't been any serious work on the interplay with spectral analysis.

There are many modern versions of spectral analysis of time series. There are remedies for all of the problems discussed in the ANOVA section. It seems impossible to survey this briefly. Anderson (1971), Brillinger (1975) and Priestly (1981) are good references; leading to others. With work, most of these ideas should be transferrable to other spectral domains when the need arises.

One final thought: real examples in hard science settings are a sparkling part of time series analysis. If progress is to be made in generalization, similar examples will have to be found.

Chapter 9. Models

Fitting models to data is a popular activity. For data taking values in a group or homogeneous space, the associated representation theory gives neat families of models. Briefly, if $P(x)$ is a probability on X, write $P(x) = e^{h(x)}$ with $h = \log P$. Then expand h in a natural basis b_i of $L(X)$: $h(x) = \Sigma\theta_i b_i(x)$. Any positive probability can be expanded in this way. Truncating the expansion to a fixed number of terms leads to families of simple measures or models; because b_i are orthogonal, θ_i are identifiable.

It is an interesting fact that many models introduced by applied workers fall into this class. The general story is presented first, then a specialization to data on spheres, then a specialization to partially ranked data. A brief review of other approaches to ranked data is followed by a section supplying the relevant exponential family theory.

A. Exponential families from representations

Let G be a group acting transitively on a compact set X. Let $L(X)$ denote the real valued continuous functions on X. Suppose X has an invariant distribution dx. The following abstracts an idea introduced by Lo (1977) and Beran (1979).

Definition. Let Θ be an invariant subspace of $L(X)$ containing the constants. Define a family of measures, one for each $\theta \in \Theta$, by specifying the densities to be

$$P_\theta(dx) = a(\theta)e^{\theta(x)}dx,$$

where $a(\theta)$ is a normalizing constant forcing $P_\theta(dx)$ to integrate to 1.

Suppose Θ is finite dimensional. Let b_0 =constant, $b_1, b_2, \ldots, b_p$ be a basis for Θ. Then the family can be parameterized as

$$(*) \qquad P_\theta(dx) = a(\theta)e^{\theta' b}dx,\ \theta \in \mathbb{R}^p, b = (b_1(x), \ldots, b_p(x)).$$

Lemma 1. *The family * is well parameterized in the sense that $P_\theta = P_{\theta'}$ if and only if $\theta = \theta'$.*

Proof. Only the forward direction requires proof. If $P_\theta = P_{\theta'}$, then

$$(\theta - \theta') \cdot b(x) = \log(a(\theta')/a(\theta)) \text{ for all } x.$$

The left side is a linear combination of $b_1, b_2, \ldots, b_p$ which is constant. But $1, b_1, b_2, \ldots, b_p$ is a basis, so $\theta = \theta'$. □

In applications there is a decomposition into invariant subspaces $L(X) = V_0 \oplus V_1 \oplus V_2 \oplus \ldots$ and Θ's are chosen as a finite direct sum of subspaces. Usually

these nest together neatly to form zeroeth order models (the uniform distribution), 1st order models, etc. The matrix entries of the irreducible representations then provide a convenient basis.

The easiest example is for data on Z_2^k. The exponential families that the group theory suggests are exactly the log-linear models that statisticians fit to $2 \times 2 \ldots \times 2$ tables (k factors). Here a person is classified via k dichotomous variables. This gives rise to a vector in Z_2^k or locates a cell in the table.

A useful entry to the statistical literature is provided by the first few chapters of Gokhole and Kullback (1978). General contingency tables can be treated similarly. Since this is such a well studied area, we will not pursue it further than mentioning the important paper of Darroch, Lauritzen and Speed (1980). This gives an elegant interpretation to setting $\theta_i = 0$ for a large class of models. It would be an important contribution to generalize their ideas to the general group case.

The measure $P_\theta(dx)$ governs a single observation. We model a sample of size n by a product measure

$$P_\theta(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P_\theta(x_i) dx_i.$$

The statistical problem becomes, given a model Θ, and observations x_1, $x_2, \ldots, x_n$, what is a reasonable guess for θ, and how sure are you about the answer.

Remark 1. Crain (1973, 1974, 1976) suggested expansion of log $P(x)$ in a basis of orthogonal functions as a route to nonparametric density estimation. He truncated the expansion at a point depending on sample size. This leads to an approximate density in a finite dimensional exponential family as in the definition above.

Crain's later papers give conditions on how large the cutoff point should be to have the maximum likelihood estimator exist. These are discussed in Section E below.

Remark 2. There is a growing literature on orthogonal series estimators – density estimators based on expanding the density directly as $P(x) = \Sigma\theta_i b_i(x)$. Hall (1986) makes noteworthy contributions providing simple useable estimators and giving sharp rates of convergence. He gives pointers to the literature. Hall's results can be carried over to problems on compact homogeneous spaces in a straightforward way.

Orthogonal series estimators suffer from the possibility of negative density estimates. This is why Crain worked with log $P(x)$. It is a worthwhile project to combine the ideas and bounds of Hall with the ideas of Crain.

Remark 3. One problem encountered with log P: it is badly behaved if $P = 0$. Consider a density on the circle. If $P(x) > 0$ outside an interval, things can be rescaled and there is no trouble. If $P(x) = 0$ on several intervals the problem can be treated as a mixture, but the log P approach is wearing out its welcome.

There are so many other density estimates possible – from histograms, through kernel estimators, through projection pursuit.

On more general homogeneous spaces, problems with vanishing density seem even less approachable.

Remark 4. The definition above is in terms of real valued functions. This works fine for the symmetric group and its homogeneous spaces and for the orthogonal group. In general, $L(X)$ may be taken as all complex functions and a model may be taken as an invariant subspace of $L(X)$. Just as any real function on Z_n can be expanded in terms of $\sin(2\pi jk/n)$ and $\cos(2\pi jk/n)$, any real function on X can be expanded as a real linear combination of the real and imaginary parts of the matrix entries of the irreducible representations that occur in the splitting of $L(X)$.

Remark 5. The models introduced here blend in nicely with the spectral theory of Chapter 8. They are the largest models which allow as sufficient statistics the ingredients of the matching spectral analysis. See E-1 below.

Remark 6. The ideas set out above can be generalized in various ways. One natural extension begins with a space X and a symmetric Markov chain $P(x, dx)$ on X. Symmetric chains can be orthogonally diagonalized, and the eigen vectors provide a convenient orthogonal basis for $L(X)$. There are chains that don't arise from groups where this basis can be written explicitly. See Banni and Ito (1986, 1987) or Diaconis and Smith (1987). It is not clear if these models can be connected to the underlying chain.

A word of caution: I find the statistical community introduces models much too easily. In some cases, there is a justification: "height is the sum of a lot of small factors, so heights should be approximately normally distributed" or "the number of accidents is the sum of a lot of roughly independent binomial variables with small parameters, so accidents should be approximately Poisson." In some cases linearity or physical justification (and repeated comparison with reality) justify models: Gauss' discovery of Ceres, Bright-Wigner distributions in particle physics or multinomial distributions in genetics are examples.

The cases where some slim justification is given seem alarmingly few to me. Usually, one is contemplating some data and a model is chosen for convenience as a way of doing data analysis. This is a curve fitting approach and is fine, except that the product model assumes independence. Further, the assumptions about $P_\theta(dx)$ may be a drastic oversimplification. One may well do better looking directly at the data using spectral analysis, or a convenient ad hoc approach.

I must admit that I too find ad hoc modeling attractive and occasionally useful – it seems like a most worthwhile project to try to isolate what good comes out of the modeling paradigm and attempt to build a theory that optimizes this good instead of behavior in a non-existent fantasy land of iid repetitions.

B. Data on spheres.

Spherical data is discussed in Chapter 5-C. One important special problem is testing for uniformity. A large number of special tests have been suggested. These are reviewed by Mardia (1972) and Watson (1983). We discuss here one of the earliest tests and follow its later developments.

Let $X_1, X_2, \ldots, X_n$ be unit vectors on the sphere S^p in p dimensions. Define the sample resultant $\overline{R}$ and sample mean direction $U(\overline{\theta})$ by

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{R}U(\overline{\theta}).$$

Intuitively, if X_i are uniform, $\overline{R}$ will be "small" because there will be a lot of cancellation. If X_i are non-uniform and cluster around some point, then $\overline{R}$ will be "large." Rayleigh (1919) worked out the distribution of $\overline{R}$ under the uniform distribution and could thus propose a test "reject uniformity if $\overline{R} > r$" where r is chosen to achieve a given proportion of false rejections. A nice derivation of Rayleigh's results is given by Feller (1971, pg. 32).

Questions and alternate tests immediately suggest themselves. Observe that Rayleigh's test is invariant: $\overline{R}$ does not change if $X_1, X_2, \ldots, X_n$, are replaced by $\Gamma X_1, \ldots, \Gamma X_n$, Γ orthogonal. On the negative side, Rayleigh's test would not be appropriate if the X_i tend to cluster either close to a point *or* its antipode. When is this test a good one? Some answers have come from statistical theory.

Independent of Rayleigh, a class of natural non-uniform distributions was developed and used by Von Mises and Fisher. These have the form

$$P_{\mu,k}(dx) = C_p(k)e^{k\mu' x}dx$$

with $x \in S^p$, dx the uniform distribution, $\mu \in S^p$, and $k \geq 0$. The normalizing constant is

$$C_p(k) = k^{(p-1)/2}/(2\pi)^{p/2}I_{(p-1)/2}(k)$$

with $I_r(k)$ the modified Bessel function of the first kind.

The $P_{\mu,k}$ have "mean direction" μ and as k increases are more and more concentrated about μ. They arise naturally from the first hitting place of a Brownian particle with drift on the sphere. Watson (1983, Chapter 3) discusses this and other justifications.

A nice result is that the likelihood ratio test of

$$H_0: k = 0 \text{ vs. } H_1: k > 0, \ \mu \text{ unknown}$$

reduces to Rayleigh's test. Further, Rayleigh's test is the uniformly most powerful invariant test of uniformity versus $P_{\mu,k}(dx)$. These results are due to Beran (1968) who discusses their analog on compact homogeneous spaces. Giné (1975), Wellner (1979), and Jupp and Spurr (1985) amplify and develop these ideas. Closely related developments in the signal processing literature are surveyed by Lo and Eshleman (1979).

These developments give a pleasing answer to the original question: when is Rayleigh's test good – it's good if data cluster about one point μ in a spherically symmetric way.

Remark. I cannot resist reporting some background on Fisher's motivation for working with the distribution discussed above. This story was told to me in 1984 by the geologist Colin B. B. Bull. Dr. Bull was a student in Cambridge in the early 1950's. One day he ran across the street in haste and knocked an old man off a bicycle! The old man seemed dazed. When asked where he was bound he replied "India." It turned out to be R. A. Fisher who was meeting a train enroute to a visit to the Indian Statistical Institute. A month later, Bull met Fisher at Cambridge and again apologized. Fisher asked what area Bull worked in. Bull explained that a group of geologists was trying to test Wegener's theory of continental drift. Wegener had postulated that our current continents used to nest together. He tested this by looking at the distribution of a wide variety of bird, animal and plant life – arguing that matching points had close distributions.

Geologists found themselves far afield in trying to really understand Wegener's arguments. They searched for data that were closer to geology. They had hit on the distribution of magnetization angle in rocks. This gave points naturally distributed on the sphere. They had two distributions (from matching points on two continents) and wanted to test if the distributions were the same.

Fisher took a surprisingly keen interest in the problem and set out to learn the relevant geology. In addition to writing his famous paper (which showed the distributions were different) he gave a series of talks at the geology department to make sure he'd got it right. Bull told me these were very clear, and remarkable for the depth Fisher showed after a few months study.

Why did Fisher take such a keen interest? A large part of the answer may lie in Fisher's ongoing war with Harold Jeffries. They had been rudely battling for at least 30 years over the foundations of statistics. Jeffries has never really accepted (as of 1987!) continental drift. It is scarcely mentioned in Jeffries' book on geophysics. Fisher presumably had some extra-curricular motivation.

The motivation for Rayleigh's and Von Mises' work seems equally fascinating! Watson (1983, Chapter 3) gives a good set of pointers.

There is a second family of probabilities on S^p that has received a good deal of attention. The *Bingham densities* are defined on S^p as

$$b_p(D)\exp\{tr[DR'xx'R]\}dx$$

where D is a $p \times p$ diagonal matrix with (p,p) entry zero, and R is a $p \times p$ orthogonal matrix.

These densities are invariant under $x \to -x$ and so are possible models for unsigned directional data – lines in $\mathbf{R}^p$ (or points in projective space). A host of properties and characterizations of these densities are known.

Beran (1979) points out that both the Fisher-Von Mises and Bingham families fit nicely with the definition of models given in Section A. Here, the group $SO(p)$ of $p \times p$ orthogonal matrices with determinant 1 operates transitively on the space $X = S^p$. Take $L(X)$ as the continuous real valued functions on X.

Let P_k be the homogeneous polynomials (in $\mathbb{R}^p$) of degree k. Let M_k be the subspace of harmonic functions in P_k: $M_k = \{f: \nabla^2 f = 0\}$ where $\nabla^2 = \sum_{i=1}^{p} \frac{\partial^2}{\partial x_i^2}$. These M_k are invariant and irreducible under the action of SO_p. Further, $L(X) = \oplus_{k=0} M_k$ as a Hilbert space direct sum. Proofs are in Dunkl and Ramirez (1971).

Following the definition, M_0 – the zero-th order model gives only the uniform distribution. $M_0 \oplus M_1$ – the first order models is obviously spanned by 1, $x_1, x_2, \ldots, x_p$ (these are all killed by ∇^2). The associated exponential family is the Fisher-Von Mises family.

A second-order model is defined by $M_0 \oplus M_1 \oplus M_2$. Beran (1979) shows these are spanned by $\{x_i x_j\} - \{x_p^2\}$, giving the Bingham distribution. In general, a basis for $\bigoplus_{k=0}^{r} M_k$ consists of all distinct monomials of degree r and $r-1$, excluding x_p^r if r is even or x_p^{r-1} if r is odd.

Some more technical discussion of estimates and their properties is given in Section E below.

C. Models for permutations and partially ranked data.

Begin with a data set on the symmetric group S_n. Say $f(\pi)$ is the proportion of the data choosing rankng π. In working with such data it seems natural to begin by looking at first order statistics: the proportions ranking each item first, or last, and more generally the proportion ranking item i in position j. The average rank given each item is a popular summary which is a mean of these first order statistics.

Paul Holland suggested working with the exponential family through the first order statistics in the early 1970's. This leads to

Holland's model. Let ρ be the $n-1$ dimensional irreducible representation of S_n. Let $\mathrm{Mat}(n-1)$ be the set of all $n-1$ by $n-1$ real matrices. Define

$$P_\theta(\pi) = c(\theta) e^{Tr[\theta \rho(\pi)]}; \text{ for } \theta \in \mathrm{Mat}(n-1),$$

with $c(\theta)^{-1} = \Sigma_\pi \, e^{Tr(\theta(\rho(\pi))}$.

Remarks. These models are well parameterized by $\theta \in \mathrm{Mat}(n-1) = \mathbb{R}^{(n-1)^2}$. To give an example, consider a simple sub family:

$$Q_\theta(\pi) = c(\theta) e^{\theta \delta_1(\pi(1))}, \ \theta \in \mathbb{R}.$$

This can be described intuitively as "there is some special chance of ranking item 1 in position 1; whether or not this is done, the rest of the permutation is chosen uniformly.

If item 1 were carefully ranked, and then the others chosen at random, the appropriate family would be

$$Q_\theta(\pi) = c(\underset{\sim}{\theta}) e^{\theta_1 \delta_1(\pi(1)) + \ldots + \theta_{n-1} \delta_{n-1}(\pi(1))}, \ \underset{\sim}{\theta} \in \mathbb{R}^{n-1}.$$

Holland's model extends these considerations to a full first order model.

Joe Verducci (1982) began with Holland's model and the observation that $(n-1)^2$ parameters is still a lot to work with and think about. He introduced some natural low dimensional subfamilies and fit them successfully to real data sets. One of his nice observations is that some of Mallows' metric models introduced in Chapter 6-A-1 are subfamilies of first order exponential families.

Consider

$$Q_\lambda(\pi) = c(\lambda)e^{\lambda H(\pi,\pi_0)}\lambda \in \mathbf{R},\ H\ =\ \text{Hamming distance.}$$

For fixed π_0, this is a subfamily of Holland's, taking $\theta = \lambda\rho(\pi_0^{-1})$. Of course, if π_0 is also treated as a parameter, the two models are different. Verducci observed that replacing H by Spearman's S^2 also gives a first order model.

Arthur Silverberg (1980) began to work with second order models using the proportion ranking i, i' in position j, j'. Verducci (1982) realized the connection with group representations could help sort out questions of when a model is full, or well parameterized.

Silverberg worked with q-permutations, where people rank their favorite q out of n. This would be data on S_n/S_{n-q} in the language of Chapter 7. Generalizing slightly, let λ be a partition of n. Let $X = S_n/S_{\lambda_1} \times S_{\lambda_2} \ldots \times S_{\lambda_k}$ be the set or partial rankings of shape λ. Using Young's rule, and notation of Chapter 7,

$$L(X) = M^\lambda = \bigoplus_{\nu \vdash n} k(\nu;\lambda)S^\nu$$

where the sum is over all partitions ν of n which are larger than λ in the partial order of majorization and $k(\nu,\lambda)$ is the multiplicity of S^ν in M^λ. See the remarks to Theorem 1 in Chapter 7A. Restricting attention to a few of the pieces in this decomposition gives models of various sorts.

If $\lambda = (\lambda_1, \ldots, \lambda_k)$, the $n-1$ dimensional representation appears $(k-1)$ times $(k(n-1,1);\lambda) = k-1)$. The direct sum of these $k-1$ dimensional subspaces has dimension $(k-1)(n-1)$ and it spans the first order model.

Let us apply Young's rule to answer a question posed by Silverberg (1980) – what is the dimension of 2nd order models for q-permutation data. The partition involved is $n-q, 1^q$. Suppose that $2 \le q \le n-q$. Second order models are associated with partitions $(n-2,1,1)$ and $(n-2,2)$. By Young's rule, the multiplicity of each in $M^{1^q,n-q}$ is $\binom{q}{2}$. By the hook length formula of Chapter 7, the dimension of $S^{n-2,1,1}$ is $(n-1)(n-2)/2$. The dimension of $S^{n-2,2}$ is $n(n-3)/2$.

If we also include the first order component, the dimension of the second order model is

$$q\,(n-1) + \binom{q}{2}\binom{n-1}{2} + \binom{q}{2}\frac{n(n-3)}{2}.$$

Of course, it is important to keep the pieces separated, both for computation and inference.

The models discussed above have not been broadly applied. At present, there are no simple processes that lead to these models, nor simple interpretations or benefits from them. Since exponential families have such a good track record in

these directions, it seems like a worthwhile project to study and develop properties of low order exponential families on partially ranked data.

Some technical and practical aspects of the models in this section are discussed in Section E of this chapter.

D. Other models for ranked data.

The models proposed for ranked data in the previous section and the metric models of Chapter 6 have a distinctly ad-hoc flavor to them. There have been energetic attempts in the psychological literature to develop models for ranked data that are grounded in some more basic processes. This section briefly describes some of the models and gives pointers to the literature.

To fix a problem, consider an experiment in which p tones are played for a subject who is to rank them in order of loudness. It is an empirical fact that even a single subject, asked to repeat this task on different days, gives different answers. To account for this variability, Thurstone introduced an unobservable "discriminal process" of the form u_1+X_1, $u_2+X_2, \ldots u_p+X_p$ where $u_1, u_2, \ldots, u_p$ are fixed constants, and $X_1, \ldots, X_p$ are random variables, independent with the same distribution. It is postulated that on a given trial, a subject rank orders tone i in position j if $u_i + X_i$ is the jth largest.

Thurstone proposed normal distributions for the X_i. With a distribution fixed, one can estimate best fitting u_i and compare data and model. There has been a lot of experimental work showing a good fit for certain tasks. An extensive, readable review of this work appears in Luce and Suppes (1965).

A second line of work stems from a simple model put forward by Luce (1959). This postulates an unobservable system of weights $w_1, w_2, \ldots, w_p$. It is proposed that a subject ranks items by choosing the first ranked item with probability proportional to W_i. This choice being I, the second ranked item is chosen with probability proportional to $\{w_j\} - w_I$, and so on.

This model has also been fit to data with some success. Holman and Marley proved that if the underlying random variables X_i in Thurstone's approach have an extreme value distribution $P\{X < t\} = e^{-e^{-t}}$, $-\infty < t < \infty$, the resulting choice probabilities are given by Luce model as well. Yellott (1977) gives references, proves a converse, and suggests some intriguing open probability problems.

Yellott's results deal with location shifts of extreme value distributions. Louis Gordon (1983) has observed a neat reformulation: consider the basic weights $w_1, \ldots w_p$ in Luce's model. Let $Y_1, Y_2, \ldots Y_p$ be independent and identically distributed standard exponential variables: $P(Y > t) = e^{-t}$. Put a probability on permutations by considering the order statistics of $Y_1/w_1, \ldots, Y_p/w_p$. Gordon shows this induces the distribution of Luce's sequential model. Since the log of an exponential variable has an extreme value distribution, this is a special case of the results described by Yellott. Gordon shows how to use the representation to give an efficient algorithm for generating random permutations from this distribution.

Independent of the literature cited above, Plackett (1975) developed a family of non-uniform probabilities on permutations. Plackett's first order models are the same as the Luce models. These are fit to some race horse data by Henery

(1981). An order statistics version of Plackett's higher order model is given by Dansie (1983). Plackett's motivation is interesting. One has available data on the chance that a horse finishes first in a race. One wants to predict the chance that the horse "shows" (finishes in the top 3). Plackett fit a model on the final permutation using the first order data. This approach is the basis of several *believable* systems for beating the races. See Zambia and Hausch (1984).

Models like Luce's have been extended, axiomatized, and tested by modern mathematical psychologists. The extensions account for practical difficulties such as the irrelevance of alternatives. If Luce's model is taken literally, one postulates a weight associated to the ith object independent of the other choices available. This easily leads to thought experiments generating data at variance with such a model. The following example is due to L. J. Savage.

Suppose you are indifferent between a trip to Paris and a trip to Rome. Thus w(Paris) $\doteq$ w(Rome). You clearly prefer Paris + \$10 to Paris. On Luce's model, if asked to choose between Paris, Paris + \$10, or Rome, you choose Rome about 1/3 of the time. Something is wrong here – it is unlikely that such a small inducement would change things so drastically. Tversky (1972) gives other examples and discussion.

One simple way around this objection is to allow the weights to depend on the problem under consideration. Going further, after the first choice is made, the second choice can be modeled by a new set of weights. But then any set of choice probabilities can be matched exactly so no test of the model is possible.

Some interesting half-way houses have been worked out. For example, Tversky (1972) describes choice by a hierarchical elimination process. Each alternative is viewed as a collection of measurable aspects. To make a choice, one selects an aspect with probability proportional to its measure. This eliminates all alternatives not possessing this aspect. The process continues until one alternative remains. For example, in choosing a restaurant for dinner, we may first choose type of food (e.g. seafood), then location, then price. Tversky and Sattath (1979) consider a subclass of these hierarchical models called preference trees which have many appealing properties.

The present state of the theory is this – no one claims to have a reasonable, believable and testable theory of how we perform ranking or choice. There *is* a list of constraints and desiderata on potential theories. These offer insight into choice behavior and rule out many naive suggestions. Thurstone's models and Luce's model are seen as straw men which triggered these investigations. Slight elaborations of these models have proven useful in horse race betting.

E. Theory and practical details.

1. *Justifying exponential families.*

Return to the setting of Section A – exponential families on a space X. One justification for these models P_θ that statisticians have developed goes as follows. Consider first a sample $X_1, X_2, \ldots X_n$ from such a family with unknown θ. The

sufficient statistics are

$$\bar{b}_i = \frac{1}{n}\sum_{j=1}^{n} b_i(X_j).$$

Any question about which $\theta \in \Theta$ generated the data can be answered as well from the averages $\bar{b}_i$ as from the full set of data. Often a working scientist, or common sense, will have reduced the data in just this way.

For example, if the data are n rankings of p items, it is natural to summarize the data by collecting together the number of people ranking item i in position j. This amounts to the first order models for permutations described in Section C above.

If summarization is deemed sensible, one may ask for the richest or fullest model for which this summarization is "legal." A classical theorem, the Koopman-Pitman-Darmois theorem, implies that this is the exponential family P_θ through these sufficient statistics.

This line of thinking has several modern versions. The Danish school of Martin-Löf-Lauritzen formalizes things as extreme point models. Lauritzen (1984) contains a clear description.

A Bayesian version is given by Diaconis and Freedman (1984). Briefly, if $X_1, X_2, \ldots, X_n$ (the data) are judged invariant under permutations (exchangeable) and more data of the same type could be collected, then de Finetti's theorem implies that the data were generated by a mixture of independent and identically distributed variables. *If* the $\bar{b}_i$ summarize the data, in the sense that given $\{\bar{b}_i\}$ all sequences $X_1, \ldots, X_n$ with these b_i are judged equally likely, then an extension of de Finetti's theorem implies the data are generated by a mixture of the exponential families introduced above. This brief description omits some technical details but is correct for the examples introduced below. Diaconis and Freedman also given versions of the Koopman-Pitman-Darmois theorem suitable for discrete data. Diaconis and Freedman (1988) give versions for continuous data.

There is a related motivation in the non Bayesian setting when x_i are iid: the maximum entropy distribution for $X_1, \ldots, X_n$ given the summaries $\{\bar{b}_i\}$ is the member $P_{\hat{\theta}}$ of the exponential family with $\hat{\theta}$ chosen so the mean of $P_{\hat{\theta}}$ equals $\bar{b}_i$. See Kullback (1968) or Posner (1975) for details.

These justifications boil down to the following: if the data are collected and it is judged reasonable to summarize by averages $\{\bar{b}_i\}$ then the exponential family P_θ gives the only probability model justifying this summary.

2. *Properties of exponential families.* Consider a sample $X_1, X_2, \ldots, X_n$ from P_θ, where it is assumed $\theta \in \mathbb{R}^p$. The maximum likelihood estimate of θ is a value $\hat{\theta}$ which maximizes $\Pi\, P_\theta(x_i)$. If X is finite this is an intuitively plausible procedure. It also has the Bayesian justification of being the (approximate) mode of the posterior distribution. Finally, it has quite a good track record in applied problems. The log-likelihood function is

$$L_n(\theta) = \theta' \sum_{i=1}^{n} b(X_i) - n\ \psi(\theta), \psi(\theta) = -\log\, a(\theta).$$

From the standard theory of maximum likelihood estimation in regular exponential families (see for example, Barndorf-Nielsen (1978) or Brown (1987)), we have

(i) $L_n(\theta)$ is strictly concave in θ.

(ii) $\psi(\theta)$ is analytic and $\nabla\ \psi(\theta) = E_\theta(b(x))$, $\nabla^2\ \psi(\theta) = \text{cov}_\theta(b(x))$, $\nabla^2\ \psi(\theta)$ is positive definite.

(iii) With probability one, there is an integer $n_0 = n_0(X_1, X_2, \ldots)$ such that the MLE $\hat{\theta}$ exists for all $n \geq n_0$. If the MLE exists, it is unique.

Crain (1974, 1976) gives results proving that, for continuous carriers,

- If the number of observations is larger than dim Θ, then the MLE exists.
- If dim Θ is allowed to grow with the sample size, then the "nonparametric density estimator" $f^*(x) = a(\theta^*)e^{\theta^* (x)}$ (θ^* the MLE) converges to the true sampling density. When X is finite this is clear, for eventually Θ becomes the set of all functions and $f^*(x)$ is then the frequency cell count for a multinomial.

(iv) A necessary and sufficient condition for the existence of the MLE is that $\bar{b}_i = \frac{1}{n}\sum_{i=1}^{n} b(X_i) \in$ int Hull (K), where K=range $\{b(x);\ x \in X\} \subset \mathbb{R}^p$.

(v) The MLE $\hat{\theta}$ exists iff the equations

$$E_\theta(b(X)) = \frac{1}{n}\sum_{i=1}^{n} b(X_i)$$

have a solution. When a solution exists it is unique and is the MLE. Thus, the MLE is that value of θ that makes the theoretical expectation of t equal its observed average.

(vi) The MLE is almost surely a consistent estimate of θ, and as n tends to infinity. Further, for large n, the difference between $\hat{\theta}$ and θ has an approximate normal distribution:

$$n^{\frac{1}{2}}(\hat{\theta} - \theta) \sim n(0,\ \nabla^2\psi(\theta)^{-1}).$$

This allows confidence intervals for θ, by using $\nabla^2\psi(\hat{\theta})^{-1}$ for the covariance matrix.

(vii) We have $P_\theta(dx) = a(\theta)e^{\theta' b}dx$. The sufficient statistics are $\bar{b}_i$. Following Crain (1974), consider a second expansion:

$$\frac{P_\theta(dx)}{dx} = \lambda_0 + \Sigma\ \lambda_i\ b_i(x).$$

If the b_i are orthogonal with respect to dx, then

$$\lambda_i = E_\theta(b_i) = E_\theta(\bar{b}_i).$$

In practice, $\hat{\theta}$ will not have a nice closed form expression. It will have to be determined numerically. There is a reasonable discussion of Newton-Raphson (called the method of scoring) in C. R. Rao's (1965) book. Beran (1979) suggests some other procedures as does Crain (1976).

There has not been a lot of work on a reasonable Bayesian analysis for these models. Consonni and Dawid (1985) develop some ideas which may generalize. A second starting place is to consider, as in Diaconis and Ylvisaker (1979), conjugate priors, and then their mixtures. There is probably some nice mathematics along the lines of Diaconis and Ylvisaker (1983), but bringing in some group theory.

3. *Introducing covariates.* A. P. Dempster (1971) has suggested a reasonable method of enlarging standard exponential families to include covariates. Suppose X is a finite homogeneous space. We observe pairs (x_i, z_i), $1 \leq i \leq n$ where $x_i \in X$ and $z_i \in \mathbb{R}^p$ is a covariate. Suppose that $b_1, b_2, \ldots, b_q$ is a basis for the model as above. The analog of Dempster's suggestion is the following family of probability densities (with respect to the uniform measure dx):

$$f(x|z) = \exp\left(\alpha + \sum_{i=1}^{q}\sum_{i=1}^{p} \phi_{ij} z_j b_i(x)\right).$$

Here, of course α is a normalizing constant and ϕ_{ij} are $p \cdot q$ parameters to be estimated. This amounts to the usual log-linear expansion

$$\exp\left(\alpha + \sum_{i=1}^{q} \alpha_i b_i(x)\right)$$

with $\alpha = \sum_{j=1}^{p} \phi_{ij} x_j$. Dempster discusses some of the calculus of such families, as well as some of the numerical and philosophical problems associated to such models. Dempster's analysis is an early version of the currently popular generalized linear model (GLM). See McCullagh and Nelder (1983). It may be that some of these analyses can be easily run in GLM.

Aldous, D. (1982). Markov chains with almost exponential hitting times. *Stochastic Proc. Appl.* **13**, 305–310.

Aldous, D. (1983a). Random walk on finite groups and rapidly mixing Markov chains. In *Seminaire de Probabilites* **XVII**, 243–297. Lecture Notes in Mathematics 986.

Aldous, D. (1983b). Minimization algorithms and random walk on the d-cube. *Ann. Prob.* **11**, 403–413.

Aldous, D. (1985). Self-intersections of random walks on discrete groups. *Math. Proc. Camb. Phil. Soc.* **98**, 155–177.

Aldous, D., and Diaconis, P. (1986). Shuffling cards and stopping times. *American Mathematical Monthly* **93**, 333–348.

Aldous, D., and Diaconis, P. (1987a). Strong uniform times and finite random walks. Advances in Appl. Math. **8**, 69–97.

Aldous, D., and Diaconis, P. (l987b). Examples of strong uniform times. Unpublished manuscript.

Anderson, T. (1971). *THE STATISTICAL ANALYSIS OF TIME SERIES.* Wiley, New York.

Anderson, S. (1975). Invariant normal models. *Ann. Statist.* **3**, 132–154.

Asimov, D. (1983). The grand tour. *SIAM Jour. Sci. Statist. Comp.* **6**, 128–143.

Athreya, K. B., and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245**, 493–501.

Athreya, K. B., McDonald, D., and Ney, P. (1978). Limit theorems for semi-Markov processes and renewal theory for Markov chains. *Ann. Prob.* **6**, 788–797.

Bailey, R. A., and Rowley, C. A. (1986). General balance and treatment permutations. Technical Report, Dept. of Statistics, Rochamsted Experimental Station, Harpenden, Herts AL5 2JQ, United Kingdom.

Bahadur, R. R. (1961). On classification based on responses to n dichotomous items. In H. Solomon (ed.) *STUDIES IN ITEM ANALYSIS.* Stanford University Press, Stanford, CA.

Bannai, E. and Ito, T. (1984). *Algebraic Combinatorics I Association Schemes*, Benjamin, Menlo Park, CA.

Bannai, E. and Ito, T. (1986). Current research on algebraic combinatorics: *Graphs and Combinatorics* **2** 287–308.

Barndorff-Nielsen, O. (1978). *INFORMATION AND EXPONENTIAL FAMILIES.* Wiley, New York.

Bartfai (1966). Grenzwertsatze auf der Kreisperispherie und auf kompakten Abelschen gruppen. *Studia Sci. Math. Hung.* **1**, 71–85.

Beran, R. (1968). Testing for uniformity on a compact homogeneous space. *Jour. Appl. Prob.* **5**, 177–195.

Beran, R. (1979). Exponential models for directional data. *Ann. Statist.* **7**, 1162–1178.

Berger, J. (1985). *STATISTICAL DECISION THEORY AND BAYESIAN ANALYSIS.* Springer-Verlag, New York.

Berger, P. (1973). On the distribution of hand patterns in bridge: man-dealt versus computer-dealt. *Canadian J. Statist.* **1**, 261–266.

Beth, T. Jungnickel, D., and Lenz, H. (1986). *Design Theory.* Cambridge University Press, Cambridge.

Bhattacharya, P. K., Chernoff, H., and Yang, S. S. (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist.* **11**, 505–514.

Bhattacharya, R. N. (1972). Speeds of convergence of the n-fold convolution of a probability measure on a compact group. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **25**, 1–10.

Biggs, N. (1974). *ALGEBRAIC GRAPH THEORY.* Cambridge University Press, London.

Bloomfield, P. (1976). *FOURIER ANALYSIS OF TIME SERIES: AN INTRODUCTION.* Wiley, New York.

Bolker, E. (1987). The finite Radon transform. *Contemporary Math.* **63**, 27–50.

Bolthausen, E. (1984). An estimate of the remainder in a combinatorial central limit theorem. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **66**, 379–386.

Bondesson, L. (1983). A simple generalization of Poincare's shuffling theorem. In A. Gut. L. Holst (eds.) *PROBABILITY AND MATHEMATICAL STATISTICS. Essays in Honor of Carl-Gustav Esseen, pg. 11–15.* Department of Mathematics, Uppsala University, Uppsala.

Borel, E. and Cheron, A. (1955). *THEORIE MATHEMATIQUE DU BRIDGE, 2nd ed.*, Gauthier Villars, Paris.

Bougerol, P. (1983). *Un Mini-Cours Sur Les Couples de Guelfand.* Pub. du Laboratoire de Statistique et Probabilites, Universite Paul Sabatier, Toulouse.

Bougerol, P. and Lacroix, J. (1985). *LIMIT THEOREMS FOR PRODUCTS OF RANDOM MATRICES.* Birkhauser, Boston.

Bourbaki, N. (1968). *GROUPES ET ALGEBRES DE LIE, Chapitres 4, 5, 6.* Herman, Paris.

Box, G., and Cox, D. (1964). An analysis of transformations. *J. R. Statist. Soc. B* **26**, 211–252.

Box, G., and Tiao, G. (1973). *BAYESIAN INFERENCE IN STATISTICAL ANALYSIS.* Addison-Wesley, Reading, Massachusetts.

Brams, S. and Fishburn, P. (1983). *Approval Voting.* Birkhauser, Boston.

Breiman, L., and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *JASA* **80**, 580–597.

Brillinger, D. (1975). *TIME SERIES, DATA ANALYSIS AND THEORY.* Holt, Rinehart, and Winston, New York.

Brillinger, D. (1987). Some statistical methods for random processes, data from seismology and neurophysiology. *Ann. Statist.* **16**, 1–54.

Brown, L. D. (1987). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Institute of Math. Statist., Hayward.

Calvin, L. (1954). Doubly balanced incomplete block designs for experiments in which the treatment effects are correlated. *Biometrices* **10**, 61–88.

Chamberlin, J. R., Cohen, G., and Coombs, C. (1981). A case study of social choice: Four Presidential Elections of The American Psychological Association. Technical Report. Institute of Public Policy Studies.

Chernoff, H. (1981). An analysis of the Massachusetts numbers game. *Mathematical Intelligencer* **3**, 166–172.

Chihara, L. (1987). On the zeros of the Askey-Wilson polynomials, with application to coding theory. *SIAM Jour. Math. Analysis* **18**, 183–207.

Chung, F., Diaconis, P., and Graham, R. L. (1987). A random walk problem arising in random number generation. *Ann. Prob.* **15**, 1148–1165.

Cochran, W., and Cox, G. (1957). *EXPERIMENTAL DESIGNS*, 2nd ed. Wiley, New York.

Cohen, A. (1982). Analysis of large sets of ranking data. *Comm. Statist.-Th. Meth.* **11**, 235–256.

Cohen, A. and Mallows, C. (1980). Analysis of ranking data. Technical memorandum, Bell Laboratories. Murray Hill, New Jersey.

Consonni, G., and Dawid, A. P. (1985). Decomposition and Bayesian analysis of invariant normal linear models. *Linear Alg. Appl.* **70**, 21–49.

Coombs, E. (1964). *A THEORY OF DATA*. Wiley, New York.

Crain, B. (1973). A note on density estimation using orthogonal expansions. *JASA* **68**, 964–965.

Crain, B. (1974). Estimation of distributions using orthogonal expansions. *Ann. Statist.* **2**, 454–463.

Crain, B. (1976). Exponential models, maximum likelihood estimation, and the Haar condition. *JASA* **71**, 737–745.

Critchlow, D. (1985). *METRIC METHODS FOR ANALYZING PARTIALLY RANKED DATA*. Lecture Notes in Statistics No. 34, Springer-Verlag, Berlin.

Critchlow, D. (1986). A unified approach to constructing nonparametric rank tests. Technical Report #376, Dept. of Statistics, Stanford University.

Csiszar, I. (1962). Informationstheoretische konvergenzbegriffe. Im Raum der Wahrscheinlichkeits-verteilingen. *Publications of Mathematical Institut, Hungarian Academy of Sciences* **VII**, 137–158.

Culbertson, E. (1934). *CONTRACT BRIDGE RED BOOK ON PLAY*. Winston, Philadelphia.

Dansie, B. R. (1983). A note on permutation probabilities. *J. R. Statist. Soc.* **45**, 22–24.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8**, 664–672.

Davis, P. (1979). *CIRCULANT MATRICES*. Wiley, New York.

de Bruijn, N. (1964). Polya's theory of counting. In E. Beckenbach (ed.) *APPLIED COMBINATORIAL MATHEMATICS*, 144–184. Wiley, New York.

Dempster, A. P. (1971). An overview of multivariate analysis. *J. Multivar. Anal.* **1**, 316–346.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1–38.

Devroye, L. (1986). *NON-UNIFORM RANDOM VARIATE GENERATION*. Springer-Verlag, New York.

Diaconis, P. (1983). Projection pursuit for discrete data. Technical Report #198, Dept. of Statistics, Stanford University. To appear in *Scandinavian Jour. Stat.*

Diaconis, P. (1985). Theories of data analysis: from magical thinking through classical statistics. In D. Hoaglin, F. Mosteller, J. Tukey (eds.) *EXPLORING DATA TABLES, TRENDS AND SHAPES*, 1–36, Wiley, New York.

Diaconis, P. (1986). Application of non-commutative Fourier analysis to probability problems. To appear in P. L. Hennequin (ed.) *Proc. 16th St. Flour Conf. in Probability.* Technical Report #275, Dept. of Statistics, Stanford University.

Diaconis, P. (1989). Spectral analysis for ranked data. To appear in *Ann. Statist.*

Diaconis, P. and Fill, J. (1988). Strong stationary times and duality. Technical Report, Dept. of Statistics, Stanford University.

Diaconis, P., and Freedman, D. (1984). Partial exchangeability and sufficiency. In J. K. Ghosh and J. Roy (eds.) *PROC. INDIAN STATIST. INST. GOLDEN JUBILEE INT'L. CONFERENCE ON STATISTICS: APPLICATIONS AND NEW DIRECTIONS*, 205–236, Indian Statistical Institute, Calcutta.

Diaconis, P. and Freedman, D. (1987). A dozen de Finetti style theorems in search of a theory. *Ann. Institut Henri Poincaré* **Sup. 23** 397–423.

Diaconis, P. and Garsia, A. (1988). Gelfand pairs: a survey. To appear in Proc. Institut. Math. Appl. Workshop on Invariant Theory and Young Tableaux.

Diaconis, P., and Graham, R. (1977). Spearman's footrule as a measure of disarray. *J. R. Statist. Soc. B* **39**, 262–268.

Diaconis, P., and Graham, R. (1985a). The Radon transform on Z_2^k. *Pacific J. of Math* **118**, 323–345.

Diaconis, P., and Graham, R. (1985b). Monotonicity properties of random walk. Unpublished manuscript.

Diaconis, P., Graham, R. L., and Kantor, W. M. (1983). The mathematics of perfect shuffles. *Advances in Applied Math.* **4** 175–196.

Diaconis, P., and Mallows, C. (1985). The trace of a random matrix. Unpublished manuscript.

Diaconis, P. and Rockmore, D. (1988). Efficient computation of the Fourier transform on finite groups. Technical Report 292, Department of Statistics, Stanford University.

Diaconis, P., and Shahshahani, M. (1981). Generating a random permutation with random transpositions. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **57**, 159–179.

Diaconis, P., and Shahshahani, M. (1986a). Products of random matrices as they arise in the study of random walks on groups. *Contemporary Mathematics* **50**, 183–195.

Diaconis, P., and Shahshahani, M. (1986b). On square roots of the uniform distribution on compact groups. *Proc. American Math'l Soc.* **98**, 341–348.

Diaconis, P., and Shahshahani, M. (1987a). The subgroup algorithm for generating uniform random variables. *Prob. in Engineering and Info. Sciences* **1**, 15–32.

Diaconis, P., and Shahshahani, M. (1987b). Time to reach stationarity in the Bernoulli-Laplace diffusion model. *SIAM J. of Math'l Analysis* **18**, 208–218.

Diaconis, P., and Smith, L. (1988). Fluctuation theory for Gelfand pairs. Unpublished manuscript.

Diaconis, P., and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269–281.

Diaconis, P., and Ylsivaker, D. (1985). Quantifying prior opinion. In J. M. Bernardo. M. H. DeGroot, D. V. Lindley, and A.F.M. Smith (eds.) *BAYESIAN STATISTICS 2, Proc. of 2nd Valencia Int'l Meeting 9-83*, 133–156, North-Holland, Amsterdam.

Diaconis, P., and Zabell, S. (1982). Updating subjective probability. *JASA* **77**, 822–830.

Dieudonné, J. (1970). *TREATISE ON ANALYSIS*, **Vol. II**, Academic Press, New York.

Dieudonné, J. (1978). *TREATISE ON ANALYSIS*, **Vol. VI**, Academic Press, New York.

Dodge, Y. (1985). *ANALYSIS OF EXPERIMENTS WITH MISSING DATA.* Wiley, New York.

Doob, J. (1953). *STOCHASTIC PROCESSES.* Wiley, New York.

Donner, J. R., and Uppuluri, V.R.R. (1970). A markov chain structure for riffle shuffling. *SIAM J. Appl. Math.* **18**, 191–209.

Doran, G. (1979). The Hare voting system is inconsistent. *Political Studies* **27**, 283–286.

Dudley, R. M. (1968). Distances of probability measures and random variables. *Ann. Math. Statist.* **31**, 1563–1572.

Duncan, O. D., and Brody, C. (1982). Analyzing n rankings of three items. In R. M. Hansen et al (eds.) *SOCIAL STRUCTURE AND BEHAVIOR.* Academic Press, New York.

Dunkl, C. F., and Ramirez, D. (1971). *TOPICS IN HARMONIC ANALYSIS.* Appleton-Century-Crofts, New York.

Efron, B. (1969). Student's t-test under symmetry conditions. *JASA* **64**, 1278–1302.

Ellis, M. (1980). On Kamai's conjecture concerning the $\bar{d}$ distance between two-state Markov processes. *Ann. Prob.* **8**, 372–376.

Epstein, R. (1977). *THE THEORY OF GAMBLING AND STATISTICAL LOGIC, Revised ed.*, Academic Press, New York.

Ewens, W. J. (1979). *MATHEMATICAL POPULATION GENETICS*, Springer-Verlag, Berlin.

Feigin, P., and Alvo, M. (1986). Intergroup diversity and concordance for ranking data: An approach via metrics for permutations. *Ann. Statist.* **14**, 691–707.

Feigin, P., and Cohen, A. (1978). On a model for concordance between judges. *Jour. Royal Statist. Soc. B* **40**, 203–213.

Feller, W. (1968). *AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS* **Vol. I**, 3rd ed., Wiley, New York.

Feller, W. (1971). *AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS*, **Vol. II**, 2nd ed., Wiley, New York.

Fienberg, S. E. (1971). Randomization and social affairs: The 1970 Draft Lottery. *Science* **172**, 255–261 (also p. 630).

Fill, J. (1987). The Radon transform on Z_n. Technical Report, Dept. of Statistics, University of Chicago. To appear in *SIAM Jour. Discrete Math.*

Finch, P. D. (1969). Linear least squares predictions in non-stochastic time series. *Adv. Appl. Prob.* **1**, 111–122.

Fishburn, P. (1973). *THE THEORY OF SOCIAL CHOICE.* Princeton University Press, Princeton, New Jersey.

Fisher, R. A. (1953). Dispersion on a sphere. *Proc. Royal Soc. London* **A 217**, 295–305.

Flatto, L., Odlyzko, A. M., and Wales, D. B. (1985). Random shuffles and group representations. *Ann. Prob.* **13**, 154–178.

Fligner, M. and Verducci, J. (1986). Distance based ranking models. *Jour. Royal Statist. Soc. B* **48**, 359–369.

Fligner, M., and Verducci, J. (1987). Aspects of two group concordance. *Comm. Statist.-Theor. Meth.* to appear.

Fligner, M., and Verducci, J. (1988a). A nonparametric test for judges' bias in an athletic competition. *Appl. Statist.* **37**, 101–110.

Fligner, M., and Verducci, J. (1988b). Multi-stage ranking models. *Jour. Amer. Statist. Assoc.* to appear.

Floyd, R. (1964). Letter to Don Knuth.

Fortini, P. (1977). Representations of groups and the analysis of variance. Ph.D. Thesis, Dept. of Statistics, Harvard University.

Fredman, M. (1975). on computing the length of longest increasing subsequences. *Discrete Math.* **11**, 29–35.

Freedman, D. (1981). Bootstrapping regression models. *Ann. Statist.* **9**, 1218–1228.

Freedman, D., and Lane, D. (1980). The empirical distribution of Fourier coefficients. *Ann. Statist.* **8**, 1244–1251.

Friedman, J. H., and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7**, 697–717.

Friedman, J. H., and Rafsky, L. C. (1983). Graph-theoretic measures of multivariate association and prediction. *Ann. Statist.* **11**, 377–391.

Gardner, M. (1977). *MATHEMATICAL MAGIC SHOW.* Knopf, New York.

Garsia, A. and McLarnan, T. (1988). Relations between Youngs natural and the Kazhdan-Lusztig representations of S_n. *Advances in Math.* **69**, 32–92.

Gilbert, E. (1955). Theory of shuffling. Technical Memorandum, Bell Laboratories.

Giné, E. (1973). Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms. *Ann. Statist.* **3**, 1243—1266.

Gleason, A. C., and McGilchrist, C. A. (1978). An introduction to relationship algebras. *Commun. Statist.-Theor. Meth.* **A 7 11**, 1053–1078.

Gokhale, D. V., and Kullback, S. (1978). *THE INFORMATION IN CONTINGENCY TABLES.* Marcel Dekker, New York.

Goldstein, S. (1979). Maximal coupling. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **46**, 193–204.

Golub, G. and Van Loan, C. (1983). *Matrix Computations.* John Hopkins Press, Baltimore.

Good, I. J. (1951). Random motion on a finite Abelian group. *Proc. Cambridge Phil. Soc.* **47**, 756–762.

Gordon, A. D. (1983). A measure of agreement between rankings. *Biometrika* **66**, 7–15.

Gordon, L. (1983). Successive sampling in large finite populations. *Ann. Statist.* **11**, 702–706.

Graham, R. L., Li, W., and Li, R. (1980). On the structure of t-designs. *SIAM J. Alg. Disc. Meth.* **1**, 8–14.

Greene, C., Nijenhuis, A., and Wilf, H. (1979). A probabilistic proof of a formula for the number of Young tableaux of a given shape. *Advances in Math.* **31**, 104–109.

Greene, C., Nijenhuis, A., and Wilf, H. (1984). Another probabilistic formula in the theory of Young tableaux. *Jour. Combin. Th. A* **37**, 127–135.

Greenhalgh, A. (1987). Random walks on groups with subgroup invariance properties. Ph.D. dissertation, Dept. of Mathematics, Stanford University.

Grenander, U. (1981). *ABSTRACT INFERENCE.* Wiley, New York.

Griffeath, D. (1975). A maximal coupling for Markov chains. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **31**, 95–100.

Griffeath, D. (1978). Coupling methods for Markov chains. In G. C. Rota (ed.) *STUDIES IN PROBABILITY AND ERGODIC THEORY*, 1–43.

Grofman, B., and Owen, G. (1986). *INFORMATION POOLING AND GROUP DECISION MAKING.* Jai Press, Greenwich, Connecticut.

Hall, P. (1986). On the rate of convergence of orthogonal series density estimators. *J. R. Statist. Soc. B* **48**, 115–122.

Hannan, E. J. (1965). Group representations and applied probability. *J. Appl. Prob.* **2**, 1–68.

Harding, E. F. (1984). An efficient, minimal storage procedure for calculating the Mann-Whitney U, generalized U, and similar distributions. *Appl. Statist.* **33**, 1–6.

Hastie, T., and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1**, 297–318.

Hazewinkel, M., and Martin, C. F. (1983). Representations of the symmetric group, the specialization order, systems, and Grassmann manifolds. *L'Enseignement Math.* **29**, 53–87.

Heavlin, W. (1980). A parametric analysis of structured and unstructured Q-sort data. Ph.D. Thesis, Dept. of Statistics, Stanford University.

Heller, A. (1965). On stochastic processes derived from Markov chains. *Ann. Math. Statist.* **36**, 1286–1291.

Henery, R. J. (1981). Permutation probabilities for horse races. *J. R. Statist. Soc. B* **43**, 86–91.

Henry, D. (1983). Multiplicative models in projection pursuit. Ph.D. Thesis, Dept. of Statistics, Stanford University.

Heyer, H. (1977). *PROBABILITY MEASURES ON LOCALLY COMPACT GROUPS.* Springer-Verlag, Berlin.

Heyer, H. (1981). Moments of probability measures on a group. *Int'l Jour. Math. Sci.* **4**, 1–37.

Ho, S. T., and Chen, L. (1978). An L_p bound for the remainder in a combinatorial limit theorem. *Ann. Prob.* **6**, 231–249.

Hoaglin, D., Mosteller, F., and Tukey, J. (1983). *UNDERSTANDING ROBUST AND EXPLORATORY ANALYSIS.* Wiley, New York.

Hoaglin, D., Mosteller, F., and Tukey, J. (1985). *EXPLORING DATA TABLES, TRENDS, AND SHAPES.* Wiley, New York.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19**, 293–325.

Hoeffding, W. (1951). A combinatorial limit theorem. *Ann. Math. Statist.* **22**, 558–566.

Houtman, A. M., and Speed, T. P. (1983). Balance in designed experiments with orthogonal block structure. *Ann. Statist.* **11**, 1069–1085.

Hsiao, C. (1986). *ANALYSIS OF PANEL DATA.* Cambridge University Press, Cambridge.

Huber, P. (1981). *ROBUST STATISTICS.* Wiley, New York.

Huber, P. (1985). Projection pursuit. *Ann. Statist.* **13**, 435–525.

Ingram, R. E. (1950). Some characters of the symmetric group. *Proc. Amer. Math. Soc.* **1**, 358–369.

Izenman, A., and Zabell, S. (1978). Babies and the blackout: The genesis of a misconception. Technical Report #38, Dept. of Statistics, University of Chicago.

James, A. T. (1957). The relationship algebra of an experimental design. *Ann. Math. Statist.* **27**, 993–1002.

James, A. T. (1982). Analysis of variance determined by symmetry and combinatorial properties of zonal polynomials. In G. Kallianpur et al (eds.) *STATISTICS AND PROBABILITY ESSAYS IN HONOR OF C. R. RAO.* North-Holland, New York.

James, G. D. (1978). *THE REPRESENTATION THEORY OF THE SYMMETRIC GROUPS. Lecture Notes in Mathematics* **682**, Springer-Verlag, Berlin.

James, G., and Kerber, A. (1981). *THE REPRESENTATION THEORY OF THE SYMMETRIC GROUP.* Addison-Wesley, Reading, Massachusetts.

Jenkins, G. (1961). General considerations in the analysis of spectra. *Technometrics* **3**, 133–166.

Jupp, P. E., and Spurr, B. D. (1985). Sobolev tests for independence of directions. *Ann. Statist.* **13**, 1140–1155.

Kac, M. (1947). Random walk and the theory of Brownian motion. *American Math. Monthly* **54**, 369–391.

Karlin, S., and McGregor, J. (1961). The Hahn polynomials, formulas and an application. *Scripta Math.* **23**, 33–46.

Kazhdan, D., and Lusztig, G. (1979). Representation of Coxeter groups and Hecke algebra. *Inven. Math.* **53**, 165–184.

Kemperman, J. (1961). *THE PASSAGE PROBLEM FOR A STATIONARY MARKOV CHAIN.* University of Chicago Press, Illinois.

Kemperman, J. (1975). Sharp bounds for discrepancies (mod 1) with applications to the first digit problems. Technical Report, Dept. of Mathematics, University of Rochester.

Kendall, D. G. (1974). Pole-seeking Brownian motion and bird navigation. *J. R. Statist. Soc. B* **36**, 365–417.

Kendall, M. G. (1970). *RANK CORRELATION METHODS, 4th ed.* Hafner, New York.

Kendall, M. G., and Stuart, A. (1967). *THE ADVANCED THEORY OF STATISTICS*, **Vol. 2**, 2nd ed. Hafner, New York.

Kerov, S., and Vershick, A. (1985). The characters of the infinite symmetric group and probability properties of the Robinson-Schensted-Knuth algorithm. *SIAM Jour. Alg. Disc. Meth.* **7**, 116–124.

Kloss, B. M. (1959). Limiting distributions on bicompact topological groups. *Th. Prob. Appl.* **4**, 237–270.

Klotz, J. (1970). Markov chain clustering of births by sex. *Proc. 6th Berkeley Symp.* **Vol. IV**, 173–185.

Knuth, D. (1981). *THE ART OF COMPUTER PROGRAMMING.* **Vol. II**, 2nd ed. Addison-Wesley, Menlo Park.

Kosambi, D. D., and Rao, U. V. R. (1958). The efficiency of randomization by card shuffling. *J. R. Statist. Soc. A* **128**, 223–233.

Kruskal, W. (1958). Ordinal measures of association. *JASA* **53**, 814–861.

Kullback, S. (1968). *INFORMATION THEORY AND STATISTICS.* Dover, New York.

Kung, J. P. (1982). *YOUNG TABLELAUX IN COMBINATORICS, INVARIANT THEORY, AND ALGEBRA.* Academic Press, New York.

Kung, J. P. (1986). Radon transforms in combinatorics and lattice theory. *Contemporary Math.* **57**, 36–74.

Lander, E. (1982). *SYMMETRIC DESIGNS: AN ALGEBRAIC APPROACH.* Cambridge University Press, Cambridge, Massachusetts.

Lander, E. (1986). Unpublished manuscript.

Lauritzen, S. (1984). Extreme point models in statistics. *Scand. Jour. Statist.* **11**, 63–91.

Ledermann, W. (1967). Representation theory and statistics. *Seminaire Dubreil-Pisot (Algebre et Theorie des nombres), 20e Anne.* **15**, 15.01–15.08.

Lehmann, E. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37**, 1137–1153.

Letac, G. (1981). Problemes classiques de probabilite sur un couple de Gelfand. In *LECTURE NOTES IN MATH*, **861**, Springer-Verlag, New York.

Letac, G. (1982). Les fonctions spheriques d'un couple de Gelfand symetrique et les chaines de Markov. *Advances Appl. Prob.*, **14**, 272–294.

Levy, P. (1939). L'addition des variables aléatoires définiés sur une circonférence. *Bull. Soc. Math. France* **67**, 1–41.

Lo, J. (1977). Exponential Fourier densities and optimal estimation and detection on the circle. *IEEE Trans. Info. Th.* **23**, 321–336.

Lo, J. and Eshleman, L. (1979). Exponential Fourier densities on SO(3) and optimal estimation and detection for rational processes. *SIAM Jour. Appl. Math.* **36**, 73–82.

Logan, B., and Shepp, L. (1977). A variational problem for random Young tableaux. *Advances in Math.* **26**, 206–222.

Lubotzky, A., Phillips, R. and Sarnak, P. (1986). Hecke operators and distributing points on the sphere I. *Comm. Pure Appl. Math.* **31**, 149–186.

Luce, R. D., and Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, and E. Galanter (eds.) *HANDBOOK OF MATHEMATICAL PSYCHOLOGY* **Vol. III**, Wiley, New York.

Macdonald, I. G. (1979). *SYMMETRIC FUNCTIONS AND HALL POLYNOMIALS.* Clarendon Press, Oxford.

MacWilliams, J., and Sloane, N. (1977). *THE THEORY OF ERROR CORRECTING CODES.* North-Holland, Amsterdam.

Major, P., and Shlossman, S. B. (1979). A local limit theorem for the convolution of a probability measure on a compact group. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **50**, 137–148.

Mallows, C. L. (1957). Non-null ranking models I. *Biometrika* **44**, 114–130.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *JASA* **56**, 878–888.

Mardia, K. V. (1972). *STATISTICS OF DIRECTIONAL DATA.* Academic Press, New York.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1978). *MULTIVARIATE ANALYSIS.* Academic Press, London.

Marshall, A. W., and Olkin, I. (1979). *INEQUALITIES: THEORY OF MAJORIZATION AND ITS APPLICATIONS.* Academic Press, New York.

Matthews, P. (1985). Covering problems for random walks on spheres and finite groups. Ph.D. Thesis, Dept. of Statistics, Stanford University.

Matthews, P. (1986b). A strong uniform time for random transpositions. Technical Report, Dept. of Statistics, Purdue University. To appear in *Theoretical Prob.*

Matthews, P. (1987). Mixing rates for a random walk on the cube. *SIAM Jour. Alg. Disc. Meth.* **8** 746–752.

Matthews, P. (1988). Covering problems for Brownian motion on spheres. *Ann. Prob.* **16**, 189–199.

Matthews, P. (1988). Some sample path properpties of a random walk on the cube. Technical Report 88–17, Dept. of Mathematics, University of Maryland, Baltimore County Campus.

McCullagh, P., and Nelder, J. (1983). *GENERALIZED LINEAR MODELS.* Chapman and Hall, London.

Mellish, M. J. (1973). Optimal card shuffling. *Eureka* **36**, 9–10.

Milnor, J. (1976). Curvatures of left invariant metrices on Lie groups. *Advances in Math.* **21**, 293–329.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and application. *JASA* **78**, 47–65.

Morrison, J. A. (1986). Weighted averages of Radon transforms on Z_2^k. *SIAM J. Alg. Disc. Meth.* **7**, 404–413.

Moses, L. E. et al (1967). Scaling data on inter-nation action. *Science* **156**, 1054–1059.

Nadler, S. B. (1978). *HYPERSPACES OF SETS.* Marcel Dekker, New York.

Narayana, T. V. (1979). *LATTICE PATH COMBINATORICS WITH STATISTICAL APPLICATIONS.* University of Toronto Press, Toronto.

Nelder, J. A. (1965). The analysis of randomized experiments with orthogonal block structure II. Treatment structure and the general analysis of variance. *Proc. Roy. Soc. A* **283**, 163–178.

von Neumann, J. (1937). Some matrix and metrization of matrix space. Tomsk. Univ. Rev. **1**, 286–300. In A. H. Traub (ed.) *JOHN VON NEUMANN COLLECTED WORKS.* **Vol. IV**, 205–218. Pregamon, Oxford, 1962.

Niederreiter, H., and Phillipp, W. (1973). Berry-Esseen bounds and a theorem of Erdos and Turren on uniform distribution mod 1. *Duke Math. Jour.* **40**, 633–649.

Olshen, R. (1967). Asymptotic properties of the periodogram of a discrete stationary process. *Jour. Appl. Prob.* **4**, 508–528.

Parzen, E. (1961). Mathematical considerations in the estimation of spectra. *Technometrics* **3** , 167–190.

Pasta, D. (1987). Robust analysis of variance. Ph.D. Thesis, Dept. of Statistics, Stanford University.

Pemantle, R. (1987). An analysis of overhand shuffles. To appear in *Theoretical Prob.*

Perlman, M. (1987). Group symmetry covariance models. *Statist. Sci.* **2**, 421–425.

Pitman, J. W. (1976). On the coupling of Markov chains. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete.* **35**, 315–322.

Plackett, R. L. (1968). Random permutations. *J. R. Statist. Soc. B* **30**, 517–534.

Plackett, R. L. (1975). The analysis of permutations. *Applied Statist.* **24**, 193–202.

Poincare, H. (1912). *CALCUL PROBABILITIES.* Gauthier-Villars, Paris.

Posner, E. C. (1975). Mutual information for constructing joint distributions. *Utilitas. Math.* **7**, 3–23.

Priestly, M. B. (1981). *SPECTRAL ANALYSIS AND TIME SERIES.* **Vols. I and II.** Academic Press, New York.

Rachev, S. T. (1986). The Monge-Kantorovich mass transference problem and its stochastic applications. *Theor. Prob. Appl.* **29**, 647–676.

Rao, C. R. (1965). *LINEAR STATISTICAL INFERENCE AND ITS APPLICATIONS.* Wiley, New York.

Rayleigh, Lord (1919). On the problem of random vibrations, and of random flights in one, two, or three dimensions. *Phil. Mag.* **37**, 321–347.

Reeds, J. (1981). Theory of riffle shuffling. Unpublished manuscript.

Riordan, J. (1950). Review *Math. Rev.*, May 1950, p. 306.

Robbins, D. P., and Bolker, E. (1981). The bias of three pseudo-random shuffles. *Aequationes Math.* **22**, 268–292.

Robinson, E. Q. (1982). A historical perspective of spectrum estimation. *Proc. I.E.E.E.* **7**, 885–907.

Roelcke, W., and Dierolf, S. (1981). *UNIFORM STRUCTURES ON TOPOLOGICAL GROUPS AND THEIR QUOTIENTS.* McGraw-Hill, New York.

Rothaus, O., and Thompson, J. G. (1966). A combinatorial problem in the symmetric group. *Pacific J. Math.* **18**, 175–178.

Rukhin, A. (1970). Some statistical and probabilistic problems on groups. *Proc. Steklov Inst. Math.* **111**, 59–129.

Rukhin, A. (1977). On the estimation of a transformation parameter of a distribution given on a finite group. In *Transaction 7th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, 439–448. Academia, Prague.

Sade, A. (1949). *SUR LES SUITES HAUTES DES PERMUTATIONS.* Marseilles.

Sawyer, S. (1978). Isotropic random walks in a tree. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **42**, 279–292.

Scheffe, H. (1959). *ANALYSIS OF VARIANCE.* Wiley, New York.

Sen, P. K. (1968). Estimates of the regression coefficient in linear regression. *JASA* **63**, 1379–1389.

Serre, J. P. (1977). *LINEAR REPRESENTATIONS OF FINITE GROUPS.* Springer-Verlag, New York.

Sevastyanov, B. A. (1972). Poisson limit law for a scheme of sums of dependent random variables. *Prob. Th. Appl* **17**, 695–698.

Silverberg, A. (1980). Statistical models for q-permutations. Ph.D. Thesis, Dept. of Statistics, Princeton University.

Silverman, B., and Brown, T. (1978). Short distances, flat triangles, and Poisson limits. *Jour. Appl. Prob.* **15**, 815–825.

Sloane, N.J.A. (1982). Recent bounds for codes, sphere packings and related problems obtained by linear programming and other methods. *Contemp. Math.* **9**, 153–185.

Sloane, N. (1983). Encrypting by random rotations. In T. Beth (ed.) *CRYPTOGRAPHY, Lecture Notes in Computer Science*, **149**, Springer-Verlag, 11–127.

Smith, L. and Diaconis, P. (1988). Honest Bernoulli excursions. *J. Appl. Prob.* **25**, 1–14.

Solomon, L. (1963). Invariants of finite reflection groups. *Nogoya Math. J.* **22**, 57–64.

Speed, T. (1987). What is an analysis of variance. With discussion. *Ann. Statist.* **15**, 885–941.

Spitzer, F. (1964). *PRINCIPLES OF RANDOM WALK.* Van Norstrand, Princeton.

Stanley, R. (1971). Theory and application of plane partitions, Parts 1, 2. *Studies in Applied Math.* **50**, 167–188, 259–279.

Stanley, R. (1979). Invariants of finite groups and their applications to combinatorics. *Bull. Amer. Math. Soc.* **1**, 475–511.

Stanley, R. (1980). Weyl groups, the hard Lefshitz theorem, and the Sperner property. *SIAM Jour. Alg. Disc. Meth.* **1**, 168–184.

Stanley, R. (1983). Factorization of permutations into n-cycles. *Disc. Math.* **37**, 255–262.

Stanley, R. (1985). *ENUMERATIVE COMBINATORICS.* Wadsworth, Belmont, California.

Stanton, D. (1984). Orthogonal polynomials and Chevalley groups. In R. Askey et al (eds.) *SPECIAL FUNCTIONS: GROUP THEORETICAL ASPECTS*

AND APPLICATIONS, 87–92. Dordrecht, Boston.

Steele, J. M. (1987). Gibbs measures on combinatorial objects and the central limit theorem for an exponential family of random trees. *Prob. Eng. Info. Sci.* **1**, 47–60.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In F. N. David (ed.) *RESEARCH PAPERS IN STATISTICS*, 351–366. Wiley, New York.

Stein, C. (1986). Poisson limit theorems on graphs. Unpublished manuscript.

Stone, C. J. (1985). Additive regression and other non-parametric models. *Ann. Statist.* **13**, 689–705.

Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36**, 423–439.

Suzuki, M. (1982). *GROUP THEORY* **Vol. I**, Springer-Verlag, New York.

Takacs, L. (1982). Random walks on groups. *Lin. Alg. Appl.* **43**, 49–67.

Takacs, L. (1981). Random flights on regular polytopes. *SIAM J. Alg. Disc. Meth.* **2**, 153–171.

Takacs, L. (1986). Harmonic analysis on Schur algebras and its application in the theory of probability. In J. A. Chao, W. Woyczynski (eds.) *PROBABILITY THEORY AND HARMONIC ANALYSIS*, 227–283, Marcel Dekker, New York.

Thorisson, H. (1986). On maximal and distributional coupling. *Ann. Prob.* **11**, 873–876.

Thorisson, H. (1987). Independent μ times. A new proof of the basic limit theorems of Markov chains. Technical Report #243, Dept. of Statistics, Stanford University.

Thorp, E. (1973). Nonrandom shuffling with applications to the game of Faro. *JASA* **68**, 842–847.

Tjur, T. (1984). Analysis of variance models in orthogonal designs. *Int. Statist. Rev.* **52**, 33–82.

Tukey, J. (1949). One degree of freedom for non-additivity. *Biometrics* **5**, 232–242.

Tukey, J. (1961). Discussion, emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics* **3**, 191–219.

Tukey, J. (1977). *EXPLORATORY DATA ANALYSIS*. Addison-Wesley, Reding, Massachusetts.

Tukey, J. (1986). *THE COLLECTED PAPERS OF JOHN TUKEY*, **Vols. III, IV**, Wadsworth, Belmont, CA.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychol. Rev.* **79**, 281–299.

Tversky, A., and Sattath, S. (1979). Preference trees. *Psychol. Rev.* **86**, 542–573.

Ulam, S. M. (1961). Monte Carlo calculations in problems of mathematical physics. In E. F. Beckenbach (ed.) *MODERN MATHEMATICS FOR THE ENGINEER, Second Series*, McGraw-Hill, New York.

Ury, H., and Kleinecke, D. (1979). Tables of the distribution of Spearman's footrule. *Appl. Statist.* **28**, 271–275.

Verducci, J. (1982). Discriminating between two probabilities on the basis of ranked preferences. Ph.D. Thesis, Dept. of Statistics, Stanford University.

Vershik, A. M., and Kerov, S. V. (1981). Asymptotic theory of characters of the symmetric group. *Functional Analysis* **15**, 246–255.

Walters, P. (1982). *AN INTRODUCTION TO ERGODIC THEORY*, Springer-Verlag, New York.

Watson, G. (1983). *STATISTICS ON SPHERES.* Wiley, New York.

Wellner, J. (1979). Permutation tests for directional data. *Ann. Statist.* **7**, 929–943.

White, D. (1983). A bijection proving orthogonality of the characters of S_n. *Adv. Math.* **50**, 160–186.

Yellott, J. (1977). The relationship between Luce's choice axiom. Thurston's theory of comparative judgement, and the double exponential distribution. *J. Math. Psychol.* **15**, 109–144.

Young, H. P., and Levenglick, A. (1978). A consistent extension of Condorcet's election principle. *SIAM J. Appl. Math.* **35**, 285–300.

Zambia, W. T., and Hausch, D. B. (1984). *BEAT THE RACETRACK.* Harcourt, Bruce, and Jovanovich, San Diego.

Zolotarev, V. M. (1983). Probability metrics. *Th. Prob. Appl.* **28**, 278–302.