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*

$$AV(\rho) = \frac{1}{3}(n^2 - 1)$$

$$Var(\rho) = \frac{1}{45}(n+1)(2n^2+7)$$

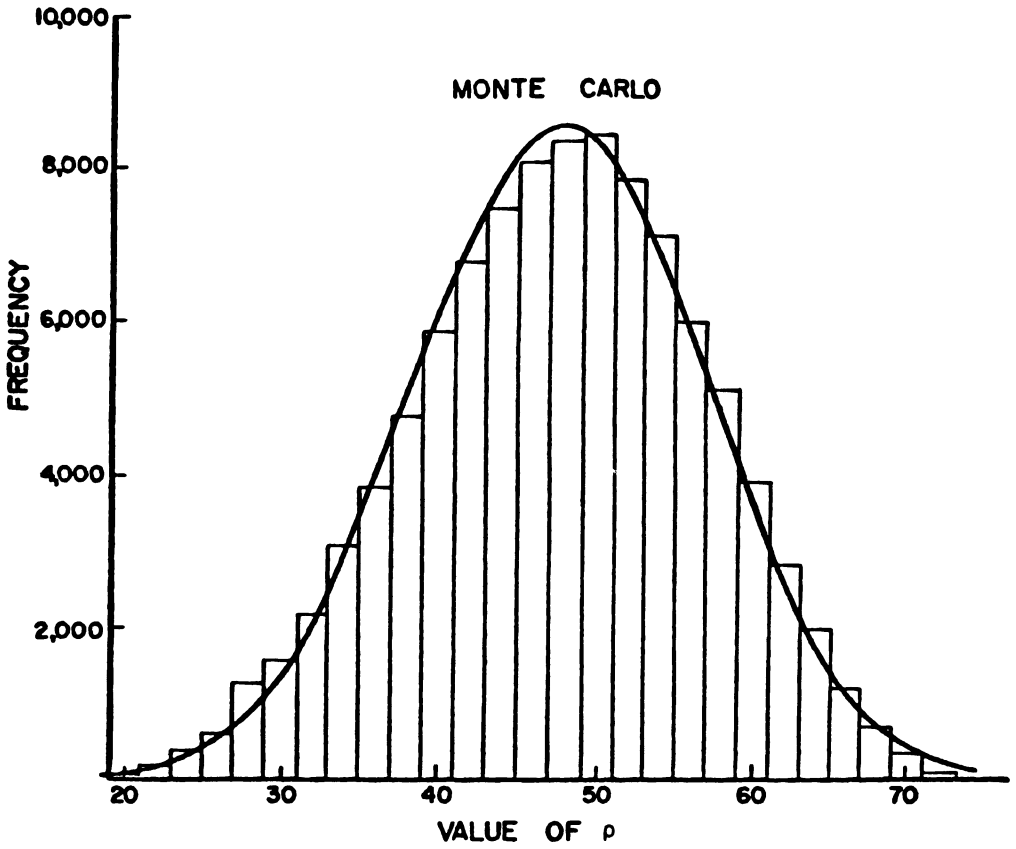
$$P\{\frac{\rho - AV}{SD} \leq t\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx + o(1).$$

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.

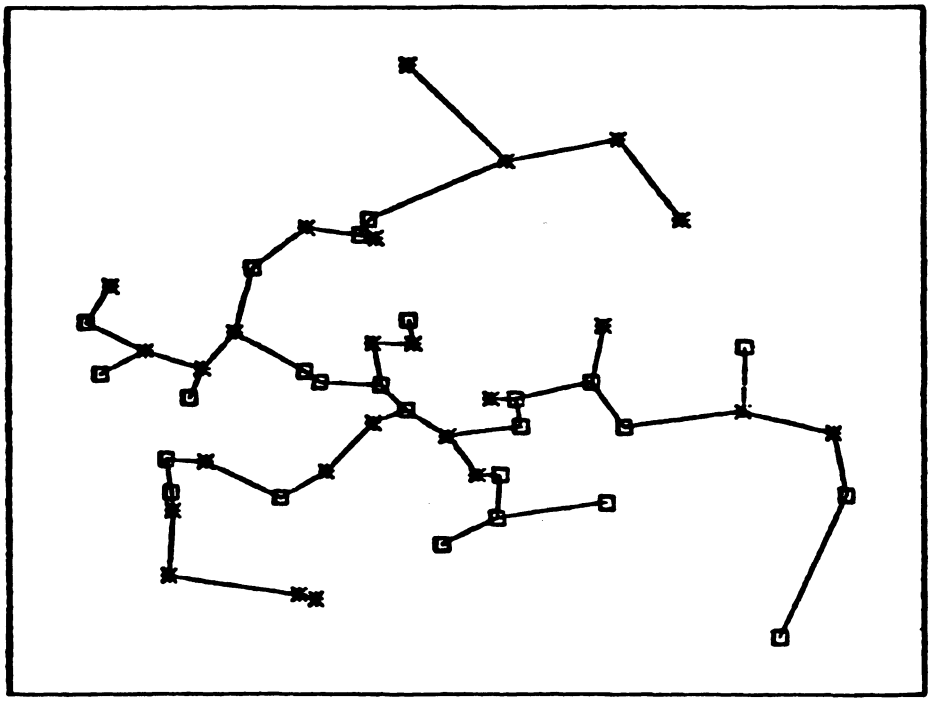
Example 4. Two-sample problems. Here is a fourth use of metrics: as a means of looking at 2 sample problems. In such problems we consider two sets of permutations $\pi_1, \ldots, \pi_n$ and $\sigma_1, \ldots, \sigma_m$ and ask about their similarities and differences. One classical question: "Can these 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

$$V(X) = \inf_{a \in T} \int_T [\arg(z\overline{a})]^2 P_X(dz)$$

(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

$$\phi = (\Sigma|w_i|^p)^{\frac{1}{p}}, \ \max|w_i|, \text{ or } \{\sum_{i_1 \leq i_2 \ldots \leq i_j} w_{i_1} w_{i_2} \ldots w_{i_j}\}^{\frac{1}{j}}.$$

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 $\bar{\ell}(w)$. It is an amazing fact that the generating function of $\bar{\ell}$ factors as

$$(*) \qquad \sum_w t^{\bar{\ell}(w)} = \prod_{i=1}^{m} (1 + e_i t).$$

COROLLARY 2. *Let (W, S) be a Weyl group with exponents e_i. Then the length $\bar{\ell}$ 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 $0 \leq a_i \leq i - 1$. 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) $E(I) = \mu_n + \ldots + \mu_{n-k+1} - \mu_k - \mu_{k-1} \ldots - \mu_2 = \frac{k(n-k)}{2}$

b) $Var(I) = \sigma_n^2 + \ldots + \sigma_{n-k+1}^2 - \sigma_k^2 - \sigma_{k-1}^2 \ldots - \sigma_2^2 = \frac{k(n+1)(n-k)}{12}$

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

$$\{\frac{I-\mu_I}{\sigma_I}\}\frac{\sigma_I}{\overline{\sigma}_{n-k}} + \{\frac{\overline{U}_k - \overline{\mu}_k}{\overline{\sigma}_k}\}\frac{\overline{\sigma}_k}{\overline{\sigma}_{n-k}} \stackrel{D}{=} \frac{\overline{U}_{n-k} - \overline{\mu}_{n-k}}{\overline{\sigma}_{n-k}}.$$

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.