METRICS ON COMPOSITIONS AND COINCIDENCES AMONG RENEWAL SEQUENCES

PERSI DIACONIS*, SUSAN HOLMES†, SVANTE JANSON‡, STEVEN P. LALLEY§, AND ROBIN PEMANTLE¶

Abstract. We study several metrics on the space $C(m,n)$ of compositions of m into at most n parts. Understanding the geometry of these spaces leads to the study of the distribution of the distance between randomly chosen compositions. This in turn leads to some non-standard probability problems. One involves pinned Wiener processes. A second leads to the following renewal theory problem: Let $X_1, X_2, \ldots X_n$; $Y_1, Y_2, \ldots Y_n$ be positive integer valued random variables. Let $C_n = |\{i, j \leq n : X_1 + \ldots + X_i = Y_1 + \cdots + Y_j\}|$ be the number of coincidences among the partial sums. We determine limiting approximations to the distribution of C_n. When X_i and Y_i are jointly independent and identically distributed; the limit is non normal. When $X_1 + \cdots + X_n = Y_1 + \cdots + Y_n$ is tied down (as in the application to compositions) the limit is normal. Our study was motivated by algorithms for careful approximation of the bootstrap.

1. Introduction. Let $C(m,n)$ be the set of compositions of m into at most n parts. Thus

$$C(m,n) = \{(h_1, h_2 \cdots, h_n) : h_i \geq 0, \text{ integer}, h_1 + \cdots + h_n = m\} \tag{1.1}$$

The familiar stars and bars argument shows $|C(m,n)| = \binom{m+n-1}{n-1}$.

For example, when $m = 4$, $n = 3$ the 15 compositions in $C(4,3)$ are

112	400	013	130	220
121	040	031	301	202
211	004	103	310	022

Compositions are a basic combinatorial object which arise in several statistical applications. For example, a class of m students given grades in $\{A, B, C, D, E\}$ give rise to a point in $C(m,5)$. Aitchison [1] gives a comprehensive treatment of compositional data.

Our motivation for careful study of compositions arose from analysis of the statistical tool known as the bootstrap. This is based on repeated samples of of n items chosen with replacement from a list of n. Each sample can be associated to a point in $C(n,n)$, with h_i being the number of times item i appears in the sample. We were seeking "well distributed" points in

* Dept. of Mathematics, Harvard University, Cambridge, MA 02138.
† Unité De Biométrie, INRA-UMII-ENSA.M, 34060 Montpellier, France; and Dept. of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305.
‡ Dept. of Mathematics, Uppsala University, P.O. Box 480, S-75106 Uppsala, Sweden.
§ Dept. of Statistics, Purdue University, West Lafayette, IN 47907.
¶ Dept. of Mathematics, University of Wisconsin, Madison, WI 53706.

$C(n, n)$ this required natural motions of distance. See Diaconis and Holmes [7] for further discussion.

In Section 2 we introduce and study a number of metrics on $C(m, n)$. As explained below, this is closely related to the study of metrics on the space of probability measures. We study the size of the balls in a given metric by studying the following problem: Pick h, h' at random in $C(m, n)$. What is the distribution of $d(h, h')$? We give fairly complete answers on $C(n, n)$. The results lead to some non-standard probability problems.

The final two sections focus on some renewal theory problems arising from the study of our subset metric. Let $X_1, X_2, \cdots; Y_1, Y_2 \cdots$ be positive integer valued random variables. Let S_j^X and S_j^Y be the corresponding partial sums. Define the coincidence number

$$C_n = |\{i, j;\ 1 \le i, j \le n : S_i^X = S_j^Y\}| \tag{1.2}$$

That is, the number of common values between the sets

$$\{S_1^X, S_2^X, \cdots, S_n^X\} \text{ and } \{S_1^Y, S_2^Y, \cdots, S_n^Y\}.$$

We study the limit distribution of C_n under two distributional assumptions. We first treat the independent and identically distributed case. In Section 3 we prove

THEOREM 1.1. *Let $\{X_i\}_{i=1}^{\infty}$, $\{Y_j\}_{j=1}^{\infty}$ be jointly independent and identically distributed positive integer valued random variables. Suppose that g.c.d.$\{h : P\{X_1 = h\} > 0\} = 1$ and that X_1 has finite mean μ and finite positive variance σ^2. Then, as $n \to \infty$*

$$\frac{C_n - n\mu^{-1}}{\sqrt{n}} \Longrightarrow \min(Z_1, Z_2)$$

where the vector (Z_1, Z_2) has a bivariate normal distribution with mean zero and non-degenerate (rank 2) covariance matrix. The limit is thus non-normal.

In Section 4 we prove

THEOREM 1.2. *Let $\{W_i\}_{i=1}^n$ be the cell counts when n balls are dropped into n boxes. Let $X_i = W_i + 1$. Let $\{Y_i\}_{i=1}^n$ be an independent copy of $\{X_i\}_{i=1}^n$. As $n \to \infty$, the C_n of (1.2) satisfies*

$$\frac{C_n - n/2}{\sqrt{n}} \Longrightarrow Z$$

where Z is normal with mean 0 and positive, finite variance.

The connection between Theorems 1.1 and 1.2 and the metrics of Section 2 is explained in the introduction to Section 3.

2. Metrics on compositions. Let $C(m,n)$ be the set of compositions of m into at most n parts. In this section we define and study several metrics on $C(m,n)$: total variation (Section 2.1), the subset metric (Section 2.2), and Vassershtein metric (Section 2.3). Basic properties and sampling distributions are developed. The sampling distributions are derived for both the uniform and multinomial distributions. The first gives a geometric feeling for the space. The second is natural for the bootstrap applications: the measure induced on $C(n,n)$ by bootstrap sampling is exactly the multinomial distribution of n balls dropped randomly into n boxes.

Compositions may be regarded as measures on $\{1,2,...,n\}$ with total mass m. This allows any distance on probabilities to be adapted to a metric on $C(m,n)$. Section 2.4 gives pointers to the relevant literature and some further examples.

Total variation emerges as our favorite metric. The others are developed because they have natural invariance properties or lead to interesting math problems.

2.1. Total variation. For $x, y \in C(m,n)$ define

$$d_{TV}(x,y) = \left\{ \begin{array}{l} \text{minimum number of } \pm 1 \text{ switches} \\ \text{needed to bring } x \text{ to } y. \end{array} \right. \tag{2.1}$$

For example, take $m=4$, $n=3$, $x=(4,0,0)$, $y=(1,1,2)$. We bring x to y by $400 \to 301 \to 202 \to 112$ so $d_{TV}(x,y)=3$.

The standard properties of the total variation distance between two probability measures (see e.g. Diaconis [6, Chapter 3]) can be translated to give the following equivalences.

LEMMA 2.1. *The total variation distance on $C(m,n)$ defined in (3.1) satisfies*

$$d_{TV}(x,y) = \frac{1}{2}\sum_{i=1}^{n} |x_i - y_i| = \frac{1}{2}\max_{|f_i|\le 1} \sum_{i=1}^{n} x_i f_i - y_i f_i = m - \sum_{i=1}^{n} \min(x_i, y_i)$$

Remark 2.2.

1. The first equality gives an easy way to calculate d_{TV}. The second equality shows that if two compositions are close then linear combinations of them are uniformly close. The third equality gives a statistically natural property of d_{TV}. In the bootstrap application, $\min(x_i, y_i)$ is the amount of overlap or redundancy between the two samples. For $m=n$, the distance is largest (equal to n) if the two compositions come from disjoint bootstrap replications.
2. The distance d_{TV} is invariant under coordinate permutations. On the other hand the number of points in a metric ball can depend on where the ball is centered. For example, when $m=4$ and $n=3$, there are two points at distance 1 from (004) and 6 points at

distance 1 from (112). Generic x have $n(n-1)$ points at distance 1 but for larger balls, all depends on how close x is to the corner of the simplex $C(m,n)$. This becomes more pronounced for larger n. For example, on $C(n,n)$, if Y has a multinomial distribution and $x = (n, 0, \cdots, 0)$, $d_{TV}(x, Y)$ is essentially constant at n while if $x = (1, 1, \cdots, 1)$ $d_{TV}(x, Y)$ has an approximate normal distribution centered at n/e. This can be proved using the argument of Lemma 2.3 which interpolates between these extremes by choosing x at random.

The next two lemmas give the approximate sampling distributions of $d_{TV}(X,Y)$ where X and Y are randomly chosen compositions. We derive limiting approximations when m and n are large with $m/n \to \lambda$. This is the domain of interest for bootstrap applications. The answers can be quite different in other zones (m small n large or vice-versa). Nowadays, one can easily simulate this distribution for any specific m, n of interest.

LEMMA 2.3. *Let X and Y be independently chosen from the multinomial distribution on $C(m,n)$. Then, for $m, n \nearrow \infty$ with $m/n \to \lambda$, $0 < \lambda < \infty$, $d_{TV}(X,Y)$ is approximately normally distributed with*

$$\text{mean} \sim \frac{n}{2}\mu(\lambda), \quad \text{var} \sim \frac{n}{4}\sigma^2(\lambda)$$

For $\mu(\lambda) = E|W - W'|$ with W, W' independent Poisson (λ) and

$$\sigma^2(\lambda) = [2\lambda - \mu(\lambda)^2]\{1 - 2[2\lambda^2 - \lambda\mu(\lambda)^2]^{-1}[E|W^2 - WW'| - \lambda\mu(\lambda)]^2\}$$

Proof. The means and the variances can be computed by elementary arguments or by using the conditioned limit arguments as below.

Without essential loss, take $\lambda = m/n$. Realize the multinomial variables X and Y as the conditional values of independent Poisson(λ) vectors X', Y', with $X_1' + \ldots + X_n' = Y_1' + \ldots + Y_n' = m$. Let normalized random variables be defined by:

$$\begin{aligned} A_n &= \frac{1}{\sqrt{n(2\lambda - \mu^2(\lambda))}} \sum_{i=1}^n \{|X_i' - Y_i'| - \mu(\lambda)\} \\ B_n &= \frac{1}{\sqrt{n\lambda}} \sum_{i=1}^n (X_i' - \lambda) \\ C_n &= \frac{1}{\sqrt{n\lambda}} \sum_{i=1}^n (Y_i' - \lambda) \end{aligned}$$

Then the conditional law of A_n given $B_n = C_n = 0$ is the law of $(2d_{TV}(X,Y) - \mu(\lambda))/\sqrt{n(2\lambda - \mu^2(\lambda))}$. From the multivariate central limit theorem,

$$\begin{pmatrix} A_n \\ B_n \\ C_n \end{pmatrix} \longrightarrow \begin{pmatrix} A \\ B \\ C \end{pmatrix}$$

with $(A,B,C)^T$ trivariate normal having mean vector 0 and covariance

matrix $\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix}$ where $\rho = [2\lambda^2 - \lambda\mu(\lambda^2)]^{-\frac{1}{2}}[E|W^2 - WW'| - \lambda\mu(\lambda)]$.
Conditioned limit theory as in Holst [12] (cor. 3.6) implies that

$$\mathcal{L}(A_n|B_n = C_n = 0) \longrightarrow \mathcal{L}(A|B = C = 0)$$

Now, if a random normal vector Z is partitioned into Z_1 and Z_2, using standard notation, $\mathcal{L}(Z_1|Z_2 = z_2)$ is normal with mean $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2)$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. In the present case $\mathcal{L}(A|B = C = 0)$ is normal with mean 0 and variance $1 - 2\rho^2$. This yields the stated result after elementary rescaling. □

Remark 2.4. Ramasubban [19,20] studied $\Delta_r = E|W - W'|^r$. He gives $\mu(\lambda) = 2\lambda e^{-2\lambda}[I_0(2\lambda) + I_1(2\lambda)]$ with $I_n(x)$ the n^{th} order modified Bessel function of the first kind. We compute

λ	.5	1	1.5	2	2.5	3
$\mu(\lambda)$	.674	1.048	1.319	1.543	1.738	1.912
$\sigma^2(\lambda)$	0.330	0.712	1.082	1.448	1.812	2.177

Feller [8] relates Poisson differences and Bessel functions.

Essentially the same proof, conditioning on geometric variables instead of Poisson variables, gives a limit theorem for the total variation distance under the uniform distribution on $C(m, n)$, in this case the variance has a closed form.

LEMMA 2.5. *Let X and Y be independently chosen from the uniform distribution on $C(m, n)$. Then, for $m, n \nearrow \infty$ with $m/n \to \lambda$, $0 < \lambda < \infty$, $d_{TV}(X, Y)$ is approximately normally distributed with*

$$\textit{mean} \sim \frac{n}{2}\mu(\lambda)$$
$$\textit{var} \sim \frac{n}{4}\sigma^2(\lambda)$$

For $\mu(\lambda) = E|W - W'| = \dfrac{2\lambda(1+\lambda)}{1+2\lambda}$ where W, W' are independent geometric $(\dfrac{1}{1+\lambda})$ variables $(P(W = j) = \theta(1-\theta)^j, 0 \le j < \infty, \theta = \frac{1}{1+\lambda})$ and

$$\begin{aligned}\sigma^2(\lambda) &= [2\lambda(\lambda+1) - \mu(\lambda)^2]\{1 - 2[2\lambda^2(\lambda+1)^2 \\ &\qquad -\lambda(\lambda+1)\mu^2(\lambda)]^{-1}[E|W^2 - WW'| - \lambda\mu(\lambda)]^2\} \\ &= 4\lambda^2(\lambda+1)^2(2\lambda^2+2\lambda+1)/(2\lambda+1)^4\end{aligned}$$

Remark 2.6. We compute

λ	.5	1	1.5	2	2.5	3
$\mu(\lambda)$	0.75	1.333	1.875	2.4	2.92	3.43
$\sigma^2(\lambda)$	0.352	0.988	1.868	2.995	4.372	5.998

Thus, typical pairs X, Y tend to be further apart under the uniform as compared with the multinomial distribution.

2.2. Subset distance. Compositions are in 1–1 correspondence with subsets of size $n-1$ in a set of $m+n-1$ elements: arrange $1, 2, \cdots, m+n-1$ in a row and circle the elements in the subset. The associated composition has n parts corresponding to the number of elements between the circles. Thus, with $m = 4$, $n = 3$, the composition (400) corresponds to 1 2 3 4 ⑤ ⑥ while (013) corresponds to ①2 ③4 5 6. The composition $(k_1, k_2, \cdots, k_n)$ corresponds to the subset $\{k_1+1, k_1+k_2+2, \cdots, k_1+\cdots+k_{n-1}+n-1\}$. Write $s(x)$ for the subset corresponding to the composition x.

There is a natural metric on subsets which induces a metric on compositions. For $x, y \in C(m, n)$ define

$$d_s(x, y) = (n-1) - |s(x) \cap s(y)|$$

Thus for $x = (400)$, $y = (013)$; $s(x) = \{5, 6\}$, $s(y) = \{1, 3\}$, $d(x, y) = 2$.

The metric d_s depends on the ordering: $d_s(50000, 11111) = 2$, $d_s(05000, 11111) = 3$. On the other hand the metric d_s has the following invariance property: the number of points in the ball $\{y : d_s(x, y) \leq \omega\}$ does not depend on x. This follows from the invariance of the metric $(n-1) - |s \cap t|$ on subsets under the action of the permutation group S_{n+m-1}. The main reason for studying d_s is because of the frequent interplay between subsets and compositions in the combinatorial literature.

The following two results give the limiting distribution of $d_s(X, Y)$ under the two distributions on $C(m, n)$.

LEMMA 2.7. *Let X and Y be independently chosen from the uniform distribution on $C(m, n)$. Then, for $m, n \nearrow \infty$ with $m/n \to \lambda$, $0 < \lambda < \infty$, d_s has a hypergeometric distribution. It is approximately normally distributed with*

$$\begin{aligned} \textit{mean} &= (n-1) - \frac{(n-1)^2}{n+m-1} \sim n\lambda/(1+\lambda) \\ \textit{var} &= \frac{(n-1)^2 m^2}{(m+n-1)^2(m+n-2)} \sim n\lambda^2/(1+\lambda)^3 \end{aligned}$$

Proof. Using the correspondence, $d_s(X, Y)$ has the same distribution as $(n-1) - |S \cap T|$ where S and T are randomly chosen subsets of size $(n-1)$ from $\{1, 2, \cdots m+n-1\}$. By invariance, S may be fixed at

$\{1, 2, \cdots, n-1\}$. Now, the distribution of $|S \cap T|$ is hypergeometric and the result stated is classical. □

LEMMA 2.8. *Let X and Y be independently chosen from the multinomial distribution on $C(n, n)$. Then, as $n \nearrow \infty$, $d_s(X, Y)$ is approximately normally distributed with*

$$\begin{aligned} \textit{mean} &\sim n/2 \\ \textit{var} &\sim \sigma^2 n \quad \textit{for some } \sigma^2,\ 0 < \sigma^2 < \infty . \end{aligned}$$

Proof. Choosing $X = (X_1, X_2, \cdots, X_n) \in C(n, n)$ from the multinomial distribution amounts to dropping n balls into n boxes with X_i the number of balls in box i. Let Y be similarly distributed. Now $|s(X) \cap s(Y)| = C_{n-1}$ as defined in (1.2) applied to the variables $X_i + 1$. The result now follows from Theorem 1.2. See Section 4 below. □

Remark 2.9. Very similar arguments give results similar to Lemma 2.8 for general m, n.

2.3. Vassershtein distance. The final metric considered in detail is the analog of a standard metrization of the weak star topology. For $x, y \in C(m, n)$ define

$$d_V(x, y) = \begin{array}{l}\text{minimum number of adjacent } \pm 1 \\ \text{switches need to bring } x \text{ to } y .\end{array} \tag{2.2}$$

Thus $d_V(400, 112) = 5$ from $(400) \to (310) \to (301) \to (211) \to (202) \to (112)$. Rachev [18] discusses the history and literature for this distance on probability measures.

From results proved there we have the following equivalent versions:

LEMMA 2.10. *The Vassershtein distance d_V on $C(m, n)$ defined by (2.2) satisfies*

$$\begin{aligned} d_V(x, y) &= \textstyle\sum_{i=1}^{n} |x_i^+ - y_i^+|; \qquad x_i^+ = x_1 + \cdots + x_i \\ &= \max_{f \in Lip_1} |\textstyle\sum x_i f_i - y_i f_i|, \quad Lip_1 = \{f_i \,|\, |f_i - f_{i+1}| \le 1 \quad , 1 \le i < n\}. \end{aligned}$$

Remark 2.11. Note that d_V depends on the coordinate ordering: $d_V(300, 030) = 3$, $d_V(300, 003) = 6$. However, in the bootstrap application, if the original sample values are ordered real numbers then adjacency has a natural meaning and the Vassershtein distance becomes interesting.

LEMMA 2.12. *Let X and Y be chosen independently from the uniform distribution on $C(m, n)$. Then, as $m, n \nearrow \infty$, with $m/n \to \lambda$, $d_V(X, Y)/(n^{\frac{3}{2}}(2\lambda(1+\lambda))^{\frac{1}{2}})$ converges in distribution to $\int_0^1 |B_0(t)|dt$ with $B_0(t)$ the standard Brownian bridge on* $[0, 1]$.

Under the multinomial distribution, $d_V(X, Y)/(n^{\frac{3}{2}}(2\lambda)^{\frac{1}{2}})$ *converges to the same limit.*

Proof. We give the proof for the multinomial distribution, the proof for the Uniform being similar.

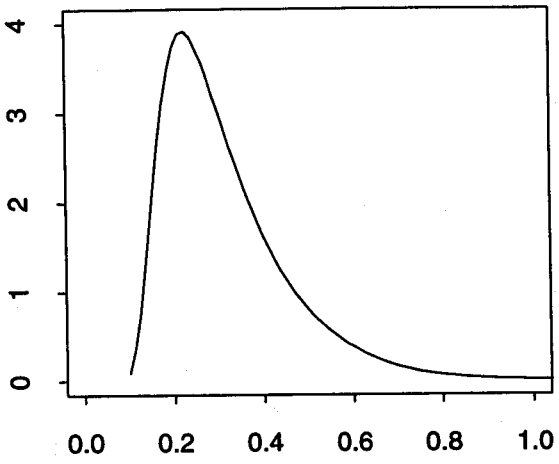
Under the multinomial distribution

$$d_V(X, Y) \stackrel{d}{=} \sum_{i=1}^{n} |X_i^+ - Y_i^+|$$

with X_i^+ and Y_i^+ independent binomial $(m, \frac{i}{n})$. Thus, for $i = \theta n$, $0 < \theta < 1$ fixed, $(X_i^+ - Y_i^+)/\sqrt{n}$ has a normal limit with mean 0 and variance $2\theta(1-\theta)\lambda$. Checking the covariances, we see that the increments have the covariances of a Brownian bridge. To make the convergence argument rigorous, we may appeal to Billingsley [3, Theorem 24.2]. This asserts that if $\xi_1, \xi_2, \cdots, \xi_n$ are exchangeable random variables with sum zero, sum of squares tending to 1, and max tending to 0, then the associated random function converges to a Brownian bridge.

Multiplying and dividing the expression for $d_V(X, Y)$ by n we get a Riemann sum for the integral. The result follows from the continuous mapping theorem. □

Remark 2.13. Following work of Cifarelli and Regazzini [4], Shepp [22] and Rice [21] carried out a careful investigation of the law of $\int_0^1 |B_0(t)|dt$. Shepp gives an elegant derivation of the Fourier transform in terms of Airy functions and a recursion for moments. Rice managed to numerically invert the Fourier transform to give highly accurate percentage points and the following graph of the density.



2.4. Other metrics. Compositions in $C(m, n)$ can be identified with probability vectors of length n by dividing by m. Thus any metric on probabilities can be carried over to compositions. Rachev [18] gives an encyclopedic survey of metrics on probabilities and Vegelius et al. [23] study measures of similarity between distributions. Two simple metrics are the Hellinger and ℓ_2 distances

$$d_{He}(x, y) = \sum_{i=1}^{n} (x_i^{\frac{1}{2}} - y_i^{\frac{1}{2}})^2 \qquad d_2^2(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$$

The first represents compositions as points on a sphere. The second has all of the advantages of Euclidian space. Each has an easily derived normal approximation under either the uniform or the multinomial distribution, using the techniques of Section 2.1.

We conclude by mentioning two further metrics which do not have natural versions on probabilities. The first is Hamming distance:

$$d_H(x, y) = |\{i : x_i \neq y_i\}|$$

This is invariant under permuting coordinates. It also has the following invariance property: $|\{y : d_H(x, y) \leq k\}|$ does not depend on x. Under both uniform distribution u and multinomial distribution m, $d_H(X, Y)$ has an approximate normal limiting distribution, when $m = n$, the means and variances are

$$E_u\{d_H(X, Y)\} = \frac{2}{3}n \qquad E_m\{d_H(X, Y)\} = \theta n,$$

$$1 - \theta = e^{-2} \sum_{j=0}^{\infty} \frac{1}{(j!)^2} \doteq .3085^+$$

$$\mathrm{var}_u\{d_H(X, Y)\} \cong \tfrac{14}{81}n \qquad \mathrm{var}_m\{d_H(X, Y)\} \cong \theta(1 - \theta)$$

$$-2 * [1 - \theta - e^{-2} \sum_{k} \frac{1}{(k-1)!k!}]^2$$

Our second metric may be called the childs metric

$$d_c(x, y) = \begin{matrix} \text{miminum number of moves} \\ \text{required to bring X to Y} \end{matrix}$$

where a move takes $(x_1, \cdots, x_n) \to (x_1 \cdots, x_i - a, \cdots, x_j + a, \cdots, x_n)$ for any pair of coordinates $i \neq j$ and any integer a chosen so all coordinates are non-negative. Thus $d_c(50000, 02300) = 2$ from $(50000) \to (32000) \to (02300)$. If a composition is thought of as n piles of blocks, a move consists of picking any number of blocks from a pile and depositing them on any other pile. This metric has good invariance properties. Alas, R.L. Graham (personal communication) has shown that computing d_c is $\#-p$ complete. Indeed, given $x, y \in C(m, n)$ form $z_i = x_i - y_i$. Let t be the maximum

number of blocks in a partition such that z_i, summed over each block, gives zero. The $d_c(x, y) = n - t$. However, deciding if $d_c(x, y) \leq n - 2$ involves computing if some non-trivial subset sum of z_i is zero. This is a well known NP complete problem.

Final Remark 2.14. Metrics on combinatorial objects offer a rich area of study and application. See Critchlow [5] and Diaconis [6, Chapter 6] for examples in the space of permutations. We hope that some of the present results will be found similarly useful for analyzing compositional data.

3. Coincidences for independent renewal sequences. This section studies the number of coincidences between independent renewal sequences. Throughout we assume that $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ are each independent and identically distributed random variables. With $\{X_i\}$ independent of $\{Y_i\}$. We do not assume X_i and Y_i have the same distribution. Assume further that all variables are strictly positive, integer valued, nonarithmetic (g.c.d. $\{h : P(X_1 = h) > 0\} = 1$), non degenerate (not almost surely equal to 1), and have finite second moments.

We let $S_i^X = X_1 + \cdots + X_i$ and S_i^Y be the partial sums. The object of study is the number of coincidences

$$C_n = \left|\{i, j \leq n : S_i^X = S_j^Y\}\right| .$$

We will prove a limit theorem for C_n when X_i and Y_i have the same law in Section 3.1. We explain what happens when X_i and Y_i have different laws in Section 3.2, which also contains a review of relevant literature.

Our motivation came from studying the metric d_s explained in Section 2.2 above. Under the multinomial distribution on $C(n, n)$, this has the same distribution as the number of coincidences between the two sequences $\{W_1 + 1, W_1 + W_2 + 2, \cdots, W_1 + \cdots + W_{n-1} + (n-1)\}$ $\{Z_1 + 1, Z_1 + Z_2 + 2, \cdots, Z_1 + \cdots + Z_{n-1} + (n-1)\}$ where $\{W_i\}_{i=1}^n$ are the number of balls in box i when n balls are dropped into n boxes according to the multinomial distribution and $\{Z_i\}_{i=1}^n$ are independent, with the same distribution. Heuristically, W_i are approximately independent Poisson (1). The limit theory in this section was developed to study such coincidences. It turns out that the heuristic is wrong: the multinomial counts are tied down and this matters, the correct results for the balls in boxes case are explained in Section 4. The present results seem of independent interest. We found a rigorous developement under minimal conditions challenging.

We conclude this introduction with an overview of the argument. With notation as above, for the remainder of this introduction take X_i and Y_i with a common law. Define

$$R_X = \{S_1^X, S_2^X, \cdots\} \qquad R_Y = \{S_1^Y, S_2^Y, \cdots\}. \tag{3.1}$$

Let the common points of R_X and R_Y be $T_1 < T_2 < T_3 < \cdots$. There are infinitely many common points from the renewal theorem (condition on

the set R_X and use the renewal theorem on the renewal process S_n^Y). The T_i form a renewal processes. Moreover, the X and Y excursions between successive T_i are independent and identically distributed. Consequently, if we define U_n, V_n by

$$S_{U_n}^X = S_{V_n}^Y = T_n \tag{3.2}$$

The successive (vector) increments of the two dimensional process

$$(U_n, V_n) \tag{3.3}$$

are *iid* so that (U_n, V_n) is a random walk which is strictly increasing in both coordinates. Theorem 1.1 of the introduction follows from a study of this walk. Indeed, C_n is just the number of points of the walk (U_j, V_j) that lie in the square $[0, n] \times [0, n]$. This is because each point of the walk corresponds to a coincidence of partial sums with (U_j, V_j) giving the times when the partial sums are equal. This coincidence is counted in C_n if and only if $U_i, V_j \leq n$.

We next explain (heuristically) why the limit law is the minimum of two correlated normal variables. The argument is based on a useful fact from renewal theory. Let $Z_1, Z_2, \ldots$ be non-negative independent and identically distributed, integer valued random variables with $E(Z_i) = \beta$, var $(Z_i) = \sigma^2$, $0 < \sigma^2 < \infty$. Let S_n^Z be the partial sum process. Define $W_h = |\{n : S_n^Z \leq h\}|$. Then, as $h \nearrow \infty$,

$$\frac{W_h - h\beta^{-1}}{\sqrt{h\beta^{-3}\sigma^2}} \Longrightarrow \xi$$

where ξ has a standard normal distribution. This is an immediate consequence of the ordinary central limit theorem since $P\{W_h < n\} = P\{S_n^Z > h\}$.

Now consider the renewal process (U_j, V_j) defined above. As noted, C_n is the number of points of this sequence in the square $[0, n] \times [0, n]$. This is clearly the same as the minimum of N_n^U, N_n^V where these are respectively the number of points in the renewal sequences U_j, V_j in $[0, n]$. Now Hunter [13] showed $(N_n^U - n\mu^{-1})/\sqrt{n}$, $(N_n^V - n\mu^{-1})/\sqrt{n}$ has a bivariate limiting normal distribution.

Thus C_n suitably normalized converges to the minimum of two normals.

To make the argument rigorous, we must study the distribution of (U_1, V_1). We show this has a non-degenerate covariance matrix in Section 3.2 which also contains further remarks and references.

3.1. Time between coincidences. Throughout we assume $\{X_i\}$, $\{Y_i\}$ satisfy the assumptions of the first paragraph of Section 3. We study the moments of T_n and (U_i, V_i) defined in (3.1, 3.2). We need a preliminary lemma for which we introduce notation.

Let $F(z) = \sum_{h=0}^{\infty} P_h z^h$, $|z| < 1$, be the probability generating function of the increments of a renewal process S_n. Let $U(z) = \sum_{h=0}^{\infty} u_h z^h$, where

$$u_h = \sum_{n=0}^{\infty} P\{S_n = h\} \tag{3.4}$$

is the renewal measure. Then $U(z) = 1/(1 - F(z))$.

LEMMA 3.1. *Assume $\mu = F'(1) < \infty$. Then, the increments have finite variance if and only if*

$$U(z) - \frac{1}{\mu(1-z)} = \sum_{h=0}^{\infty} (u_h - \frac{1}{\mu}) z^h$$

stays bounded as $z \to 1$. In this case, $U(z) - \frac{1}{\mu(1-z)} \to \frac{\sigma^2+\mu^2-\mu}{2\mu^2}$ with σ^2 the variance of the increment.

Proof. $U(z) - \frac{1}{\mu(1-z)} = \frac{1}{1-F(z)} - \frac{1}{\mu(1-z)} = \frac{\mu(1-z)-(1-F(z))}{\mu(1-z)(1-F(z))} = \frac{F(z)-1-(z-1)F'(1)}{\mu(1-z)(1-F(z))}$. Since $\frac{1-F(z)}{1-z} \to F'(1) = \mu$ as $z \nearrow 1$, $\sup_{0 \le z < 1} |U(z) - \frac{1}{\mu(1-z)}| < \infty$ if and only if $F(z) - 1 - (z-1)F'(1) = O\left((1-z)^2\right)$ as $z \nearrow 1$. Now

$$\frac{F(z) - 1 - (z-1)F'(1)}{(1-z)^2} = \int_0^1 F''(1 - t(1-z))\,(1-t)dt$$

By monotone convergence, the right hand side stays bounded as $z \to 1$ if and only if $F''(1-) < \infty$ if and only if the increments have finite variance. □

The next result shows that the increments of the renewal process T_i defined at (3.1, 3.2) have finite variance.

PROPOSITION 3.2. *Let $\{X_i\}_{i=1}^{\infty}$, $\{Y_i\}_{i=1}^{\infty}$, be independent and identically distributed non-arithmetic, integer valued random variables with finite means μ_X, μ_Y and finite, positive variances σ_X^2, σ_Y^2. Let $\{T_i\}_{i=1}^{\infty}$ be the intersection places (cf. 3.2) of the partial sum processes. Then, $\{T_i\}$ is a renewal process with iid increments $Z_i = T_{i+1} - T_i$ having $E(Z_i) = \mu_X \mu_Y$ and var $(Z_i) = \sigma^2(Z_i) < \infty$.*

Proof. Let u_h^Z be the renewal measure defined at (3.4). Clearly $\mu_h^Z = u_h^X u_h^Y$. Hence $u_h \to \mu_X^{-1}\mu_Y^{-1}$ as $h \nearrow \infty$. So $\mu_Z = \mu_X \mu_Y < \infty$. Now, with notation as in Lemma 3.1,

$$u_h^Z - \frac{1}{\mu_Z} = u_h^X u_h^Y - \frac{1}{\mu_X \mu_Y} = \frac{1}{\mu_Y}(u_h^X - \frac{1}{\mu_X}) + \frac{1}{\mu_X}(u_h^Y - \frac{1}{\mu_Y}) + (u_h^X - \frac{1}{\mu_X})(u_h^Y - \frac{1}{\mu_Y}).$$

thus

$$U^Z(z) - \frac{1}{\mu_Z(1-z)} = \frac{1}{\mu_Y}\left[U^X(z) - \frac{1}{\mu_X(1-z)}\right] + \frac{1}{\mu_X}\left[U^Y(z) - \frac{1}{\mu_Y(1-z)}\right] + \sum_{h=0}^{\infty}\left(u_h^X - \frac{1}{\mu_X}\right)\left(u_h^Y - \frac{1}{\mu_Y}\right) z^h.$$

By Lemma 3.1, the first two terms on the right hand side stay bounded as $z \nearrow 1$. So, using Lemma 3.1 again, it suffices to show that the third term stays bounded.

Using the Cauchy-Schwartz inequality it suffices to show

$$\lim_{r \nearrow 1} \sum_{h=0}^{\infty} \left(u_h^X - \frac{1}{\mu_X} \right)^2 r^{2h} = \sum_{h=0}^{\infty} \left(\mu_h^X - \frac{1}{\mu_X} \right)^2 < \infty \tag{3.5}$$

and the same with Y replacing X. Using the Plancherel Theorem, the first sum in (3.5) equals

$$\frac{1}{2\pi} \int_0^{2\pi} \left| U^X(re^{it}) - \frac{1}{\mu_X(1 - re^{it})} \right|^2 dt \leq \sup_{|z|<1} \left| U^X(z) - \frac{1}{\mu_X(1-z)} \right|^2$$

This last is finite since

$$U^X(z) - \frac{1}{\mu_X(1-z)} = \frac{1}{1 - F(z)} - \frac{1}{\mu_X(1-z)}$$

is continuous on $|z| \leq 1$: The only singularity could be at $z = 1$ but the argument in Lemma 3.1 shows convergence for $z \to 1$ as long as $|z| \leq 1$. □

Remark 3.3. We do not know how to relate higher moments of X and Y to those of Z. We have shown that if X and Y have finite moment generating functions then Z does as well.

The next result of this section shows that the bivariate vector (U_1, W_1) of (3.2) has a finite, nondegenerate covariance matrix.

PROPOSITION 3.4. *Let $\{X_i\}_{i=1}^{\infty}$, $\{Y_i\}_{i=1}^{\infty}$ satisfy the assumptions of Proposition 3.2. Let (U_i, V_i) be the times of intersection of the two partial sum processes (cf. 3.2). Then (U_i, V_i) are iid with finite, rank two covariance matrix and $E(U_i) = \mu_X$, $E(V_i) = \mu_Y$.*

Proof. By Proposition 3.2, $ET_1^2 < \infty$. But $U_1 \leq T_1$ and $V_1 \leq T_1$. Thus the second moments are finite. For the means, condition on $\{Y_i\}_{i=1}^{\infty}$. Then, U_1 is a stopping time relative to $\{X_i\}_{i=1}^{\infty}$. Since $T_1 = \sum_{i=1}^{U_1} X_i$, $E(T_1|\{Y_i\}) = E(U_1|\{Y_i\})\mu_X$ by Walds Lemma (see e.g. Gut [10, Theorem I.5.3(i)]). Hence $E(T_1) = E(U_1)\mu_X$ but $E(T_1) = \mu_X\mu_Y$ by Proposition 3.2 so $E(U_1) = \mu_X$. A parallel argument gives $E(V_1) = \mu_Y$. □

The difficult part of the argument is showing that the covariance matrix of (U_1, V_1) is rank 2. If the covariance matrix is singular, then $aU_1 + bV_1 + c = 0$ a.s. for some a, b, c not all zero. We may take a, b, c integral. Now $a = b = 0$ implies $c = 0$ which is excluded. Hence without loss $a \neq 0$. We may further assume $a > 0$ and that a and b have no common divisors > 1. The argument proceeds in two cases:

Case 3.5. $b \geq 0$. Then, $aU_1 \leq -c$ so $U_1 \leq -c/a$ almost surely. This is impossible since $P\{U_1 > M\} > 0$ for any integer M by the following

argument. We will assign values to $X_1, X_2, \cdots, X_M$ and $Y_1, Y_2, \cdots, Y_n$ (for some n) which are taken with positive probabilities and are such that $S_i^X \neq S_j^Y$ for all $i \leq M$, $j \leq n$ and $S_M^X < S_n^Y$. Thus $S_i^X \neq S_j^Y$ for all $i \leq M, j \leq n$ and so $U_1 > M$. To do the assignment, start with any X_1 and $Y_1 \neq X_1$. If say $Y_1 < X_1$, choose Y_2 such that $S_2^Y \neq S_1^X = X_1$. If $S_2^Y < S_1^X$, Choose Y_3 but if $S_2^Y > S_1^X$, choose X_2. Continue, always choosing the next value in the sequence with the smallest sum. This is always possible since there are at least two allowed values for each variable and only one of these is excluded by the construction. The construction stops when X_{M+1} is to be chosen.

Case 3.6. $b < 0$. Let $\beta = -b > 0$. The solutions to the diophantine equation $aU - \beta V + c = 0$ are of the form $U = u_h = u_0 + \beta h$, $V = v_h = v_0 + ah$ for some u_0, $v_0 > 0$ and $h = 0, 1, 2, \cdots$. Consider the sequence $\{S_{u_h}^X - S_{v_h}^Y\}_{h \geq 0}$. This is a random walk (starting at $S_{u_0}^X - S_{v_0}^Y$) with increments distributed as $S_\beta^X - S_a^Y$. Since the starting point and the increments take on at least three possible values with positive probability, there exists, for each $M > 0$, a possible realization of $\{S_{u_h}^X - S_{v_h}^Y\}_{h=0}^n$ for some $n \geq 0$, with all terms $\neq 0$ and $|S_{u_n}^X - S_{v_n}^Y| > M$. This corresponds to a realization of $\{X_i\}_{i=1}^{u_n}$ and $\{Y_i\}_{i=1}^{v_n}$ such that $(U_1, V_1) \neq (u_h, v_h)$ for every $h \leq n$. Thus, by the assumption $aU_1 + bV_1 + c = 0$, $U_1 \geq u_{n+1}$ and $V_1 \geq v_{n+1}$ for a.e. continuation of $\{X_i\}$, $\{Y_i\}$. If M is chosen large enough, every integer larger than M is a sum of possible values of X_i and a sum of possible values of Y_i. Hence if say $S_{u_n}^X - S_{v_n}^Y < -M$, there are possible continuations $\{X_i\}_{i=u_n+1}^{u_n+b}$, with $b \geq 1$, and $X_{u_n+1} + \cdots + X_{u_n+b} = S_{v_n}^Y - S_{u_n}^X$. This gives a realization $\{X_i\}_{i=1}^{u_n+b}$, $\{Y_i\}_{i=1}^{v_n}$ with $S_{u_n+b}^X = S_{v_n}^Y$, and thus $V_1 \leq v_n$ which is a contradiction.

Remark 3.7. It is the nondegeneracy of the covariance matrix which gives a non-normal limit in Theorem 1.1. We have no real hold on any of the elements of the covariance matrix.

Proof of Theorem 1.1. Suppose now that $\{X_i\}$, $\{Y_j\}$ have common distributions. From Proposition 3.4, the bivariate random walk (U_i, V_i) has $E(U_1) = E(V_1) = \mu = E(X_1)$ and finite rank two covariance matrix

$$\sum = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

Hunter [13] shows

$$\lim_{n \to \infty} P\left\{ \frac{N_n^U - n\mu^{-1}}{\sigma(n/\mu^3)^{\frac{1}{2}}} \leq \alpha, \quad \frac{N_n^V - n\mu^{-1}}{\sigma(n/\mu^3)^{\frac{1}{2}}} \leq \beta \right\} = \Phi_\rho(\alpha, \beta)$$

with Φ_ρ a bivariate normal distribution with mean 0, variances 1, and correlation ρ.

From this, $(C_n - n\mu^{-1})/\sqrt{n}$ converges to $\min(Z_1, Z_2)$ where (Z_1, Z_2)

is bivariate normal with mean 0 and covariance matrix

$$\begin{pmatrix} \sigma^2\mu^{-3} & \rho\sigma^2\mu^{-3} \\ \rho\sigma^2\mu^{-3} & \sigma^2\mu^{-3} \end{pmatrix}$$

□

Remark 3.8. Under the assumptions of Theorem 1.1, standard renewal theory (e.g. [10, Theorem II.5.1]) shows $\frac{1}{n}N_n^U \to \mu^{-1}$ a.s. and $\frac{1}{n}N_n^V \to \mu^{-1}$ a.s. hence $\frac{1}{n}C_n \to \mu^{-1}$ a.s.

3.2. Remarks and related literature.

1. Consider the number of coincidences under the conditions of Theorem 1.1 when X_1 and Y_1 have different distributions. Proposition 3.4 and Hunter's Central Limit Theorem [13] can be used to show that if the means differ then C_n has a normal limit. If the means are equal, then C_n has a non-normal limit as above.
2. Under the assumptions of Theorem 1.1 the number of coincidences among the partial sums of j *iid* random walks is distributed as the minimum of a j-variate normal (with suitable norming).
3. Things change radically if the increments are allowed to take negative values. Then, the sequence of partial sums will tend to fill out an interval and stochastic fluctuations take place at the fringes.
4. Proposition 3.2 of Section 3 is closely related to the regenerative phenomena studied by Kendall and Kingman [14] [16]. They study the Abelian semi-group of renewal sequences $\{u_n\}_{n=1}^{\infty}$ under the coordinate wise product. A masterful summary of this work is in Kingman [16]. Fristed [9] gives recent developments. A survey of its extensions to the theory of delphic semi-groups can be found in the work of Kendall and Harding [15]. This includes a survey of the work of Rollo Davidson which has connections to several of the authors of the present paper.

   Specialize the set up of Proposition 4.2 to the case where X_1 and Y_1 have the same distribution. There is then a map from measures to measures (from the Law of X_1 to the Law of Z_1) this corresponds to squaring the renewal measure: $u_n^Z = (u_n^X)^2$. Evidently this map is one to one and continuous. To see that it is not onto, we observe that if $P(Y = 1) = P(Y = 2) = \frac{1}{2}$, the corresponding renewal sequence does not have a square-root. Kendall [14] classified the infinitely divisible renewal sequences, showing they form a convex set with a countable set of extreme points.
5. The problem of coincidences among renewal sequences arises in the analysis of coupling arguments (see e.g. Lindvall [17]). Here the focus is on the occurrence of the first coincidence.

4. Coincidences with tied down sequences. This section gives a proof of Theorem 1.2, restated here for the readers convenience. Consider n balls, placed uniformly and independently in n boxes. For $1 \le j \le n$, let S_j equal j plus the number of balls in boxes $1, \ldots, j$. Let $\{S_j^X : 1 \le j \le n\}$ and $\{S_j^Y : 1 \le j \le n\}$ each be jointly distributed as $\{S_j : 1 \le j \le n\}$ and be independent of each other. Define

$$C_n = C_n(\{S_j^X\}, \{S_j^Y\}) = \#\{(i, j) : S_i^X = S_j^Y\} \tag{4.1}$$

to be the number of coincidences among the partial sums.

THEOREM 4.1.

$$\frac{C_n - (n/2)}{\sqrt{n}} \xRightarrow{D} Z \quad \text{as } n \to \infty, \tag{4.2}$$

where Z is normal with zero mean and positive finite variance.

The steps in proving Theorem 4.1 are:

(1) reduce to a one-sided local Central Limit Theorem (CLT);
(2) Poissonize;
(3) use the Poisson representation to embed in a renewal problem;
(4) apply a known local CLT to the renewal problem to prove the one-sided local CLT.

We begin with step 1, a reduction to Theorem 4.1′.

THEOREM 4.1.′ There exist $\sigma > 0$ and ϕ satisfying $\sup_k \phi(n, k) \to 0$ as $n \to \infty$, such that for all integers n and k,

$$\mathbf{P}(C_n = k) \ge \frac{1}{\sigma\sqrt{2\pi n}} \left(e^{-(n-2k)^2/(2n\sigma^2)} - \phi(n, k)\right). \tag{4.3}$$

To see that this implies Theorem 4.1, let $f : \mathbf{R} \to [0, 1]$ be any continuous function, and let $W_n = (C_n - n/2)/\sqrt{n}$. Then

$$\begin{aligned}
\liminf_n \mathbf{E}f(W_n) &\ge \lim_{A\to\infty} \liminf_{n\to\infty} \sum_{k:|k-n/2|<A\sqrt{n}} \frac{1}{\sigma\sqrt{2\pi n}} e^{-(n-2k)^2/(2n\sigma^2)} \\
&\qquad f\left((k - n/2)/\sqrt{n}\right) \\
&= \lim_{A\to\infty} \int_{-A}^{A} f(x)dG(x) \\
&= \int f(x)dG(x),
\end{aligned}$$

where G is the distribution of $\sigma/2$ times a standard normal. Replacing f by $1 - f$ shows that $\mathbf{E}f(W_n) \to \int f(x)dG(x)$ for all bounded continuous f, establishing the implication.

For step 2, we construct a version of the process $\{S_j\}$ from a Poisson point process.

PROPOSITION 4.2. *Let $\{N_t : t \geq 0\}$ be a rate 1 Poisson point process. Then the conditional law of the process $\{N_j + j : j = 1, 2, 3, \ldots\}$ given $N_n = n$ is the same as the law of the process $\{S_j\}$ above.*

Proof. One way to assign n balls uniformly and independently into n boxes is to generate n IID random variables $\{T_j\}$, uniform on the real interval $[0, n)$, and then to place ball j in box $\lfloor T_j \rfloor$. The values of $\{S_j\}$ arising in this way will be unaffected if the sequence $T_1, \ldots, T_n$ is replaced by its ascending rearrangement, $T_{(1)}, \ldots, T_{(n)}$. But the times of the Poisson process $\{N_t\}$ conditioned on $N_n = n$ are distributed as the order statistics of n independent draws from $[0, n)$, and therefore $\{N_j + j\} \stackrel{\mathcal{D}}{=} \{S_j\}$. □

Step 3 is to construct a version of C_n via the Poisson representation. Let $\{N_t^X\}$ and $\{N_t^Y\}$ be independent rate 1 Poisson processes. Let $T_0 = 0$ and inductively define T_n to be the least integer $k > T_{n-1}$ such that

$$N_i^X + i = N_j^Y + j = k \text{ for some integers } i \text{ and } j .$$

In other words, T_1, $T_2, \ldots$ are the joint renewal times of the independent renewal processes $\{N_j^X + j\}$ and $\{N_j^Y + j\}$, whose increments are each 1 plus a Poisson of mean 1. Let $R_0^X = R_0^Y = 0$ and define R_k^X (respectively R_k^Y) to be the integer i for which $N_i^X + i = T_k$ (respectively $N_i^Y + i = T_k$). Then the triples $\{(U_k, V_k, Z_k) : k \geq 1\}$ defined by

$$(U_k, V_k, Z_k) = \left(R_k^X - R_{k-1}^X,\ R_k^Y - R_{k-1}^Y,\ T_k - T_{k-1}\right) \tag{4.4}$$

are IID. In words, the blocks between joint renewals are IID, and these blocks contain the information: number of renewals in first process, number of renewals in second process, total time from last joint renewal.

This is an example of the situation studied in Section 3. In particular, Propositions 3.2 and 3.4 show that U_k, V_k and Z_k have finite variances and $EU_k = EV_k = E(N_1 + 1) = 2$ and $EZ_k = 4$.

Let $\tilde{C}_n = \sup\{k : T_k \leq 2n\}$ denote the number of joint renewals before time $2n$. Let G denote the event $\{N_n^X = N_n^Y = n\}$. Proposition 4.2 implies that the conditional law of $\{N_i^X + i,\ N_j^Y + j : 1 \leq i,\ j \leq n\}$ given G is the same as the law of $\{S_i^X,\ S_j^Y : 1 \leq i,\ j \leq n\}$. Observe that on the intersection of G and $\{\tilde{C}_n = k\}$, one has $R_k^X = R_k^Y = n$. Denoting

$$p_{n,k} = \mathbf{P}(R_k^X = R_k^Y = n,\ T_k = 2n),$$

we then have

$$\mathbf{P}(C_n = k) = \mathbf{P}(\tilde{C}_n = k | G) = \frac{p_{n,k}}{\sum_j p_{n,j}}, \tag{4.5}$$

Finally, step 4 consists of applying a local CLT to $\{(U_j, V_j, Z_j)\}$ to derive Theorem 4.1′ from (4.5). This is done in three substeps. First we show that

$$p_{n,k} = ck^{-3/2} \exp\left(-\frac{Q(n-2k,\ n-2k,\ 2n-4k)}{2k}\right) + g(n, k), \tag{4.6}$$

where Q is the quadratic form obtained by inverting the covariance matrix of the triple (U_1, V_1, Z_1) and

$$\lim_{k\to\infty} \sup_n k^{3/2}\left(1 + \frac{(n-2k)^2}{k}\right)|g(n, k)| = 0.$$

Secondly we show that

$$p_{n,k} = c2^{3/2}n^{-3/2} \exp\left(-\frac{(n-2k)^2}{2n\sigma^2}\right) + h(n, k), \tag{4.7}$$

where

$$\lim_{k\to\infty} \sup_n k^{3/2}\left(1 + \frac{(n-2k)^2}{k}\right)|h(n, k)| = 0. \tag{4.8}$$

Thirdly, we show that (4.7), (4.8) and (4.5) imply Theorem 4.1′.

Step 4a is an application of the local CLT from Bhattacharya and Rao [2, Theorem 22.1, Corollary 22.3]. The conclusion (4.6) follows immediately once the following hypotheses are verified:

(*i*) $\mathbf{E}(U_1, V_1, Z_1) = (2, 2, 4)$;
(*ii*) $\mathbf{E}(U_1^2 + V_1^2 + Z_1^2) < \infty$;
(*iii*) $\mathbf{P}(Z_1 > t) < C_1 e^{-C_2 t}$ for some $C_1, C_2 > 0$;
(*iv*) (U_1, V_1, Z_1) generates a truly 3-dimensional lattice.

Of these, (*iv*) is obvious, (*iii*) is a consequence of an easy coupling argument, and (*i*) and (*ii*) are consequences of a remark above.

Step 4b, the derivation of (4.7) from (4.6), begins with the easy observation that $Q(x, x, 2x)$ is nondegenerate, and can be written as $x^2/(2\sigma^2)$ for some positive finite σ. [For if this fails, summing (4.6) in k contradicts $\lim_{n\to\infty} \mathbf{P}(R_n^X = R_n^Y = n) \to 0$.] Letting (4.7) define the quantity $h(n, k)$, it remains to establish (4.8). In other words, we must show that

$$\mathcal{E}(n, k) := \left|(2k)^{-3/2} \exp\left(-\frac{(n-2k)^2}{4k\sigma^2}\right) - n^{-3/2} \exp\left(-\frac{(n-2k)^2}{2n\sigma^2}\right)\right| \tag{4.9}$$

satisfies $\sup_n \mathcal{E}(n, k)k^{3/2}(1 + (n-2k)^2/k) \to 0$. Fix k, and consider first the case $|n-2k| > k^{0.55}$. The exponential terms in (4.9) are then at most $e^{-ck^{0.1}}$, which establishes (4.8). On the other hand, when $|n-2k| \le k^{0.55}$, then the mean value theorem gives

$$|n^{-3/2} - (2k)^{-3/2}| \le Ck^{-5/2}|n-2k| \le Ck^{-1.95}.$$

Also, for sufficiently large k,

$$\left|\exp\left(-\tfrac{(n-2k)^2}{2n\sigma^2}\right) - \exp\left(-\tfrac{(n-2k)^2}{4k\sigma^2}\right)\right| \leq \left|\tfrac{(n-2k)^2}{2n\sigma^2} - \tfrac{(n-2k)^2}{4k\sigma^2}\right| = \tfrac{|n-2k|^3}{4nk\sigma^2}$$

which gives a contribution to (4.8) of

$$(1 + \frac{(n-2k)^2}{k})\frac{|n-2k|^3}{4nk\sigma^2} \leq C\frac{|n-2k|^3}{k^2} + C\frac{|n-2k|^5}{k^3}$$

This tends to 0 if $|n-2k| < k^\delta$, for any δ strictly less than 0.6, this yields (4.8) in the case $|n-2k| \leq k^{0.55}$.

The final substep, 4c, is accomplished by showing

$$\lim_{A\to\infty} \limsup_{n\to\infty} \frac{\sum_{k:|k-n/2|>A\sqrt{n}} p_{n,k}}{\sum_k p_{n,k}} = 0\,. \tag{4.10}$$

For this shows that given $\epsilon > 0$, we can choose A such that for sufficiently large n, all but mass ϵ of the conditional law of $\tilde{C}_n$ given G is contained in the range $[-A\sqrt{n}, A\sqrt{n}]$, from which is is immediate that (4.7) and (4.5) imply Theorem 4.1′.

Firstly, from (4.7),

$$\sum_k p_{n,k} \geq \sum_{k:|k-n/2|\leq\sqrt{n}} p_{n,k} \geq 2\sqrt{n}\left[cn^{-3/2}\exp\left(-2/(\sigma^2)\right) - \sup_k h(n,k)\right],$$

and this is at least $c'n^{-1}$ for some constant c' and sufficiently large n. Secondly, let $c_0 = \sup_{n,k} |h(n,k)k^{3/2}(1 + (n-2k)^2/k)|$. Thus $|h(n,k)| < c_0 k^{-1/2}(n-2k)^{-2}$. Estimating the numerator of (4.10) gives

$$\begin{aligned}
&\sum_{k:|k-n/2|>A\sqrt{n}} p_{n,k} \\
&\quad = \sum_{k:|k-n/2|>A\sqrt{n}} cn^{-3/2}\exp\left(-\frac{(n-2k)^2}{2n\sigma^2}\right) + h(n,k) \\
&\quad = \sum_{k:|k-n/2|>A\sqrt{n}} cn^{-3/2}\exp\left(-\frac{(n-2k)^2}{2n\sigma^2}\right) + \sum_{k<n/3} h(n,k) + \\
&\qquad \sum_{k\geq n/3,\,|k-n/2|>A\sqrt{n}} h(n,k)\,.
\end{aligned}$$

The first of these terms is $2n^{-1}(1 - \Phi(2A/\sigma))$ plus an error term going to zero as $n \to \infty$ (uniformly in A, though we don't need the uniformity). The second term is at most

$$c_0 \sum_{k<n/3} k^{-1/2}(n-2k)^{-2} \leq 9c_0 n^{-3/2}\,.$$

And the third term is at most

$$2c_0 \sum_{j>A\sqrt{n}} (n/3)^{-1/2}(2j)^{-2} \leq \frac{c_0}{An}.$$

These three estimates establish (4.10), thus finishing step 4 and the proof of Theorem 4.1'. □

Acknowledgement: This work was carried out in the halls of the IMA workshop during the session on Monte Carlo and Markov chains. We thank the staff of the IMA for its terrific, supportive environment. We thank Bert Fristedt, Allan Gut, Greg Lawler, Jim Pitman and Larry Shepp for their help.

REFERENCES

[1] J. AITCHISON, *The Statistical Analysis of Compositional Data*, Chapman and Hall, New York 1986.
[2] R.N. BHATTACHARYA AND R.R. RAO, *Normal Approximation and Asymptotic Expansions*, Wiley, New York 1976.
[3] P. BILLINGSLEY, *Convergence of Probability Measures*, Wiley, New York 1968.
[4] D. CIFARELLI AND E. REGAZZINI, *On the asymptotic distribution of a statistic arising in testing the homogeneity of two samples*, Giornole Degli Economisie (1975), pp. 233–249.
[5] D. CRITCHLOW, *Metric methods for analyzing partially ranked data* in *Lecture Notes in Statistics No. 34*, Springer-Verlag, Berlin 1985.
[6] P. DIACONIS, *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, Hayward, California 1988.
[7] P. DIACONIS AND S. HOLMES, *Gray codes for randomization procedures*, Statistics and Computing, no 4, 1994.
[8] W. FELLER, *An Introduction to Probability Theory and its Applications Vol II*, (second edition) Wiley, New York 1971.
[9] B. FRISTEDT, *The central limit problem, for infinite products of, and Lévy processes of renewal sequences*, Z. Wahr. Verw. Gebiete **58** (1981), pp. 479–507.
[10] A. GUT, *Stopped Random Walks*, Springer-Verlag, Berlin 1987.
[11] L. HOLST, *Two conditional limit theorems with applications*, Ann. Stat. **7** (1979), pp. 551–557.
[12] L. HOLST, *Some conditional limit theorems in exponential families*, Ann. Prob. **9** (1981), pp. 818–830.
[13] J. HUNTER, *Renewal theory in two dimensions: asymptotic results*, Adv. Appl. Prob. **6** (1974), pp. 546–562.
[14] D. KENDALL, *Renewal sequences and their arithmetic* in *Symp. on Probability Methods in Analysis*, Springer Lecture Notes in Math. **31** (1967), pp. 147–175.
[15] D. KENDALL AND E. HARDING, *Stochastic Analysis*, Wiley, London 1973.
[16] J. KINGMAN, *Regenerative Phenomena*, Wiley, London 1972.
[17] T. LINDVALL, *Coupling Methods*, Cambridge University Press, Cambridge 1993.
[18] S. RACHEV, *The Monge-Kantorovich mass transference problem and its stochastic applications*, Theor. Prob. Appl. **29** (1986), pp. 647–676.
[19] T. RAMASUBBAN, *The mean difference and the mean deviation of some discontinuous distributions*, Biometrika **45** (1958), pp. 549–556.
[20] T. RAMASUBBAN, *The generalized mean differences of the binomial and Poisson distributions*, Biometrika **46** (1959), pp. 223–229.

[21] S. RICE, *The integral of the absolute value of the pinned Wiener processes—calculation of the probability density by numerical integration*, Ann. Prob. **10** (1982), pp. 240–243.

[22] L. SHEPP, *On the integral of the absolute value of the pinned Wiener processes*, Ann. Prob. **10** (1982), pp. 234–239. (Acknowledgement of priority Ann. Prob. **19**, p. 1397.)

[23] J. VEGELIUS, S. JANSON and F. JOHANSSON, *Measures of similarity between distributions*, Quality and Quantity **20** (1986), pp. 437–441.