COMBINATORICS AND GRAPH THEORY
BANACH CENTER PUBLICATIONS, VOLUME 25
PWN-POLISH SCIENTIFIC PUBLISHERS
WARSAW 1989

ON ENTROPIES OF OCCUPANCY DISTRIBUTIONS

OVE FRANK

Department of Statistics, University of Stockholm
Stockholm, Sweden

KRZYSZTOF NOWICKI

Department of Mathematical Statistics, University of Lund
Lund, Sweden

Occupancy models for equal or unequal objects distributed into equal or unequal cells can be represented by random partitions of integers or sets and also by randomly colored and labeled graphs. Applications of occupancy models to statistical disclosure control and to cluster inference in random graphs motivate an interest in entropies of occupancy distributions. We determine and compare entropies of various distributions encountered in some classical and recent occupancy models.

1. Introduction

Occupancy models refer to random distributions of equal or unequal objects into equal or unequal cells. Such models are useful in different kinds of applications, and they are also referred to as urn models (Johnson and Kotz, 1977) and allocation models (Kolchin, Sevast'yanov and Chistyakov, 1978).

If n equal objects are distributed into c equal (unequal) cells, the occupancy pattern can be represented by an unordered (ordered) partition of n into c nonnegative integers. If n unequal objects are distributed into c equal (unequal) cells, the occupancy pattern can be represented by an unordered (ordered) partition of an n-set into c disjoint subsets.

Our purpose is to investigate entropies of random partitions of integers or sets. The entropy of a random partition is a convenient measure of variation that has applications, for instance, in statistical disclosure control and in cluster

Key words and phrases: random partitions, random graphs, random allocations, random equivalence relations, cluster models, disclosure control, information and entropy.

analysis of random graphs. The next section comments briefly on these applications in order to provide some motivation for considering entropies of random partitions. Section 3 surveys some classical and recent occupancy models. Section 4 contains results on various entropies for different kinds of uniform occupancy models, and Section 5 contains results on entropies for the general multinomial occupancy model.

2. Applications to disclosure control and clustering

We consider briefly two fields of applications for entropies of occupancy distributions. We refer to Gallager (1968) for general information theory and basic facts about entropy that are needed.

General statistical disclosure is discussed by Fellegi (1972) and Dalenius (1974). Frank (1976) investigates disclosure in frequency tables. Consider n individuals classified according to a categorical distribution into c categories, and let $x_1, \dots, x_n$ be the category labels of the individuals and $y_1, \dots, y_c$ the frequencies in the categories. A particular disclosure control problem is to measure how protected the individual data $x_1, \dots, x_n$ are if the frequencies $y_1, \dots, y_c$ are released. By setting up a stochastic model for the data, i.e. by assuming that $x_1, \dots, x_n$ are observations on random variables $X_1, \dots, X_n$, a possible approach is to consider the entropy of $X_1, \dots, X_n$ conditioned by the given frequencies $y_1, \dots, y_c$ as a measure of individual privacy. Frank (1978, 1979) discusses disclosure in frequency tables when prior knowledge is also taken into account. Prior knowledge may for instance be used to estimate which individuals belong to the same categories. This specifies a partition of the n-set of individuals without specifying the categories. Disclosure then amounts to matching the sizes of the subsets of the partition to the frequencies $y_1, \dots, y_c$. This problem depends on the frequencies of different frequencies among $y_1, \dots, y_c$ only, say on $z_0, \dots, z_n$ where z_k is the number of categories containing exactly k individuals. The relevant entropy could be the entropy of $X_1, \dots, X_n$ conditioned both by the frequencies $y_1, \dots, y_c$ and the partition of the n-set.

Another field of application is to statistical inference about clusters. Data can for instance be similarity measurements for all pairs of objects in an n-set. In cluster analysis it is desired to use the similarities to group the objects together into an unknown number c of subsets (clusters) of similar objects. A possible model for the similarities could be a random graph generated by some latent clusters. Frank and Harary (1982) discuss such an approach to clustering with applications to linguistics (McNeil, 1973) and ecology (Engen, 1978). Inference problems concern the number of clusters, their expected sizes, their entropies, etc. Even some very simple random graph models lead to complicated probabilistic problems. Simple models are, for instance, the so-called random transitive graphs which have the objects as vertices and all pairs of objects in the same clusters connected by edges. Such graphs represent unordered partitions of the object set. Information about the clusters inherent in an observed graph could be measured by the entropy of the graph.

3. Occupancy distributions

Consider n equal or unequal objects distributed at random into c equal or unequal cells. Let the objects and cells be labeled by integers $1, \ldots, n$ and $1, \ldots, c$, respectively. Let $X = (X_1, \ldots, X_n)$ denote the cell labels of the n objects, $Y = (Y_1, \ldots, Y_c)$ the frequencies of objects in the c cells, and $Z = (Z_0, \ldots, Z_n)$ the frequencies of cells occupied be exactly $0, \ldots, n$ objects, respectively. Let G be the graph on the n-set with an edge between objects i and j if and only if $X_i = X_j$.

In partition terminology (Andrews, 1976) X represents an ordered partition of an n-set into c disjoint •subsets, G represents an unordered partition of the n-set into at most c nonempty disjoint subsets, Y represents an ordered partition of n into c nonnegative integers, and Z represents an unordered partition of n into at most c positive integers. It follows that the numbers of outcomes of X, Y, Z, and G are c^n, $\binom{n+c-1}{n}$, $\sum_{k=1}^{c} P(n, k)$, and $\sum_{k=1}^{c} S(n, k)$, respectively, where $P(n, k)$ denotes the number of unordered partitions of n into k positive integers, and $S(n, k)$ the number of unordered partitions of an n-set into k nonempty disjoint subsets. $S(n, k)$ is a Stirling number of the second kind.

In graph terminology (Harary, 1969), X represents a colored labeled graph (c colors and n labels), Y a colored unlabeled graph isomorphic to X, Z an uncolored unlabeled graph isomorphic to X, and G an uncolored labeled graph isomorphic to X. Figure 1 shows the number of ways to color and label the graph Z.

According to the classical Maxwell-Boltzmann model, X has a uniform distribution on its c^n outcomes. This means that equal probabilities are given to all distinct occupancy patterns for n unequal objects into c unequal cells. Y has

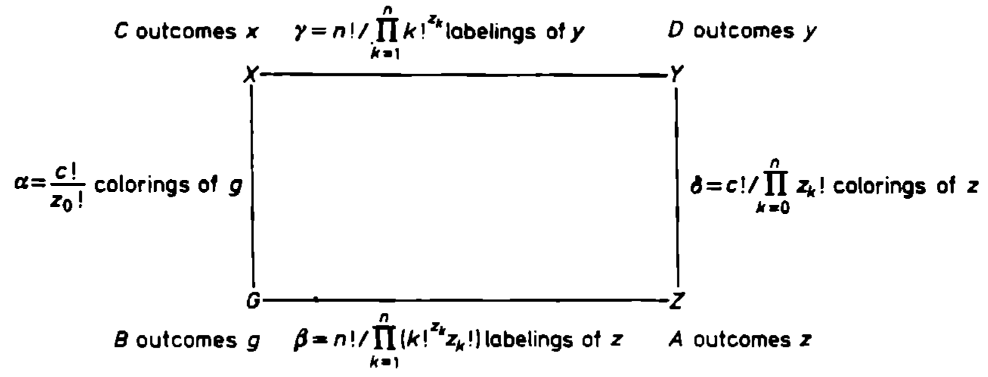
C outcomes x
$\gamma = n!/\prod_{k=1}^{n} k!^{z_k}$ labelings of y
D outcomes y
X
Y
$\alpha = \frac{c!}{z_0!}$ colorings of g
$\delta = c!/\prod_{k=0}^{n} z_k!$ colorings of z
G
Z
B outcomes g
$\beta = n!/\prod_{k=1}^{n} (k!^{z_k} z_k!)$ labelings of z
A outcomes z

Fig. 1. The number of ways to color and label a graph

a multinomial distribution with integer parameter n and with c equal category probabilities $1/c$. The probability function of Z is given by

$$(1) \qquad P(Z=z) = \frac{c!n!}{c^n \prod_{k=0}^{n} (k!^{z_k} z_k!)}$$

for all $z = (z_0, \ldots, z_n)$ satisfying

$$(2) \qquad z_0, \ldots, z_n \geqslant 0, \quad z_0 + \ldots + z_n = c, \quad z_1 + 2z_2 + \ldots + nz_n = n.$$

The number $K = Z_1 + \ldots + Z_n$ of occupied cells has probability function

$$(3) \qquad P(K=k) = \frac{c!S(n,k)}{(c-k)!c^n} \quad \text{for } k = 1, \ldots, c.$$

This distribution is sometimes called *Arfwedson's distribution* (Johnson and Kotz, 1969, p. 251). Finally, the Maxwell–Boltzmann model implies that G has probability function

$$(4) \qquad P(G=g) = \frac{c!}{(c-k)!c^n}$$

for each labeled graph g having k components.

According to the classical Bose–Einstein model, Y is supposed to have a uniform distribution on its $\binom{n+c-1}{n}$ outcomes. This model gives equal probabilities to all distinct occupancy patterns for n equal objects into c unequal cells. It follows that Z has probability function

$$(5) \qquad P(Z=z) = \frac{c!}{\binom{n+c-1}{n} \prod_{k=0}^{n} z_k!}$$

for all z satisfying (2). The number K of occupied cells has a hypergeometric distribution:

$$(6) \qquad P(K=k) = \frac{\binom{c}{k}\binom{n-1}{n-k}}{\binom{n+c-1}{n}} \quad \text{for } k = 1, \ldots, c.$$

The Bose–Einstein model together with the assumption that all labelings of Y are equally probable implies that G has probability function

$$(7) \qquad P(G=g) = \frac{c!(c-1)! \prod_{k=1}^{n} k!^{z_k}}{(n+c-1)!z_0!}$$

for each labeled graph g isomorphic to the unlabeled graph z, and X has probability function

$$(8) \qquad P(X = x) = \frac{(c-1)! \prod_{k=1}^{n} k!^{z_k}}{(n+c-1)!}$$

for each x isomorphic to z.

Another classical model, the Fermi–Dirac model, assumes that $n \leqslant c$ and that Y is uniformly distributed on the $\binom{c}{n}$ outcomes that correspond to the ordered partitions of n into c binary digits. The Fermi–Dirac model permits no multiple occupancy, and with this restriction it gives equal probabilities to all distinct occupancy patterns for n equal objects into c unequal cells. The frequencies in Z degenerate into $Z_0 = c-n$, $Z_1 = n$. The graph G degenerates to the empty graph on n vertices.

A random partition model investigated by Harper (1967), Haigh (1972) and Recski (1976) assumes that $n \leqslant c$ and that G has a uniform distribution on the $B_n = \sum_{k=1}^{n} S(n, k)$ unordered partitions of the vertex set (B_n is the so-called Bell number). According to this model, it follows that Z has probability function

$$(9) \qquad P(Z = z) = \frac{n!}{B_n \prod_{k=1}^{n} (k!^{z_k} z_k!)}$$

for z satisfying (2), and the number K of components of the graph G has probability function

$$(10) \qquad P(K = k) = \frac{S(n, k)}{B_n} \quad \text{for } k = 1, \ldots, n.$$

A generalization of this model without the restriction $n \leqslant c$ is obtained by assuming that G is uniformly distributed on the

$$(11) \qquad B(n, c) = \sum_{k=1}^{c} S(n, k)$$

unordered partitions of the n-set into at most c nonempty disjoint subsets. Then (9) holds with B_n replaced by $B(n, c)$, and (10) is replaced by

$$(12) \qquad P(K = k) = \frac{S(n, k)}{B(n, c)} \quad \text{for } k = 1, \ldots, c.$$

The uniform G-distribution, together with the assumption that different colors are assigned at random to the components of each outcome of G, impies that

$$P(X = x) = \frac{z_0!}{c!B(n, c)} \tag{13}$$

for each x with $c - z_0$ components, and

$$P(Y = y) = \frac{n!z_0!}{c!B(n, c) \prod_{k=1}^{n} k!^{z_k}} \tag{14}$$

for each y isomorphic to z.

The Maxwell-Boltzmann model has a uniform X-distribution, the Bose-Einstein model has a uniform Y-distribution, and the generalized Harper-Haigh-Recski model has a uniform G-distribution. If we now consider a model with a uniform Z-distribution, together with the assumption that each outcome of Z is labeled at random and independently colored at random, then, for each x, y, and g isomorphic to z,

$$P(X = x) = \frac{\prod_{k=0}^{n} (k!^{z_k} z_k!)}{n!c!A(n, c)}, \tag{15}$$

$$P(Y = y) = \frac{\prod_{k=0}^{n} z_k!}{c!A(n, c)}, \tag{16}$$

$$P(G = g) = \frac{\prod_{k=1}^{n} (k!^{z_k} z_k!)}{n!A(n, c)}, \tag{17}$$

where

$$A(n, c) = \sum_{k=1}^{c} P(n, k). \tag{18}$$

All these distributions are easily obtained by referring to Fig. 1. In order to simplify notation we shall use the following notation for the four numbers of choices given in Fig. 1:

$$\begin{aligned} \alpha &= \alpha_z = c!/z_0!, & \gamma &= \gamma_z = n!/\prod_{k=1}^{n} k!^{z_k}, \\ \beta &= \beta_z = n!/\prod_{k=1}^{n} (k!^{z_k} z_k!), & \delta &= \delta_z = c!/\prod_{k=0}^{n} z_k!. \end{aligned} \tag{19}$$

The total numbers of outcomes of Z, G, X and Y are denoted by

$$\begin{aligned} A &= A(n, c) = \sum_{k=1}^{c} P(n, k), \\ B &= B(n, c) = \sum_{k=1}^{c} S(n, k), \\ C &= C(n, c) = c^n, \\ D &= D(n, c) = \binom{c+n-1}{n}. \end{aligned} \tag{20}$$

We note that $\alpha\beta = \gamma\delta$ and

$$\sum_z \beta_z = B, \quad \sum_z \delta_z = D, \quad \sum_z \alpha_z \beta_z = \sum_z \gamma_z \delta_z = C, \tag{21}$$

where all sums are over A values of z.

4. Entropies under different kinds of uniform occupancy

In this section we consider the four models specified by assuming uniform distribution of X, Y, G, or Z and, if appropriate, also uniform colorings and labelings of their outcomes. Table 1 summarizes the probability functions of X, Y, G and Z under these four models. The models are referred to as MB, BE, HHR, and FN. For all four models we have

Table 1

Probability functions of X, Y, G and Z under four different uniform occupancy models

	X	Y	G	Z
MB	$\frac{1}{C}$	$\frac{\gamma}{C}$	$\frac{\alpha}{C}$	$\frac{\alpha\beta}{C}$
BE	$\frac{1}{\gamma D}$	$\frac{1}{D}$	$\frac{\alpha}{\gamma D}$	$\frac{\delta}{D}$
HHR	$\frac{1}{\alpha B}$	$\frac{\beta}{\delta B}$	$\frac{1}{B}$	$\frac{\beta}{B}$
FN	$\frac{1}{\alpha\beta A}$	$\frac{1}{\delta A}$	$\frac{1}{\beta A}$	$\frac{1}{A}$

$$\begin{aligned} P(X = x) &= P(Z = z)/\alpha\beta, \\ P(Y = y) &= P(Z = z)/\delta, \\ P(G = g) &= P(Z = z)/\beta, \end{aligned} \tag{22}$$

if x, y, and g are isomorphic to z. In order to give formulas for the entropies we introduce the function

$$\varphi(p) = \begin{cases} -p\log p & \text{if } p > 0, \\ 0 & \text{if } p = 0, \end{cases} \tag{23}$$

so that the entropy of Z is

$$H(Z) = \sum_z \varphi(P(Z = z)). \tag{24}$$

Table 2 gives the entropies of X, Y, G, and Z under the four models MB, BE, HHR, and FN.

The information about X gained by observing $Y = y$ is equal to the entropy difference

$$I(X \mid Y = y) = H(X) - H(X \mid Y = y), \tag{25}$$

where

$$H(X \mid Y = y) = \sum_x \varphi(P(X = x \mid Y = y)) \tag{26}$$

Table 2

Entropies of X, Y, G and Z under four different uniform occupancy models

	X	Y	G	Z
MB	$\log C$	$\sum_z \delta\varphi(\gamma/C)$	$\sum_z \beta\varphi(\alpha/C)$	$\sum_z \varphi(\alpha\beta/C)$
BE	$\sum_z \alpha\beta\varphi(1/\gamma D)$	$\log D$	$\sum_z \beta\varphi(\alpha/\gamma D)$	$\sum_z \varphi(\delta/D)$
HHR	$\sum_z \alpha\beta\varphi(1/\alpha B)$	$\sum_z \delta\varphi(\beta/\delta B)$	$\log B$	$\sum_z \varphi(\beta/B)$
FN	$\sum_z \alpha\beta\varphi(1/\alpha\beta A)$	$\sum_z \delta\varphi(1/\delta A)$	$\sum_z \beta\varphi(1/\beta A)$	$\log A$

is a conditional entropy. Positive information corresponds to reduced entropy, and negative information to increased entropy. The expected information is always nonnegative and equal to

$$EI(X \mid Y) = H(X) - EH(X \mid Y), \tag{27}$$

where $I(X \mid Y)$ and $H(X \mid Y)$ denote the information and conditional entropy considered as random variables of Y. Now, generally we have

$$EI(X \mid Y) = EI(Y \mid X), \tag{28}$$

and, since Y is a function of X, $EH(Y \mid X) = 0$. It follows that the expected information about X provided by Y is given by the entropy $H(Y)$. From (25) it also follows that the variance of the information $I(X|Y)$ is equal to the variance of the conditional entropy $H(X|Y)$, i.e.

$$VI(X|Y) = VH(X|Y). \tag{29}$$

(It is common in information theory not to consider variances but only expected values of conditional entropies; our notation $EH(X|Y)$ is then often simplified to $H(X|Y)$.)

Since G is a function of X, and since Z is a function of either one of Y and G, we find analogously that the expected informations about X provided by G and Z are equal to the entropies $H(G)$ and $H(Z)$, respectively.

LEMMA 1. *Under each of the models MB, BE, HHR, FN we have*

$$H(Z|Y=y) = H(Z|G=g) = H(Y|X=x) = H(G|X=x) = 0,$$

$$H(X|G=g) = \log\alpha_z \quad \textit{where } g \textit{ is isomorphic to } z,$$

$$H(G|Z=z) = H(G|Y=y) = \log\beta_z \quad \textit{where } y \textit{ is isomorphic to } z,$$

$$H(X|Y=y) = \log\gamma_z \quad \textit{where } y \textit{ is isomorphic to } z,$$

$$H(Y|Z=z) = H(Y|G=g) = \log\delta_z \quad \textit{where } g \textit{ is isomorphic to } z,$$

$$H(X|Z=z) = \log(\alpha_z\beta_z),$$

$$H(X|Y=y, G=g) = \log(\alpha_z/\delta_z) \quad \textit{where } y \textit{ and } g \textit{ are isomorphic to } z.$$

The conditional entropies in this lemma are easily obtained from Table 1. Using Lemma 1 and basic properties of entropies we can readily prove the following theorem. We can also prove it directly from Table 1 by using (23).

THEOREM 2. *Under each of the models MB, BE, HHR, FN we have*

$$H(G) = H(Z) + E\log\beta,$$

$$H(Y) = H(Z) + E\log\delta,$$

$$H(X) = H(Z) + E\log(\alpha\beta).$$

THEOREM 3. *Under each of the models MB, BE, HHR, FN we have*

$$H(Z) \leqslant H(Y) \leqslant H(X), \qquad H(Z) \leqslant H(G) \leqslant H(X),$$

$$-\log E(1/\alpha) \leqslant H(X) - H(G) \leqslant \log E\alpha,$$

$$-\log E(1/\beta) \leqslant H(G) - H(Z) \leqslant \log E\beta,$$

$$-\log E(1/\gamma) \leqslant H(X) - H(Y) \leqslant \log E\gamma,$$

$$-\log E(1/\delta) \leqslant H(Y) - H(Z) \leqslant \log E\delta,$$

$$-\log E(\delta/\beta) \leqslant H(G) - H(Y) \leqslant \log E(\beta/\delta).$$

Proof. The first four inequalities follow from Theorem 2 since α, β, γ, and δ are $\geqslant 1$. The last inequalities give the logarithms of thè harmonic and arithmetic means as bounds to the logarithm of the geometric mean of α, β, γ, δ, and β/δ, respectively.

COROLLARY 4. *The following holds under each of the models MB, BE, HHR, FN. If $E(\beta/\delta) < 1$, then $H(G) < H(Y)$. If $E(\delta/\beta) < 1$, then $H(Y) < H(G)$. In particular, under the BE-model, $B < D$ implies that $H(G) < H(Y)$, and under the HHR-model, $D < B$ implies that $H(Y) < H(G)$.*

Remark. Any one of B and D can be the largest. For instance, $n = 5$, $c = 3$ yield $B = 41$, $D = 21$, and $n = 5$, $c = 4$ yield $B = 51$, $D = 56$.

Let $P = P_z$ denote the probability function of Z under any one of the models MB, BE, HHR or FN. Then $H(Z) = -E \log P$, and by applying the harmonic-geometric-arithmetic-mean inequalities to the expressions for $H(G)$, $H(X)$ and $H(Y)$ given by Theorem 2, we find the following bounds, which are sometimes sharper than the bounds that can be obtained from Theorem 3.

THEOREM 5. *Under each of the models MB, BE, HHR, FN we have*

$$\begin{aligned} 0 &\leqslant H(Z) \leqslant \log A, && \text{with equality to the right iff FN holds,}\\ -\log E(P/\beta) &\leqslant H(G) \leqslant \log B, && \text{with equality to the right iff HHR holds,}\\ -\log E(P/\alpha\beta) &\leqslant H(X) \leqslant \log C, && \text{with equality to the right iff MB holds,}\\ -\log E(P/\delta) &\leqslant H(Y) \leqslant \log D, && \text{with equality to the right iff BE holds.} \end{aligned}$$

COROLLARY 6. *Under each of the models MB, BE, HHR, FN we have*

$$\begin{aligned} &H(Z) = O(\sqrt{n}) \quad \text{for } n \to \infty,\\ &H(G) = O(n\log c) \quad \text{for } c = O(n),\ n \to \infty,\\ &H(G) = O(n\log n) \quad \text{for } n \to \infty,\\ &H(Y) = O(n) \quad \text{for } c = O(n),\ n \to \infty,\\ &H(X) = O(n\log c) \quad \text{for } n \to \infty. \end{aligned}$$

In particular, under the FN-model with $c = n$

$$H(Z) = \pi\sqrt{2n/3} - \log(4\sqrt{3}n) + o(1),$$

under the HHR-model with $c = n$

$$H(G) = \log B_n = O(n\log n),$$

under the BE-model with $c = n$

$$H(Y) = (2n-1)\log 2 - \log\sqrt{n\pi} + o(1),$$

under the BE-model with c fixed

$$H(Y) = (c-1)\log n - \log(c-1)! - (c-1) + o(1),$$

and under the MB-model

$$H(X) = n\log c.$$

Proof. We note that

$$A(n, c) \leqslant A(n, n) \sim \frac{1}{4\sqrt{3}n} e^{\pi\sqrt{2n/3}} \tag{30}$$

(see, e.g., Andrews, 1976, p. 70), and the results for $H(Z)$ follow from Theorem 5. The results for $H(G)$ are based on the following upper bounds for B:

$$B(n, c) \leqslant B(n, n), \tag{31}$$

$$B(n, c) = \sum_{k=1}^{c} S(n, k) \leqslant \sum_{k=1}^{c} \binom{n-1}{k-1} k^{n-k} \leqslant c^n \sum_{k=1}^{c} \binom{n-1}{k-1} \leqslant c^n 2^{n-1}. \tag{32}$$

The results for $H(X)$ are immediate from $C = c^n$. Finally, the results for $H(Y)$ follow from an investigation of $D = \binom{c+n-1}{n}$; for finite c

$$D \sim (n/e)^{c-1}/(c-1)! \tag{33}$$

and for $c \to \infty$ and $n \to \infty$

$$D \sim (n+c-1)^{n+c-1/2}/n^{n+1/2}(c-1)^{c-1/2}\sqrt{2\pi}. \tag{34}$$

5. Entropies under general multinomial occupancy

Throughout this section we assume that X consists of independent identically distributed random variables with probability function $P(X_i = j) = p_j$ satisfying $p_1, \ldots, p_c > 0$ and $p_1 + \ldots + p_c = 1$. This model implies that Y is multinomial $(n, p_1, \ldots, p_c)$, and it is usually referred to as the general multinomial occupancy model. Kolchin, Sevast'yanov and Chistyakov (1978, Chapter 3) investigate the asymptotic behavior of the probability distribution of Z when n and c tend to infinity and various restrictions are imposed on n, c, $p_1, \ldots, p_c$. For $p_1 = \ldots = p_c = 1/c$ we recover the MB-model.

Most of the assertions of Lemma 1 are still valid for multinomial occupancy. The following lemma is easily proved.

LEMMA 7. *Under multinomial $(n, p_1, \ldots, p_c)$ occupancy Z and G have probability functions*

$$P(Z = z) = \gamma_z \sum_{y} \prod_{j=1}^{c} p_j^{y_j}, \quad P(G = g) = (\alpha_z/\delta_z) \sum_{y} \prod_{j=1}^{c} p_j^{y_j},$$

where the sums are over y isomorphic to z, and z is isomorphic to g. Conditional entropies satisfy

$$\begin{aligned} &H(Z|Y=y)=H(Z|G=g)=H(Y|X=x)=H(G|X=x)=0,\\ &H(G|Z=z)=H(G|Y=y)=\log\beta_z,\\ &H(X|Y=y)=\log\gamma_z,\\ &H(X|G=g)\leqslant\log\alpha_z,\\ &H(Y|Z=z)=H(Y|G=g)\leqslant\log\delta_z,\\ &H(X|Z=z)\leqslant\log(\alpha_z\beta_z),\\ &H(X|Y=y,\ G=g)\leqslant\log(\alpha_z/\delta_z),\end{aligned}$$

where y and g are isomorphic to z.

Using this lemma, we readily obtain the following theorem.

THEOREM 8. *Under multinomial* $(n, p_1, \ldots, p_c)$ *occupancy we have*

$$H(X)=n\sum_{i=1}^{c}\varphi(p_i),\qquad H(Y)=H(X)-E\log\gamma,\qquad H(G)=H(Z)+E\log\beta,$$

$$H(Z)\leqslant H(Y)\leqslant H(X),\qquad H(Z)\leqslant H(G)\leqslant H(X).$$

It is less evident whether or not a general inequality holds between $H(Y)$ and $H(G)$. Theorems 9 and 12 below show that $H(Y)\leqslant H(G)$ is true for sufficiently large n if $c\log c=o(n)$. In fact, $H(X)\sim H(G)$ and $H(Y)\sim H(Z)$ for $c\log c=o(n)$ and $n\to\infty$.

THEOREM 9. *Under multinomial* $(n, p_1, \ldots, p_c)$ *occupancy we have*

$$H(Y)=n\sum_{j=1}^{c}\varphi(p_j)-\log n!+\sum_{k=0}^{n}\sum_{j=1}^{c}\binom{n}{k}p_j^k(1-p_j)^{n-k}\log k!.$$

For $\sum_{j=1}^{c}(1/p_j)=O(c^2)$, $c^2=o(n)$ *and* $n\to\infty$

$$H(Y)=\frac{c-1}{2}\log(2\pi en)+\tfrac{1}{2}\sum_{j=1}^{c}\log p_j+O(c^2/n).$$

Proof. According to Lemma 7 and Theorem 8 it follows that

$$H(Y)=H(X)-EH(X|Y)=n\sum_{j=1}^{c}\varphi(p_j)-\log n!+\sum_{j=1}^{c}E\log Y_j!. \tag{35}$$

Multinomial occupancy implies that Y_j is binomial (n, p_j), and the first formula for $H(Y)$ follows readily. To prove the asymptotic formula we use a refinement of Stirling's approximation (cf. Feller, 1957, p. 52) and deduce that $H(X|Y)$ has upper bound

(36) $$(n+\tfrac{1}{2})\log n-\sum_{j=1}^{c}(Y_j+\tfrac{1}{2})\log Y_j-\frac{c-1}{2}\log(2\pi)+\frac{1}{12n}-\sum_{j=1}^{c}\frac{1}{12Y_j+1}$$

and lower bound

(37) $$(n+\tfrac{1}{2})\log n-\sum_{j=1}^{c}(Y_1+\tfrac{1}{2})\log Y_j-\frac{c-1}{2}\log(2\pi)+\frac{1}{12n+1}-\sum_{j=1}^{c}\frac{1}{12Y_j-1}$$

provided we define $\log 0 = 0$. Therefore

(38) $$H(X\,|\,Y) = n\sum_{j=1}^{c}\varphi(Y_j/n)-\tfrac{1}{2}\sum_{j=1}^{c}\log(Y_j/n)-\frac{c-1}{2}\log(2\pi n)+O(c^2/n).$$

Basharin (1959) uses $\sum\varphi(Y_j/n)$ as an estimator of $\sum\varphi(p_j)$ and shows that its bias for finite c and $n\to\infty$ is $-(c-1)/2n+O(1/n^2)$. This method of proof can be used to show that

(39) $$E\sum_{j=1}^{c}\varphi(Y_j/n)=\sum_{j=1}^{c}\varphi(p_j)-(c-1)/2n+O(c^2/n^2).$$

Furthermore,

(40) $$E\log(Y_j/n)=\log p_j+O(1/n).$$

Substitution of (38), (39) and (40) into (35) yields the desired result.

COROLLARY 10. *For $c = 2$, $p_1 = p$, $p_2 = q$, and $n\to\infty$*

(41) $$H(Y)=\tfrac{1}{2}\log(2\pi enpq)+O(1/n).$$

Remark. The entropy of a normal distribution with probability density function $f(x)$ is given by Kullback (1959, p. 32) as

(42) $$-\int_{-\infty}^{\infty}f(x)\log f(x)\,dx=\tfrac{1}{2}\log(2\pi e\sigma^2),$$

where σ^2 is the variance. Corollary 10 implies that the entropy of a binomial (n, p) distribution is asymptotically equal to the entropy of a normal distribution with variance npq.

THEOREM 11.

$$H(Z)\geqslant n\sum_{j=1}^{c}\varphi(p_i)-\log n!-\log c!+\sum_{k=0}^{n}\sum_{j=1}^{c}P(Z_k\geqslant j)\log(k!j)$$

with equality iff $p_1 = \ldots = p_c = 1/c$. For $\sum(1/p_i) = O(c^2)$, $c = o(n)$ and $n\to\infty$,

$$H(Z)=\frac{c-1}{2}\log n+O(c\log c).$$

Proof. According to Lemma 7, it follows that

$$(43)\qquad H(Z) = H(Y) - EH(Y|Z) \geqslant H(Y) - \log c! + \sum_{k=0}^{n} E\log Z_k! \geqslant H(Y) - \log c!,$$

which together with $H(Y) \geqslant H(Z)$ yields the asymptotic result. The iiequality of Theorem 11 follows from (43) and (35) by noting that

$$(44)\qquad \sum_{j=1}^{c} E\log Y_j! = \sum_{k=0}^{n} (EZ_k)\log k!,$$

$$(45)\qquad EZ_k = \sum_{j=1}^{c} P(Z_k \geqslant j),$$

$$(46)\qquad E\log Z_k! = \sum_{j=1}^{c} P(Z_k \geqslant j)\log j.$$

THEOREM 12.

$$H(G) \geqslant n \sum_{j=1}^{c} \varphi(p_j) - \log c! + \sum_{k=1}^{c} \frac{c!S(n,k)\log(c-k)!}{(c-k)!c^n}$$

with equality iff $p_1 = \ldots = p_c = 1/c$. *For* $c\log c = o(n)$ *and* $n \to \infty$,

$$H(G) = n \sum_{j=1}^{c} \varphi(p_j) + o(n).$$

Proof. According to Lemma 7, $(X|G = g)$ is uniform iff $p_1 = \ldots = p_c = 1/c$, and it follows that

$$(47)\qquad H(G) = H(X) - EH(X|G) \geqslant H(X) - \log c! + E\log Z_0!$$

with equality iff $p_1 = \ldots = p_c = 1/c$. Substituting $Z_0 = c - K$ where K has probability function (3) yields the first part of Theorem 12. The second part follows from the inequality

$$(48)\qquad n \sum_{j=1}^{c} \varphi(p_j) - \log c! \leqslant H(G) \leqslant n \sum_{j=1}^{c} \varphi(p_j).$$

We conclude with a result about the variance of the information $I(X|Y)$. Recall that according to (27), $EI(X|Y) = H(Y)$, and according to (29), $VI(X|Y) = VH(X|Y)$.

THEOREM 13. *Under multinomial* $(n, p_1, \ldots, p_c)$ *occupancy with* $\sum(1/p_j) = O(c^2)$, $c^2\log^2 c = o(n)$ *and* $n \to \infty$, *we have*

$$VH(X|Y) = n\Big[\sum_{j=1}^{c} p_j\log^2 p_j - \Big(\sum_{j=1}^{c} \varphi(p_j)\Big)^2\Big] + O(c^2\log^2 c).$$

Proof. According to (35), (38) and Theorem 9 we obtain

$$H(X|Y)-EH(X|Y)=n\sum_{j=1}^{c}\left[\varphi\left(\frac{Y_j}{n}\right)-\varphi(p_j)\right]+\frac{c-1}{2}\tag{49}$$
$$-\frac{1}{2}\sum_{j=1}^{c}\log\frac{Y_j}{np_j}+O(c^2/n).$$

Now the geometric-arithmetic means inequality together with the assumption $\sum(1/p_j)=O(c^2)$ imply that

$$\sum_{j=1}^{c}\log\frac{Y_j}{np_j}=O(c\log c),\tag{50}$$

and it follows from (49) that

$$VH(X|Y)=n^2V\sum_{j=1}^{c}\varphi\left(\frac{Y_j}{n}\right)+O\left(\frac{c^4}{n^2}\right)+O(c^2\log^2 c)+O\left(\frac{c^3\log c}{n}\right).\tag{51}$$

Here the first term can be handled by the method used by Basharin (1959), and this leads to the formula in Theorem 13.

It is possible to improve this result by expansion of (49) without using (50). Then after tedious work it is possible to show that the remainder term of order $O(c^2\log^2 c)$ can be replaced by $O(c^2)$ where we need to assume $c^2=o(n)$ only.

References

[1] G. Andrews (1976), *The Theory of Partitions*, Addison-Wesley, Reading, Mass.

[2] G. P. Basharin (1959), *On a statistical estimate for the entropy of a sequence of independent random variables*, Theory Probab. Appl. 4, 333–336.

[3] T. Dalenius (1974), *The invasion of privacy problem and statistics production—an overview*, Statistisk Tidskrift 3, 213–225.

[4] S. Engen (1978), *Stochastic Abundance Models*, Chapman and Hall, London.

[5] I. P. Fellegi (1972), *On the question of statistical confidentiality*, J. Amer. Statist. Assoc. 67, 7–18.

[6] O. Frank (1976), *Individual disclosures from frequency tables*, in: Proc. Sympos. on Personal Integrity and the Need for Data in the Social Sciences, T. Dalenius and A. Klevmarken (eds.), Swedish Council for Social Science Research, Stockholm, 175–187.

[7] O. Frank (1978), *An application of information theory to the problem of statistical disclosure*, J. Statist. Plann. Inference 2, 143–152.

[8] O. Frank (1979), *Inferring individual information from released statistics*, invited paper for the 42nd Session of the International Statistical Institute, Manila, Philippines.

[9] O. Frank and F. Harary (1982), *Cluster inference by using transitivity indices in empirical graphs*, J. Amer. Statist. Assoc. 77, 835–840.

[10] R. Gallager (1968), *Information Theory and Reliable Communication*, Wiley, New York.

[11] J. Haigh (1972), *Random equivalence relations*, J. Combin. Theory Ser. A 13, 287–295.

[12] F. Harary (1969), *Graph Theory*, Addison-Wesley, Reading, Mass.

[13] L. H. Harper (1967), *Stirling behavior is asymptotically normal*, Ann. Math. Statist. 38, 410–414.

[14] N. L. Johnson and S. Kotz (1969), *Discrete Distributions*, Houghton Mifflin, Boston.

[15] N. L. Johnson and S. Kotz (1977), *Urn Models and Their Applications*, Wiley, New York.

[16] V. F. Kolchin, B. A. Sevast'yanov and V. P. Chistyakov (1978), *Random Allocations*, Winston, Washington, D.C.

[17] S. Kullback (1959), *Information Theory and Statistics*, Wiley, New York.

[18] D. R. McNeil (1973), *Estimating an author's vocabulary*, J. Amer. Statist. Assoc. 68, 92–96.

[19] A. Recski (1976), *On random partitions*, Discrete Math. 16, 173–177.

Presented to the Semester
Combinatorics and Graph Theory
September 14 – December 17, 1987