PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 107, Number 2, October 1989

AN ENTROPY INEQUALITY FOR THE BI-MULTIVARIATE HYPERGEOMETRIC DISTRIBUTION

FRED KOCHMAN, ALAN MURRAY AND DOUGLAS B. WEST

(Communicated by Andrew Odlyzko)

ABSTRACT. Given parameters $\bar{r} = r_1, \ldots, r_m$ and $\bar{c} = c_1, \ldots, c_n$ with $\sum r_i = \sum c_j = N$, the *bi-multivariate hypergeometric distribution* is the distribution on nonnegative integer $m \times n$ matrices with row sums $\bar{r}$ and column sums $\bar{c}$ defined by $\text{Prob}(A) = \prod r_i! \prod c_j! / (N! \prod a_{ij}!)$. It is shown that the entropy of this distribution is a Schur-concave function of the block-size parameters.

1. INTRODUCTION

The multivariate hypergeometric distribution is well studied in probability theory. Given nonnegative integer parameters r and $c_1, \ldots, c_n$, the distribution is defined by $P(x_1, \ldots, x_n) = \prod \binom{c_i}{x_i} / \binom{\sum c_i}{r}$ if $\sum x_i = r$, and 0 otherwise (for example, see [1, p. 308]). The combinatorial interpretation of this is an experiment in which an urn contains c_i balls of type i, and r balls are to be drawn. If the balls are drawn randomly, then the probability of getting x_i balls of type i is $P(x_1, \ldots, x_n)$. The special case $n = 2$ is known as the hypergeometric distribution.

There is also a natural combinatorial interpretation for a further generalization in which the parameter r is replaced by a vector of parameters $r_1, \ldots, r_m$. Given parameters $\bar{r} = r_1, \ldots, r_m$ and $\bar{c} = c_1, \ldots, c_n$ with $\sum r_i = \sum c_j = N$, we define the *bi-multivariate hypergeometric distribution* to be the distribution on nonnegative integer $m \times n$ matrices with row sums $\bar{r}$ and column sums $\bar{c}$ defined by $\text{Prob}(A) = \prod r_i! \prod c_j! / (N! \prod a_{ij}!)$. Note the symmetry of the probability function and the fact that it reduces to multivariate hypergeometric distribution when $m = 2$.

The combinatorial interpretation arises from the following experiment. Partition the rows and columns of an $N \times N$ matrix into blocks of sizes $r_1, \ldots, r_m$ and $c_1, \ldots, c_n$, respectively. Then the probability that a random permutation matrix has a_{ij} ones in block i, j is given by $\text{Prob}(A)$ as defined above (we will show this shortly). For a given permutation matrix, call the matrix whose i, jth entry is the number of 1's in the i, jth block the *collapsed matrix*.

Received by the editors April 30, 1988 and, in revised form, January 30, 1989.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 60E15, 94A17.

The entropy of this distribution is a measure of the information retained in the collapsed matrix. For fixed N, m, n, the entropy is a function $H(\bar{r}, \bar{c})$ of the block-size parameters. As with many parametrized discrete distributions, one may intuitively expect that the entropy is maximized when the r_i's and c_j's are chosen as close as possible to N/m and N/n, respectively; i.e., when no pair of r_i's or c_j's differ by more than one.

In fact, we prove a stronger result from which this follows immediately. We prove that the entropy increases whenever we move two of the r_i's or two of the c_j's closer to each other by the same integral amount. More formally, a function $f(x_1, \ldots, x_n)$ is said to be *Schur-convex* [1] if moving any two arguments closer together by the same amount (while staying within the domain of definition) decreases f, and f is *strictly Schur-convex* if this decrease is always strict. If $-f$ is Schur-convex, then f is *Schur-concave.* We prove the following.

Theorem. *For $m, n > 1$, the entropy $H(\bar{r}, \bar{c})$ of the generalized multivariate hypergeometric distribution is a strictly Schur-concave function of its row-parameters $r_1, \ldots, r_m$ or of its column parameters $c_1, \ldots, c_n$.*

It is not difficult to reduce this theorem to the case $m = 2$, i.e., to prove the Schur-concavity of the entropy of the multivariate hypergeometric distribution. We do not take this approach, because it does not materially shorten the rest of our proof. Nevertheless, it is worth considering this approach, because known results on Schur-concavity may apply to that problem.

The book by Marshall and Olkin [1] contains a thorough discussion of Schur-convexity, including the following theorem on [1, p. 309] (paraphrased to our notation): "If $X_1, \ldots, X_n$ have a multivariate hypergeometric distribution (with parameters $c_1, \ldots, c_n$) then $\psi(c_1, \ldots, c_n) = E\phi(X_1, \ldots, X_n)$ is a Schur-concave function of the parameters $c_1, \ldots, c_n$ whenever ϕ is a Schur-concave function." Entropy is the expectation of the negative logarithm of the probability, so this theorem would seem to solve the problem. Unfortunately, it does not apply, because the negative log probability need not be a Schur-concave function of the *arguments* to the multivariate hypergeometric function.

In particular, consider the case $n = 2$, with parameters a, b and distribution $\text{Prob}_{a,b}(i, j) = \binom{a}{i}\binom{b}{j}/\binom{a+b}{i+j}$. The function $\phi(X, Y) = -\log \text{Prob}_{a,b}(X, Y)$ is not Schur-concave when $(a, b) = (2, 6)$, for example. We then have $\phi(0, 2) = \log(28/15)$, $\phi(1, 1) = \log(28/12)$, $\phi(2, 0) = \log(28/1)$. The value of $\phi(1, 1)$ is intermediate, so ϕ is not Schur-concave.

There are some distributions for which the entropy is known to be Schur-concave in the parameters. Marshall and Olkin discuss these on [1, p. 405–407]. Let S be the sum of independent Bernoulli random variables with probability parameters $p_1, \ldots, p_n$ chosen between 0 and 1. The multinomial distribution is specified by probability parameters $p_1, \ldots, p_t$ summing to 1, with the probability of the outcome $k_1, \ldots, k_t$ being $\binom{\sum k_i}{\bar{k}} \prod p_i^{k_i}$. Both the distribution of S and the multinomial distribution have entropy that is Schur-concave in the

parameters. Both of these results were obtained by Mateev [2] and by Shepp and Olkin [3]. Our result can be viewed as adding the bimultivariate hypergeometric distribution to this list. Our proof uses only elementary methods.

Although we do not use the reduction to the case $m = 2$, our proof does proceed by a sequence of reductions. The first is to note symmetry. We will only need to prove $H(\bar{r},\bar{c}') > H(\bar{r},\bar{c})$ when $c_j + 1 < c_{j'}$ for some j, j' and $\bar{c}'$ is obtained from $\bar{c}$ by increasing c_j by 1 and decreasing $c_{j'}$ by 1, because $H(\bar{r},\bar{c})$ is symmetric in $\bar{r}$ and $\bar{c}$. In particular, consider the interpretation of $p(A)$ using permutations. Given $\bar{r},\bar{c}$ with $\sum r_i = \sum c_j = N$, computing $p(A)$ is equivalent to counting the $N \times N$ permutation matrices meeting the block constraints. Permutation matrices that have block structure $(\bar{c},\bar{r})$ are the transpose of those with block structure $(\bar{r},\bar{c})$, so the distribution for $(\bar{c},\bar{r})$ and for $(\bar{r},\bar{c})$ will be the same.

We close the introduction by proving that our combinatorial interpretation of the probability distribution agrees with the formula. Fix N, m, n and nonnegative integer parameters $\bar{r} = r_1, \ldots, r_m$ and $\bar{c} = c_1, \ldots, c_n$ with $\sum r_i = \sum c_j = N$. For each $m \times n$ integer matrix $A = (a_{ij})$, there is some number $q(A)$ of $N \times N$ permutation matrices with a_{ij} 1's in the i,jth block, as described above. Unless the row sums and the column sums of A are $\bar{r}$ and $\bar{c}$ and the entries are all nonnegative, $q(A) = 0$. We define a probability distribution on the set of $m \times n$ integer matrices by putting $p(A) = (1/N!)q(A)$.

We need a formula for $q(A)$, the number of $N \times N$ permutation matrices meeting the constraints. Consider the jth block of columns. These c_j columns contain c_j 1's. These 1's must be assigned to row blocks, with a_{ij} of them in the ith row block. The ways to do this are counted by the multinomial coefficient $\binom{c_j}{a_{1j},\ldots,a_{mj}}$. Having done this for all j, the row constraint on the a_{ij}'s implies we have assigned 1's from r_i columns into the ith row block, but we have not assigned specific columns to specific rows. These 1's may be located in the ith block of rows in $r_i!$ ways, independently of all other choices made. Doing this for each i, we have constructed a permutation matrix, and all permutation matrices satisfying the constraints are obtained in this way. Thus $q(A) = \prod_{j=1}^{n} \binom{c_j}{a_{1j},\ldots,a_{mj}} \prod_{i=1}^{m} r_i!$, and

$$p(A) = \frac{1}{N!} q(A) = \frac{\prod_{j=1}^{n} \binom{c_j}{a_{1j},\ldots,a_{mj}}}{\binom{N}{\bar{r}}}.$$

This formula makes it clear that this distribution generalizes the hypergeometric distribution. Although this formula does not display symmetry in rows and columns, the symmetry can be exhibited by writing the formula in terms of factorials: $p(A) = \prod_{i=1}^{m} r_i \prod_{j=1}^{n} c_j!/N! \prod_{ij} a_{ij}!$, as originally claimed.

The expression in terms of multinomial coefficients is the more useful one for an inductive proof of Schur-concavity. To simplify notation for the multinomial coefficients, we write the lower parameters in vector notation, letting $\bar{a}_j$ denote

the jth column of A. Thus there are $\prod_{j=1}^{n} \binom{c_j}{a_j}$ ways to assign the 1's from each column block to row blocks, and $p(A) = \prod_{j=1}^{n} \binom{c_j}{a_j} / \binom{N}{r}$.

2. Lemmas and reduction

We begin by stating several elementary lemmas that will be useful both for reducing the theorem to a more convenient statement and for proving that statement.

Lemma 1 (Log-sum Formula). *If $S = \sum s_i$ with $s_i \geq 0$ for all i, then*

$$S \log S = \sum s_i \log s_i + \sum s_i \log(S/s_i). \quad (1)$$ □

Lemma 2 (Generalized Pascal Identity). *If $\bar{a}$ is an integer vector and $\{\bar{e}_i\}$ are the canonical unit vectors, then*

$$\binom{c}{\bar{a}} = \sum_{i=1}^{n} \binom{c-1}{\bar{a}-\bar{e}_i}. \quad (2)$$

Proof. When $c = \sum a_j$, $\binom{c}{\bar{a}}$ counts the n-dimensional lattice point walks from the origin to $\bar{a}$ by unit steps in positive coordinate directions. The last step may come from any one of the coordinate directions. □

Lemma 3 (Generalized Vandermonde Convolution). *Let $A_n(\bar{r})$ denote the collection of n tuples of m-dimensional integer vectors $A = (\bar{a}_1, \ldots, \bar{a}_n)$ that sum to $\bar{r}$. Then*

$$\sum_{A \in A_n(\bar{r})} \binom{c_1}{\bar{a}_1} \cdots \binom{c_n}{\bar{a}_n} = \binom{\sum c_j}{\bar{r}}. \quad (3)$$

Proof. If $\sum r_i = \sum c_j$, then the right side counts the m-dimensional lattice walks from the origin to $\bar{r}$. The left side partitions these according to their locations after c_1 steps, $c_1 + c_2$ steps, $\ldots, (\sum c_j)$ steps. There are $\prod_{j=1}^{n} \binom{c_j}{a_j}$ walks that pass through $\bar{a}_1, \bar{a}_1 + \bar{a}_2, \ldots, \sum \bar{a}_j$, if each c_j is the sum of the coordinates of $\bar{a}_j$. If any $\bar{a}_j$ fails to sum to c_j, then the contribution is 0. □

A useful special case is $n = 2$, where the convolution reduces to $\sum_{\bar{a}} \binom{c}{\bar{a}} \binom{d}{\bar{r}-\bar{a}} = \binom{c+d}{\bar{r}}$.

Finally, after all the rearrangements and reductions, our inequality will follow from an inequality of calculus.

Lemma 4 (Compound Interest Inequality). *For $x > 0$, $x \log(1 + 1/x)$ is strictly increasing function of x.*

Proof. The derivative of $x \log(1 + 1/x)$ is $\log(1 + 1/x) - 1/(x + 1)$. Writing $\log(1+1/x)$ as $-\log(1-1/(x+1))$ and expanding in powers of $(x+1)$ expresses the derivative as a series with all terms positive. □

In the remainder of this section, we reduce the convexity of $H(\bar{r}, \bar{c})$ to an inequality for a simpler function. First, note that we can ignore the nonnegativity

and column-sum constraints and compute $H(\bar{r},\bar{c})$ by summing $p(A)\log p(A)$ over all integer matrices A whose columns $\bar{a}_1,\dots,\bar{a}_n$ sum to $\bar{r}$, because $\binom{c_j}{\bar{a}_j} = 0$ when the other constraints are violated in column j. Also, recall that $H(\bar{r},\bar{c})$ is symmetric in $\bar{r}$ and $\bar{c}$. Hence it suffices to fix $\bar{r}$ and prove that the function $H(\bar{c})$ defined by

$$H(\bar{c}) = -\sum_{A\in A_n(\bar{r})} \frac{\prod_{j=1}^n \binom{c_j}{\bar{a}_j}}{\binom{N}{\bar{r}}} \log \frac{\prod_{j=1}^n \binom{c_j}{\bar{a}_j}}{\binom{N}{\bar{r}}}$$

is strictly Schur-concave.

The normalization by $\binom{N}{\bar{r}}$ is an inconvenience, and a further reduction is helpful. If we define $K(\bar{c}) = \sum_{A\in A_n(\bar{r})} \prod_{k=1}^n \binom{c_k}{\bar{a}_k} \log \prod_{j=1}^n \binom{c_j}{\bar{a}_j}$, then

$$H(\bar{c}) = -\frac{K(\bar{c})}{\binom{N}{\bar{r}}} + \sum_{A\in A_n(\bar{r})} \frac{\prod_{k=1}^n \binom{c_k}{\bar{a}_k}}{\binom{N}{\bar{r}}} \log \binom{N}{\bar{r}}.$$

By (3), the summation reduces to $\log\binom{N}{\bar{r}}$. Hence the only dependence on $\bar{c}$ is in $K(\bar{c})$, and it suffices to show that $K(\bar{c})$ is strictly Schur-convex in $\bar{c}$. For our final reduction, we will express $K(\bar{c})$ in terms of the function $f(c) = \sum_{\bar{a}} \binom{N-c}{\bar{r}-\bar{a}} \binom{c}{\bar{a}} \log \binom{c}{\bar{a}}$. Rewriting the log product as a sum of logs, interchanging the summations over A and j, and grouping the sum over $A_n(\bar{r})$ by the possible values of $\bar{a} = \bar{a}_j$, we obtain

$$\begin{aligned} K(\bar{c}) &= \sum_{A\in A_n(\bar{r})} \sum_j \log \binom{c_j}{\bar{a}_j} \prod_{k=1}^n \binom{c_k}{\bar{a}_k} \\ &= \sum_j \sum_{\bar{a}} \binom{c_j}{\bar{a}} \log \binom{c_j}{\bar{a}} \sum_{A\in A_{n-1}(\bar{r}-\bar{a})} \prod_{k\neq j} \binom{c_k}{\bar{a}_k}. \end{aligned}$$

In the innermost sum over $A_{n-1}(\bar{r}-\bar{a})$, the indices of the vectors in A run from 1 to n but omit j. Applying (3) to this sum, we obtain

$$K(\bar{c}) = \sum_j \sum_{\bar{a}} \binom{N-c_j}{\bar{r}-\bar{a}} \binom{c_j}{\bar{a}} \log \binom{c_j}{\bar{a}} = \sum_j f(c_j). \tag{4}$$

It is well known that a function $G(c_1,\dots,c_n) = \sum_i g(c_i)$ is Schur-convex if and only if $g(c)$ is convex. The argument used to prove this yields our final reduction of the problem:

Lemma 5. *If the first difference $f(c+1)-f(c)$ is strictly increasing for $0 \le c \le N-1$, then $H(\bar{c})$ is strictly Schur-convex and $H(\bar{r},\bar{c})$ is strictly Schur-concave.*

Proof. Given j, j' such that $c_j < c_{j'}' - 1$, it suffices to show that $K(\bar{c})$ decreases when c_j is replaced by c_j+1 and $c_{j'}'$ is replaced by $c_{j'}'-1$ to obtain $\bar{c}'$. By (4),

$$K(\bar{c}') - K(\bar{c}) = f(c_j+1) - f(c_j) - [f(c_{j'}') - f(c_{j'}'-1)],$$

which by hypothesis is negative. □

3. Proof of the theorem

By Lemma 5, we need only show that $f(c)$ has a positive second difference. Let $F(c) = f(c+1) - f(c)$, i.e.,

$$F(c) = \sum_{\overline{a}} \binom{c+1}{\overline{a}} \binom{N-c-1}{\overline{r}-\overline{a}} \log \binom{c+1}{\overline{a}} - \sum_{\overline{a}} \binom{c}{\overline{a}} \binom{N-c}{\overline{r}-\overline{a}} \log \binom{c}{\overline{a}}.$$

In order to combine these sums termwise, we want to apply (2) to $\binom{c+1}{\overline{a}}$ and to $\binom{N-c}{\overline{r}-\overline{a}}$. In the latter case, note that $\sum_{\overline{a}} \binom{c}{\overline{a}} \binom{N-c}{\overline{r}-\overline{a}} = \sum_{i=1}^{m} \sum_{\overline{a}} \binom{c}{\overline{a}} \binom{N-c-1}{\overline{r}-\overline{a}-\overline{e}_i}$. For each fixed i, the inner summation over $\overline{a}$ is over all integer vectors $\overline{a}$. Replacing $\overline{a}$ by $\overline{a} - e_i$ in the ith sum (and then interchanging the order of summation again) yields

$$F(c) = \sum_{\overline{a}} \sum_{i=1}^{m} \binom{c}{\overline{a}-\overline{e}_i} \binom{N-c-1}{\overline{r}-\overline{a}} \log \binom{c+1}{\overline{a}} - \sum_{\overline{a}} \sum_{i=1}^{m} \binom{c}{\overline{a}-\overline{e}_i} \binom{N-c-1}{\overline{r}-\overline{a}} \log \binom{c}{\overline{a}-\overline{e}_i}.$$

Factoring the constant $\binom{N-c-1}{\overline{r}-\overline{a}}$ out of the inner summation and applying (1) to $\binom{c+1}{\overline{a}}$ turns the first summation into

$$\sum_{\overline{a}} \binom{N-c-1}{\overline{r}-\overline{a}} \sum_{i=1}^{m} \binom{c}{\overline{a}-\overline{e}_i} \left(\log \binom{c}{\overline{a}-\overline{e}_i} + \log \left[\binom{c+1}{\overline{a}} \Big/ \binom{c}{\overline{a}-\overline{e}_i} \right] \right).$$

With $\overline{a} = (a_1, \ldots, a_m)$, the ratio of the multinomial coefficients inside the logarithm is $(c+1)/a_i$. Hence cancellation with the second summation reduces the formula to

$$F(c) = \sum_{\overline{a}} \sum_{i=1}^{m} \binom{c}{\overline{a}-\overline{e}_i} \binom{N-c-1}{\overline{r}-\overline{a}} \log \frac{c+1}{a_i}.$$

Now consider $F(c+1) - F(c)$, given by

$$F(c+1) - F(c) = \sum_{\overline{a}} \sum_{i=1}^{m} \binom{c+1}{\overline{a}-\overline{e}_i} \binom{N-c-2}{\overline{r}-\overline{a}} \log \frac{c+2}{a_i} - \sum_{\overline{a}} \sum_{i=1}^{m} \binom{c}{\overline{a}-\overline{e}_i} \binom{N-c-1}{\overline{r}-\overline{a}} \log \frac{c+1}{a_i}.$$

As before, we apply (2) to $\binom{c+1}{\overline{a}-\overline{e}_i}$ and to $\binom{N-c-1}{\overline{r}-\overline{a}}$ to turn this into

$$\sum_{\overline{a}} \sum_{i=1}^{m} \sum_{j=1}^{m} \binom{c}{\overline{a}-\overline{e}_i-\overline{e}_j} \binom{N-c-2}{\overline{r}-\overline{a}} \log \frac{c+2}{a_i} - \sum_{\overline{a}} \sum_{i=1}^{m} \sum_{j=1}^{m} \binom{c}{\overline{a}-\overline{e}_i} \binom{N-c-2}{\overline{r}-\overline{a}-\overline{e}_j} \log \frac{c+1}{a_i}.$$

This time, we shift indices in the summation over j in the first term by replacing $\bar{a}$ by $\bar{a}+\bar{e}_j$. We can then combine the summations to obtain

$$F(c+1)-F(c)=\sum_{\bar{a}}\sum_{i=1}^{m}\binom{c}{\bar{a}-\bar{e}_i}\sum_{j=1}^{m}\binom{N-c-2}{\bar{r}-\bar{a}-\bar{e}_j}\left[\log\frac{c+2}{a_i+\delta_{ij}}-\log\frac{c+1}{a_i}\right].$$

The factor involving the logarithms can be replaced by

$$\log\frac{c+2}{c+1}-\delta_{ij}\log\frac{a_i+1}{a_i}.$$

This yields two triple summations. Since $\log(c+2)/c+1$ is independent of i and j, we can factor it out of the first summation. The second summation, involving δ_{ij} has nonzero terms only for $i=j$. Using these facts and applying (2) and (3), we obtain

$$F(c+1)-F(c)=\binom{N}{\bar{r}}\log\frac{c+2}{c+1}-\sum_{\bar{a}}\sum_{i=1}^{m}\binom{c}{\bar{a}-\bar{e}_i}\binom{N-c-2}{\bar{r}-\bar{a}-\bar{e}_i}\log\frac{a_i+1}{a_i}.$$

Make the replacement $\binom{c}{\bar{a}-\bar{e}_i}=\binom{c+1}{\bar{a}}a_i/(c+1)$ in the summation. Multiply both sides of the equality by $c+1$, and use the compound interest inequality to increase $a_i\log\frac{a_i+1}{a_i}$ to $(c+1)\log\frac{c+2}{c+1}$ in each term of the summation. If $m>1$, there are non-zero terms in the summation for which $a_i\neq c+1$, so this replacement gives a strict inequality. The quantity $c\log\frac{c+1}{c}$ now factors out of the summation, and application of (2) and (3) again evaluates the multinomial coefficient sum as $\binom{N}{\bar{r}}$. The resulting inequality is

$$(c+1)[F(c+1)-F(c)]>\left[(c+1)\log\frac{c+2}{c+1}\right]\left[\binom{N}{\bar{r}}-\binom{N}{\bar{r}}\right]=0,$$

which completes the proof. □

Note. The most general result implying Schur-concavity that we know of is [1, Chapter 3, Theorem J.2], which asserts the Schur-concavity of a certain class of functions. However, our inequality does not seem to be implied by that theorem. Perhaps both follow from some more general result.

References

1. A. W. Marshall and I. Olkin, *Inequalities: theory of majorization and its applications*, Academic Press, New York City, 1979.
2. P. S. Mateev, *The entropy of the multinomial distribution* (Russian. English summary.), Teor. Verojatnost. i Primenen **23** (1978), 196–198.
3. L. A. Shepp and I. Olkin, *Entropy of the sum of independent Bernoulli random variables and of the multinomial distribution*, in Contributions to Probability, Academic Press, New York City, 1981, 201–206.

Institute for Defense Analyses, Princeton, New Jersey 08540

Department of Operations Research, Stanford University, Stanford, California 94305

Department of Mathematics, University of Illinois, Urbana, Illinois 61801

Current address (Alan Murray): Institute for Defense Analyses, Princeton, New Jersey 08540