ENTROPY OF THE SUM OF INDEPENDENT BERNOULLI RANDOM VARIABLES AND OF THE MULTINOMIAL DISTRIBUTION

BY

L. A. SHEPP and I. OLKIN

TECHNICAL REPORT NO. 131
JULY 1978

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION GRANT
MPS 75-09450

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ENTROPY OF THE SUM OF INDEPENDENT BERNOUlli RANDOM
VARIABLES AND OF THE MULTINOMIAL DISTRIBUTION

by

L. A. SHEPP and I. OLKIN

TECHNICAL REPORT NO. 131
JULY 1978

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MPS 75-09450

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Entropy of the Sum of Independent Bernoulli Random Variables and of the Multinomial Distribution

L. A. Shepp and I. Olkin
Bell Laboratories and Stanford University

1. Introduction.

Among other properties of the entropy, \( h = -\sum_k \pi_k \log \pi_k \), of a discrete probability distribution \( \pi_k \geq 0, \sum_k \pi_k = 1 \), it is part of the intuition of information theorists that \( h \) gives a measure of the degree of uniformness of \( \pi \), i.e., the larger is \( h \), the more uniform is \( \pi \).

In particular, if \( X_1, \ldots, X_n \) are independent Bernoulli random variables with

\[
P(X_k = 1) = p_k, \quad P(X_k = 0) = 1 - p_k = q_k, \quad 0 \leq p_k \leq 1, \quad k = 1, \ldots, n,
\]

then it is intuitive that the distribution

\[
(1.1) \quad \pi_k^n = P(X_1 + \cdots + X_n = k)
\]

should be most uniform in some sense when \( p_1 = \cdots = p_n = 1/2 \), and we in fact prove that this choice of \( p \)'s maximizes the entropy.

Similarly, in the case of the multinomial distribution,

\[
(1.2) \quad \pi_k^n = \binom{n}{k_1, \ldots, k_r} \theta_1^{k_1} \cdots \theta_r^{k_r}, \quad \sum_1^n = 1,
\]

we expect and prove the most uniform case to be the one with \( \theta_1 = \cdots = \theta_r = 1/r \).

These ideas are made more precise in the following sections.
We note that the "most uniform case" arises in connection with other inequalities for the sum of Bernoulli random variables (see e.g., Hoeffding (1956), Gleser (1975)) and for the multinomial distribution (see e.g., Rinott (1973)).

2. Entropy of sum of independent Bernoulli random variables.

We show that the entropy is a Schur concave function in $p_1, \ldots, p_n$. For relevant definitions, results, and references concerning Schur functions see Marshall and Olkin (1974), (1979).

**Theorem 1.** The entropy function $h(p_1, \ldots, p_n)$ of the sum $s_n = X_1 + \cdots + X_n$ of independent Bernoulli random variables with probabilities $p_1, \ldots, p_n$ is a Schur concave function of $p_1, \ldots, p_n$.

In particular,

\[(2.1) \quad h(p_1, \ldots, p_n) \leq h(\bar{p}, \ldots, \bar{p}) , \quad \bar{p} = \Sigma p_k / n .\]

**Proof.** Since $h(p_1, \ldots, p_n)$ is permutation symmetric, to prove Schur concavity we need to show that $(p_1 - p_2) \left( \frac{\partial h}{\partial p_1} - \frac{\partial h}{\partial p_2} \right) < 0$. Write

\[(2.2) \quad \pi_k^n = p_1 p_2 \pi_{k-2}^{n-2} + (p_1 q_2 + q_1 p_2) \pi_{k-1}^{n-2} + q_1 q_2 \pi_{k-2}^{n-2} ,\]

where $\pi_j^{n-2} = \pi_j^{n-2}(p_2', \ldots, p_n')$. Then

\[(2.3) \quad \frac{\partial \pi_k^n}{\partial p_1} - \frac{\partial \pi_k^n}{\partial p_2} = -(p_1 - p_2) [\pi_k^{n-2} - 2 \pi_{k-1}^{n-2} + \pi_{k-2}^{n-2}] ,\]

\[(2.4) \quad \frac{\partial h}{\partial p_1} - \frac{\partial h}{\partial p_2} = - \sum_k \left( 1 + \log \pi_k^n \right) (\frac{\partial \pi_k^n}{\partial p_1} - \frac{\partial \pi_k^n}{\partial p_2}) .\]
Putting (2.3) into (2.4) and changing the index of summation from \( k-1 \) to \( k \) in the second term and from \( k-2 \) to \( k \) in the third term, we obtain

\[
(p_1 - p_2) \left( \frac{\partial h}{\partial p_1} - \frac{\partial h}{\partial p_2} \right) = (p_1 - p_2)^2 \sum_k \pi^n_k \log \frac{\pi^n_k \pi^n_{k-2}}{(\pi^n_{k-1})^2},
\]

where the sum is over only those \( k \) for which the argument of the logarithm is finite and nonzero.

That (2.5) is negative follows from the fact that

\[
\pi^n_k \pi^n_{k-2} \leq (\pi^n_{k-1})^2.
\]

To show (2.6), let

\[
s^n_k = s^n_k(x_1, \ldots, x_n) = \sum_{1 \leq i_1 < \cdots < i_k \leq n} x_{i_1} \cdots x_{i_k}
\]

denote the \( k \)-th elementary symmetric function. For \( k < 0 \) or \( k > n \), set \( s^n_k = 0 \). It is well-known [see e.g., Hardy, Littlewood, and Pólya (1952) p. 52] that \( s^n_k \) is log concave in \( k \), i.e.,

\[
s^n_{k-1} s^n_{k+1} \leq (s^n_k)^2.
\]

Consequently, with \( x_i = p_i / q_i \), \( s^n_k = \pi^n_k / (q_1, \ldots, q_n) \), which yields (2.6) and completes the proof. \( \|
\]

Since a symmetric concave function is Schur concave, a stronger result than Schur concavity is that \( h(p_1, \ldots, p_n) \) is concave on the
cube $0 \leq p_k \leq 1$, $k = 1, \ldots, n$. It is also intuitive that $h(p_1, \ldots, p_n)$ is increasing in each $p_k$ on the cube $0 \leq p_k \leq 1/2$, $k = 1, \ldots, n$. On the basis of numerical calculations and verification in the special cases $n = 2, 3$, it seems likely that both the concavity and monotonicity hold. However, these assertions remain as conjectures. A partial result is given by

**Theorem 2.** $h(p_1, \ldots, p_n)$ is concave in each $0 \leq p_k \leq 1$, $k = 1, \ldots, n$.

**Proof.** Recall from the definition that $\pi_k^n$ is linear in $p_i$:

$$\pi_k^n = \Pr(X_1 + \cdots + X_n = k) = b_k + p_1 a_k,$$

where $b_k = \Pr(X_2 + \cdots + X_n = k)$, $a_k = \Pr(X_2 + \cdots + X_n = k - 1) - b_k$. Furthermore, $\Sigma b_i = 1$, $\Sigma a_i = 0$. Then

$$h(p_1, \ldots, p_n) = - \sum_{0}^{n} (b_i p_i a_i) \log(b_i + p_i a_i),$$

from which

$$\frac{\partial h}{\partial p_1} = - \sum a_i \log(b_i + p_i a_i), \quad \frac{\partial^2 h}{\partial p_1^2} = - \sum a_i^2 / (b_i + p_i a_i) < 0.$$

3. **Entropy of the Multinomial Distribution.**

For the multinomial distribution (1.2) we obtain that the entropy function $H(\theta_1, \ldots, \theta_r)$ is concave on the simplex $0 \leq \theta_k$, $k = 1, \ldots, r$, $\theta_1 + \cdots + \theta_r = 1$. Since $H(\theta_1, \ldots, \theta_r)$ is permutation symmetric, it is Schur concave, so that
\[ H(\theta_1, \ldots, \theta_r) \leq H\left(\frac{1}{r}, \ldots, \frac{1}{r}\right). \]

In particular, when \( r = 2 \), the multinomial distribution reduces to the binomial distribution, in which case the entropy function \( h_n(p) = h_n(p, 1-p) \) is concave on \( 0 \leq p \leq 1 \), and

\[ h_n(p) \leq h_n\left(\frac{1}{2}\right). \]

**Theorem 3.** Let \( \pi_n^k \) denote the multinomial distribution (1.2). Then

\[ H_n(\theta_1, \ldots, \theta_r) = -\Sigma \pi_n^k \log \pi_n^k \]

\[ = -\log \Gamma(n+1) + E \Sigma \log \Gamma(k_i+1) - h \Sigma \theta_i \log \theta_i \]

is concave on the simplex \( 0 \leq \theta_k, k = 1, \ldots, r, \theta_1 + \cdots + \theta_r = 1 \).

Before proving Theorem 3 we state a concavity result for the binomial distribution that may have some intrinsic interest. The motivation for this result stems from the representation of the entropy in (3.3).

**Theorem 4.** If \( X \) has a binomial distribution with \( EX = np \), then

\[ g(p) = -np \log p + E \log \Gamma(x+1) \]

is a concave function in \( p \).

Note that \( g(p) + g(1-p) = h_n(p) + \log \Gamma(n+1) \), so that concavity of \( h_n(p) \) follows from Theorem 4.

**Proof.** Write

\[ g(p) = -np \log p + \Sigma_{x=0}^{n} \binom{n}{x} p^x q^{n-x} \log \Gamma(x+1), \]

where \( q = 1-p \). Differentiating with respect to \( p \) and collapsing terms yields
\[ g'(p) = -n - n \log p + n \sum_{x=1}^{n-1} \binom{n-1}{x-1} p^{x-1} q^{n-x} \log \Gamma(x+1) - n \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j} \log \Gamma(j+1) \]

\[ = -n - n \log p + n \sum_{y=0}^{n-1} \binom{n-1}{y} p^y q^{n-1-y} \log(y+1). \]

A second differentiation yields

\[ g''(p) = -\frac{n}{p} + n(n-1) \sum_{y=1}^{n-2} \binom{n-2}{y-1} p^{y-1} q^{n-1-y} \log(y+1) - n(n-1) \sum_{j=0}^{n-2} \binom{n-2}{j} p^j q^{n-2-j} \log(j+1) \]

\[ = -\frac{n}{p} + n(n-1) \sum_{z=0}^{n-2} \binom{n-2}{z} p^z q^{n-2-z} \log \frac{z+2}{z+1}. \]

Since \( \log(1+u) < u \) for \( u > 0 \) and

\[ \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \frac{1}{j+1} = \frac{1-q^{m+1}}{(m+1)p}, \]

\[ g''(p) < -\frac{n}{p} + n(n-1) \frac{1-q^{n-1}}{(n-1)p} = -\frac{np^{n-1}}{p} < 0. \]

Consequently, \( g(p) \) is strictly concave. 

\textbf{Proof of Theorem 3.} First note that by a direct computation, \( H_n(\theta_1, \ldots, \theta_r) \) can be written as

\[ H_n(\theta_1, \ldots, \theta_r) = -\log \Gamma(n+1) + \sum_{l=1}^{r} [n\theta_l \log \theta_l - E \log \Gamma(K_l+1)] \]

\[ = -\log \Gamma(n+1) + \sum_{l=1}^{r} g(\theta_l), \]

where each \( K_i \) has a binomial distribution with \( EK_i = n\theta_i, \ i = 1, \ldots, r. \)

The result now follows by invoking theorem 4. 

For fixed $p$ as $n \to \infty$, use of Stirling's approximation shows that $h_n(p)$ is nearly independent of $p$ and is about $\frac{1}{2} \log n$, or $\frac{1}{2}$ the maximum possible entropy $(\frac{1}{2} \log(n+1))$ of an $(n+1)$-valued random variable, namely

$$h_n(p) = \frac{1}{2} \log n + \frac{1}{2}(1 + \log 2npq) + o(1), \quad n \to \infty.$$  

Curiously, this agrees up to $o(1)$ with the entropy $-\int p \log p \, dx$ for the normal density $p$ with the same variance as $S_n$. Figure 1 is a graph of $h_{30}(p)$, $0 \leq p \leq 1$.

Acknowledgements.

The authors are grateful to A. D. Wyner for raising the questions that led to (3.1) and (3.2), and to P. Diaconis for his suggestions in an earlier version of the paper.
References


