Economics Letters 31 (1989) 27–30
North-Holland

COMPARATIVE PROPERTIES OF SUMS OF INDEPENDENT BINOMIALS WITH DIFFERENT PARAMETERS

Raaj Kumar SAH *

Yale University, 27 Hillhouse Avenue, New Haven, CT 06520, USA

Received 18 January 1989
Accepted 27 February 1989

This paper presents some results on how certain kinds of changes in parameters induce a first-order stochastic change in the sums of independent binomials with different parameters. These results are useful for qualitative economic analysis. For example, in situations in which a decision-making unit receives observations from a variety of sources or subsamples, these results help predict the effects of changes in the composition and sizes of subsamples on the economic variables under consideration.

1. Introduction

Let x_j denote a binomial variable with parameters (N_j, p_j). Define $y = \sum_{j=1}^{J} x_j$ as the sum of J mutually independent variables. Define vectors $N \equiv (N_1, \ldots, N_J)$ and $p = (p_1, \ldots, p_J)$. Let $F(y, N, p)$ denote the distribution function of y.

The following question is of interest for qualitative economic analysis. What effects do specific kinds of changes in N have on the properties of F? Can we identify some systematic differences between $F(y, N, p)$ and $F(y, N+\epsilon, p)$ for specific types of vectors of integers $\epsilon \equiv (\epsilon_1, \ldots, \epsilon_J)$? For reasons described below, one might particularly wish to identify the kinds of ϵ's which induce a first-order stochastic change in the distribution of y. For later use, not that ϵ induces a first-order stochastic improvement (FOSI), if

$$F(y, N+\epsilon, p) - F(y, N, p) < 0, \quad \text{for} \quad R - 1 \geq y \geq 0, \tag{1}$$

where $R = \max\{\sum_j (N_j + \epsilon_j), \sum_j N_j\}$ is the upper end of the relevant support of y. Obviously, $F(R, N+\epsilon, p) = F(R, N, p) = 1$. I do not separately discuss first-order stochastic worsening, because if ϵ induces a stochastic improvement, then $-\epsilon$ induces a stochastic worsening.

An example of the economic contexts in which such results are useful is as follows. Consider an economic unit (e.g., an individual or a firm) which receives observations from several different sources, such that the number of observations from each source is an independent binomial variate with parameters which are different across sources. That is, observations come from many subsamples; the size of subsample j is N_j; and the number of observations from this subsample is a binomial variate with parameters (N_j, P_j). Now suppose an economic variable (e.g., the probability of making a particular choice, the utility of an individual, or the profit of a firm) is an increasing function of the

* I thank Jacques Cremer for comments and Susan Mattern and Jingang Zhao for research assistance.

total number of observations, y. Then, any ϵ that induces a first-order stochastic improvement in the distribution of y also increases the expected value of the economic variable under consideration. [See Fishburn and Vickson (1978) and Lippman and McCall (1986), among others, for results on stochastic dominance.] Knowing the kinds of the ϵ's that have the preceding property will allow the analyst to make qualitative predictions for the economic variable under consideration.

The main theorem and its two corollaries (which have transparent implications) are described in section 2. Brief proofs are presented in section 3. Since the results of this paper are useful primarily for economic analysis, it is not surprising that these or similar results are not available in the statistics literature [see Johnson and Kotz (1971, Ch. 3), Patil et al. (1984), and relevant references therein].

2. Results

I assume that the p_j's are different, because if two subsamples have parameters (N_j', p_j) and (N_j'', p_j) then, for our purposes, one can define an overall subsample with parameters $(N_j \equiv N_j' + N_j'', p_j)$. The convention concerning indices that $i = (1, \ldots, J)$, and $j = (1, \ldots, J)$ is used, unless stated otherwise. The indices are ordered such that $1 > p_J > p_{J-1} > \cdots > p_1 > 0$. An index set I is defined such that $j \in I$ if $\epsilon_j < 0$. I assume throughout that $\epsilon \neq 0$; that is , at least one element of ϵ is non-zero. Also $N_j \geq 0$, and $N_j + \epsilon_j \geq 0$. Recall that FOSI is a short-hand expression for eq. (1). The results are as follows.

Theorem 1. FOSI is induced if

$$\sum_{i=j}^{J} \epsilon_i \geq 0 \quad \text{for} \quad \text{all} \quad j \in I. \tag{2}$$

Corollary 1. FOSI is induced if $\epsilon_j \geq 0$ for all j (that is, if the set I is empty).

Corollary 2. FOSI is induced if

$$\sum_{i=1}^{j} \epsilon_i \leq 0 \quad \text{for} \quad \text{all} \quad \text{j}, \quad \text{and} \quad \sum_{i} \epsilon_i = 0. \tag{3}$$

These corollaries are intuitive. Corollary 1 simply says that if the size of at least one subsample increases and the size of no subsample decreases then there is a first-order stochastic improvement in the distribution of y.

To interpret Corollary 2, define $g_j \equiv (N_j + \epsilon_j)/\Sigma_i\ (N_i + \epsilon_i)$ to be the fraction of the total sample that comes from the jth subsample. If g_j is viewed as a probability density, then eq. (3) means that ϵ induces a first-order stochastic improvement in this density. Corollary 2 thus says that such a change in the composition of subsamples induces a first-order stochastic improvement in the distribution of y.

3. Proofs

I begin with two identities and a lemma. The theorem and corollaries are proven later. Note that the argument p of F is suppressed for brevity. Also, the range of y for which a relationship holds is suppressed, because the range is obvious from the context.

Define a $(1 \times J)$ vector e_j of which the jth element is one and other elements are zero. Thus

$$\epsilon = \sum_j \epsilon_j e_j. \tag{4}$$

Let $f(y, N, p)$ denote the probability density of y. That is, $F(y, N) \equiv \sum_{k=0}^{y} f(k, N, p)$, and

$$f(k, N, p) \equiv \sum \prod_j \binom{N_j}{k_j} p_j^{k_j} (1-p_j)^{N_j - k_j}, \tag{5}$$

where the outer summation in the above right-hand side is for all $k_1 \geq 0, \ldots, k_J \geq 0$, such that $\Sigma_j k_j = k$. Then

$$f(k, N + e_j, p) = p_j f(k-1, N, p) + (1-p_j) f(k, N, p). \tag{6}$$

This identity, which can be proven using eq. (5), arises because a binomial variate is the sum of independent Bernoulli variates. Identity (6) and the definition of F yield another identity:

$$F(y, N + e_j) = p_j F(y-1, N) + (1-p_j) F(y, N). \tag{7}$$

Lemma 1.

$$F(y, N + \epsilon) - F(y, N) < 0, \tag{8}$$

if there is a value of j, denoted as j_k, such that: (i) $\epsilon_j = 0$ for $j \leq j_k - 1$, (ii) $\epsilon_{j_k} < 0$, (iii) $\epsilon_j \geq 0$ for $j \geq j_k + 1$, and (iv) $\sum_{j=j_k}^{J} \epsilon_j = 0$.

Proof of Lemma 1. Expression (7) yields

$$F(y, N - e_j + e_i) - F(y, N) = (p_j - p_i) f(y, N - e_j, p) < 0 \quad \text{if} \quad i > j.$$

Using the preceding relationship repeatedly, and eq. (4), the lemma is proven.

Proof of Corollary 1. Expression (7) yields

$$F(y, N + e_j) - F(y, N) = p_j [F(y-1, N) - F(y, N)] = -p_j f(y, N, p) < 0.$$

Using the preceding relationship repeatedly, and eq. (4), the corollary is proven.

Proof of Theorem 1. Recall that $j \in I$ if $\epsilon_j < 0$. Assume I is non-empty (otherwise Theorem 1 is the same as Corollary 1 proven above). Let the elements of I be $\{j_1, \ldots, j_k\}$, where $j_1 < \cdots < j_k$.

Construct two $(1 \times J)$ vectors ϵ^k and δ^k which have the following properties:

(i) $\epsilon_j = \epsilon_j^k + \delta_j^k$ for all j,

(ii) $\delta_j^k = 0$ for $j \leq j_k - 1$,

(iii) $\delta_{j_k}^k = \epsilon_{j_k}$, and

(iv) $\delta_j^k = \min\{\epsilon_j, -\sum_{i=j_k}^{j-1} \delta_i^k\}$ for $j \geq j_k + 1$.

Then, δ^k satisfies the conditions for eq. (8). That is: (i) $\delta_j^k = 0$ for $j \leq j_k - 1$, (ii) $\delta_{j_k}^k < 0$, (iii) $\delta_j^k \geq 0$ for $j \geq j_k + 1$, (iv) $\sum_{j=j_k}^{J} \delta_j^k = 0$. From Lemma 1, therefore follows:

$$F(k, N + \epsilon) = F(k, N + \epsilon^k + \delta^k) < F(k, N + \epsilon^k). \tag{9}$$

Also, note that vector ϵ^k satisfies property (2); that is: $\sum_{i=j}^{J} \epsilon_i^k \geq 0$ for $j \in \{j_1, \ldots, j_{k-1}\}$.

Repetition of the above argument yields

$$F(y, N + \epsilon) < F(y, N + \epsilon^k) < F(y, N + \epsilon^{k-1}) < \ldots < F(y, N + \epsilon^1).$$

Further, $F(y, N + \epsilon^1) \leq F(y, N)$ from Corollary 1, because ϵ^1 is a non-negative vector. Theorem 1 follows.

Proof of Corollary 2. A special case of eq. (2) is where: (i) $\sum_i \epsilon_i = 0$, and (ii) $\sum_{i=j}^{J} \epsilon_j \geq 0$ for $j = (2, \ldots, J)$. Subtracting the preceding inequality from $\sum_i \epsilon_i = 0$, one obtains eq. (3).

References

Fishburn, P.C. and R.G. Vickson, 1978, Theoretical foundations of stochastic dominance, in: G.A. Whitmore and M.C. Findlay, eds., Stochastic dominance: An approach to decision making under risk (Heath, Lexington, MA) 37–113.

Johnson, N.L. and S. Kotz, 1971, Discrete distributions (Wiley, New York).

Lippman, S.A. and J.H. McCall, 1986, The economics of uncertainty, in: K.J. Arrow and M.D. Intriligator, eds., Handbook of mathematical economics, Vol. I (North-Holland, Amsterdam).

Patil, G.P., et al., 1984, Dictionary and classified bibliography of statistical distributions in scientific work, Vol. 1 (International Co-operative Publishing House, Fairland, MD).