AN ELEMENTARY CHARACTERIZATION OF THE MULTINOMIAL AND THE MULTIVARIATE HYPERGEOMETRIC DISTRIBUTIONS

J.Panaretos (U.S.A., Columbia)

1. Introduction.

The multinomial distribution is one of the most important multivariate distributions, especially because of its wide use in testing the fit of a particular distribution to a given set of data. For this reason, properties which are characteristic for the multinomial distribution have always attracted interest. One of the most useful characterizytions of the multinomial distribution has been given by Bol'shev [1] . He showed that the independent, non-negative, integer valued random variables $X_1, X_2, \ldots, X_n$ have non-degenerate Poisson distributions if and only if (iff) the conditional distribution of these variables for a fixed sum $K = X_1 + X_2 + \cdots + X_n$ is a non-degenerate multinomial distribution. He also provided a number of examples where this characterization can be proved useful in simplifying testing of hypotheses problems. Characterizations of the Poisson distribution along similar lines have been studied more recently by Rao and Srivastava [7] and Gerber [2,3] . Another characterization of the multinomial distribution in a slightly different set up was obtained by Janardan [4] . He extended an earlier result of Patil and Seshadri [6] by showing that two independent random vectors $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ have multinomial distributions with the same parameter vector the other parameters being respectively m and n iff the conditional distribution of $\underset{\sim}{Y}$ given the sum $\sum (Y_i + Z_i) = N$ is a multivariate hypergeometric distribution with parameters $m+n, N$.

In the present paper we obtain a characterization of the multinomial distribution without requiring $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ to be independent. The condition involved is also used to characterize the multivariate hypergeometric distribution as the conditional distribution. Both characterizations are obtained as corollaries of a general lemma concerning multivariate finite distributions.

In the sequel, we use the notation $\underset{\sim}{X} = (X_1, X_2, \ldots, X_s)$, and $\underset{\sim}{Y} = (Y_1, Y_2, \dots, Y_s)$ to denote two random vectors, where $X_i, Y_i, \quad i = 1, 2, \dots, s$ are non-negative, integer valued random variables with $\underset{\sim}{X} \geq \underset{\sim}{Y}$ and we denote the vector $\underset{\sim}{X} - \underset{\sim}{Y}$ by $\underset{\sim}{Z}$. We will say that the conditional distribution of $\underset{\sim}{Y}$ given $\underset{\sim}{X}$ is the multivariate hypergeometric if

$$P(\underset{\sim}{Y} = \underset{\sim}{r} \mid \underset{\sim}{X} = \underset{\sim}{n}) = \binom{n_0}{r_0}\binom{n_1}{r_1}\cdots\binom{n_s}{r_s} \Big/ \binom{n}{r} \tag{1}$$

$$r, n, r_i, n_i \geq 0 \quad \text{for all } i; \quad \sum_{i=1}^{s} n_i \leq n, \quad \sum_{i=1}^{s} r_i \leq r,$$

$$n_0 = n - \sum_{i=1}^{s} n_i, \qquad r_0 = r - \sum_{i=1}^{s} r_i.$$

We will also say that $\underset{\sim}{X}$ follows a multinomial distribution if

$$g_{\underset{\sim}{n}} = P(\underset{\sim}{X} = \underset{\sim}{n}) = \frac{n!}{n_0!\, n_1! \dots n_s!}\, p_0^{n_0} p_1^{n_1} \dots p_s^{n_s}, \tag{2}$$

$$0 < p_i < 1; \quad \sum_{i=1}^{s} n_i \leq n, \quad \sum_{i=1}^{s} p_i < 1; \quad n_0 = n - \sum_{i=1}^{s} n_i, \quad p_0 = 1 - \sum_{i=1}^{s} p_i.$$

Moreover, by $\{m_1, m_2, \dots, m_s\}$ we denote an integer partition of the integer r (i.e., $m_1 + m_2 + \dots + m_s = r$).
Finally, $\{h_1, h_2, \dots, h_w\}$, $w \leq s$ designates an arbitrary subset of size w of $\{1, 2, \dots, s\}$.

2. Characterization of the Multinomial and the Multivariate Hypergeometric Distributions.

Theorem 1. (Characterization of the Multinomial Distribution) Suppose that the random vectors $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ defined above are such that the conditional distribution of $\underset{\sim}{Y}$ on $\underset{\sim}{X}$ is multivariate hypergeometric as in (1). Then $\underset{\sim}{X}$ follows a multinomial distribution as in (2) iff

$$P(\underset{\sim}{Y}=\underset{\sim}{r}\mid \underset{\sim}{Z}=\underset{\sim}{0})=P(\underset{\sim}{Y}=r\mid Z_{h_1}=\ell_{h_1}m_{h_1}+1,$$

$$Z_{h_2}=\ell_{h_2}m_{h_2}+1,\ldots,Z_{h_j}=\ell_{h_j}m_{h_j}+1;\ Z_{h_{j+1}}=0,\ Z_{h_{j+2}}=0,\ldots,Z_{h_s}=0) \quad (3)$$

for all $j=1,2,\ldots,s$, where ℓ_i, $i=1,2,\ldots,s$ are non-negative integers such that $\sum_{i=1}^{s}\ell_i m_i \leq n-r-s$.

Note: One may observe that when $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ are independent, then condition (3) is automatically satisfied. It is also true that when (1) and (2) hold than $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ are independent.

Before stating our second theorem characterizing the hypergeometric distribution we introduce some further notation. We denote by

$$\{(a_{\underset{\sim}{n}},b_{\underset{\sim}{n}}):\underset{\sim}{n}=(n_1,n_2,\ldots,n_s),\ n_i=0,1,\ldots;\ i=1,2,\ldots,s\}$$

a sequence of real vectors such that $a_{\underset{\sim}{r}}>0$ for those $\underset{\sim}{r}=(r_1,r_2,\ldots,r_s)$ for which $\sum_{i=1}^{s}r_i\leq r$ with r a positive integer, $r\leq n$. Also, $b_{\underset{\sim}{n}}>0$ for all $\underset{\sim}{n}=(n_1,n_2,\ldots,n_s)$:

$$n_i=0,1,m_i+1,\ 2m_i+1,\ldots,\ell_i m_i+1$$

where m_i,ℓ_i, $i=1,2,\ldots,s$ are defined as in theorem 1.
Also the sequence $\{c_{\underset{\sim}{n}}\}$ is defined as

$$c_{\underset{\sim}{n}}=\sum_{\underset{\sim}{r}=\underset{\sim}{0}}^{\underset{\sim}{n}}a_{\underset{\sim}{r}}b_{\underset{\sim}{n}-\underset{\sim}{r}},\ \text{with}\ \sum_{\underset{\sim}{r}=\underset{\sim}{0}}^{\underset{\sim}{n}}\ \text{denothing}\ \sum_{r_1=0}^{n_1}\ \sum_{r_2=0}^{n_2}\cdots\sum_{r_s=0}^{n_s}.$$

Theorem 2. (Characterization of the Multivariate Hypergeometric Distribution).

Suppose that the random bector $(\underset{\sim}{X},\underset{\sim}{Y})$ defined in the introduction is such that $\underset{\sim}{X}$ follows a multinomial distribution as in (2). Let the conditional distribution of $\underset{\sim}{Y}$ on $\underset{\sim}{X}$ be of the structural form

$$P(\underset{\sim}{Y}=\underset{\sim}{r}\,|\,\underset{\sim}{X}=\underset{\sim}{n})=\frac{a_{\underset{\sim}{r}}\,b_{\underset{\sim}{n}-\underset{\sim}{r}}}{c_{\underset{\sim}{n}}} \qquad (4)$$

with $a_{\underset{\sim}{n}}, b_{\underset{\sim}{n}}, c_{\underset{\sim}{n}}$ as defined above. Then, condition (3) is necessary and sufficient for the conditional distribution of $\underset{\sim}{Y}\,|\,\underset{\sim}{X}=\underset{\sim}{n}$ to be multivariate hypergeometric as in (1). The proofs of theorem 1 and theorem 2 follow as a consequence of the following lemma.

<u>Lemma</u>. Consider a random vector $(\underset{\sim}{X},\underset{\sim}{Y})$ where $\underset{\sim}{X}=(X_1,X_2,\dots,X_s)$, $\underset{\sim}{Y}=(Y_1,Y_2,\dots,Y_s)$ and X_i, Y_i, $i=1,2,\dots,s$ are non-negative, integer-valued random variables. Let $P(\underset{\sim}{X}=\underset{\sim}{n})=$ $=P(X_1=n_1, X_2=n_2,\dots,X_s=n_s)=g_{\underset{\sim}{n}}$, $\underset{\sim}{n}=(n_1,\dots,n_s)$. Suppose that $g_{\underset{\sim}{n}}>0$ for those $\underset{\sim}{n}=(n_1,n_2,\dots,n_s)$ for which $\sum n_i\le n$, where n is a positive integer. Assume that $g_{\underset{\sim}{0}}<1$, $\underset{\sim}{X}\geqslant\underset{\sim}{Y}$ and let $\underset{\sim}{Z}=X-Y$.
Define $\{(a_{\underset{\sim}{n}},b_{\underset{\sim}{n}}): \underset{\sim}{n}=(n_1,n_2,\dots,n_s),\ n_i=0,1,\dots;\ i=1,2,\dots,s\}$ to be a sequence of real vectors such that $a_{\underset{\sim}{r}}>0$ for those $\underset{\sim}{r}$ for which $\sum r_i\le r$ where r is a positive integer, $r\le n$. Also $b_{\underset{\sim}{n}}>0$ for all $\underset{\sim}{n}: n_i=0, 1, m_i+1, 2m_i+1, \dots, \ell_i m_i+1$, $i=1,2,\dots,s$ where $\{m_1, m_2,\dots,m_s\}$ is a set on non-negative integers such that $m_1+m_2+\dots+m_s=r$ (i.e.,an integer partition of r) and ℓ_i, $i=1,2,\dots,s$ are non-negative integers such that $\sum_{i=1}^{s}\ell_i m_i\le n-r-s$.

Define the sequence $\{c_{\underset{\sim}{n}}\}$ as $c_{\underset{\sim}{n}}=\sum_{\underset{\sim}{r}=\underset{\sim}{0}}^{\underset{\sim}{n}} a_{\underset{\sim}{r}}\,b_{\underset{\sim}{n}-\underset{\sim}{r}}$ with $\sum_{\underset{\sim}{r}=\underset{\sim}{0}}^{\underset{\sim}{n}}$ denoting $\sum_{r_1=0}^{n_1}\sum_{r_2=0}^{n_2}\cdots\sum_{r_s=0}^{n_s}$.
Suppose further that, whenever $g_{\underset{\sim}{n}}>0$

$$P(\underset{\sim}{Y}=\underset{\sim}{r}\,|\,\underset{\sim}{X}=\underset{\sim}{n})=\frac{a_{\underset{\sim}{r}}\,b_{\underset{\sim}{n}-\underset{\sim}{r}}}{c_{\underset{\sim}{n}}},\ r_i=0,1,\dots;n_i=0,1,\dots; \qquad (5)$$

$$\sum r_i\le r,\quad \sum n_i\le n.$$

Let $\{h_1, h_2, \ldots, h_w\}$, $w \le s$ denote any subset of size w of $\{1, 2, \ldots, s\}$. Then, the conditions

$$P(\underset{\sim}{Y}=\underset{\sim}{r} \mid \underset{\sim}{Z}=\underset{\sim}{0}) = P(\underset{\sim}{Y}=\underset{\sim}{r} \mid Z_{h_1}=\ell_{h_1} m_{h_1}+1, \tag{6}$$

$$Z_{h_2}=\ell_{h_2} m_{h_2}+1, \ldots, Z_{h_j}=\ell_{h_j} m_{h_j}+1,\ Z_{h_{j+1}}=0,\ Z_{h_{j+2}}=0, \ldots, Z_{h_s}=0)$$

for all $j=1,2,\ldots,s$, and all ℓ_i such that $\sum_{i=1}^{s} \ell_i m_i \le n-r-s$ hold iff

$$\frac{g_{\underset{\sim}{n}}}{c_{\underset{\sim}{n}}} = \frac{g_{\underset{\sim}{0}}}{c_{\underset{\sim}{0}}} \prod_{i=1}^{s} \theta_i^{n_i} \tag{7}$$

for some $\theta_i > 0$, $i=1,2,\ldots,s$ and for all $\underset{\sim}{n}: \sum n_i \le n$.

(The same problem in the univariate case has been studied in Panaretos (1981)).

Proof. (An outline of the proof is given for $s=2$. The general case $(s>2)$ is then straightforward). The "if" part follows directly from the fact that when (5) and (7) hold then $P(\underset{\sim}{Y}=\underset{\sim}{r} \mid \underset{\sim}{Z}=\underset{\sim}{j})$ is independent of $\underset{\sim}{j}$. To prove the "only if" part we first observe that form

$$P(\underset{\sim}{Y}=\underset{\sim}{r} \mid Z_1=\ell_1 m_1+1, Z_2=0) \quad \text{we have}$$

$$\frac{g_{r_1+\ell_1 m_1+1,\, r_2}}{c_{r_1+\ell_1 m_1+1,\, r_2}} = \frac{g_{0,r_2}}{c_{0,r_2}} \theta_1^{r_1+\ell_1 m_1+1}$$

for all non-negative integers r_1, r_2, ℓ_1 such that $r_1+r_2 \le r$ and $\ell_1 m_1 \le n-r-1$ or, equivalently

$$\frac{g_{n_1,r_2}}{c_{n_1,r_2}} = \frac{g_{0,r_2}}{c_{0,r_2}} \theta_1^{n_1}, \quad n_1 \le n,\ r_2 \le r,\ n_1+r_2 \le n. \tag{8}$$

Similarly, from

$P(\underset{\sim}{Y}=\underset{\sim}{r}\mid \underset{\sim}{Z}=\underset{\sim}{0})=P(\underset{\sim}{Y}=\underset{\sim}{r}\mid Z_1=0, Z_2=\ell_2 m_2+1)$ we get

$$\frac{g_{r_1, r_2+\ell_2 m_2+1}}{c_{r_1, r_2+\ell_2 m_2+1}} = \frac{g_{r_1,0}}{c_{r_1,0}}\theta_2^{r_2+\ell_2 m_2+1}$$

for all r_1, r_2, ℓ_2 such that $r_1+r_2 \le r$, $\ell_2 m_2 \le n-r-1$

which is equivalent to

$$\frac{g_{r_1,n_2}}{c_{r_1,n_2}} = \frac{g_{r_1,0}}{c_{r_1,0}}\theta_2^{n_2} \qquad (9)$$

for all $r_1, n_2 : r_1 \le r, \ n_2 \le n_1, \ r_1+n_2 \le n$.

Finally, form

$$P(\underset{\sim}{Y}=\underset{\sim}{r}\mid Z_1=\ell_1 m_1+1, Z_2=\ell_2 m_2+1)=P(\underset{\sim}{Y}=\underset{\sim}{r}\mid Z_1=\ell_1 m_1+1, Z_2=0),$$

for fixed ℓ_1 and for all $\ell_2 : \ell_1 m_1+\ell_2 m_2 \le n-r-2$.

we have

$$\frac{g_{r_1+\ell_1 m_1+1,\, r_2+\ell_2 m_2+1}}{c_{r_1+\ell_1 m_1+1,\, r_2+\ell_2 m_2+1}} = \frac{g_{r_1+\ell_1 m_1+1,\,0}}{c_{r_1+\ell_1 m_1+1,\,0}}\theta_2^{r_2+\ell_2 m_2+1}$$

for all $r_1, r_2, \ell_1, \ell_2 : r_1+r_2 \le r, \ \ell_1 m_1+\ell_2 m_2 \le n-r-2$

(since ℓ_1 was fixed but arbitrary).

This implies that

$$\frac{g_{n_1,n_2}}{c_{n_1,n_2}} = \frac{g_{n_1,0}}{c_{n_1,0}}\theta_2^{n_2} \qquad (10)$$

for all $n_1, n_2 : n_1 + n_2 \le n$ and for some $\theta_2 > 0$.

Combining (8) and (10) we get

$$\frac{g_{n_1,n_2}}{c_{n_1,n_2}} = \frac{g_{0,0}}{c_{0,0}} \theta_1^{n_1} \theta_2^{n_2}$$ for some

$\theta_1, \theta_2 > 0$ and for all n_1, n_2 such that $n_1 + n_2 \le n$.

This completes the proof of the lemma.

Note 1. The proof of the "if" part of the lemma indicates that when (5) and (7) hold then $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ are independent.

Using the lemma we have just proved we can now prove theorems 1 and 2.

Proof of theorem 1. Consider the following sequences

$$a_{\underset{\sim}{r}} = \begin{cases} \dfrac{r!}{r_0!\, r_1! \dots r_s!} \; ; & r_i \ge 0, \quad \sum_{i=1}^{s} r_i \le r, \quad r_0 = r - \sum_{i=1}^{s} r_i \\ 0 \quad , & \text{otherwise} \end{cases} \tag{11}$$

and

$$b_{\underset{\sim}{n}} = \frac{n!}{n_0!\, n_1! \dots n_s!}, \quad n_i \ge 0, \quad i = 1, 2, \dots, s; \qquad \sum_{i=1}^{s} n_i \le n, \quad n_0 = n - \sum_{i=1}^{s} n_i . \tag{12}$$

Then, we can easily see that

$$c_{\underset{\sim}{n}} = \frac{n!}{n_0!\, n_1! \dots n_s!}, \quad n_i \ge 0, \; i = 1, 2, \dots, s; \qquad n_0 = n - \sum_{i=1}^{s} n_i . \tag{13}$$

Clearly $a_{\underset{\sim}{r}}, b_{\underset{\sim}{n}}, c_{\underset{\sim}{n}}$ from (11), (12), (13) satisfy all the conditions set by the lemma and they can be used to express (2) in the form $a_{\underset{\sim}{r}}\, b_{\underset{\sim}{n}-\underset{\sim}{r}} / c_{\underset{\sim}{n}}$. So, using the lemma we find that (3) holds iff

$$g_{\underset{\sim}{n}} = g_0 \frac{n!}{n_0!\, n_1! \ldots n_s!} \theta_1^{n_1} \theta_2^{n_2} \ldots \theta_s^{n_s}$$

for some $\theta_1, \theta_2, \ldots, \theta_s > 0$.

Furthermore, we have that $\sum_{\underset{\sim}{n}} g_{\underset{\sim}{n}} = 1$. Hence, for some $\theta_0 > 0$ $(g_{\underset{\sim}{0}})^{-1} = \left(\sum_{i=0}^{s} \theta_i\right)^n / \theta_0^{n_0}$. This completes the proof of theorem 1 if we take $p_j = \theta_j / \left(\sum_{i=0}^{s} \theta_i\right)$, $j = 0, 1, \ldots, s$.

<u>Proof of theorem 2</u>. From the lemma we have that (3) holds if

$$c_{\underset{\sim}{n}} = c_{\underset{\sim}{0}} \frac{g_{\underset{\sim}{n}}}{g_{\underset{\sim}{0}}} \theta_1^{-n_1} \theta_2^{-n_2} \ldots \theta_s^{-n_s}, \quad \theta_i > 0, \quad i = 1, 2, \ldots, s.$$

But $g_{\underset{\sim}{n}}$ was assumed to be multinomial as in (2). Consequently,

$$c_{\underset{\sim}{n}} = c_{\underset{\sim}{0}} \frac{n!}{n_0!\, n_1! \ldots n_s!} \left(\frac{p_1}{p_0 \theta_1}\right)^{n_1} \ldots \left(\frac{p_s}{p_0 \theta_s}\right)^{n_s}.$$

This implies that $c_{\underset{\sim}{n}}$ is proportional to a multinomial distribution. Moreover, from the note following the lemma we have that when (1) and (2) hold, then $\underset{\sim}{Y}$ and $\underset{\sim}{X} - \underset{\sim}{Y}$ are independent. Consequently a_r and $b_{\underset{\sim}{n}-\underset{\sim}{r}}$ are independent. Shanbhag and Basawa (1974) have shown that the multinomial distribution is uniquely decomposable into two multinomials i.e., if the convolution of two s-dimensional non-negative independent random vectors is multinomial then each of the components is multinomial with the same set of probabilities as their convolution. Hence $a_{\underset{\sim}{r}}$ and $b_{\underset{\sim}{n}}$ are also proportional to multi-

nomial distributions with probabilities $p_i / p_0 \theta_i$, $i = 1, 2, \ldots, s$.
A direct substitution in (4) completes the proof of theorem 2.

Note 2. Extensions of theorems 1 and 2 providing characterizations for the multiple multinomial and the multiple multivariate hypergeometric distributions are easily obtainable.

Note 3. It would be interesting to investigate whether theorems 1 and 2 can be used in problems of testing hypotheses in ways similar to those suggested by Bol'shev [1] .

REFERENCES

1. Bol'shev L.N. On a Characterization of the Poisson Distribution and its Statistical Applications. - Theory Probab.App., 1965. v.10, No 2, p.446-456.

2. Gerber H.U. A Characteristic Property of the Poisson Distribution.- Amer. Statist, 1979, v.33, p.85-86.

3. Gerber H.U. A Characterization of Certain Families of Distributions via Esscher Transforms and Independence.-J.Amer. Statist. Assoc., 1980, v.75, p.1015-1018.

4. Janardan K.G. A Characterization of Multinomial and Negative Multinomial Distributions. - Scand. Actuar. J., 1974, v.57, p.58-62.

5. Panaretos J. On the Joint Distribution of Two Discrete Random Variables. - Annals of the Institute of Statistical Mathematics, A (Theory and Methods), 1981, v.33(2), p.191-198.

6. Patil G.P., Seshadri V. Characterization Theorems for Some Univariate Probability Distributions. - J.Roy. Statist. Soc. Ser.B, 1964, v.26, p.286-292.

7. Rao C.R., Srivastava R.C. Some Characterizations Based on Multivariate Splitting Model. - Sankhyā A, 1979, v.41, p.124-128.

8. Shanbhag D.N., Basawa I.V. On a characterization property of the Multinomial Distribution. - Trabajos Estadist. Investigation Oper, 1974, v.25, p.109-112.