PROPERTIES OF THE EXTENDED HYPERGEOMETRIC DISTRIBUTION

By W. L. Harkness

The Pennsylvania State University

1. Introduction. Let X and Y be independent binomial random variables with parameters (n_1, p_1) and (n_2, p_2), $0 < p_i < 1$, $i = 1, 2$. A statistical problem of great practical significance is that of testing the equality of the two proportions p_1 and p_2, that is, the hypothesis $H_0 : p_1 = p_2$. For alternative hypotheses, one usually considers $p_1 \neq p_2$, $p_1 < p_2$, or $p_1 > p_2$. In any case, the usual test of the null hypothesis is a conditional test, based on the tails of the conditional distribution of X for fixed values r of $X + Y$. The probability density of X, conditional on the fixed sum $X + Y = r \varepsilon \{0, 1, \cdots, n_1 + n_2 = n\}$, is given by the "extended hypergeometric" function

$$f(x; t) = g(x)t^x/P(t), \qquad a \leqq x \leqq b, \tag{1}$$

where $a = \max(0, r - n_2)$, $b = \min(n_1, r)$, $t = p_1q_2/p_2q_1$, $q_i = 1 - p_i$, $g(x) = \binom{n_1}{x}\binom{n_2}{r-x}/\binom{n}{r}$, and $P(t) = \sum_a^b g(y)t^y$ is the factorial generating function of the ordinary hypergeometric distribution. If $p_1 = p_2$, then $t = 1$, and $P(1) = 1$, so that $f(x; t)$ reduces to $g(x)$. More generally, we observe that the density $f(x; \mathrm{t})$ is of fundamental importance in considerations of power functions for tests of independence in 2×2 contingency tables ([1], [5], [6], [9]). The parameter t is often interpreted as a measure of dependence or association in such contingency tables; $t = 1$ indicates independence and $t < 1$ and $t > 1$ correspond to positive and negative dependence respectively. As pointed out by Lehmann ([7], p. 145), t is equivalent to Yule's measure of association given by $Q = (1 - t)/(1 + t)$. Goodman and Kruskal [3] have discussed these and other measures of association.

In Sections 2 and 3, we discuss moments, moment inequalities, and maximum likelihood estimation of t, all for finite samples. In Section 4, we obtain approximations for the density $f(x; t)$, taking full advantage of the corresponding results for the particular case when $t = 1$, as given for example in [2] and considered by Van Eeden [12]. In the last two sections we discuss the asymptotic distribution of the maximum likelihood estimator and construct confidence intervals for t.

2. Moments. In terms of the hypergeometric series

$$F(\alpha; \beta; \gamma; t) = \sum_{j=0}^{\infty} (\alpha)_j(\beta)_j t^j/j!(c)_j,$$

where, for example, $(\alpha)_j = \prod_{s=1}^{j} (\alpha - s + 1)$, it is easily seen that

$$\sum_{j=0}^{b} \binom{n_1}{j}\binom{n_2}{r-j}t^j = \binom{n_2}{r}F(\alpha; \beta; \gamma; t)$$

Received 6 April 1964; revised 23 December 1964.