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 $f(x; t)$ 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.

with $\alpha = -n_1$, $\beta = -r$, and $\gamma = n_2 - r + 1$, where we assume here that $n_1 + r \leqq n$. It readily follows that the kth factorial moment $\mu'_{[k]}$ of the hypergeometric distribution is given by $\mu'_{[k]} = (\alpha)_k(\beta)_k t^k F(k)/(\gamma)_k F(0)$, where, for brevity, $F(k) \equiv F(\alpha + k, \beta + k; \gamma + k; t)$, for $k = 0, 1, \cdots$. Using the recurrence relation ([11], p. 31)

$$(2)\quad t(1-t)(\alpha+1)(\beta+1)F_2 + (\gamma+1)[\gamma - (\alpha+\beta+1)t]F_1 = \gamma(\gamma+1)F_0,$$

we find that the first and second ordinary moments μ_1' and μ_2' about the origin are related by $2(1-t)\mu_2' = n_1 r t - c_n \mu_1'$, with $c_n = n - (n_1 + r)(1-t)$, so that the variance σ^2 may be expressed in terms of the mean μ_1' as $(1-t)\sigma^2 = n_1 r t - c_n \mu_1' - (1-t)(\mu_1')^2$. Alternatively,

$$(3)\quad \mu_1' = \{-c_n + [c_n^2 + 4t(1-t)n_1 r - 4(1-t)^2\sigma^2]^{\frac{1}{2}}\}\{2(1-t)\}^{-1}.$$

Further, it is easily shown that the derivative of μ_1', as a function of t, is equal to σ^2/t, which shows that the mean is an increasing function of t. Recurrence relations for higher moments may be obtained similarly, but appear to be of little value. The calculation of μ_1' requires the evaluation of the ratio of two polynomials in t of degree $b = \min(n_1, r)$ and is tedious. We proceed therefore to obtain upper and lower bounds on μ_1'.

Denoting by $\mu_1'(-1)$ the mean value when the parameters n_1, r, and n are each reduced by one and using (2), we find that

$$(4)\quad \mu_1' = n_1 r t\{(1-t)\mu_1'(-1) + c_n + (1-t)\}^{-1}$$

$$(5)\quad \mu_2' = \mu_1'\{\mu_1'(-1) + 1\}$$

$$(6)\quad \sigma^2 = \mu_1'\{\mu_1'(-1) - \mu_1' + 1\}.$$

Since $\sigma^2 \geqq 0$, it follows from (3) and (4) that, for $0 < t < 1$,

$$(7)\quad \mu_1' \leqq \lambda_n n_1 r/n, \qquad \mu_1'(-1) \leqq \lambda_{n-1}(n_1 - 1)(r-1)/n - 1$$

where λ_n is the unique positive real number satisfying

$$(8)\quad \lambda_n\{1 - (n_1/n) - (r/n) + \lambda_n(n_1 r/n^2)\} = t\{1 - \lambda_n(n_1/n)\}\{1 - \lambda_n(r/n)\},$$

and λ_{n-1} is defined as λ_n with n_1, r, and n replaced by $n_1 - 1$, $r - 1$, and $n - 1$. Upper and lower bounds $d_U(t)$ and $d_L(t)$ on μ_1', for $0 < t < 1$, can now be obtained using (4) in conjunction with (7). It is found that

$$(9)\quad d_L(t) = n_1 r t\{[(1-t)(n_1 - 1)(r-1)\lambda_{n-1}/n - 1] + [c_n + 1 - t]\}^{-1} \leqq \mu_1' \leqq \lambda_n n_1 r/n.$$

Finally, for $t > 1$, it can be shown that

$$(10)\quad \mu_1' = E(X \mid n_1, r, n; t) = n_1 - E(X \mid n_1, n - r, n; t^{-1}),$$

TABLE I

x	0	1	2	3	4	5	6
$f(x; \frac{1}{2})$	.00677	.08120	.29914	.37219	.20936	.04785	.00349

TABLE II

x	<3	3	4	5	6	7	8	9	10	11	12
$f(x; 6)$	0	.00001	.00016	.00209	.01625	.07521	.20511	.31906	.26507	.10327	.01377

from which upper and lower bounds for μ_1' may be obtained using the bounds for a mean value when $t < 1$.

EXAMPLE 1. Let $n_1 = 6$, $r = 12$, $n = 20$, $t = .5$. The probability distribution of X is given in Table I. We find that $\mu_1' = 2.854$, $\sigma^2 = 1.072$, and the bounds for μ_1' [using (8)] are $2.850 \leqq \mu_1' \leqq 2.892$.

EXAMPLE 2. Let $n_1 = 12$, $r = 15$, $n = 30$, $t = 6$. In Table II we give the probability distribution of X. Here, $\mu_1' = 9.09946$, $\sigma^2 = 1.48202$, and the bounds for μ_1', obtained by first finding bounds on $E(X \mid 12, 15, 30, \frac{1}{6})$ using (9) and then noting (10), are $9.000 \leqq \mu_1' \leqq 9.121$.

3. Estimation of t. Let $X_1, \cdots, X_N$ be a random sample of size N from a population having a probability density $f(x; t)$ of the form given in (1). We observe that a random variable X having this density has a generalized power series distribution, as defined by Patil [8] and that the parameter t does not have a minimum variance unbiased estimator (see [8], p. 1052), for any finite sample size N. We consider the estimation of t by the method of maximum likelihood.

On setting the first derivative of the likelihood function equal to zero, we find that the maximum likelihood estimator $\hat{t}$ of t must satisfy the equation $\mu_1'(\hat{t}) = \bar{X}$, where $\mu_1'(\hat{t}) = E(X; t)\,|_{t=\hat{t}}$ and $\bar{X} = N^{-1}\sum_{i=1}^{N} X_i$. If $\bar{X} = 0$, then $\hat{t} = 0$, while if $\bar{X} = \min(n_1, r)$, $\hat{t} = +\infty$. We note also that if $\bar{X} = n_1 r/n$, then $\hat{t} = 1$. For all other values of $\bar{X}$, the computation of $\hat{t}$ is tedious, so we consider finding the upper and lower bounds on $\hat{t}$. This is accomplished by using the upper and lower bounds derived for μ_1' as given by (9) together with (10).

If $\bar{X} \leqq n_1 r/n$, then $\hat{t} \leqq 1$, so that

$$d_L(\hat{t}) \leqq \mu_1'(\hat{t}) = \bar{X} \leqq d_U(\hat{t}) \tag{11}$$

where $d_L(\hat{t})$ and $d_U(\hat{t})$ are obtained from $d_L(t)$ and $d_U(t)$ by replacing t by $\hat{t}$ wherever t appears in these expressions. Inverting the inequalities in (11), we find that, for $\bar{X} \leqq n_1 r/n$,

$$\check{t} \leqq \hat{t} \leqq \hat{t} + \bar{X}(n_1 r - n\bar{X})/n_1 r(n_1 - \bar{X})(r - \bar{X}) \tag{12}$$

where $\check{t} = \bar{X}(n - n_1 - r + \bar{X})/(n_1 - \bar{X})(r - \bar{X})$. If $\bar{X} \geqq n_1 r/n$, we find that

$$\hat{t}[1 + \{(n_1 - \bar{X})(n - \bar{X} - n_1 r)/n_1(n - r)(n_1 - \bar{X})(r - \bar{X})\}]^{-1} \leqq \hat{t} \leqq \check{t}. \tag{13}$$

EXAMPLE 3. Let $n_1 = 12$, $r = 15$, $n = 30$. In Table III we give the interval in which $\hat{t}$ lies for various values of $\bar{X}$. If $\bar{X} = 6$, $\hat{t} = 1$, while for $\bar{X} = 0$, $t = 0$

TABLE III

$\bar{X}$	3	9	10	11
Interval	(.167, .181)	(5.54, 6.00)	(11.4, 13.0)	(31.86, 38.5)

and for $\bar{X} = 12$, $\hat{t} = +\infty$. Note that the bounds for $\hat{t}$ do not depend on the sample size N, so that all the results hold for the special case $N = 1$.

4. Limiting distributions. It is well-known that the ordinary hypergeometric distribution is asymptotically binomial, Poisson, or normal, depending upon the growth to infinity of n_1, r, and n. In this section, we shall derive binomial and Poisson approximations to the density $f(x; t)$—a normal approximation has been given previously in [4], which we shall state below for completeness.

THEOREM 1. (Binomial Approximation). *Let n_1, $n \to +\infty$ in such a way that $n_1/n \to p$, $0 < p < 1$. Then for each fixed (but arbitrary) value of $X \in \{0, 1, \cdots, r\}$ and $r \in \{0, 1, \cdots\}$,*

$$\lim f(x; t) = \binom{r}{x}\theta^x(1-\theta)^{r-x}, \qquad \theta = pt/(q+pt), \qquad q = 1-p, \quad t > 0. \tag{14}$$

PROOF. This result follows easily from inequalities ([2], p. 57) on the ordinary hypergeometric distribution.

THEOREM 2. (Poisson Approximation). *Let n_1, r, and $n \to \infty$ in such a way that $n_1 r/n \to \mu$. Then for each fixed value of $X \in \{0, 1, 2, \cdots\}$,*

$$\lim f(x; t) = e^{-\mu t}(\mu t)^x/x!, \qquad t > 0. \tag{15}$$

PROOF. We consider first the case when $t \in (0, 1]$. According to the Poisson limit theorem for the hypergeometric distribution, $\lim g(x) = e^{-\mu}\mu^x/x!$ Using Feller's continuity theorem ([2], p. 262) we see also that $\lim P(t) = \exp\{\mu(t-1)\}$, so that (15) follows, for $\theta < t \leqq 1$.

Suppose next that $t > 1$. Since $P(t)$ is the factorial generating function of the hypergeometric distribution, we have

$$P(t) = \sum_{k=0}^{b} \mu'_{[k]}(t-1)^k/k!, \quad \text{where} \quad \mu'_{[k]} = P^{(k)}(1) = (n_1)_k(r)_k/(n)_k$$

is the kth factorial moment, and $b = \min(n_1, r)$. From the inequality

$$\mu'_{[k+1]} = (n_1 - k)(r-k)\mu'_{[k]}/(n-k) \leqq n_1 r \mu'_{[k]}/n \leqq (n_1 r/n)^{k+1},$$

it follows that

$$\lim P(t) \leqq \lim \sum_{k=0}^{\infty} (n_1 r/n)^k (t-1)^k/k! = \lim \exp\{n_1 r(t-1)/n\} = \exp\{\mu(t-1)\}. \tag{16}$$

Now let $\epsilon > 0$ be arbitrary and let M be an integer such that $\sum_{k=0}^{M} \{\mu(t-1)\}^k/k! > e^{\mu(t-1)} - \epsilon$. Then we have that

$$\lim P(t) \geqq \lim \sum_{k=0}^{M} \mu'_{[k]}(t-1)^k/k! = \sum_{k=0}^{M} \{\mu(t-1)\}^k/k! > e^{\mu(t-1)} - \epsilon, \tag{17}$$

since $\lim \mu'_{[k]} = \mu^k$ for $k = 0, 1, \cdots, M$, for any fixed integer M. Therefore,

since $\epsilon > 0$ is arbitrary in (17), in view of (16) we see that $\lim f(x; t) = e^{-\mu}(\mu t)^x e^{-\mu(t-1)}/x! = e^{-\mu t}(\mu t)^x/x!$.

As for the normal approximation to (1), let P_i, $0 < P_i < 1$, $i = 1, 2$, be the unique numbers satisfying

$$P_1Q_2 = tP_2Q_1, \qquad n_1P_1 + n_2P_2 = r. \tag{18}$$

Then the normal approximation given in [4] (modified slightly) is given by.

THEOREM 3. *Let* $H_1 = (n_iP_iQ_i)^{-1/2}$, $H^2 = H_1^2 + H_2^2$, *and* $X_k = H(k - n_1P_1)$ *Then* $f(x; t) \sim H\phi(X_k)$ *as* H, $HX_k^3 \to 0$, *and* $\sum_{k=\alpha}^{\beta} f(k; t) \sim \Phi(X_{\beta+\frac{1}{2}}) - \Phi(X_{\alpha-\frac{1}{2}})$ *as* H, HX_α^3, $HX_\beta^3 \to 0$, *where* $\phi(x) = (2\pi)^{-\frac{1}{2}} \exp\{-x^2/2\}$, $\Phi(x) = \int_\infty^x \phi(y)\, dy$, *and "*$a \sim b$*" means the ratio of* a *and* b *tends to one.*

It is relatively easy to show that the moments of the extended hypergeometric distribution converge to the corresponding moments of the binomial or Poisson distributions, depending on the mode of convergence of the parameters. For the normal case, it is more difficult to prove such convergence. We now show that the mean and variance, when divided by n, converge to the mean and variance of the appropriately normalized limiting normal distribution. In the following two theorems, we let n_1, r, and $n \to +\infty$ such that $n_1/n \to \eta_1$, $r/n \to \eta_2$.

THEOREM 4. $\lim \mu_1'/n = \lambda\eta_1\eta_2$, *where* λ *is the unique positive real number satisfying the equation*

$$\lambda(1 - \eta_1 - \eta_2 + \lambda\eta_1\eta_2) = t(1 - \lambda\eta_1)(1 - \lambda\eta_2). \tag{19}$$

PROOF. It is easily verified that $\lambda_n \to \lambda$, where λ_n is defined by (8). For $0 < t < 1$, from (9) it follows that $\lim \mu_1'/n \leqq \lambda\eta_1\eta_2$. On the other hand, using the lower bound for μ_1' as given in (9), we see that

$$\lim \mu_1'/n \geqq \eta_1\eta_2 t\{(1 - t)\lambda\eta_1\eta_2 + 1 - (\eta_1 + \eta_2)(1 - t)\}^{-1},$$

which is easily shown to be equal to $\lambda\eta_1\eta_2$, using (19). For $t > 1$, we have

$$\lim \mu_1'/n = \lim \{n_1/n - E(X \mid n_1, n - r, n; t^{-1})/n\}$$
$$= \eta_1 - \nu\eta_1(1 - \eta_2) = \eta_1\{1 - \nu(1 - \eta_2)\},$$

where ν satisfies the equation $\{1 - \nu\eta_1\}\{1 - \nu(1 - \eta_2)\} = \nu t\{1 - \eta_1 - (1 - \eta_2) + \nu\eta_1(1 - \eta_2)\}$. However, from (19) and this last equation we find that $1 - \lambda\eta_2 = \nu(1 - \eta_2)$, from which the conclusion follows.

THEOREM 5. Lim $\sigma^2/n = (\sum_{i=1}^4 \pi_i^{-1})^{-1} = t\eta_1\eta_2(d\lambda/dt)$, *where* $\pi_1 = \lambda\eta_1\eta_2$, $\pi_2 = \eta_1 - \pi_1$, $\pi_3 = \eta_2 - \pi_1$, *and* $\pi_4 = 1 - \pi_1 - \pi_2 - \pi_3$, *and* λ *satisfies* (19).

PROOF. Since $\sigma^2 = \mu_1'\{\mu_1'(-1) - \mu_1' + 1\}$ and $\mu_1'/n \to \lambda\eta_1\eta_2$, we must show that $\mu_1'(-1) - \mu_1' + 1 \to \lambda^{-1}t(d\lambda/dt) = \lambda\eta_1\eta_2(\sum_{i=1}^4 \pi_i^{-1})^{-1}$. The validity of this last limit is established by using the various bounds for μ_1' and $\mu_1'(-1)$ to obtain upper and lower bounds for $\mu_1'(-1) - \mu_1' + 1$ and then passing to the limit. We omit the details.

Alternatively, we observe that μ_1' is an infinitely differentiable function of t. Simple calculations show that $\sigma^2 = t(d\mu_1'/dt) > 0$, so that μ_1' is a strictly in-

creasing function of t. Theorem 5 can then be obtained by showing that

$$\lim \sigma^2/n = t \lim (d\mu_1'/dt) = t\, d\{\lim \mu_1'/n\}/dt = \eta_1\eta_2 t(d\lambda/dt).$$

Theorems 4 and 5 suggest that the exact mean μ_1' and variance σ^2, for moderate values of n_1, r, and n, may be approximated by $\mu_1' \simeq \lambda_n n_1 r/n$ and $\sigma^2 \simeq n\{\sum_{i=1}^{4} (1/\pi_i^*)\}^{-1}$, where π_i^* is π_i with n_1/n and r/n replacing η_1 and η_2 respectively. It is not hard to show that $\lambda_n n_1 r/n = n_1 P_1$ and that the approximate variance is also equal to $H^{-2} = \{(n_1 P_1 Q_1)^{-1} + (n_2 P_2 Q_2)^{-1}\}^{-1}$, where P_1 and P_2 satisfy (18). If $n_1 = 12$, $r = 15$, $n = 30$, and $t = 6$, we find that the approximation to $\mu_1' = 9.09946$ is 9.00, while $\sigma^2 = 1.48202$ is approximated by $1.44 = H^{-2}$.

5. Asymptotic properties of $\hat{t}$ and $\tilde{t}$. In considering the asymptotic properties of $\hat{t}$ there are several cases which can be examined, depending on the growth to infinity of n_1, r, n and N. However, some of these cases present no real problems. For example, if $n_1/n \to p$, with r and N fixed, it is easily seen that the asymptotic distribution of $\hat{t}$ is the same as the distribution of $\bar{X}(1-p)/p(r-\bar{X})$, where $N\bar{X}$ has a binomial distribution with parameters Nr and $\theta = pt/(q+pt)$. (See Section 4, Theorem 1.) Similarly, if $n_1 r/n \to \mu$ as n_1, r, and $n \to +\infty$, with N fixed, then the asymptotic distribution of $\hat{t}$ is the same as the distribution of $\bar{X}/\mu$, with $N\bar{X}$ having a Poisson distribution with parameter $N\mu t$ (see Theorem 2). In the binomial case, the maximum likelihood estimator (m.l.e.) of t is obtained by first finding the m.l.e. $\hat{\theta} = \bar{X}$ for θ; then the m.l.e. for t is obtained by solving the equation $\theta = pt/(q+pt)$ for t and then replacing θ by $\hat{\theta}$. Confidence intervals for t may be obtained by finding a confidence interval for θ and then making use of the 1-1 correspondence between θ and t. Similar arguments apply in the Poisson case.

If the parameters n_1, r, and n are fixed, then by well-known theorems on the asymptotic distribution of maximum likelihood estimators, it follows that, as $N \to +\infty$, $\hat{t}$ converges in probability to t and $N^{\frac{1}{2}}(\hat{t}-t)$ converges in distribution to a random variable which is normally distributed with mean 0 and variance $[E\{(\partial/\partial t) \ln f(x;\, t)\}^2]^{-1}$. In this case, $E\{(\partial/\partial t) \ln f(x;\, t)\}^2 = \sigma^2/t^2$, so that for large values of N, the maximum likelihood estimator $\hat{t}$ is approximately normal with mean t and variance $t^2/N\sigma^2$. Finally, we consider the case when $f(x;\, t)$ may be approximated by a normal distribution. In particular, we shall examine the asymptotic distribution of $\hat{t}$ as n_1, r, and n each become large in such a way that $n_1/n \to \eta_1$, $r/n \to \eta_2$, $0 < \eta_1, n_2 < 1$. These conditions ensure the validity of Theorem 3. To simplify the exposition, we assume $N = 1$; the results for $N > 1$ follow immediately from this special case. We shall need the following lemma, which is easily proved.

LEMMA. $X/n \to_{\Pr} \lambda\eta_1\eta_2$.

The next theorem asserts that $\tilde{t}$ is a consistent estimator of t under the assumptions just stated.

THEOREM 6. $\tilde{t} \to_{\Pr} t$.

PROOF. Let $\phi_{n_1,r,n}(y) = y[1 - (n_1/n) - (r/n) + y]/[(n_1/n) - y][(r/n) - y]$.

Then $\phi_{n_1,r,n}(y)$ converges to $y(1 - \eta_1 - \eta_2 + y)/(\eta_1 - y)(\eta_2 - y) = \phi(y)$ uniformly in y over the closed interval $[\gamma, \delta]$, where $\max(0, \eta_1 + \eta_2 - 1) < \gamma \leqq \delta < \min(\eta_1, \eta_2)$. Since $Y_n = X/n \to_{\Pr} \lambda\eta_1\eta_2$ by the lemma, the desired conclusion follows on applying a result in [13], Problem 4.16, p. 112, which guarantees that $\varphi_{n_1,r,n}(Y_n) \to_{\Pr} \varphi(\lambda\eta_1\eta_2) = t$.

Since $t^* = X(n_1r - nX)/n_1r(n_1 - X)(r - X)$ converges in probability to zero as $n_1/n \to \eta_1$, $r/n \to \eta_2$, $\dot{t} \leqq \hat{t} \leqq \dot{t} + t^*$, it follows immediately from Theorem 6 that $\hat{t}$ also converges in probability to t, for $0 < t \leqq 1$. The result for $t > 1$ follows similarly.

In order to find the asymptotic distribution of $\dot{t}$ (and hence of $\hat{t}$), let $\dot{t} = \psi(X) = X(n - n_1 - r + X)/(n_1 - X)(r - X)$. Then from the previous results obtained here (in particular, Theorem 3) and a theorem of Rao ([10], pp. 207–208), it readily follows that $\dot{t}$ is asymptotically normal with mean t and variance $\sigma_{\dot{t}}^2 = H^{-2}[\psi'(n_1P_1)]^2$, where it can be shown that $\sigma_{\dot{t}}^2 = H^2t^2$.

6. Confidence intervals. We now combine Theorems 3, 4, 5, and 6 to obtain an approximate $(1 - \alpha)$ confidence interval for the parameter t. Let $Z_{\alpha/2}$ be such that the probability is $\alpha/2$ that a standardized normally distributed random variable Z exceeds $Z_{\alpha/2}$. According to Theorem 3, $\lim P\{|X - n_1P_1| < Z_{\alpha/2}H^{-1}\} = 1 - \alpha$. In the notation of Theorem 3, $n_1P_1 = \lambda_n n_1 r/n$, where λ_n is defined by (8). Since $\hat{t} \to_{\Pr} t$, it follows easily that $\hat{P}_1 = \hat{\lambda}_n r/n \to_{\Pr} \lambda\eta_2$ and $\hat{P}_2 = (r - n_1\hat{P}_1)/(n - n_1) \to_{\Pr} \eta_2(1 - \lambda\eta_1)/(1 - \eta_1)$, where $\hat{\lambda}_n$ is defined in the same manner as was λ_n, i.e., by (8), but with t replaced by $\hat{t}$ in (8). It then follows from Slutsky's theorem that

$$\lim P\{|X - n_1P_1| < Z_{\alpha/2}\hat{H}^{-1}\} = 1 - \alpha \tag{20}$$

where $\hat{H}^2 = \{n_1\hat{P}_1(1 - \hat{P}_1\}^{-1} + \{n_2\hat{P}_2(1 - \hat{P}_2)\}^{-1}$ and explicitly $\hat{P}_1 = [-\hat{c}_n + \{\hat{c}_n^2 + 4\hat{t}(1 - \hat{t})n_1r\}^{\frac{1}{2}}]/2n_1(1 - \hat{t})$, $\hat{P}_2 = (r - n_1\hat{P}_1)/(n - n_1)$, and $\hat{c}_n = n - (n_1 + r)(1 - \hat{t})$. If $\hat{t} = 1$, $\hat{P}_1 = r/n = \hat{P}_2$. In summary, H depends on t, but Slutsky's theorem permits us to allow the dependence of H on t to be replaced by a dependence on $\hat{t}$. This greatly simplifies the confidence interval. Inverting the inequalities in (20) for t, we obtain the approximate $(1 - \alpha)$ confidence interval for t

$$\frac{(x - Z_{\alpha/2}\hat{H}^{-1})(n - n_1 - r + x - Z_{\alpha/2}\hat{H}^{-1})}{(n_1 - x + Z_{\alpha/2}\hat{H}^{-1})(r - x + Z_{\alpha/2}\hat{H}^{-1})} < t < \frac{(x + Z_{\alpha/2}\hat{H}^{-1})(n - n_1 - r + x + Z_{\alpha/2}\hat{H}^{-1})}{(n_1 - x - Z_{\alpha/2}\hat{H}^{-1})(r - x - Z_{\alpha/2}\hat{H}^{-1})}. \tag{21}$$

The actual computation of this confidence interval can be simplified by estimating t by the consistent estimator $\dot{t}$. If this estimator is used, we find that $\hat{P}_1 = x/n_1$ and $\hat{P}_2 = (r - x)/(n - n_1)$, and the resulting confidence interval will continue to have a confidence coefficient of approximately $(1 - \alpha)$. We illustrate the calculation of 90 per cent confidence intervals for the parameter t. Let $n_1 = 12$, $r = 15$, $n = 30$ and $x = 3$. Then $\hat{P}_1 = x/n_1 = \frac{1}{4}$, $\hat{P}_2 = (r - x)/(n - n_1) = \frac{2}{3}$,

$\tilde{t} = \hat{P}_1\hat{Q}_2/\hat{P}_2\hat{Q}_1 = \frac{1}{6}$, $n_1\hat{P}_1\hat{Q}_1 = (12)(\frac{1}{4})(\frac{3}{4}) = \frac{9}{4}$, $n_2\hat{P}_2\hat{Q}_2 = (18)(\frac{2}{3})(\frac{1}{3}) = 4$, $\hat{H}^2 = \frac{4}{9} + \frac{1}{4} = (\frac{5}{6})^2$, $\hat{H}^{-1} = 1.2$, $Z_{.05} = 1.645$. Hence, the upper confidence bound, as given by (21), is found to be 0.563 and the lower confidence bound is 0.029. For $n_1 = 120$, $r = 150$, $n = 300$, $x = 30$, the approximate 90 per cent confidence interval for t is (0.105, 0.252), while for $n_1 = 60$, $r = 120$, $n = 200$, and $x = 40$, the interval for t is (0.919, 2.64).

One could also consider constructing a confidence interval for t by using the asymptotic distribution of $\tilde{t}$. If confidence intervals for t are constructed following this procedure, one obtains the interval $(\tilde{t}\{1 - Z_{\alpha/2}\hat{H}\}, \tilde{t}\{1 + Z_{\alpha/2}\hat{H}\})$. However, the first procedure for constructing approximate $(1 - \alpha)$ confidence intervals for t [leading to (21)] seems to give better results than this second method. This is probably due to the fact that the rate of convergence to the normal distribution of the distribution of $H(X - n_1P_1)$ is "faster" than that of $(\tilde{t} - t)/\sigma_{\tilde{t}}$.

Finally, we observe that a confidence interval for Yule's measure of association, $Q = (1 - t)/(1 + t)$, is readily obtained using (21).

6. Acknowledgments. I wish to express sincere and grateful thanks to Professor Leo Katz for suggesting a problem leading to a consideration of the present topic. Part of the results given here are based on the author's doctoral dissertation, written under Professor Katz supervision. I would also like to acknowledge the valuable assistance of Mr. George Hoetzl in carrying out many of the calculations.

REFERENCES

[1] BENNETT, P. M. and HSU, P. (1960). On the power function of the exact test for the 2×2 contingency table. *Biometrika* **47** 393–398.

[2] FELLER, WILLIAM (1957). *An Introduction to Probability Theory and Its Applications*, **1**(2nd ed.). Wiley, New York.

[3] GOODMAN, L. and KRUSKAL, W. H. (1954). Measures of association for error classifications. *J. Amer. Statist. Assoc.* **49** 732–764.

[4] HANNAN, J. and HARKNESS, W. (1963). Normal approximation to the distribution of two independent binomials, conditional on fixed sum. *Ann. Math. Statist.* **34** 1593–1595.

[5] HARKNESS, WILLIAM LEONARD (1959). An investigation of the power function for the test of independence in 2×2 contingency tables. Ph.D. Thesis, Michigan State Univ.

[6] HARKNESS, W. L. and KATZ, L. Comparisons of the power functions for the test of independence in 2×2 contingency tables. *Ann. Math. Statist.* **35** 1115–1127.

[7] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

[8] PATIL, G. P. (1963). Minimum variance unbiased estimation and certain problems of additive number theory. *Ann. Math. Statist.* **34** 1050–1056.

[9] PATNAIK, P. B. (1948). The power function of the test for the difference between two proportions in a 2×2 table. *Biometrika* **35** 157–175.

[10] RAO, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. Wiley, New York.

[11] SNOW, C. (1952). *Hypergeometric and Legendre Functions with Applications to Integral Equations of Potential Theory*. Nat. Bur. Standards Appl. Math. Ser. 19.

[12] VAN EEDEN, CONSTANCE and RUNNENBURG, J. TH. (1960). Conditional limit distributions for the entries in a 2×2 table. *Statistica Neerlandica* **14** 111–126.

[13] WILKS, S. S. (1962). *Mathematical Statistics*. Wiley, New York.