stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SN 1/70 AUGUSTUS

W. MOLENAAR
ON HEMELRIJK'S RESULTS CONCERNING APPROXIMATIONS TO THE HYPERGEOMETRIC DISTRIBUTION

2e boerhaavestraat 49 amsterdam

On Hemelrijk's results concerning approximations to the hypergeometric distribution

by W. Molenaar

0. Summary

An empirical study of 2×2 tables with grand total $N \leq 35$ has led HEMELRIJK (1967) to a footrule for the choice between the normal and the χ^2 approximation to the hypergeometric tail probabilities. For tails of less than .07, χ^2 was found to be generally better when the sum of the two smaller marginals $n+r$ exceeds $\frac{1}{2}N$, and worse otherwise. The present note offers an explanation for this phenomenon. It will be shown that little accuracy is lost by using χ^2 throughout. Moreover, a square-root type normal approximation will be presented which is usually superior to both classical approximations (normal and χ^2). For derivations not given here the reader is referred to a recent monograph (MOLENAAR, 1970, section IV.2).

1. Introduction

The hypergeometric distribution function *)

$$(1.1) \qquad P[\underline{a} \leq a] = \sum_{j=0}^{a} \binom{r}{j} \binom{N-r}{n-j} / \binom{N}{n}$$

is connected with the 2×2 table

$$(1.2) \qquad \begin{array}{cc|c} \underline{a} & \underline{b} & n \\ \underline{c} & \underline{d} & m = N-n \\ \hline r & s = N-r & N \end{array} \quad .$$

*) Random variables will be denoted by underlined symbols.

It is no restriction to assume that $0 \leq a < n \leq r \leq \frac{1}{2}N$ (the table can be re-ordered such that n and r become the two smallest marginals, the distributions of $\underline{b}$, $\underline{c}$, $\underline{d}$ are determined by that of $\underline{a}$, and $P\lceil \underline{a} \leq \underline{n} \rceil = 1$). We introduce the notation

$$\Phi(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} \exp(-\tfrac{1}{2}t^2)dt;$$

$$\mu = nr\,N^{-1};\ \tau^2 = mnrs\,N^{-3};\ \sigma^2 = \tau^2 N(N-1)^{-1}; \tag{1.3}$$

$$u = (a+\tfrac{1}{2}-\mu)\sigma^{-1};\ \chi = (a+\tfrac{1}{2}-\mu)\tau^{-1}.$$

Then the classical normal approximation is $P\lceil \underline{a} \leq \underline{a} \rceil \approx \Phi(u)$, and it is not difficult to see that the classical χ^2 approximation is equivalent to $P\lceil \underline{a} \leq \underline{a} \rceil \approx \Phi(\chi)$, with a trivial exception for values of a which are very close to μ.

2. Asymptotic expansions

For any given integers a, r, n, N satisfying $0 \leq a < n \leq r \leq \frac{1}{2}N$, there exists a unique *exact normal deviate* $\xi = \xi(a,n,r,N)$ defined by $\Phi(\xi) = P\lceil \underline{a} \leq \underline{a} \rceil$. Explicit solution of ξ from this transcendental equation is not possible. However, under the assumptions

$$N \to \infty,\quad \mu \to \infty,\quad \tau \to \infty,\quad \xi \text{ bounded} \tag{2.1}$$

an asymptotic expansion for ξ in powers of τ^{-1} can be given (MOLENAAR, 1970). Inversion of this expansion leads to

$$\begin{aligned}
&\chi = \xi + T_1 + T_\chi + o(\tau^{-2}),\\
&u = \xi + T_1 + T_u + o(\tau^{-2}),\\
&T_1 = \tau^{-1}(m-n)(s-r)N^{-2}(\xi^2-1)/6,\\
&T_\chi = \tau^{-2}\{\xi^3(-1-2mnN^{-2}-2\,rsN^{-2}+26mnrsN^{-4}) +\\
&+ \xi(-2+14mnN^{-2}+14rsN^{-2}-74mnrsN^{-4})\}/72,\\
&T_u = T_\chi - \tfrac{1}{2}\,\xi N^{-1} = T_\chi - \tau^{-2}\,\xi mnrsN^{-4}/2.
\end{aligned} \tag{2.2}$$

3. Comparison of normal and χ^2

Following HEMELRIJK (1967), we shall suppose that $|\xi| > 1.5$, i.e. the tail probability $\min(P[\underline{a} \leq a], P[\underline{a} \geq a+1])$ is less than .07. It follows that $T_1 \geq 0$, whereas $T_u < T_\chi < 0$ for $\xi > 1.5$ and $0 < T_\chi < T_u$ for $\xi < -1.5$. Let τ be fixed, and large enough to make the $o(\tau^{-2})$ terms in the expansions of χ and u negligibly small. We consider two cases.

(i) Let $(m-n)(s-r)$ be so large that $T_1 > |T_u| > |T_\chi|$. The normal approximation is now better than χ^2 for $\xi > 1.5$ (righthand tails): the negative contribution of T_u gives more compensation for the positive error present in T_1 (although not enough), and $\chi-\xi > u-\xi > 0$. For $\xi < -1.5$ (lefthand tails), however, T_χ adds less to the error present in T_1 than does T_u, and now χ^2 is more accurate.

As the difference between χ and u is $O(\tau^{-2})$, whereas the errors $\chi-\xi$ and $u-\xi$ are both $O(\tau^{-1})$, the difference between the approximations becomes negligible for large values of τ. This follows also from (1.3), where the "variances" τ^2 and σ^2 are asymptotically equivalent for $N \to \infty$.

(ii) Let $(m-n)(s-r)$ be so small that $|T_u| > |T_\chi| > T_1$. Now χ^2 is better in both tails (giving less overcompensation for $\xi > 1.5$, and adding less to the error present in T_1 for $\xi < -1.5$). The difference between the two approximations becomes more important if s-r approaches zero (we recall that $n \leq r \leq \frac{1}{2}N$, and thus $m-n \geq s-r \geq 0$). In the extreme case $r = s$, we have $T_1 = 0$; then it is the leading term of the expansion of the error which is smaller for $\chi-\xi$ than for $u-\xi$.

In HEMELRIJK's (1967) investigation for $10 \leq N \leq 35$, situation (i) is nearly always found when $n+r \leq \frac{1}{2}N$ and $|\xi| > 1.5$. However, the distribution is then so skew that small left hand tails simply do not exist: with only two exceptions for $N = 34$ and $N = 35$, even $P[\underline{a} = 0]$ corresponds to a value $\xi > -1.5$.

Therefore the normal approximation is indeed better than χ^2 for $|\xi| > 1.5$, $n+r \leq \frac{1}{2}N$ and $N \leq 35$, because $|\xi| > 1.5$ then means $\xi > 1.5$. On the other hand, situation (ii) will prevail when $n+r > \frac{1}{2}N$ and $10 \leq N \leq 35$, and χ^2 is then indeed more accurate.

The general situation, illustrated by some numerical examples in Table 1, is roughly as follows:
for $r \approx s$, χ^2 is much better than normal;
for $r << s$ and left hand tails, both are very bad, but χ^2 is slightly better;
for $r << s$ and right hand tails, both are rather bad, but χ^2 is slightly worse.
The boundary between $r \approx s$ and $r << s$ depends in an intricate way on the values of a, n and N: for small N (small τ), situation (ii) may be found e.g. when $rN^{-1} = .25$, whereas for large N (large τ) a case with e.g. $rN^{-1} = .4$ should be considered as skew, not as symmetric.

In agreement with a remark by ORD (1968), the asymptotic and numerical results indicate that little is lost by using the χ^2 approximation throughout. Compared to the normal it is easier, frequently much better, and hardly ever much worse.

4. A new approximation

Square root type normal approximations to the binomial and Poisson distribution function were proposed by FREEMAN & TUKEY (1950). Their asymptotic error is generally minus one half of the corresponding error for the classical normal approximations.

For the hypergeometric distribution function $P[\underline{a} \leq a]$, one finds in MOLENAAR (1970) a derivation of the approximations

$$(4.1) \qquad \Phi(2\{N-1\}^{-\frac{1}{2}} \{(a+1)^{\frac{1}{2}}(N-n-r+a+1)^{\frac{1}{2}} - (n-a)^{\frac{1}{2}}(r-a)^{\frac{1}{2}}\})$$

and

$$(4.2) \qquad \Phi(2N^{-\frac{1}{2}}\{(a+\tfrac{3}{4})^{\frac{1}{2}}(N-n-r+a+\tfrac{3}{4})^{\frac{1}{2}}-(n-a-\tfrac{1}{4})^{\frac{1}{2}}(r-a-\tfrac{1}{4})^{\frac{1}{2}}\}),$$

which share this property of inverting and halving the asymptotic error of the classical ones. Numerical evaluation of (4.1) is easy; it is designed to be especially accurate for tails between .01 and .05. One could use (4.2) when more accuracy is desired for values of $P[\underline{a} \leq a]$ between .05 and .95. For the doubly symmetric case $n = r = \frac{1}{2}N$, formula (4.2) becomes identical to $\Phi(\chi)$.

Example: evaluation of $P[\underline{a} \leq 11]$ in the 2×2 table $\begin{matrix} 11 & 39 \\ 43 & 68 \end{matrix}$. Here $N = 11+39+43+68 = 161$, and approximation (4.1) becomes $\Phi(2(161-1)^{\frac{1}{2}}\{(12\times69)^{\frac{1}{2}} - (39\times43)^{\frac{1}{2}}\}) = \Phi(-1.9252) = .0271$. The exact probability is .0269; the classical approximations lead to $\Phi(u) = .0286$ and $\Phi(\chi) = .0290$ respectively.

In this example the absolute error of approximation (4.1) is roughly one tenth of the absolute error of the classical approximations. The difference in accuracy is not always so spectacular, but one finds in most cases that the square root type (4.1) and (4.2) are indeed superior to the classical normal and χ^2 approximations. Table 1 shows that the 1 : 2 ratio of the asymptotic errors is hardly detectable in the numerical values. A really good agreement between numerical results and asymptotic properties can only be found for large values of the parameter τ. We expand in powers of τ^{-1}, and e.g. $n = r = 20$, $N = 200$ means $\tau = 1.3$, or $n = r = 50$, $N = 100$ means $\tau = 2.5$. In such cases terms of higher asymptotic order cannot be neglected.

When cumulative Poisson or binomial tables are available, hypergeometric tails can also be evaluated by means of a suitable Poisson or binomial approximation. Especially accurate, cf. WISE (1954) and MOLENAAR (1970), is the binomial approximation of n experiments with a success probability p chosen dependent on a, n, r and N, such as

$$(4.3) \qquad p = (2r-a)(2N-n+1)^{-1} + n(2nrN^{-1}-2a-1)(2N-n+1)^{-1}/3.$$

5. Numerical examples

Table 1 presents some numerical values of the relative tail error for some normal approximations.

TABLE 1. Event $\underline{a} \leq a$ or $\underline{a} \geq a+1$, exact hypergeometric probability and relative tail error in per cent. for some normal approximations. Example: $P[\underline{a} \leq 0] = .1085$ for $n = 20$, $r = 20$, $N = 200$; approximation (4.1), with relative tail error +38.16 per cent., gives $1.3816 \times .1085 = .1399$.

Event	proba-bility	normal $\Phi(u)$	chi-squared $\Phi(\chi)$	square root type (4.1)	(4.2)
$n = 20$	$r = 20$	$N = 200$			
$\underline{a} \leq 0$	.1085	+10.45	+9.91	+38.16	-1.01
$\underline{a} \leq 1$	.3782	-8.10	-8.19	+17.30	+3.83
$\underline{a} \geq 3$	.3213	+8.17	+8.06	-12.90	-1.96
$\underline{a} \geq 4$	.1222	-1.93	-2.41	-9.17	+4.75
$\underline{a} \geq 5$	.0345	-27.50	-28.32	+3.17	+21.39
$\underline{a} \geq 6$	.0073	-58.42	-59.29	+28.99	+54.41
$\underline{a} \geq 7$	.0012	-82.06	-82.65	+79.03	+117.66
$n = 20$	$r = 80$	$N = 200$			
$\underline{a} \leq 2$	.0024	+70.40	+67.10	-1.47	-17.45
$\underline{a} \leq 4$	.0425	+9.34	+8.38	+5.80	-1.74
$\underline{a} \leq 6$	.2377	-.79	-1.02	+4.02	+.85
$\underline{a} \geq 9$	.4005	+1.17	+1.11	-2.03	-.40
$\underline{a} \geq 11$	.1152	-.12	-.62	-.66	+1.26
$\underline{a} \geq 13$	.0160	-3.58	-4.89	+4.57	+6.07
$n = 50$	$r = 50$	$N = 200$			
$\underline{a} \leq 6$	.0092	+29.77	+27.87	-.68	-9.15
$\underline{a} \leq 8$	.0625	+5.86	+5.09	+3.96	-1.56
$\underline{a} \leq 10$	.2276	-.75	-1.00	+3.99	+.65
$\underline{a} \geq 13$	.4936	+1.29	+1.29	-2.46	-.60
$\underline{a} \geq 15$	.2234	+1.13	+.87	-3.14	-.21
$\underline{a} \geq 17$	.0678	-2.34	-3.05	-1.93	+2.26
$\underline{a} \geq 19$	.0134	-10.33	-11.64	+2.67	+8.46
$n = 100$	$r = 100$	$N = 200$			
$\underline{a} \leq 39$	.0014	+5.76	+3.24	+.76	+3.24
$\underline{a} \leq 41$	.0080	+3.00	+1.32	-.34	+1.32
$\underline{a} \leq 43$	.0329	+1.44	+.40	-.62	+.40
$\underline{a} \leq 45$	.1015	+.61	+.05	-.50	+.05
$\underline{a} \leq 47$	.2398	+.21	-.02	-.25	-.02
$\underline{a} \leq 49$	.4438	+.03	-.01	-.04	-.01
$n = 5$	$r = 8$	$N = 20$			
$\underline{a} \leq 0$	.0511	+20.68	+11.43	+6.38	-.16
$\underline{a} \leq 1$	.3065	-.90	-2.42	+2.83	+.66
$\underline{a} \geq 3$	.2962	+2.55	+.98	-3.59	-.23
$\underline{a} \geq 4$	.0578	+6.67	-1.50	-3.75	+4.79
$\underline{a} \geq 5$	.0036	+41.39	+16.39	+15.98	+32.46

Acknowledgement

Many thanks are due to Professors HEMELRIJK and VAN ZWET, for encouragement and valuable advice.

References

FREEMAN, M.F. & J.W. TUKEY (1950), Transformations related to the angular and the square root, Ann. Math. Stat. 21, 607-611.

HEMELRIJK, J. (1967), The hypergeometric, the normal and chi-squared, Stat. Neerl. 21, 225-231.

MOLENAAR, W. (1970), Approximations to the Poisson, Binomial and hypergeometric distribution functions, M.C. Tract 31, Mathematisch Centrum, Amsterdam.

ORD, J.K. (1968), Approximations to distribution functions which are hypergeometric series, Biometrika 55, 243-248.

WISE, M.E. (1954), A quickly convergent expansion for cumulative hypergeometric probabilities, direct and inverse, Biometrika 41, 317-329.