The hypergeometric, the normal and chi-squared*)

by J. Hemelrijk**)

Summary

This note describes a numerical investigation of the normal and χ^2-approximations to the hypergeometric distribution, which leads to a surprisingly simple footrule. If n and r are the two smaller marginal totals, then for the tails of the distribution up to about a probability of 0.07, the normal approximation will in nearly all cases be better than the χ^2 if $n + r \leqslant \frac{1}{2} N$ (where N is the grand marginal total) and worse otherwise. Although the two approximations are nearly equivalent, this footrule is so simple that it seems worth publishing.

The hypergeometric distribution is usually obtained from a 2×2-table:

$$\begin{array}{cc|c} \underline{a} & \underline{b} & n \\ \underline{c} & \underline{d} & m \\ \hline r & s & N \end{array} \qquad (1)$$

where the underlined symbols denote random variables and n, m, r, s and N are marginal totals. Under certain well-known conditions the random variables $\underline{a}$, $\underline{b}$, $\underline{c}$ and $\underline{d}$ have hypergeometric distributions. They are completely dependent, because of the fixed marginal totals, and we only need to consider one of them. It is always possible to arrange the table such that

$$n \leqslant r \leqslant s \leqslant m \qquad (2)$$

and we shall use this arrangement. The hypergeometric distribution of $\underline{a}$ is given by

$$P(\underline{a} = a) = \binom{n}{a}\binom{m}{c} \Big/ \binom{N}{r} \qquad (a + c = r) \qquad (3)$$

with

$$\mu = nr/N \quad \text{and} \quad \sigma^2 = \frac{m\,n\,r\,s}{N^2(N-1)}. \qquad (4)$$

The normal approximation consists of using a random variable $\underline{a}_1$ instead of $\underline{a}$, $\underline{a}_1$ being normally distributed with mean μ and variance σ^2 from (4) and applying a correction for continuity, i.e.

---

*) Report S 376 (SP 102) of the Department of Mathematical Statistics of the Mathematical Centre, Amsterdam.

**) Professor University of Amsterdam.

$$\left.\begin{aligned} P(\underline{a} \leqslant a) &\simeq P(\underline{a}_1 \leqslant a + \tfrac{1}{2}) \\ P(\underline{a} \geqslant a) &\simeq P(\underline{a}_1 \geqslant a - \tfrac{1}{2}). \end{aligned}\right\} \tag{5}$$

The χ^2-approximation, with correction for continuity, can be written as follows:

$$\frac{\{|\,\underline{a}\underline{d} - \underline{b}\underline{c}\,| - \tfrac{1}{2}N\}^2 N}{m\,n\,r\,s} = \frac{\{|\,\underline{a} - \mu\,| - \tfrac{1}{2}\}^2}{\frac{N-1}{N}\sigma^2} \tag{6}$$

has approximately a χ^2-distribution with one degree of freedom (μ and σ^2 again from (4)). The χ^2-distribution with one degree of freedom being the distribution of the square of a standard normally distributed variable, this comes down to using a random variable $\underline{a}_2$ instead of $\underline{a}$, $\underline{a}_2$ being normally distributed with mean μ and variance $\frac{N-1}{N}\sigma^2$.

The problem of this note – which of the two approximations is best – therefore boils down to the question whether it is better to use σ^2 unchanged in the normal approximation or modified by putting N^3 in the denominator. The latter is usually somewhat more convenient and, of course, for large N the difference is negligable.

A numerical investigation of hypergeometric distributions up to $N = 35$ yielded a curiously simple footrule:

for $n + r \leqslant \frac{1}{2}N$ *and* $p < 0.07$, $\underline{a}_1$ *is nearly always better than* $\underline{a}_2$
(i.e. $N^2(N-1)$ in the numerator is better than N^3)
and otherwise $\underline{a}_2$ *is mostly better than* $\underline{a}_1$. (7)

(Note that n and r are the smaller marginal totals, cf. (2)).

This result was arrived at in the following way. Let, for given a,

$$\left.\begin{aligned} p &= \min\{P(\underline{a} \leqslant a),\ P(\underline{a} \geqslant a)\} \\ p_1 &= \min\{P(\underline{a}_1 \leqslant a),\ P(\underline{a}_1 \geqslant a)\} \\ p_2 &= \min\{P(\underline{a}_2 \leqslant a),\ P(\underline{a}_2 \geqslant a)\} \end{aligned}\right\} \tag{8}$$

then $\underline{a}_1$ is said to be better than $\underline{a}_2$ if

$$p \geqslant p_1 > p_2 \tag{9}$$

and the other way around if

$$p_1 > p_2 \geqslant p. \tag{10}$$

(Note that always $p_1 > p_2$, because the variance used for $\underline{a}_1$ is larger than the one used for $\underline{a}_2$).

The intermediate cases where

$$p_1 > p > p_2 \tag{11}$$

are, for the moment, left out of consideration and only those values of a were considered for which

$$p \leqslant 0{,}0668, \tag{12}$$

(corresponding to a value of 1.5 for the standard normal variate), these being the most interesting values of p.

For $N = 10$, 15, 20 and 30 the results, for all n and r, are given in fig. 1 and 2. A ⊐ with the number k inside denotes that for the relevant values of N, n and r, there are k values of a satisfying (12) and (9). A D means that (12) and (10) are satisfied. A very distinct pattern emerges and the footrule (7) is indicated by a line separating ⊐'s and D's rather nicely. For $N = 30$ there is some overlap and the separation is not complete any more. This will grow worse with increasing N, but then the difference between the two approximations diminishes.

Remarks

The footrule is purely experimental, no explanation is offered. The intermediate cases, satisfying (11) have also been investigated, defining $\underline{a}_1$ to be better than $\underline{a}_2$ if $p_1 - p < p - p_2$. As might be expected from the above result, they follow the same pattern but with somewhat more overlap; mainly they are also distinguished nicely by the footrule. The other values of N investigated (12, 14, 16 and 35) showed exactly the same kind of behaviour. Thus it seems safe enough to use the footrule also for larger values of N, as far as this is at all necessary.

Acknowledgement

The computations for this investigation were executed by C. VISSER on the Electrologica X1 of the Mathematical Centre, Amsterdam.

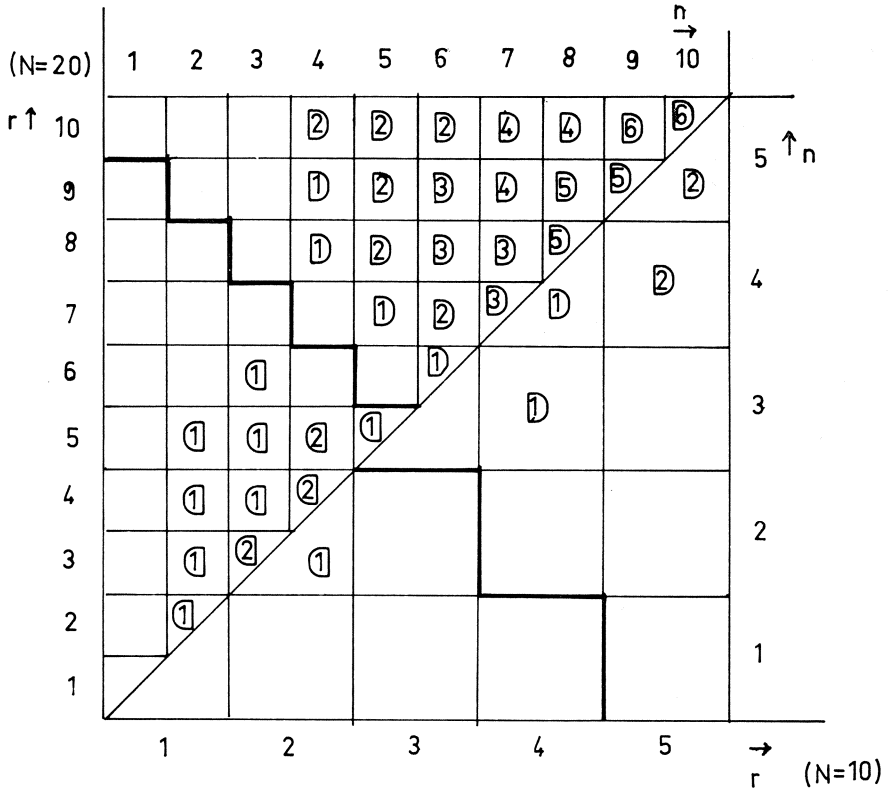

Fig. 1. □: normal approx. better
D: χ^2 approx. better
N = 10, 20

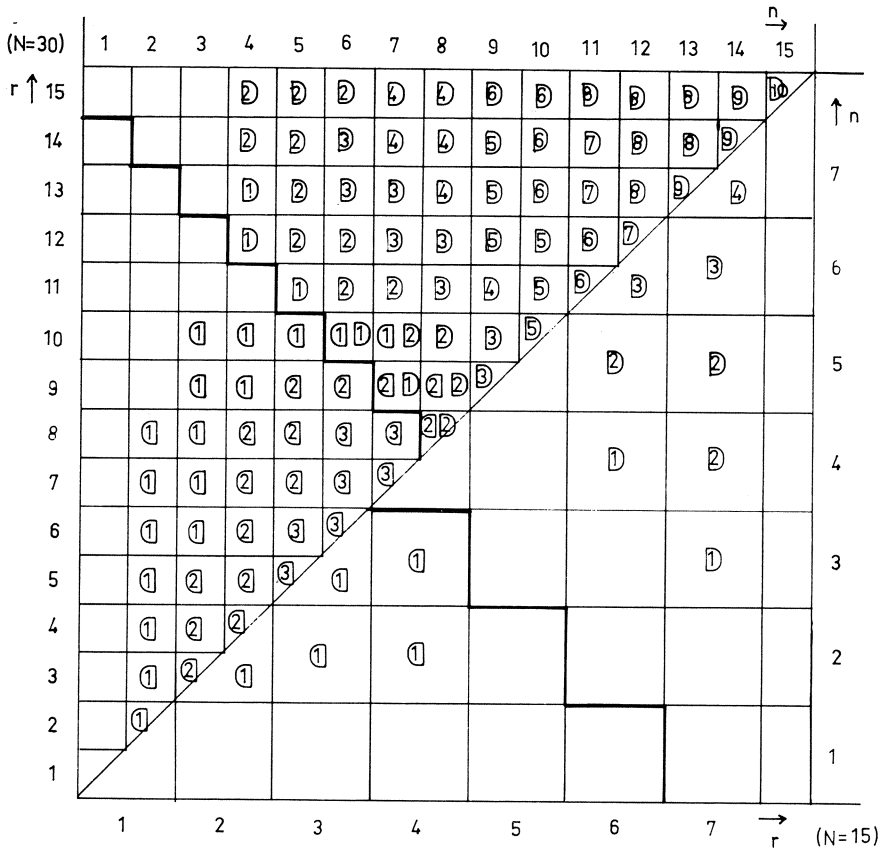

Fig. 2: C : normal approx. better
D : χ^2 approx. better
N = 15, 30