Hypergeometric tail probabilities

MARY C. PHIPPS

University of Sydney, February 2001

Abstract

This paper improves an existing inequality between the tail probabilities of two related hypergeometrics. The inequality is made more general by the removal of restrictive conditions. It is also shown to be strict in all but certain boundary cases and to have implications in significance testing.

Keywords: P-value, Fisher's Exact Test, hypergeometric, binomial.

1 Introduction

The hypergeometric variable $X \sim HG(z, m, n)$, with probability distribution

$$P(X = x) = \frac{\binom{m}{x}\binom{n}{z-x}}{\binom{m+n}{z}},$$

for integer x satisfying $\max(0, z-n) \leq x \leq \min(m, z)$, is a model for the number of Type I items in a sample of z taken without replacement from a population of size $m+n$ of which exactly m are Type I items.

To simplify notation we shall write $l = \max(0, z-n)$ and $u = \min(z, m)$ for the lower and upper bounds respectively for x. We shall also denote the upper tail probability, $P(X \geq a)$ by

$$p(a; z, m, n) = P(X \geq a) = \sum_{x=a}^{u} \frac{\binom{m}{x}\binom{n}{z-x}}{\binom{m+n}{z}}.$$

This paper investigates relationships between the tail probabilities of related hypergeometrics, $HG(z, m, n)$ and $HG(z+1, m+1, n)$.

Hypergeometric tail probabilities are used to find P-values for Fisher's (1934) Exact Test for equal success probabilities in two binomials. In the context of Fisher's Exact Test, m and n are the binomial sample sizes, z is the observed value of Z, the total number of successes in the two samples and a is the observed number of successes in the first sample. Under the assumption of equal success probabilities, the number of successes in the first sample conditional on $Z = z$ is $X \sim HG(z, m, n)$, a hypergeometric variable. For an upper one-sided test for example, the Fisher-exact P-value is $p(a; z, m, n)$, a conservative, but frequently applied, measure in small sample applications. This is a conditional measure, since z is treated as fixed rather than as an experimentally observed value of the total number of successes.

2 Established inequalities

The use of this hypergeometric tail probability $p(a+1; z+1, m+1, n+1)$ of another related hypergeometric, $HG(z+1, m+1, n+1)$ precedes Fisher's Exact test in the two binomial context (see Seneta (1994) for historical background and derivation).

Seneta and Phipps (2001) prove for all integer a satisfying $l \le a \le u$ that tail probabiliies of $HG(z, m, n)$ and $HG(z+1, m+1, n+1)$ are linked by

$$p(a+1; z, m, n) < p(a+1; z+1, m+1, n+1) < p(a; z, m, n)\,, \tag{1}$$

indicating that $p(a+1; z+1, m+1, n+1)$ is a less conservative measure than $p(a; z, m, n)$, that it has suitable properties for a significance measure and that it is possibly a reasonable approximation to the (unconditional) P-value.

Hájek and Havránek (1978, p164) proved a similar inequality relating the tail probabilities of $HG(z+1, m+1, n)$ and $HG(z, m, n)$, showing, for $a > \frac{zm}{m+n}$, that

$$p(a+1; z+1, m+1, n) \le p(a; z, m, n)\,. \tag{2}$$

The mean of $HG(z; m, n)$ is $\frac{zm}{m+n}$ and we show that the condition $a > \frac{zm}{m+n}$ is unnecessarily restrictive and that (2) holds for all integer a satisfying $l \le a \le u$. Further, we prove that inequality (2) is actually a strict inequality for all integer a satisfying $l < a \le u$ and that it is also strict at the boundary $a = l$ under certain conditions. In addition we extend the result to the double inequality of the next section.

3 A strict inequality

The focus of this section is the double inequality:

$$p(a+1; z, m, n) < p(a+1; z+1, m+1, n) < p(a; z, m, n)\,, \tag{3}$$

which indicates that $p(a+1; z+1, m+1, n)$ might be considered as an alternative to the Fisher-exact P-value in the manner of Seneta and Phipps (2001). Unlike (1), which holds for all integer a satisfying $l \leq a \leq u$, the above inequality holds for all integer a satisfying $l < a \leq u$ and can be extended to include the boundary $a = l$ only when $z < n$. In contrast, $p(a+1; z, m, n) < p(a+1; z+1, m+1, n) = p(a; z, m, n) = 1$ for the boundary $a = l$ when $z \geq n$.

We first establish the range of values of a for which

$$p(a+1; z+1, m+1, n) < p(a; z, m, n) \tag{4}$$

is valid, and because l and u coincide when $HG(z, m, n)$ is degenerate (ie when $z = 0$ or $z = m+n$), it is convenient to discuss the degenerate and non-degenerate cases separately.

The degenerate cases: $z = m + n$ **and** $z = 0$**.**

- When $z = m+n$, both $HG(z, m, n)$ and $HG(z+1, m+1, n)$ are degenerate and $l = u = m$. Therefore there are no values of a satisfying $l < a \leq u$, and at $a = m$ (the common value of l and u) it is obvious that $p(m+1; m+n+1, m+1, n) = 1 = p(m; m+n, m, n)$.

- When $z = 0(< n)$ (a special case of $z < n$), $HG(z, m, n)$ is degenerate and $l = u = 0$. Therefore there are no values of a satisfying $l < a \leq u$, and at $a = 0$ (the common value of l and u) the strict inequality (4) holds since $p(1, 1, m+1, n) = \frac{m+1}{m+n+1} < 1$, but $p(0; 0, m, n) = 1$.

The non-degenerate cases: $0 < z < m + n$

For the non-degenerate cases, $0 < z < m+n$, we note that $p(a+1; z+1, m+1, n)$ and $p(a; z, m, n)$ both have the same number of positive summands since

$$p(a; z, m, n) = \sum_{x=a}^{u} \frac{\binom{m}{x}\binom{n}{z-x}}{\binom{m+n}{z}}$$

and

$$p(a+1;z+1,m+1,n) = \sum_{s=a+1}^{u+1} \frac{\binom{m+1}{s}\binom{n}{z+1-s}}{\binom{m+n+1}{z+1}} = \sum_{x=a}^{u} \frac{\binom{m+1}{x+1}\binom{n}{z-x}}{\binom{m+n+1}{z+1}}.$$

(on putting $x = s-1$)

We therefore focus on the question of which sufficiently high x (if any) satisfy

$$\frac{\binom{m+1}{x+1}\binom{n}{z-x}}{\binom{m+n+1}{z+1}} < \frac{\binom{m}{x}\binom{n}{z-x}}{\binom{m+n}{z}}$$

$$\text{i.e. } \frac{(m+1)(z+1)}{(x+1)(m+n+1)} < 1 \tag{5}$$

Clearly the LHS of (5) decreases as x increases. Further, there is at least one value of x for which (5) holds. This is established by considering the upper boundary value $x = u = \min(m,z)$, noting that

– when $m \geq z$ we have $u = \min(m,z) = z$
  and putting $x = z$ in (5), $\frac{(m+1)(z+1)}{(x+1)(m+n+1)} = \frac{m+1}{m+n+1} < 1$ as required.

– when $m < z$, we have $u = \min(m,z) = m$
  and putting $x = m$ in (5), $\frac{(m+1)(z+1)}{(x+1)(m+n+1)} = \frac{z+1}{m+n+1} < 1$ as required, since $z < m+n$ for all non-degenerate cases.

Thus (5) holds and hence inequality (4) holds at least for $x = u$ in all non-degenerate cases and the *smallest* integer value of x (satisfying $l \leq x \leq u$) for which (5) holds is well-defined. We will call this value l'. Since $\frac{(m+1)(z+1)}{(x+1)(m+n+1)}$ decreases as x increases and since (5) holds for $x = l'$, it follows that (5) holds for $x = l', l'+1, \cdots, u$. Further, since $p(a+1;z+1,m+1,n)$ and $p(a;z,m,n)$ have the same number of summands, it follows that (4) holds for $a = l', l'+1, \cdots, u$.

To show that (4) also holds for all integer a satisfying $l < a < l'$ (and also at $a = l(=0)$ when $z < n$) we define

$$q(a;z,m,n) = 1 - p(a;z,m,n) = \sum_{x=l}^{a-1} \frac{\binom{m}{x}\binom{n}{z-x}}{\binom{m+n}{z}}. \tag{6}$$

Thus

$$q(a+1;z+1,m+1,n) = \sum_{s=max(0,z-n+1)}^{a} \frac{\binom{m+1}{s}\binom{n}{z+1-s}}{\binom{m+n+1}{z+1}} \tag{7}$$

Consider the two non-degenerate cases $0 < z < n$ and $n \leq z \leq m+n$ separately:

- For $0 < z < n$, we have $l = 0$ and also $max(0, z-n+1) = 0$. Therefore (6) can be written

$$q(a+1; z+1, m+1, n) = \sum_{s=0}^{a} \frac{\binom{m+1}{s}\binom{n}{z+1-s}}{\binom{m+n+1}{z+1}} = \frac{\binom{m+1}{0}\binom{n}{z}}{\binom{m+n+1}{z}} + \sum_{x=l}^{a-1} \frac{\binom{m+1}{x+1}\binom{n}{z-x}}{\binom{m+n+1}{z+1}}. \qquad (8)$$

(on putting $x = s-1$)

At $a = 0$, we see from (6) that $q(0; z, m, n) = 0$ and from (8) that $q(1; 1, m+1, n) > 0$. Hence (4) is satisfied as required for $a = l(= 0)$.

More generally we see from (6) and (8) that $q(a+1; z+1, m+1, n)$ has one more positive term than $q(a; z, m, n)$ for *all* integer a satisfying $0 \leq a < l'$, and since l' is the *smallest* integer x satisfying (5), it follows that $\frac{\binom{m+1}{x+1}\binom{n}{z-x}}{\binom{m+n+1}{z+1}} \geq \frac{\binom{m}{x}\binom{n}{z-x}}{\binom{m+n}{z}}$ for $x = l'-1$. Also, $\frac{(m+1)(z+1)}{(x+1)(m+n+1)}$ increases strictly as x decreases, and therefore the remaining summands (if any) of $q(a+1; z+1, m+1, n)$ are greater than or equal to the corresponding summands of $q(a; z, m, n)$. Adding over these summands gives $q(a+1; z+1, m+1, n) > q(a; z, m, n)$ and it follows immediately that $p(a+1; z+1, m+1, n) < p(a; z, m, n)$ for all integer a satisfying $l \leq a \leq u$.

- For $n \leq z < m+n$, we have $l = z-n$ and $max(0, z-n+1) = z-n+1$. Therefore from (7), putting $x = s-1$,

$$q(a+1; z+1, m+1, n) = \sum_{x=l}^{a-1} \frac{\binom{m+1}{x+1}\binom{n}{z-x}}{\binom{m+n+1}{z+1}} \qquad (9)$$

For the boundary case, $a = l$, $q(a+1; z+1, m+1, n) = q(a; z, m, n) = 0$ and so $p(a+1; z+1, m+1, n) = p(a; z, m, n) = 1$ as was also seen for the degenerate case $z = m+n$.

More generally we see from (6) and (9) that $q(a+1; z+1, m+1, n)$ and $q(a; z, m, n)$ have the same number of positive summands for $0 \leq a < l'$. At $a = l+1(= z-n+1)$, there is exactly one summand each and $q(a+1; z+1, m+1, n) = \frac{\binom{m+1}{z-n+1}}{\binom{m+n+1}{z+1}}$ which is strictly greater than the corresponding summand for $q(a; z, m, n)$, namely $\frac{\binom{m}{z-n}}{\binom{m+n}{z}}$. For $a > l+1$, the remaining summands (if any) of $q(a+1; z+1, m+1, n)$ are greater than or equal to $q(a; z, m, n)$ since they do not satisfy (5), using an argument similar to the argument for the case $0 < z < n$. Adding over these summands gives $q(a+1; z+1, m+1, n) > q(a; z, m, n)$ and it follows immediately that $p(a+1; z+1, m+1, n) < p(a; z, m, n)$ for all integer a satisfying $l < a \leq u$.

A parallel argument shows that $p(a+1; z, m, n) < p(a+1 : z+1, m+1, n)$ for all integer a satisfying $l \leq a \leq u$, and taken together with (4) imply the double inequality (3).

4 Further results

It is possible to establish the inequality (2) of Seneta and Phipps(2000) directly from (4). To do this we rewrite (2) as two separate inequalities for integer a satisfying $l \leq a \leq u$:

$$p(a+1; z+1, m+1, n+1) < p(a; z, m, n) \tag{10}$$

$$\text{and} \quad p(a+1; z, m, n) < p(a+1; z+1, m+1, n+1) \tag{11}$$

In order to prove inequality (10) we use the identity

$$p(a; z, m, n) = p(d, v, n, m) \tag{12}$$

where $v = n + m - z$ and $d = n + a - z$. Results (12) and (4) lead to

$$p(a+1; z+1, m+1, n) = p(d; v, n, m+1) > p(d+1; v+1, n+1, m+1) = p(a+1; z+1, m+1, n+1),$$

for integer d satisfying $L < d \leq U$, where $L = max(0, v - m + 1)$ and $U = min(v, n)$ and also at the boundary $d = L$ when $v < m + 1$, with an identity instead at $d = L$ when $v \geq m + 1$. This is equivalent to

$$p(a+1; z+1, m+1, n+1) < p(a+1; z+1, m+1, n) \tag{13}$$

for all integer a satisfying $l < a \leq u$ and also at the boundary $a = l$ when $z \geq n$, with an identity instead at $a = l$ when $z < n$.

Combining (4) and (13) yields, for *all* integer a in $l \le a \le u$:

(a) provided $z \ge n$

$$p(a+1; z+1, m+1, n+1) < p(a+1; z+1, m+1, n) \le p(a; z, m, n),$$

(b) provided $z < n$

$$p(a+1; z+1, m+1, n+1) \le p(a+1; z+1, m+1, n) < p(a; z, m, n).$$

Thus, regardless of the value of z it follows immediately that (10) holds, ie

$$p(a+1; z+1, m+1, n+1) < p(a; z, m, n)\,.$$

In order to prove inequality (11), we observe that

$$p(a+1; z, m, n) = \sum_{r \ge a+1} \frac{\binom{m}{r}\binom{n}{z-r}}{\binom{m+n}{z}}$$

$$= 1 - \sum_{t \ge z-a} \frac{\binom{n}{t}\binom{m}{z-t}}{\binom{m+n}{z}} \quad \text{(on putting } t = z - r\text{)}.$$

This establishes the identities

$$p(a+1; z, m, n) = 1 - p(z-a; z, n, m) \tag{14}$$

and

$$p(a+1; z+1, m+1, n+1) = 1 - p(z-a+1; z+1, n+1, m+1). \tag{15}$$

From (10), we have $p(z-a+1; z+1, n+1, m+1) < p(z-a; z, n, m)$ for integer a satisfying $(z-a) \in [\max(0, z-m),\, \min(n, z)]$, and from the identities (14) and (15) it follows that

$$p(a+1; z, m, n) < p(a+1; z+1, m+1, n+1) \quad \text{for} \quad a \in [\max(0, z-n),\, \min(m, z)]\,.$$

This establishes inequality (11), and together (10) and (11) imply that inequality (2) holds as required for all integer a satisfying $l \le a \le u$.

References

FISHER,R.A., 1934: *Statistical Methods for Research Workers*, 5th Edition. Oliver & Boyd, Edinburgh.

HAJEK,P. and HAVRANEK,T., 1978: *Mechanizing Hypothesis Formation* Springer Verlag: Berlin, Heidelberg, New York.

SENETA,E., 1994: Carl Liebermeister's Hypergeometric Tails. *Historia Mathematica* **21**, 453-462.

SENETA,E. and PHIPPS,M., 2001: On the Comparison of Two Observed Frequencies. *Biometrics* To appear.