JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 7(3), 165–174

Inequalities Between Hypergeometric Tails

MARY C. PHIPPS† maryp@maths.usyd.edu.au
School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia

Abstract. A special inequality between the tail probabilities of certain related hypergeometrics was shown by Seneta and Phipps [19] to suggest useful 'quasi-exact' alternatives to Fisher's [5] Exact Test. With this result as motivation, two inequalities of Hájek and Havránek [6] are investigated in this paper and are generalised to produce inequalities in the form required. A parallel inequality in binomial tail probabilities is also established.

Keywords: P-value, conservativeness, quasi-exact, Fisher's Exact Test, Lancaster's mid-P, Liebermeister's P

1. Introduction

The hypergeometric variable $U \sim HG(z, m, n)$ has probability distribution

$$P(U = u) = \frac{\binom{m}{u}\binom{n}{z-u}}{\binom{m+n}{z}}$$

for integer u satisfying $\max(0, z-n) \le u \le \min(m, z)$. We shall denote the upper tail probability, $P(U \ge a)$, by

$$p(a; z, m, n) = P(U \ge a) = \sum_{u=a}^{\min(m,z)} \frac{\binom{m}{u}\binom{n}{z-u}}{\binom{m+n}{z}}.$$

A standard result for independent binomial variables X and Y, where $X \sim B(m, p_1)$ and $Y \sim B(n, p_2)$ with $p_1 = p_2$ (common success probability) is that the distribution of X, conditional on $Z(= X + Y) = z$, is hypergeometric, $HG(z, m, n)$. This result is exploited in Fisher's Exact Test, the commonly used approach for testing the hypothesis of common success probability ($H_0 : p_1 = p_2 = p$) in independent binomials when the sample sizes, m and n are small. In this context, X and Y represent the number of successes in the two independent samples, and the observed success and failure frequencies may be summarized in a 2×2 table. The fixed values are m and n:

† Requests for reprints should be sent to Mary C. Phipps, School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia.

	Success	Failure	Total
Sample 1	a	b	m
Sample 2	c	d	n
	z	v	$m+n$

Based on these empirically observed values of (X, Y), the Fisher-exact P-value for an upper one-sided test (with $H_1 : p_1 > p_2$) is $P(X \geq a|Z = z) = p(a; z, m, n)$, which we shall denote by the generic p_F. The corresponding test procedure at nominal level $\alpha \in (0, 1)$ is: "Reject H_0 if $p_F \leq \alpha$", and the test is known as Fisher's Exact Test.

This test is conditional since it treats z as fixed, rather than as an observed value of the variable $Z(= X + Y)$. The use of p_F as P-value cleverly avoids the theoretical and computational problems involved in calculating an unconditional P-value, since it is free of the nuisance parameter, p, and it also avoids the problems of 'ordering' the 2×2 tables. It is generally agreed however that p_F is conservative. The difference of opinion about the reason (discreteness or conditioning) for this conservativeness is well documented, and a comprehensive overview of these opinions is presented by Sahai and Khurshid [17]. Fisher's test is obviously α-level in the unconditional setting where the variable corresponding to p_F is $p(X; Z, m, n)$. Clearly, $P_{H_0}(p_F \leq \alpha) = \sum\sum_{\{(x,z):p_F \leq \alpha\}} \binom{m}{x}\binom{n}{z-x} p^z (1-p)^{m+n-z} \leq \alpha$ for any $p \in (0, 1)$ and for any nominal level $\alpha \in (0, 1)$. Fisher's test is however very conservative, and it is not unusual to find that $P_{H_0}(p_F \leq \alpha) < \frac{1}{2}\alpha$, as demonstrated by Boschloo [3].

This excessive conservativeness of p_F suggests that a less conservative measure may be preferable, provided it is also easily calculated. In §2 we give a brief summary of the findings of Seneta and Phipps [19], concerning the properties of two measures based on hypergeometric tails. These measures, $p(.)$, not only have some statistical justification as significance measures in the two binomial problem, but also satisfy the strict double inequality (1). This means that they are less conservative than $p(a; z, m, n) = p_F$ and yet not as liberal as $p(a + 1; z, m, n)$:

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

Motivated by this result, we generalise two inequalities due to Hájek and Havránek [6], and show that there are more related hypergeometric tails, $p(.)$, satisfying (1). This is followed by a numerical example, comparing the measures $p(.)$. A parallel inequality in binomial tails is established in §3 and some implications are discussed.

2. Some Alternatives to Fisher's Exact Test

We begin by discussing two measures which are of historical significance in the two-binomial context, and which also satisfy (1).

2.1. Lancaster's mid-P, p_M

A measure which has gained acceptance as an alternative to Fisher's P-value (see for example Hirji, Tan and Elashoff [7]) is an adjustment for discrete P-values due to Lancaster [8], [9]. The adjustment is called the mid-P and will be denoted by p_M.

Lancaster's mid-P adjustment of p_F is defined by

$$p_M = \frac{1}{2}[P(X \geq a|Z = z) + P(X > a|Z = z)] = \frac{1}{2}\,[p(a; z, m, n) + p(a + 1; z, m, n)].$$

Since p_M is the average of $p(a; z, m, n)$ and $p(a+1; z, m, n)$ it is clear that (1) is satisfied by $p(.) = p_M$, and therefore that p_M is less conservative than Fisher's p_F but does not err too far in the other direction. Barnard [1] suggests that p_F and p_M should both be quoted when testing equality of success probability for small samples because of the conservativeness of p_F. Further, Berry and Armitage [2] point out that p_M has mean $\frac{1}{2}$ and variance close to $\frac{1}{12}$, in line with the properties of uniformly distributed P-values (based on continuous test statistics) and that p_M has some justification as a significance measure on these grounds. (We note here that all other weighted averages of $p(a; z, m, n)$ and $p(a + 1; z, m, n)$ also satisfy (1), but that they do not have the stated desirable properties of p_M.)

The corresponding mid-P test procedure at arbitrary nominal significance level α is "Reject H_0 when $p_M \leq \alpha$." In contrast with Fisher's Exact Test, this procedure is not strictly α-level since there is no guarantee that $P_{H_0}(p_M \leq \alpha) \leq \alpha$ for arbitrary $\alpha \in (0, 1)$. Hirji, Tan and Elashoff [7] describe the procedure as *quasi-exact*. Their extensive empirical assessment reveals the excessive conservativeness of p_F when compared with p_M. They also demonstrate that in the unconditional setting p_M is occasionally (but only mildly) anti-conservative, ie $P_{H_0}(p_M \leq \alpha) \approx \alpha$ even though α is occasionally exceeded. It is worth mentioning that this is true also of the Pearson χ^2-statistic used for large samples in this context (*loc.cit.*).

Hirji *et al.* [7] argue that *closeness to nominal levels* with only rare exceedance is an important criterion for assessing a test procedure. They conclude that although not strictly a P-value, p_M can be regarded as an approximation in the unconditional setting, just as the chi-squared approximation is used in the large-sample case.

2.2. Liebermeister's measure, p_L

We now turn to a different hypergeometric, $HG(z+1, m+1, n+1)$. The use of its tail probability, $p(a+1; z+1, m+1, n+1)$, in the two binomial setting dates back to Liebermeister [10]; Seneta [18] shows the Bayesian derivation and historical background to this tail probability, which we shall denote by p_L. We note that Overall [11], [12] also recommends the use of p_L, purely on the basis of worked numerical examples.

Seneta and Phipps [19] prove that, in addition to the Bayesian origins of p_L, inequality (1) is satisfied by $p(.) = p_L$, ie

$$p(a+1; z, m, n) < p(a+1; z+1, m+1, n+1) < p(a; z, m, n)\,, \qquad (2)$$

From (2), it is seen that Liebermeister's measure, p_L is less conservative than p_F but not too anticonservative and so, like the mid-P, p_L is *quasi-exact* and can be interpreted as an approximation to the unconditional P-value in the sense that $P_{H_0}(p_L \leq \alpha) \approx \alpha$ for arbitrary $\alpha \in (0, 1)$. A comparison of the degree of anti-conservatism and also power comparisons are carried out by Seneta and Phipps [19] for the measures p_F, p_M and p_L. The point is also made that the calculations required for p_L are no more complicated than for p_F. In fact existing software for p_F can be used simply by adding unity to the diagonals a and d in the 2×2 table of frequencies, as the numerical example in §2.4 demonstrates.

2.3. Further inequalities in hypergeometic tails

Other promising related hypergeometrics are $HG(z+1, m+1, n)$ and $HG(z, m-1, n)$. Hájek and Havránek [6] proved two inequalities involving their tail probabilities. They showed, subject to $a > \frac{zm}{m+n}$, that (in our notation):

$$p(a+1; z+1, m+1, n) \leq p_F \quad \text{and also} \quad p(a; z, m-1, n) \leq p_F.$$

We shall write $p(a+1; z+1, m+1, n)$ as p_{Ha} and $p(a; z, m-1, n)$ as p_{Hb}. In the context of an upper tail test, it is only the cases $a > \frac{zm}{m+n}$ which are of interest since the mean of $HG(z, m, n)$ is $\frac{zm}{m+n}$. Nevertheless we show that $a > \frac{zm}{m+n}$ is unnecessarily restrictive and also that the inequalities can actually extend to double inequalities like (1), which means that p_{Ha} and p_{Hb} are both less conservative than p_F, but not as liberal as $p(a+1; z, m, n)$.

2.3.1. *The inequality for* $p_{Ha} = p(a+1; z+1, m+1, n)$

The inequality:

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

holds for $l < a \leq u$, where $l = \max(0, z-n)$ and $u = \min(z, m)$ are the lower and upper bounds respectively of $HG(z, m, n)$.

The boundary case $a = l$ is of no interest in significance testing, but we note here for completeness that (3) *does* also hold for $a = l$ when $z < n$. The right hand inequality '$<$' needs to be replaced by '$\leq$' *only* for case $a = l$ when $z \geq n$, and in that case $p(a+1; z+1, m+1, n) = p(a; z, m, n) = 1$.

Since $HG(z, m, n)$ is degenerate when $z = 0$ or $z = m+n$, statistical interest is in the case $0 < z < m+n$ only. A brief outline of the proof of (3) for this case now follows. The complete proof, including a discussion of the degenerate cases $z = 0$ and $z = m+n$, is in Phipps [14].

Outline of the proof The right hand inequality of (3), which is the strict version of the inequality of Hájek and Havránek [6], is considered first, namely:

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

Clearly the two tails $p(a+1; z+1, m+1, n)$ and $p(a; z, m, n)$ have the same number of summands. It can easily be seen that all the summands of $p(a+1; z+1, m+1, n)$ are strictly smaller than the corresponding summands of $p(a; z, m, n)$ when $a > \frac{(m+1)(z+1)}{(m+n+1)} - 1$, but not otherwise. Hence (4) is satisfied for $a \geq l'$, where l' is the integer part of $\frac{(m+1)(z+1)}{(m+n+1)}$.

To prove that (4) is also satisfied for $a < l'$, we focus on the summands of the lower tails: $1 - p(a+1; z+1, m+1, n)$ and $1 - p(a; z, m, n)$.

Treating the cases $z < n$ and $n \leq z < m+n$ separately, Phipps [14] proves the strict inequality $1 - p(a+1; z+1, m+1, n) > 1 - p(a; z, m, n)$ and it follows immediately that $p(a+1; z+1, m+1, n) < p(a; z, m, n)$ as required.

A parallel argument gives $p(a+1; z, m, n) < p(a+1 : z+1, m+1, n)$ for all integer a satisfying $l \leq a \leq u$. Taking this inequality together with (4), the double inequality (3) is proved for $l < a \leq u$, with a weaker inequality at $a = l$.

2.3.2. The inequality for $p_{Hb} = p(a; z, m-1, n)$

For l and u defined as in §2.3.1, the following inequality holds for $l < a \le u$:

$$p(a+1; z, m, n) \le p(a; z, m-1, n) < p(a; z, m, n)\,. \qquad (5)$$

The proof is not given here, but follows similar arguments to those given for p_{Ha}. Notice that the left hand inequality of (5) is not strict at $a = m$ since both $p(m+1; z, m, n)$ and $p(m; z, m-1, n)$ are identically zero. This means that an outcome with frequencies:

m	0	m
$z-m$	$n+m-z$	n
z	$n+m-z$	$n+m$

has positive probability, and yet $p_{Hb} = 0$. This is an unacceptable approximation to a positive P-value and so p_{H_b} is not suitable as a significance measure. Nevertheless we include p_{Hb} for completeness in the following numerical example.

2.4. A numerical example

One of the examples discussed in Seneta and Phipps [19] is this 2×2 table of observed frequencies which arose from a study by Di Sebastiano *et al.* [4] on rumbling appendix pain (success) in independent samples of non-acute and acute appendix cases. An upper tail test for success probability was required.

	Success	Failure	Total
Sample 1	5	10	15
Sample 2	1	15	16
	6	25	31

- The Fisher-P measure is $p_F = p(5; 6, 15, 16) = \sum_{x=5}^{6} \frac{\binom{15}{x}\binom{16}{6-x}}{\binom{31}{6}} = 0.072.$

- The Liebermeister-P is $p_L = p(6 : 7, 16, 17) = \sum_{x=6}^{7} \frac{\binom{16}{x}\binom{17}{7-x}}{\binom{33}{7}} = 0.035$ which is equivalent to finding p_F for the table below, where unity has been added to the diagonals of the previous table:

6	10	16
1	16	17
7	26	33

- Lancaster's mid-P is $p_M = \frac{1}{2}[p(5;6,15,16) + p(6;6,15,16)] = 0.039$
- The final two measures are $p_{Ha} = 0.0415$ and $p_{Hb} = 0.0590$.
- The frequencies are too small for the Pearson χ^2-statistic to be appropriate, but the approximate P-value calculated from its positive square root is Chi-P = 0.028. The Yates' corrected value is 0.073.

Figure 1 shows a plot of the unconditional P-value for this example:

$$P(p) = \sum_C \binom{m}{x}\binom{n}{z-x} p^z (1-p)^{m+n-z}$$

as p varies. We have used p_F as the criterion for 'ordering' the 2×2 tables, ie the region of summation used was $C = \{(x,z) : p_F(x;z,m,n) \leq 0.072\}$. Other criteria for ordering the tables, such as p_L, lead to almost identical curves. (Pierce and Peters [15] give reasons for such phenomena in a more general context.)

Superimposed on the plot of $P(p)$ in Figure 1 are horizontal lines corresponding to the Fisher-P ($p_F = 0.072$), the mid-P ($p_M = 0.039$), the Liebermeister-P ($p_L = 0.035$) and the P-value from the chi-squared test (Chi-P= 0.028). The values for the two measures, $p_{Ha} = 0.0415$ and $p_{Hb} = 0.0590$ are also superimposed. We observe that the maximum likelihood estimate of p is $6/31 \approx 0.2$ and it is clear from the diagram that the Liebermeister-P is closer to $P(p)$ for all $p \in (0.2, 0.8)$.

This numerical example is typical of 2×2 tables with small sample sizes. The two measures p_{Ha} and p_{Hb} are 'closer' than p_F to the unconditional P-value, but typically they are more conservative than either the mid-P or the Liebermeister-P. As a result, it is only p_M and p_L which are seriously considered as useful quasi-exact alternatives to Fisher's Exact Test. In their comparison of p_M, p_L and p_F as suitable easily calculated approximations to the unconditional P-value, Seneta and Phipps [19] include plots of the Type I error probability at various significance levels and for various combinations of m and n. With the exception of very unbalanced tables for which p_L behaves erratically (the example used is $m = 80, n = 40, \alpha = 0.05$) the comparisons support the computational use of p_L, but for very unbalanced tables, the use of p_M is recommended instead.

3. The Binomial Tail Analogue

An inequality corresponding to (1), for tails from the binomial $\mathcal{B}(z,p)$, is:

$$b(a+1;z,p) < b(.) < b(a;z,p). \tag{6}$$

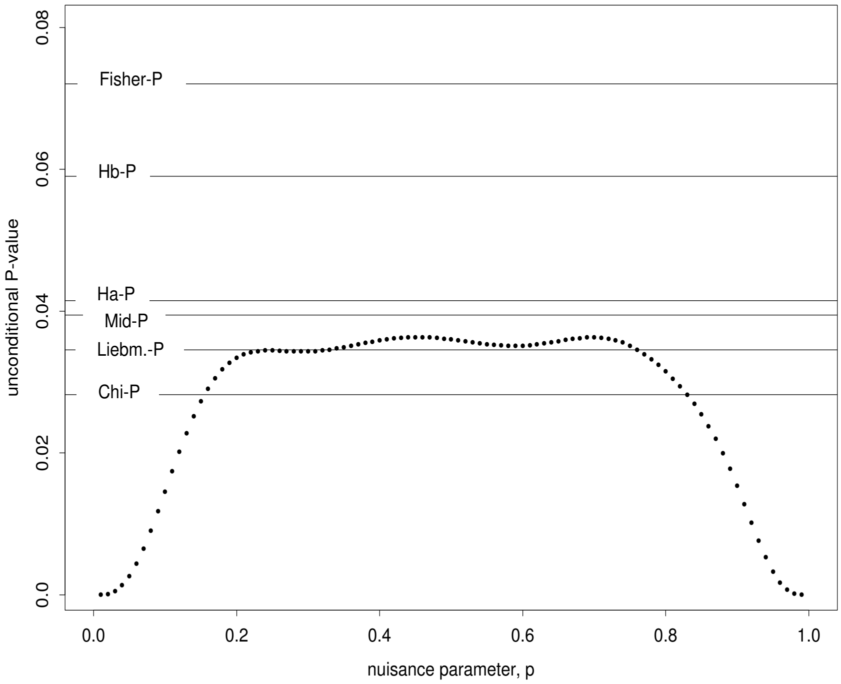

Figure 1. A plot of $P(p)$, the unconditional P-value as p varies, for the numerical example of §2.4. Approximations to $P(p)$ for this example are superimposed on the plot: p_F (Fisher-P), p_M (Mid-P), p_L (Liebm.-P), p_{Ha}(Ha-P), p_{Hb} (Hb-P) and Chi-P.

where $b(a; z, p) = \sum_{x=a}^{z} \binom{z}{x} p^x (1-p)^{z-x}$ for integer a satisfying $0 \leq a \leq z$.

Inequality (6) is satisfied by $b(.) = b(a+1; z+1, p)$. This can be proved using elementary combinatorial algebra, since it is not difficult to show that $b(a+1; z+1, p)$ can be expressed as follows:

$$b(a+1; z+1, p) = p\,[b(a+1; z, p)] + (1-p)\,[b(a; z, p)]$$

This is simply a weighted average of $b(a+1; z, p)$ and $b(a; z, p)$ and therefore inequality (6) is satisfied by $b(.) = b(a+1; z+1, p)$. The particular case $p = 0.5$ is $b(.) = b(a+1, z+1, 0.5)$ and is the mid-P in the following two tests.

3.1. Exact test for Poisson means

It is well known that if X and Y are independent Poisson variables with common parameter λ, the distribution of X conditional on $X + Y = z$ is binomial, $\mathcal{B}(z, 0.5)$. The 'exact' (upper-tail) test for common mean in the Poisson is based on this conditional distribution (see for example Robinson [16]). For an empirically observed value $(a, z-a)$ for (X, Y), the P-value for an upper tail 'exact' test is $b(a; z, 0.5)$. The less conservative mid-P, $b(a+1; z+1, 0.5)$, has some justification as an alternative measure on the grounds that it more closely resembles the uniform distribution. Seneta and Phipps [19] show that this measure is also justified on Bayesian grounds. They use uniform priors to obtain $b(a+1, z+1, 0.5)$, by analogy with the method used to derive the Liebermeister p_L. It is not difficult to show that the same result is obtained using exponential priors with arbitrary positive, finite mean. It is curious that the resulting measure, $b(a+1, z+1, 0.5)$, is identical to the mid-P, in contrast to the two measures p_L and p_M discussed in §2.

3.2. The sign test

Suppose we want an upper one-tail test of the hypothesis (H_0) of equal probability of positive and negative counts in a small sample of n counts, some of which may be zero (or ties in a sample of n pairs). Let X, Y, W be the number of positive, negative and zero (or tied) counts and write $Z = X + Y$. The variable (X, Y, W) is trinomial, and if H_0 is true, conditional on $Z(= X + Y) = z$, the distribution of X is binomial $\mathcal{B}(z, 0.5)$. The 'exact' test is therefore the usual sign test and if (a, z) is the observed value of (X, Z), the P-value is $P_{H_0}(X \geq a | Z = z) = b(a; z, 0.5)$. The parallel with Fisher's Exact Test is immediate, and the corresponding *quasi-exact* test is the test based on the mid-P. Phipps [13], in discussing the sign test, demonstrates the superiority of the mid-P, $b(a+1; z+1, 0.5)$, over the conditional P-value, $b(a; z, 0.5)$, from the sign test.

References

1. G. Barnard. On alleged gains in power from lower P-values. *Statistics in Medicine*, 8:1469–1477, 1989.
2. G. Berry and P. Armitage. Mid-P confidence intervals: a brief review. *The Statistician*, 44:417–423, 1995.
3. R. D. Boschloo. Raised conditional level of significance for the 2×2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24:1–35, 1970.
4. P. Di Sebastiano, T. Fink, F. F. Di Mola, E. Weihe, P. Innocenti, H. Freiss, and M. Büchler. Neuroimmune appendicitis. *The Lancet*, 354(9177):461–466, 1999.
5. R. A. Fisher. *Statistical Methods for Research Workers*, 5th Ed. Oliver & Boyd, Edinburgh, 1934.

6. P. Hájek and T. Havránek. *Mechanizing Hypothesis Formation.* Springer Verlag: Berlin, Heidelberg, New York, 1978.
7. K. F. Hirji, S. Tan and R. M. Elashoff. A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine*, 10:1137–1153, 1991.
8. H. O. Lancaster. The combination of probabilities arising from data in discrete distributions. *Biometrika*, 36:370–382, 1949.
9. H. O. Lancaster. Significance tests in discrete distributions. *Journal of the American Statistical Association*, 58:223–234, 1961.
10. C. Liebermeister. Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. *Sammlung Klinischer Vorträge*, (Innere Medicin No. 31-64) 110:935–962, 1877.
11. J. E. Overall. Continuity correction for Fisher's exact probability test. *Journal of Educational Statistics*, 5:177–190, 1980.
12. J. E. Overall. Comment. *Statistics in Medicine*, 9:379–382, 1990.
13. M. C. Phipps. Exact tests and the mid-P. *Eighth International Scientific Kravchuk Conference. Conference Materials*, 471–475, Kyiv. (ISBN:5-7707-2384-X), 2000.
14. M. C. Phipps. Hypergeometric tail probabilities. *Research Report of the School of Mathematics and Statistics*, 01–2, 2001.
15. D. A. Pierce and C. Peters. Improving on exact tests by approximate conditioning. *Biometrika*, 86:265–277, 1999.
16. J. Robinson. Optimal tests of significance. *The Australian Journal of Statistics*, 21:301–310, 1979.
17. H. Sahai and A. Khurshid. On analysis of epidemiological data involving (2×2) contingency tables: an overview of Fisher's Exact Test and Yates' correction for continuity. *Journal of Biopharmaceutical Statistics*, 5:43–70, 1995.
18. E. Seneta. Carl Liebermeister's Hypergeometric Tails. *Historia Mathematica*, 21:453–462, 1994.
19. E. Seneta and M. C. Phipps. On the Comparison of Two Observed Frequencies. *Biometrical Journal*, 43(1):23–43, 2001.