

Liebermeister's quasi-exact test for two binomials

Mary C Phipps

University of Sydney

School of Mathematics and Statistics

[maryp@maths.usyd.edu.au](mailto:maryp@maths.usyd.edu.au)

## Liebermeister's quasi-exact test for two binomials

In 1877, Carl Liebermeister derived an elegant mathematical result which has relevance in statistics today. Liebermeister had been particularly interested in the statistical problem of testing for equal success probability in two binomial populations when the sample sizes are small.

Using a Bayesian approach and succinct mathematical arguments he proposed a procedure based on the tail probabilities of a hypergeometric distribution. This predates by more than half a century Fisher's Exact Test, the commonly applied procedure for small sample sizes. Fisher's approach was frequentist rather than Bayesian and was conditioned on the observed total number of successes. Given the difference in approach it is curious that both procedures involve the tails of hypergeometric distributions, but this means that Liebermeister's procedure can be applied either by using a hand calculator or by using computer software already available for Fisher's Exact Test.

Although Fisher's Test is technically 'exact', it is well known that it can be excessively conservative in the sense that the rejection rate for the claim of equal success probability is well below the nominal significance level. In contrast, it has recently been proved that the earlier remarkable procedure due to Liebermeister is considerably less conservative, without erring too far in the other direction and it is consequently described as *quasi-exact*, giving results which are close to the more computer-intensive unconditional exact methods. Liebermeister's procedure therefore deserves serious attention in this context, and its adoption is strongly recommended in applications where typically the sample sizes are small.

# Liebermeister's quasi-exact test for two binomials

Mary C. Phipps<sup>1</sup>

## Summary

A procedure for testing equal success probabilities in two independent binomial experiments was proposed by Carl Liebermeister in 1877, predating by more than half a century Fisher's Exact Test, the commonly applied procedure with small sample sizes. Liebermeister's procedure is considerably less conservative, is easily calculated and gives results close to those which use computer-intensive methods. Adoption of Liebermeister's procedure is therefore strongly recommended in applications where typically the sample sizes are small. Mention is also made of Lancaster's mid-P which has already gained acceptance as an easily calculated alternative in this context.

*Keywords:* P-value; quasi-exact tests; Fisher's Exact Test; Liebermeister's measure; Lancaster's mid-P; conservativeness.

## Introduction

This paper discusses a little known test for equal success probabilities ( $H_0 : p_1 = p_2$ ) for two independent binomial variables,  $X \sim \mathcal{B}(m, p_1)$  and  $\mathcal{B}(n, p_2)$ , where  $m$  and  $n$  are small, with observed frequencies as set out in this fourfold table:

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

Fisher's (1934) Exact Test is the commonly applied small sample test in this context. For an upper-tail test against  $H_1 : p_1 > p_2$  based on these frequencies, the Fisher Exact (conditional) P-value,  $p_F(a; z, m, n)$ , is obtained by conditioning on the observed total successes,  $z (= a + c)$ , thereby removing dependence on  $p$ , the unknown common success probability under  $H_0$ :

$$p_F(a; z, m, n) = P_{H_0}(X \geq a | X + Y = z) = \sum_{r \geq a} \frac{\binom{m}{r} \binom{n}{z-r}}{\binom{m+n}{z}}$$

---

<sup>1</sup>School of Mathematics and Statistics,  
University of Sydney, NSW2006, Australia  
email: maryp@maths.usyd.edu.au

The hand calculations are simple and the P-value,  $p_F(a; z, m, n)$ , written as  $p_F$  when there is no ambiguity, is also readily available in all popular statistical software packages. To perform the corresponding test at significance level  $\alpha$ , Fisher's Exact Test, is 'Reject  $H_0$  if  $p_F \leq \alpha$ '. This test is excessively conservative, since the Type I error probability is often as small as  $\frac{1}{2}\alpha$  for small  $m$  and  $n$ , a fact which is well documented in the literature. Equivalently,  $p_F$  is often twice the value of the unconditional P-value as Boschloo (1970) observed.

Computer-intensive unconditional approaches may be used to avoid this conservativeness and consequent lack of power, and are available in some commercial software packages.

A different approach which also avoids the excessive conservativeness of Fisher's Exact Test uses instead Lancaster's (1961) mid-P (written as  $p_M$  if there is no ambiguity):

$$p_M(a; z, m, n) = p_F - \frac{1}{2} \frac{\binom{m}{a} \binom{n}{z-a}}{\binom{m+n}{z}},$$

which has gained increasing acceptance as a less conservative and easily calculated alternative to Fisher's Exact Test. The corresponding mid-P test is: 'Reject  $H_0$  if  $P_M \leq \alpha$ '. This test is not strictly  $\alpha$ -level, but is described as *quasi-exact* by Hirji, Tan and Elashoff (1991) and its properties are discussed by Seneta and Phipps (2001).

A lesser known procedure, based on Bayesian arguments and due to Liebermeister (1877), is likewise shown by Seneta and Phipps (2001) to be quasi-exact and to deserve serious consideration as an alternative to Fisher's Exact Test. Liebermeister's measure (written as  $p_L$  if there is no ambiguity) is the focus of this paper:

$$p_L(a; z, m, n) = \sum_{r \geq a+1} \frac{\binom{m+1}{r} \binom{n+1}{z+1-r}}{\binom{m+n+2}{z+1}}.$$

The corresponding *quasi-exact*  $\alpha$ -level test is: 'Reject  $H_0$  if  $p_L \leq \alpha$ '.

Although the conceptual approaches differ, the two measures  $p_F$  and  $p_L$  both involve the tail probability of different hypergeometrics. The appeal of these measures is that their values are close to those obtained by computer-intensive approaches but they can be performed on a hand calculator very simply, especially when the sample sizes are small. Also, the similarity between  $p_L$  and  $p_F$  can be exploited to find  $p_L$  by adjusting the observed table as follows:

$$\begin{array}{|c|c|c|} \hline a & b & m \\ \hline c & d & n \\ \hline z & v & m+n \\ \hline \end{array} \quad \longrightarrow \quad \begin{array}{|c|c|c|} \hline a+1 & b & m+1 \\ \hline c & d+1 & n+1 \\ \hline z+1 & v+1 & m+n+2 \\ \hline \end{array} .$$

Then use readily accessible statistical software to calculate the upper-tail  $p_F$  for the adjusted table. This will give  $p_L$  for the observed data in the original table.

### Appendix pain example

The data from a study on appendix pain by Di Sebastiano et al. (1999) is used as a numerical example. The study involved two independent binomials and resulted in 5 ‘successes’ from 15 trials and only one ‘success’ from an independent set of 16 trials. The fourfold table of observed

frequencies is:

5	10	15
1	15	16
6	25	31

To test whether there is a significantly higher success rate for the first set of trials, Fisher’s (upper-tail) P-value is  $p_F = 0.072$ , which is non-significant at the usual significance level,  $\alpha = 0.05$ . In contrast, if Liebermeister’s measure is used, the result is significant at this level. To calculate  $p_L$  replace the diagonal entries 5 and 15 by 6 and 16 (adjusting the totals accord-

ingly) and find Fisher’s upper-tail on the new table:

6	10	16
1	16	17
7	26	33

, resulting in  $p_L = 0.035$ .

Thus at level 0.05, Liebermeister’s measure indicates a significantly higher success rate. Since  $p_L (= 0.035) < p_F (= 0.072)$ , it is clear that the test based on  $p_L$  is less conservative than the test based on  $p_F$ . (Note that Lancaster’s mid-P is also less conservative, as  $p_M = 0.039$ .) #

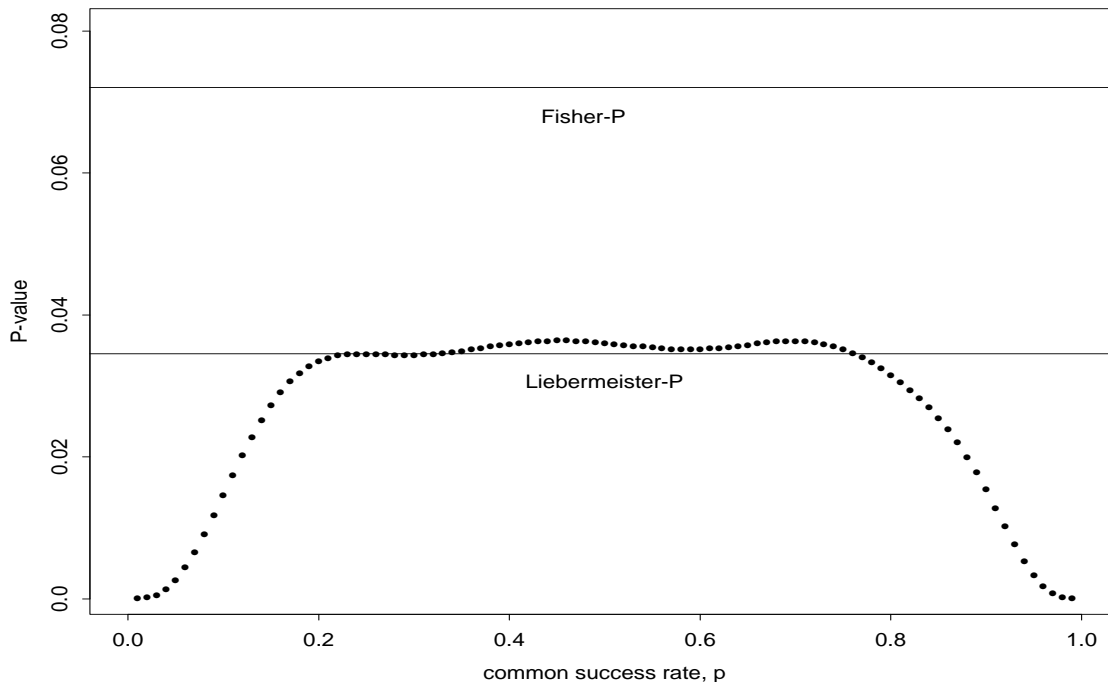
### Comparison of the two measures $p_F$ and $p_L$

Given that the measures  $p_F$  and  $p_L$  are so different in this example, it is natural to ask which measure is recommended. This question can be addressed by comparing  $p_F$  and  $p_L$  with  $\mathcal{P}(p)$ , the (unconditional) P-value as a function of  $p$ , the unknown common success probability under  $H_0$ .  $\mathcal{P}(p)$  is calculated by summing the joint binomial probabilities of  $x$  and  $z - x$  successes for the set,  $\mathcal{S}$ , of fourfold tables deemed to be more extreme (by some suitable ordering criterion) than the observed table: :

$$\mathcal{P}(p) = \sum_{\{(x,z) \in \mathcal{S}\}} \sum_{x=0}^m \binom{m}{x} \binom{n}{z-x} p^z (1-p)^{m+n-z}.$$

Technical difficulties arise in evaluating  $\mathcal{P}(p)$  because  $p$  is unknown and also because  $\mathcal{S}$  is not unique. The extensive survey by Sahai and Khurshid (1995) gives an indication of the

many different criteria which have been proposed in the literature for ordering the tables and hence identifying  $\mathcal{S}$ . Figure 1 shows how  $\mathcal{P}(p) = \sum_{\{(x,z) \in \mathcal{S}\}} \binom{m}{x} \binom{n}{z-x} p^z (1-p)^{m+n-z}$  varies with  $p$  when for example  $p_F$  is used as criterion for identifying  $\mathcal{S}$ . Superimposed on this curve are  $p_F = 0.072$  and  $p_L = 0.035$ , both of which avoid the problem of unknown  $p$ , and are therefore constant as  $p$  varies.



**Figure 1** The dotted line is the unconditional P-value,  $\mathcal{P}(p)$ , for the appendix pain example. Superimposed are  $p_F$  (Exact Fisher-P) and  $p_L$  (Liebermeister-P).

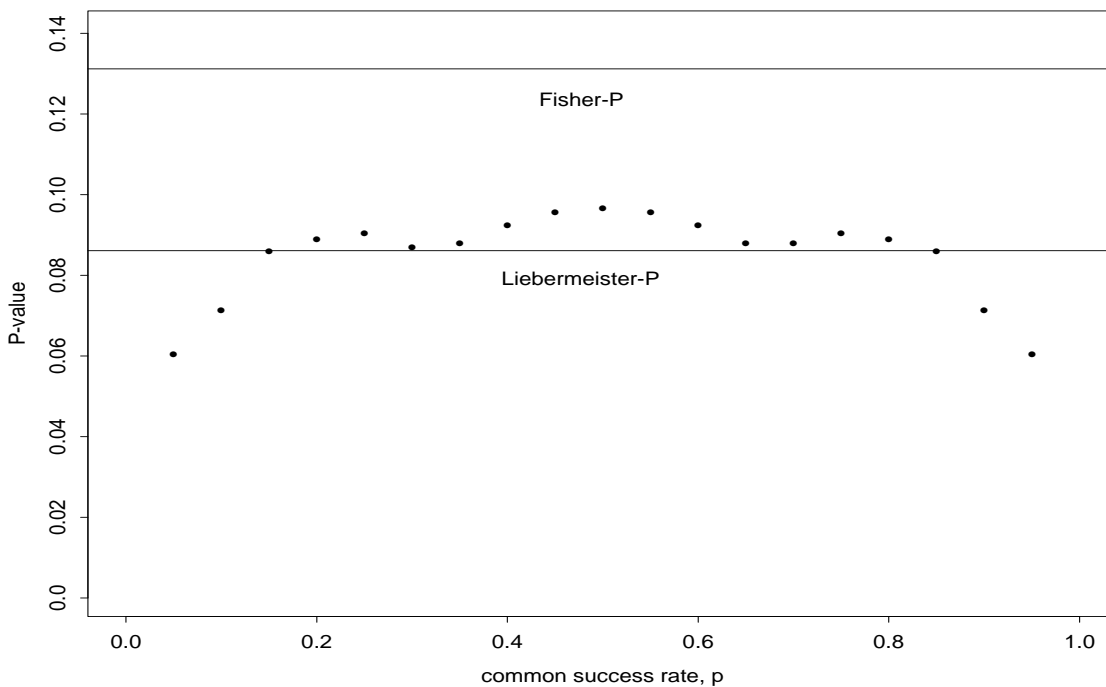
It is clear from the diagram that Fisher’s Exact P-value,  $p_F$ , is very conservative regardless of the true value of  $p$ , the common, unknown success probability. In contrast the Liebermeister-P is very close to the P-value for all  $p$  in the range 0.18 to 0.80, and in particular for  $p = 6/31 \approx 0.2$ , the maximum likelihood estimate of  $p$ .

Barnard’s (1945) suggestion for removing dependence on  $p$  from the unconditional P-value was to use  $p_B = \sup_{0 < p < 1} \mathcal{P}(p)$ . There still remains however a wide choice of criteria for ordering the tables and hence for identifying the region of summation,  $\mathcal{S}$ . In the above numerical example, when  $p_F$  is chosen as the ordering criterion the value of  $p_B$  is 0.0363. In contrast, the StatXact software package (1989), which uses a standardized difference of proportions, as discussed by Suissa and Shuster (1985), results in a value for  $p_B$  of 0.0362. The difference is

slight, which accords with Pierce and Peters (1999), who maintain (in the context of *approximate conditioning*) that natural choices of criteria are “virtually equivalent for practical purposes”.

Rice’s (1988) CBET method for avoiding the problem of unknown  $p$  is to use a Bayesian estimate  $\hat{p}$  of  $p$  and to evaluate  $p_C = \mathcal{P}(\hat{p})$ , using the difference of proportions as ordering criterion, but without standardizing the difference. For comparison,  $p_C = 0.039$  for this numerical example.

The pattern in Figure 1 (with  $p_L < p_F$ , and  $p_L \approx \mathcal{P}(p)$  for most  $p$ ) is not specific to small sample sizes. For illustration we refer to an example used by Pierce and Peters (1999) to demonstrate ‘approximate conditional P-values’. In their example, two independent samples, both of size 50, resulted in 10 and 5 successes respectively. Figure 2 plots the unconditional P-value against  $p$  and shows that even though  $m$  and  $n$  are large, Fisher’s P-value ( $p_F = 0.1312$ ) is very conservative, regardless of the true value of  $p$ , and that Liebermeister’s P ( $p_L = 0.0861$ ) is clearly superior for values of  $p$  in the range 0.15 to 0.85.



**Figure 2** The dotted line is the unconditional P-value,  $\mathcal{P}(p)$ , for the case  $m = n = 50$ ,  $a = 10$ ,  $z = 15$ . Superimposed are  $p_F$  (Exact Fisher-P) and  $p_L$  (Liebermeister-P).

A question which naturally arises is whether  $p_F$  is *always* greater than  $p_L$  as in these two numerical examples. The simple answer is ‘yes’. Seneta and Phipps (2001) proved that it is

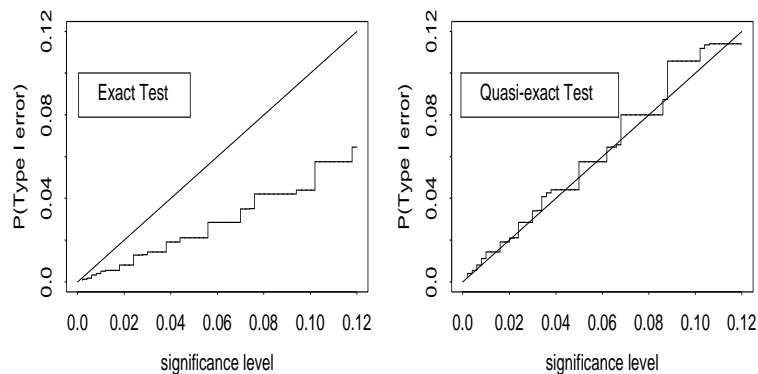
always true that  $p_F > p_L$ , and also that  $p_L$  has a well defined lower bound,  $p_F(a + 1; z, m, n)$ , so that like the mid-P,  $p_L$  is also *quasi-exact*, with suitable properties for a significance measure.

### Exact and Quasi-exact tests

It is natural to ask why one would contemplate an alternative test like Liebermeister's when an exact test is available. The answer is that although the term 'exact test' sounds very appealing, the technical meaning and the dictionary meaning of 'exact' are not the same. Recall that the significance level,  $\alpha$ , of a test is the nominated rejection rate when  $H_0$  is true, and that the Type I error rate,  $P_{H_0}(\text{Type I error})$ , of a test is the rejection rate of  $H_0$  when  $H_0$  is true.

- An 'exact' test guarantees that  $\alpha$  is not exceeded by the Type I error rate of the test, ie  $P_{H_0}(\text{Type I error}) \leq \alpha$  and the test is conservative if  $P_{H_0}(\text{Type I error}) < \alpha$ .
- A test without this guarantee of non-exceedance but for which  $P_{H_0}(\text{Type I error}) \approx \alpha$ , is 'quasi-exact'.

Figure 3 demonstrates this graphically. The step-functions are plots of the Type I error rate against  $\alpha$  for a typical conservative exact test and for a corresponding quasi-exact test. The example used here is a test for equal success probability for samples of size  $m = 20$  and  $n = 10$  when the true success rate is  $p = 0.7$ . The diagonal line is the ideal, exact relationship  $P_{H_0}(\text{Type I error}) = \alpha$ . Other combinations of  $m, n, p$  produce graphs similar in appearance.



**Figure 3** The Type I error rate plotted against nominal significance level  $\alpha$ , for a conservative exact test and also for a corresponding quasi-exact test, for  $m = 20, n = 10, p = 0.7$ .

- The step-function for Fisher's Exact Test in Figure 3 lies completely below the ideal

diagonal but it is a long way from this ideal line. The test is therefore exact in the technical sense at the expense of being excessively conservative for all  $\alpha$ .

- In contrast, the step function for Liebermeister’s ‘quasi-exact’ test, although not wholly below the diagonal, is generally much closer to this line for all  $\alpha$ . It is slightly conservative for some  $\alpha$  but also errs slightly in the other direction for other  $\alpha$ . It is not technically ‘exact’ but is a good approximation to the ideal, in the same sense that the  $\chi^2$  test is a good approximation when  $m$  and  $n$  are large.

### Power comparisons

The StatXact software package (1989) indicates the increase in power of Barnard’s  $p_B$  over Fisher’s Exact Test by comparing the sample sizes needed to achieve 80% power at level  $\alpha = 0.025$  for various choices of alternatives for  $p_1$  and  $p_2$ , when  $m = n$ . The ordering criterion chosen for use in StatXact is the difference in proportions, standardized. The following table shows that the sample sizes required (for alternative  $p_1$  and  $p_2$  used as illustrations in StatXact) are large enough for the  $\chi^2$  test to be appropriate and also to be just as powerful:

Alternative		Common sample size required for 80% power			
$p_1$	$p_2$	$p_F$	$p_B$	$p_L$	$p_{\chi^2}$
0.8	0.55	61	52	54	53
0.7	0.55	175	163	164	163

(Note for example that although  $p_L$  requires  $n = m = 164$  to achieve 80% power when  $p_1 = .7$  and  $p_2 = .55$ , the power at  $n = m = 163$  is 79.8%.)

Of more interest is the case of small (possibly unequal) sample sizes. For such cases, Seneta and Phipps (2001) demonstrate the superiority of Liebermeister’s  $p_L$  on the basis of power, strongly recommending the adoption of  $p_L$ .

### Lower-tail and two-tail tests

A lower-tail test against  $H_1 : p_1 < p_2$  based on the fourfold table in the introduction is obviously identical to an upper tail test with the two rows interchanged. The lower-tail Liebermeister measure is therefore  $p_L(c; z, n, m)$  and clearly  $p_L(c; z, n, m) = 1 - p_L(a; z, m, n)$ .

For two-tail tests in discrete distributions, a variety of approaches is found in the literature, and no agreement about which should be used. The approaches for Fisher's Exact two-tail test (with observed table as in the introduction) include:

- (a) Evaluate  $\min(1, 2P_1)$  where  $P_1$  is the smaller of the two one-tail P-values.
- (b) Adhere strictly to Neyman-Pearson ideology, adding the hypergeometric terms  $\frac{\binom{m}{x}\binom{n}{z-x}}{\binom{m+n}{z}}$  (for fixed  $m, n$  and  $z$ ) which do not exceed  $\frac{\binom{m}{a}\binom{n}{z-a}}{\binom{m+n}{z}}$ . This method is used in many statistical packages and can lead to a result which is more than twice the observed tail.
- (c) Add to the smaller one-tail P-value ( $P_1$  say) the largest opposing tail  $P_2$ , subject to  $P_2 \leq P_1$ . This is often the same answer as (b), (but never exceeds twice the observed tail) and is discussed in Blaker (2000) in the context of confidence intervals.

Our preferred approach for the two-tail Liebermeister measure is (c). Calculations can be conveniently effected using existing software for Fisher's Exact P-value,  $p_F(\cdot)$ , and this procedure follows as an algorithm.

**Algorithm for two-tail version of Liebermeister's measure,  $p_L(Two)$ .**

1. Calculate  $p_L(a; z, m, n) = p_F(a + 1; z + 1, m + 1, n + 1)$ .

- If  $p_L(a; z, m, n) = 0.5$ , then<sup>2</sup> the tails are equal and  $p_L(Two) = 1.0$
- If  $p_L(a; z, m, n) < 0.5$ , this upper tail is the smaller measure, so  $p_{L_1} = p_L(a; z, m, n)$ 
  - To find  $p_{L_2}$ , put  $t = \min(z, n)$ , and **define**  $p_L(t + 1; z, n, m) = 0$ .
  - For  $i = 0, 1, 2, \dots$  (until loop stops), calculate  $p_L(t - i; z, n, m)$ , stopping when  $p_L(t - i; z, n, m) > p_{L_1}$ . This tail is too large, so  $p_{L_2} = p_L(t - i + 1; z, n, m)$
- If  $p_L(a; z, m, n) > 0.5$ , the lower tail,  $p_L(c; z, n, m) = 1 - p_L(a; z, m, n)$ , is the smaller measure, so  $p_{L_1} = 1 - p_L(a; z, m, n)$ 
  - To find  $p_{L_2}$ , put  $t = \min(z, m)$ , and **define**  $p_L(t + 1; z, m, n) = 0$ .

---

<sup>2</sup>Equivalently: If  $a = c$  **and**  $b = d$  then the tails are equal and  $p_L(Two) = 1.0$ . The weaker (perhaps intuitive) condition,  $ad = bc$ , (or equivalently  $\frac{a}{a+b} = \frac{c}{c+d}$ ), does **not** imply that  $p_L(Two) = 1.0$ .

- For  $i = 0, 1, 2, \dots$ , (until loop stops), calculate  $p_L(t - i; z, m, n)$ , stopping when  $p_L(t - i; z, m, n) > p_{L_1}$ . This gives:  $p_{L_2} = p_L(t - i + 1; z, m, n)$

2. Evaluate  $p_L(Two) = p_{L_1} + p_{L_2}$

### Calculations for a numerical example:

Consider the fourfold table: 

5	10	15
1	15	16
6	25	31

, for which  $a = 5, z = 6, m = 15, n = 16$ .

1.  $p_L(5; 6, 15, 16) = p_F(6; 7, 16, 17) = 0.0345446$

- Since  $p_L(5; 6, 15, 16) < 0.5$ , this is the smaller tail, so  $p_{L_1} = 0.0345446$ .
  - $t = \min(z, n) = \min(6, 16) = 6$ , so **define**  $p_L(7; 6, 15, 16) = 0$ .
  - For  $i = 0$ , consider  $p_L(6; 6, 15, 16) = p_F(7; 7, 16, 17) = 0.0045524$ .  
Since  $0.0045524 < p_{L_1}$ , increase  $i$  by 1.
  - For  $i = 1$ , consider  $p_L(5; 6, 15, 16) = p_F(6; 7, 16, 17) = 0.0509039$ .  
Since  $0.0509039 > p_{L_1}$ , the appropriate tail is  $p_{L_2} = p_L(6; 6, 15, 16) = 0.0045524$ .

2. Hence,  $p_L(Two) = p_{L_1} + p_{L_2} = 0.039$ .

### Comparisons for this numerical example:

Method	2-tail Fisher-Exact	2-tail Liebermeister
(a) doubling one tail	0.144	0.069
(b) Neyman-Pearson ideology	0.083	0.039
(c) preferred approach	0.083	0.039

### Acknowledgments

I wish to thank Professor E Seneta for introducing me to this problem and for the many discussions we have had about quasi-exactness.

### References

1. Barnard, G.A., 1945: A new test for  $2 \times 2$  tables. *Nature* **3954**, 177.

2. Blaker, H. 2000: Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics* **28** (4), 783-798.
3. Boschloo R.D., 1970: Raised conditional level of significance for the 2x2-table when testing the equality of two probabilities. *Statistica Neerlandica* **24**, 1-35.
4. Di Sebastiano, P., Fink, T., Di Mola, F.F., Weihe, E., Innocenti, P., Freiss, H. and Büchler, M. 1999: Neuroimmune appendicitis. *The Lancet* **354** (9177), 461-466.
5. Fisher, R.A., 1934: *Statistical Methods for Research Workers*, 5th Ed. Oliver & Boyd, Edinburgh.
6. Hirji, K.F., Tan, S. and Elashoff, R.M.,1991: A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine* **10**, 1137-1153.
7. Lancaster, H.O. 1961: Significance tests in discrete distributions. *Journal of the American Statistical Association* **58**, 223-234.
8. Liebermeister, C., 1877: Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. *Sammlung Klinischer Vorträge* (Innere Medicin No. 31-64) **110**, 935-962.
9. Pierce, D.A. and Peters, D., 1999: Improving on exact tests by approximate conditioning. *Biometrika* **86**, 265-277.
10. Rice, W.R., 1988: A new probability model for determining exact p-values for  $2 \times 2$  contingency tables when comparing binomial proportions. *Biometrics* **44**, 1-22.
11. Sahai, H. and Khurshid, A., 1995: On analysis of epidemiological data involving  $(2 \times 2)$  contingency tables: an overview of Fisher's Exact Test and Yates' correction for continuity. *Journal of biopharmaceutical Statistics* **5**, 43-70
12. Seneta, E. and Phipps, M.C., 2001: On the comparison of two observed frequencies. *Biometrical Journal* **43** (1), 23-43.
13. StatXact, 1989: *A Statistical Package for Exact Nonparametric Inference*, Cytel Software Corporation, Cambridge, MA.
14. Suissa, S. and Shuster, J.J., 1985: Exact unconditional sample sizes for the  $2 \times 2$  binomial trial. *Journal of the Royal Statistical Society. A* **148**, 317-327.