Author/s:
Hepworth, G;Katholi, CR

Title:
Mid-P confidence intervals for group testing based on the total number of positive groups

Date:
2019-05-01

Citation:
Hepworth, G. & Katholi, C. R. (2019). Mid-P confidence intervals for group testing based on the total number of positive groups. Biometrical Journal, 61 (3), pp.688-697. https://doi.org/10.1002/bimj.201700190.

Persistent Link:
https://hdl.handle.net/11343/285249

Mid-P confidence intervals for group testing based on the total number of positive groups

Graham Hepworth[1,*] and Charles Katholi[2]

The University of Melbourne and University of Alabama at Birmingham

Summary

In the estimation of proportions by group testing, unequal sized groups results in an ambiguous ordering of the sample space, which complicates the construction of exact confidence intervals. The total number of positive groups is shown to be a suitable statistic for ordering outcomes, provided its ties are broken by the MLE. We propose an interval estimation method based on this quantity, with a mid-P correction. Coverage is evaluated using group testing problems in plant disease assessment and virus transmission by insect vectors. The proposed method provides good coverage in a range of situations, and compares favourably with existing exact methods.

Key words: coverage; estimation of proportions; exact confidence intervals; group testing; mid-P

* Corresponding author

[1] School of Mathematics and Statistics, The University of Melbourne, Victoria, 3010, Australia.

email: hepworth@unimelb.edu.au

[2] Department of Biostatistics, University of Alabama at Birmingham, 35294, USA

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/bimj.201700190

1 Introduction

Group testing (or pooled testing) occurs when individuals are pooled together and tested as a group for the presence of an attribute. A positive test result indicates that at least one individual in the group is positive. The purpose of group testing may be to identify the positive individuals, as introduced by Dorfman (1943) in relation to blood testing, or to estimate the proportion p of individuals in the population with the attribute. Our focus is on estimation of p, as that has been of primary interest in the practical situations the authors have encountered. Group testing for estimation has been applied in many areas other than blood testing, including plant disease (e.g. Fletcher et al. 1999), prevalence of mycobacteria in rodents (Durnez et al. 2008), transmission of disease by insect vectors (e.g. Crockett et al. 2012), and prevalence of parasitic worms in fecal samples (Mitchell and Pagano 2012). The obvious advantage of group testing, especially when p is small, is the potential for considerable saving of resources.

This paper is concerned with interval estimation of p based on exact methods, when groups are of unequal size. Some estimation studies employ a single group size k; the reason may be administrative convenience, compatibility with testing procedures (e.g. equipment may be set up to test a fixed number of units), or desire for simplicity of calculation. But in many situations, equal group sizes is not a practical option. For example, in studying the transmission of West Nile Virus by mosquitoes, the natural group is a trap, in which the number of units (mosquitoes) collected can range greatly (Biggerstaff 2008).

In other situations, several group sizes are used deliberately because of considerable uncertainty about the prevalence. If a single group size is used and it is too large, there is a high probability that all groups will be positive, which is uninformative (see Hepworth and Watson (2009) for a discussion of this problem). If a single group size is too small,

there is a high probability that all groups will be negative, unless a very large number of groups are tested, thus increasing the expense and thereby calling into question the point of employing group testing. A range of group sizes increases the chance of being in the required zone of information. This was recognized in very early group testing work; for example, Chiang and Reeves (1962) recommended dividing the units into "two groups of pools with the pool size in one group being markedly different from that in the other, thus increasing the chance of getting a sufficient number of negative and positive pools in at least one of the two groups". Variations on this basic design have been proposed and used since. For example, Gu et al. (2004) examined a mosquito testing scheme with six group sizes $(5, 10, 20, 30, 40, 50)$ and found in simulations that it needed only half as many units as a scheme with a single group size ($k = 50$).

Another way of producing groups of unequal size (potentially, at least) is to use a sequential (or adaptive) testing scheme with each stage having a different group size. Hughes-Oliver and Swallow (1994) suggested a two-stage procedure in which the group size at the second stage is determined by the maximum likelihood estimate (MLE) of p from the first stage. Rodoni et al. (1994) employed a three-stage sequential design with group sizes of 25, 5 and 1, in which the next stage was used if all groups at the previous stage were positive. Hardwick (1998), in the setting of reliability testing, proposed a multi-stage adaptive procedure in which at each stage either a group or an individual is tested, depending on whether the estimate of p exceeds a cut-off. Liu et al. (2012), in discussing adaptive designs, concluded that "it is desirable to develop procedures that are less sensitive to the specified values of the unknown parameters".

If groups are of equal size, an exact confidence interval (CI) for p can be found by calculating an exact interval for the proportion of positive groups θ in the population, and applying the monotone mapping that relates p to θ (Tebbs and Bilder 2004). In this case there is no ambiguity in the ordering of outcomes in the sample space; a larger number

of positive groups gives stronger evidence against a smaller hypothesized value of p. But for groups of unequal size the multidimensional sample space results in ambiguity, and an ordering of outcomes must be determined for exact intervals to be derived. Hepworth (1996) developed exact CIs for p based on ordering the MLEs ("MLE ordering"), and showed that they compared favorably with intervals based on ordering outcomes by their probability (the method for discrete distributions proposed by Sterne (1954)). A mid-P correction was found to give coverage closer to the nominal level, though at the expense of strict conservatism of intervals. Hepworth (2004) proposed an exact method based on ordering outcomes according to a likelihood ratio statistic, and with a mid-P correction, found its coverage to be even closer to the nominal level than MLE ordering. The likelihood ratio method was, however, more complicated, and allowed the possibility of disjoint sets. Tebbs and Bilder (2004) applied the technique of Blaker (2000) to group testing, but considered only equal group sizes. Dres et al. (2015) extended Blaker's method to unequal group sizes, and found it to have good coverage properties.

In general, the greater the number of different group sizes, the more large and complicated the sample space, and the more intensive the computation. It would be advantageous to simplify the sample space, especially in situations involving a large number of outcomes. One way to achieve this is to use as a statistic the total number of positive groups across all group sizes. Although it provides less information than the number of positive groups for each group size, it produces a simpler ordering of outcomes. In this paper we examine its performance in exact interval estimation with a mid-P correction, and compare it with the performance of MLE ordering. Dres et al. (2015) evaluated nine exact methods, three of which involved the total number of positive groups in the ordering. We extend their work by including a mid-P correction, which changes the framework of evaluation away from a strictly conservative one. We also provide methodological motivation for using this statistic, in contrast to their primarily empirical

work. Finally, we apply the methods to much larger group testing examples than theirs.

The paper begins with a statement of the problem, assumptions made, and key quantities defined. We then discuss the search for a sufficient statistic and a locally most powerful test. On theoretical grounds, the total number of positive groups is shown to be a suitable choice of statistic for ordering outcomes to find exact confidence intervals. This statistic (T) and the MLE are assessed using a small group testing example, and it is found that intervals based on T are improved by breaking ties using the MLE. Based on coverage, the proposed method is found to be promising, and very similar to ordering by the MLE. Three larger group testing problems are then used to further evaluate the method, with excellent results.

2 Inference on p

2.1 Assumptions and notation

The following assumptions are made about group testing for estimation in this paper:

(a) The outcomes for the units in each group follow independent and identically distributed Bernoulli(p) distributions.

(b) The units are randomly assigned to groups.

(c) The testing is conducted without error.

These assumptions were made in all the interval estimation studies referenced above, and they are very reasonable in many situations. Tebbs and Bilder (2004) discussed them in some detail, and Hepworth (2005) showed their validity in the context of testing plants in groups for viruses. Assumption (c) has been relaxed in some group testing studies (e.g. Tu, Litvak, and Pagano, 1995), while others have considered both perfect and imperfect testing (Toribio and Sergeant, 2007). Assumption (b) has been relaxed in work on group testing regression models (e.g. Bilder and Tebbs 2009).

Suppose that there are m different group sizes, and that for $i = 1, \ldots, m$, n_i groups of size k_i are tested. If the random variable $X_i = x_i$ represents the number of positive groups, X_i follows a binomial distribution with parameters n_i and $1 - (1-p)^{k_i}$, and the combined likelihood is

$$L(\boldsymbol{x}; p) = \prod_{i=1}^{m} \binom{n_i}{x_i} (1 - (1-p)^{k_i})^{x_i} (1-p)^{k_i(n_i - x_i)} \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)$. The MLE of p, $\hat{p}$, is obtained by solving the score equation

$$S(\boldsymbol{x}; p) = \frac{1}{1-p} \sum_{i=1}^{m} \left(\frac{k_i x_i}{1 - (1-p)^{k_i}} - k_i n_i \right) = 0\,. \tag{2}$$

Let $T = \sum_{i=1}^{m} X_i$ be the random variable representing the total number of positive groups across all group sizes, and let $t = \sum_{i=1}^{m} x_i$ be a realization of T. Let $n = \sum_{i=1}^{m} n_i$ be the total number of groups tested.

2.2 Sufficiency

In considering inference on p, it is good practice to look for a minimal sufficient statistic. If the likelihood ratio for two observation vectors $\boldsymbol{y}$ and $\boldsymbol{x}$ is independent of the parameter, i.e.

$$\frac{L(\boldsymbol{y}; p)}{L(\boldsymbol{x}; p)} = h(\boldsymbol{y}, \boldsymbol{x}) \quad \text{for all } p,$$

then $\boldsymbol{x}$ and $\boldsymbol{y}$ will have the same value of the minimal sufficient statistic (Cox and Hinkley 1974, p. 24). The likelihood ratio is

$$\frac{L(\boldsymbol{y}; p)}{L(\boldsymbol{x}; p)} = \frac{\prod_{i=1}^{m} \binom{n_i}{y_i}}{\prod_{i=1}^{m} \binom{n_i}{x_i}} \prod_{i=1}^{m} \left[\frac{1 - (1-p)^{k_i}}{(1-p)^{k_i}} \right]^{y_i - x_i}. \tag{3}$$

Taking logs of (3) and differentiating with respect to p gives

$$\sum_{i=1}^{m} (y_i - x_i) \frac{k_i}{(1-p)\,(1 - (1-p)^{k_i})} \tag{4}$$

which if equated to zero is equivalent to writing (3) as $h(\boldsymbol{y}, \boldsymbol{x})$, independent of p. Expression (4) is equal to zero only if x_i and y_i are equal for all i. As a result, the minimal

sufficient statistic must be simply the observations themselves, $\boldsymbol{x}$. Therefore T is not sufficient except when $m = 1$ (equal group sizes).

One major class of distributions which admits sufficient statistics of dimension less than that of $\boldsymbol{x}$ is the exponential family (Lehmann 1986, p. 57). Although the binomial distribution is well known to belong to the exponential family, we show now that the likelihood given in (1) does not in general belong to this class. The probability density function for a random variable $\boldsymbol{X}$ with parameter p belongs to an exponential family if it can be written

$$f(\boldsymbol{x}; p) = \exp\left[a(p)\, b(\boldsymbol{x}) + c(p) + d(\boldsymbol{x})\right]$$

(see, for example, Kendall et al., p. 618). Expression (1) can be written

$$\exp\left[\sum_{i=1}^{m} \log\binom{n_i}{x_i} + \sum_{i=1}^{m} x_i \log\left[\frac{(1-(1-p)^{k_i}}{(1-p)^{k_i}}\right] + \sum_{i=1}^{m} k_i n_i \log(1-p)\right].$$

The first expression in the square brackets is a function of the x_is only, and so can be written $d(\boldsymbol{x})$. The third expression is a function of p only, and so matches $c(p)$. The middle expression, however, cannot be written as $a(p)\, b(\boldsymbol{x})$, except when $m = 1$. So in general, $f(\boldsymbol{x}; p)$ does not belong to an exponential family.

2.3 Most powerful tests

Although a sufficient statistic of dimension less than m is not available, a likelihood ratio test for this problem can still be derived. Hepworth (2004) showed that the hypothesis H_0: $p = p_0$ is rejected in favour of H_1: $p = p_1 > p_0$ if

$$\log\left(\frac{L(\boldsymbol{x}; p_0)}{L(\boldsymbol{x}; p_1)}\right) = \sum_{i=1}^{m} x_i \log\left(\frac{(1-(1-p_0)^{k_i})(1-p_1)^{k_i}}{(1-(1-p_1)^{k_i})(1-p_0)^{k_i}}\right) + \sum_{i=1}^{m} n_i k_i \log\left(\frac{1-p_0}{1-p_1}\right) < c_1$$

which is effectively

$$\sum_{i=1}^{m} x_i \log\left(\frac{\dfrac{1}{(1-p_0)^{k_i}} - 1}{\dfrac{1}{(1-p_1)^{k_i}} - 1}\right) < c_2\,.$$

It follows that a general form of the test involves a weighted sum of the number of positive groups for each group size. Except in the case $m = 1$, the weights depend on p_1, and so the test is not uniformly most powerful. For each pool size, the weight is the log-odds ratio of the proportion of positive pools under H_0 to that under H_1. Gao et al. (2014) examined the finite and large-sample properties of this test statistic, and compared its performance with a test based on the score and an exact test based on T.

A reasonable next step is to look for a locally most powerful (LMP) test. To achieve maximum power for local alternatives to H_0 of the form $p = p_0 + \delta$, the appropriate test statistic is the score evaluated at p_0, with large values being significant for $\delta > 0$ (Cox and Hinkley 1974, p. 113). For small p and small to moderate k_i, $1 - (1 - p)^{k_i} \approx k_i p$ and so from (2) an LMP test would be approximately of the form

$$T = \sum_{i=1}^{m} X_i > c\,. \tag{5}$$

It is therefore reasonable to construct exact CIs by inverting this test, i.e. by using $T = t$ to order the outcomes. T has the monotone likelihood ratio property with respect to p (Gao et al. 2016). Gao et al. (2014) performed a large simulation with group sizes generated from a discrete uniform distribution over $[25, 50]$ with a test of $H_0 : p = 0.0005$ vs $H_1 : p < 0.0005$. They found the test involving T to perform best with regard to type I error among the five tests they considered, and recommended it as the best choice for small samples.

The closeness of the approximation of the LMP test to (5) depends on how variable the $k_i s$ are. For example, with similar group sizes of $k_1 = 20$ and $k_2 = 25$, $p = 0.01$ results in

$$\sum \frac{k_i x_i}{1 - (1 - p)^{k_i}} \approx 110 x_1 + 113 x_2$$

which strongly supports the choice of T as the test statistic. For $p = 0.1$, the result is $22.8x_1 + 26.9x_2$, for which the choice is not as clear-cut, but it is still reasonable, especially

given that $p = 0.1$ is a large prevalence for most circumstances in which group testing is employed, and therefore unlikely to occur very often. Even for very dissimilar group sizes of $k_1 = 30$ and $k_2 = 5$, $p = 0.01$ results in $115x_1 + 102x_2$, which again suggests T as a reasonable choice. It is only when widely variable group sizes and larger p are combined that T may not be the best choice. Overall, given its simplicity, T is worth assessing as a statistic for ordering outcomes in constructing exact CIs.

3 Construction of confidence intervals

To construct $1 - \alpha$ exact CIs for p, we define the following sum of probabilities:

$$Q(\boldsymbol{G}; p) = \sum_{\boldsymbol{X} \in \boldsymbol{G}} \prod_{i=1}^{m} \binom{n_i}{x_i} (1 - (1-p)^{k_i})^{x_i} (1-p)^{k_i(n_i - x_i)} \tag{6}$$

where $\boldsymbol{G}$ is the rejection region determined by the test statistic. The selected interval estimation method then adopts Q in specific inequalities involving the test statistic (Dres et al., 2015). For MLE ordering, the statistic is $\hat{p}$, and the lower (upper) confidence limit is found by solving $Q(\boldsymbol{G}; p) = \frac{1}{2}\alpha$, where G consists of all outcomes with MLE at least as large (small) as $\hat{p}$. If the statistic is T, G consists of all outcomes for which $T \geq t$ or for which $T \leq t$ respectively.

Both of these are "twice the smaller tail" intervals (Hirji 2006, p. 59), and tend to be highly conservative, giving coverage well above the nominal level for some p. To modify this conservatism, a mid-P correction can be applied, in which the probability of the observed outcome in (6) is halved in the summation on each side. Mid-P CIs have been recommended in several group testing studies (e.g. Hepworth, 1996), the only issue being whether there is a requirement for strict conservatism (i.e. always attaining at least the nominal level). Berry and Armitage (1995) argued that the significance level from a discrete variable should closely share the properties of one based on a normally distribued variable. They showed that the mid-P value satisfies this, with an expectation of

$\frac{1}{2}$ and a slightly smaller variance than a variable uniformly distributed between 0 and 1. Agresti and Gottard (2007) reviewed the mid-P approach, confirming its advantages in a range of discrete problems, and labelling it a "quasi-exact approach". For this study we henceforth adopt the mid-P correction for the exact methods considered.

3.1 Small example: plant disease assessment

We now illustrate the construction of CIs using a small group testing study described by Hepworth and Watson (2009). The aim of that study was to estimate the prevalence of viruses in a carnation population from which 200 plants were sampled. The plants were tested in 8 groups of 20 and 8 groups of 5. The first two main sections of Table 1 show 95% confidence intervals calculated by using either statistic to order outcomes, for a range of outcomes selected to give a spread of values of $\hat{p}$. Most of them have $x_1 > x_2$ to avoid outcomes which are very unlikely.

TABLE 1 ABOUT HERE

The intervals constructed using T are generally wider than those constructed using $\hat{p}$, and unpredictable in regard to being shifted left or right. One aspect of ordering by T which cannot occur with ordering by $\hat{p}$, is to have identical CIs for different outcomes. For example, both $(3, 7)$ and $(6, 4)$ have the same interval because both result in $T = 10$. This phenomenon is more pronounced in small problems such as this, where dissimilar group sizes give rise to outcomes with equal T having quite discrepant MLEs. It lends support to our comment above that T may not be the best choice for widely variable k.

3.2 Breaking ties

Even when group sizes are not highly variable, and p is small, it is generally desirable to form a less discrete sample space by breaking ties. Chan and Zhang (1999) did this suc-

cessfully in finding exact CIs for the difference of two binomial proportions, by dividing the MLE of the difference by its estimated variance. For group testing problems, Dres et al. (2015) suggested using $\hat{p}$ to break ties for outcomes with equal T. Because T is an integer and $0 \leq p \leq 1$, this effectively uses $T + \hat{p}$ as the test statistic. They considered only strictly conservative intervals, and so did not employ a mid-P correction. This meant that the breaking of ties always produced intervals as least as short as T on its own, and as a result, with coverage at least as close to the nominal level. This is no longer the case when the mid-P correction is employed; the primary effect of breaking ties is to produce a better ordering, though it may also result in narrower intervals.

The bottom section of Table 1 shows 95% CIs calculated by using $T+\hat{p}$ as the test statistic to order outcomes. For small p the intervals are very similar to those constructed using $\hat{p}$. In some cases the two statistics give almost identical CIs because the ordering is almost the same, and any differences in the ordering involve outcomes with very small probability. For larger p, intervals found using $T + \hat{p}$ are at times wider and shifted to the right compared to intervals constructed using $\hat{p}$, but less so than those constructed using T alone. For $T + \hat{p}$ there are of course no longer any identical CIs for different outcomes.

4 Evaluation of confidence intervals

4.1 Evaluation criteria

Our main assessment criterion in comparing interval estimation methods is coverage probability, because it concerns the very nature and definition of a CI. The coverage probability is given by

$$\sum_{\boldsymbol{X}} c(\boldsymbol{x}; p) L(\boldsymbol{x}; p)$$

where L is as defined in (1), and $c(\boldsymbol{x}; p) = 1$ if p is included in the CI for outcome $\boldsymbol{x}$, and 0 otherwise. We see interval width as a secondary criterion, to be used if choosing

between methods of similar coverage. Other desirable properties of a method include the avoidance of having to deal with disjoint sets in constructing an interval, and different CIs for different outcomes, the latter of which is violated by using T alone as the test statistic. Disjoint sets are not possible with any of the test statistics considered, but they are with others such as the exact likelihood ratio (Hepworth, 2004), even if they occur for outcomes which are not very probable.

Although p theoretically ranges from 0 to 1, it needs to be small for group testing to be of practical value. Rather than evaluating coverage for the full range of p, it is better to concentrate on values consistent with the design of the procedure. Hepworth and Watson (2009), in studying bias of estimators, restricted p to values for which the probability of all positive groups is no more than 0.05, because this outcome is highly uninformative, and best avoided. We adopt this cut-off, which we denote ψ, as an upper bound on p for purposes of evaluation. ψ also equals the lower 95% confidence limit for p when all groups are positive. For the plant disease example described in Section 3.1, $\psi = 0.211$.

4.2 Coverage comparisons

The coverage probability was calculated for 1000 equally spaced values of $p \leq 1.1\psi$ and plotted against p for each ordering method in Figure 1. Some features are common to all three plots: the discontinuities, which occur at the lower and upper confidence limits for each of the 81 outcomes; the excessive coverage for very small p, very common for exact methods near the boundary of the parameter space; and the excessive coverage for $p > \psi = 0.211$, which continues as p increases (this feature of group testing is explained by Hepworth (2004)).

FIGURE 1 ABOUT HERE

Except for very small p, the coverage for ordering by the MLE ($\hat{p}$) alone or by $T + \hat{p}$ is satisfactory for a small group testing problem, and almost the same for the two methods. It fluctuates around the nominal 0.95 level, which is expected and acceptable for methods using the mid-P correction. The coverage for ordering by T alone has larger fluctuations and more spikes, though it might still be quite acceptable to practitioners.

Table 2 gives summary statistics for the coverage probability for each method, for $p \leq \psi$. MAD denotes the median absolute deviation about 0.95, and $p_{0.95}$ denotes the proportion of points with coverage probability below 0.95. The mean is not presented as it is consistently very close to the median. Obviously it is desired that the MAD be as small as possible, and we consider it to be the most useful single indicator of the adequacy of coverage of a CI method.

TABLE 2 ABOUT HERE

The summary statistics in Table 2 confirm our impressions from the plots. For ordering by $\hat{p}$ alone or by $T + \hat{p}$, the coverage is acceptable, especially for a small problem. The two methods give virtually identical coverage, consistent with the similar ordering. Although a few of the CIs differ noticeably, most of these intervals are for unlikely outcomes, and so they have minimal effect on the coverage probability. For ordering by T alone, the larger MAD reflects the greater fluctuations observed in the plot.

A reasonable practical summary from the results in Table 2 would be along the lines of the following statement: "A nominal 95% CI is likely to be closer to an actual 95.5% CI, and at least half the time will be within 1% of 95%. It will not drop below 92% and will be above 95% about $\frac{2}{3}$ of the time". In using T as a test statistic to determine ordering of outcomes, it is clearly beneficial to break ties by adding $\hat{p}$. For our evaluation using larger group testing problems, we will no longer consider T on its own, but will restrict our attention to $T + \hat{p}$.

4.3 Larger examples

We now examine the coverage probability for three larger group testing problems, listed in Table 3. The first two are artificial, but realistic in the context of transmission of viruses by insect vectors. They involve three different group sizes and more groups than the example above. Problem 1 involves group sizes which are similar to each other, and Problem 2 has widely varying group sizes. Problem 3 comes from a study of yellow fever virus transmitted by mosquitoes, described by Walter et al. (1980). It has very large group sizes and widely varying numbers of groups, with several of the n_is equal to 1. The three problems have widely varying numbers of outcomes, as shown in Table 3.

TABLE 3 ABOUT HERE

Figure 2 shows plots of the coverage probability vs p ($p \leq 1.1\psi$) for the three problems. Table 4 shows the summary statistics for $p \leq \psi$. The corresponding plots and summary statistics for ordering by $\hat{p}$ alone are extremely similar, and are not shown.

FIGURE 2 and TABLE 4 ABOUT HERE

For all three problems, the coverage probability is very close to 0.95 for most of the range of p considered, and in the case of problem 1, extremely so. Compared to the small plant disease example considered above, the coverage has tightened markedly, due to the increased number of groups (effectively, the sample size). The quartiles are closer to the median, and the MAD is very small — to have coverage well within 0.5% of the nominal level at least half the time, would likely be acceptable to most practitioners.

The outstanding performance of the intervals for problem 1 is consistent with the small variability in group sizes. The very good performance for problem 2 is consistent with the large number of groups, which substantially compensates for the large variability in group sizes. The very good performance for problem 3 (the real scenario) is consistent

with the relatively small variability in group sizes, which compensates for the modest number of groups ($n = 23$). For example, if we consider $p = 0.01$ (close to the midpoint of $(0, \psi)$), $\sum \left[(k_i x_i/(1-(1-p)^{k_i})\right] = 145x_1 + 165x_2 + 194x_3$, which suggests T as a reasonable ordering statistic for construction of intervals.

The only summary statistic which appears worse for the larger examples is $p_{0.95}$, which is close to 0.5 for all three problems. This is inevitable when the coverage converges to the nominal level, which is happening to a considerable degree here. Another aspect of the coverage which is less than desirable, whether the problem be large or small, is the centering of the coverage probability on 0.975 for very small p. Blaker (2000) pointed out that Clopper-Pearson intervals have actual coverage of $1 - \alpha/2$ rather than $1 - \alpha$ for p near 0 or 1. The CIs we have constructed are all versions of the Clopper-Pearson method, even with the mid-P correction, so it is not surprising that they follow suit in this aspect. The point at which the coverage suddenly drops, and then recovers towards 0.95, occurs when p equals the upper confidence limit for the outcome of zero positive groups.

Overall, this evaluation using larger examples confirms the benefits of using $T + \hat{p}$ as a statistic for ordering outcomes, and strengthens the argument for using it for exact or quasi-exact CI calculation in a wide range of group testing problems.

5 Discussion

Group testing for estimation of a population proportion with groups of unequal size results in an ambiguous ordering of the sample space. This requires that a method of ordering outcomes be determined in order to derive exact confidence intervals for the proportion. We have proposed the total number of positive groups as a conceptually simple and theoretically sound statistic to use in ordering outcomes. On its own this statistic produces intervals which are too conservative because of the large number of ties, but

these can be broken by adding the MLE. We have shown that the coverage produced by the resulting confidence intervals is acceptable for small group testing problems, and very good for larger problems.

Our proposed method also avoids the issues which arise with some exact interval estimation methods. In particular, it cannot produce disjoint sets, and the breaking of ties prevents the same confidence interval occurring for different outcomes. Both of these properties make it more acceptable to practitioners. In performing an evaluation of our method, we did not use expected interval width as a criterion, because we did not have to choose between methods with similar coverage.

All the methods we evaluated incorporated a mid-P correction in the construction of intervals. If there was a requirement for strict conservatism of intervals, the method of Blaker (2000) would be a good option, as Dres et al. (2015) showed. Blaker intervals are more prone to asymmetry of coverage than the methods we examined here, which may be seen as a disadvantage.

In implementing our proposed method, computation speed can be an issue with larger group testing problems. Dres et al. (2015) found the computation time of Blaker's method to be satisfactory as long as the number of outcomes in the sample space did not exceed about 10,000. This is roughly what we also found with the methods we evaluated. Note that an individual confidence interval is not slow to compute; it is only when intervals need to be found for all outcomes (e.g. for examining coverage) that computation speed can be an issue.

For very large problems, where the number of outcomes makes exact methods prohibitive, reasonable asymptotic methods are available, such as the skew-corrected score method, which Hepworth (2005) found to have satisfactory properties.

References

Agresti, A. and Gottard, A. (2007) Nonconservative exact small-sample inference for discrete data. *Computational Statistics & Data Analysis* 51, 6447–6458.

Berry, G. and Armitage, P. (1995) Mid-P confidence intervals: a brief review. *The Statistician* 44, 417–423.

Biggerstaff, B.J. (2008) Confidence intervals for the difference of two proportions estimated from pooled samples. *Journal of Agricultural, Biological, and Environmental Statistics* 13, 478–497.

Bilder, C.R. and Tebbs, J.M. (2009) Bias, efficiency, and agreement for group-testing regression models," *Journal of Statistical Computation and Simulation* 79, 67–80.

Blaker, H. (2000) Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* 28, 783–798.

Chan, I.S. and Zhang, Z. (1999) Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* 55, 1202–1209.

Chiang, C.L. and Reeves, W.C. (1962) Statistical estimation of virus infection rates in mosquito vector populations. *American Journal of Hygiene* 75, 377–391.

Cox, D.R. and Hinkley, D.V. (1974) Theoretical Statistics. London: Chapman and Hall.

Crockett, R.K., Burkhalter, K., Mead, D., Kelly, R., Brown, J., Varnado, W., Roy, A., Horiuchi, K., Biggerstaff, B.J., Miller, B. and NasciCulex, R. (2012) Culex flavivirus and West Nile Virus in *Culex quinquefasciatus* populations in the southeastern United States. *Journal of Medical Entomology* 49, 165–174.

Dorfman, R. (1943) The detection of defective members of large populations. *Annals of Mathematical Statistics* 14, 436–440.

Dres, K.A., Hepworth, G. and Watson, R.K. (2015) Exact confidence intervals for proporitons estimated by group testing with different group sizes. *Australian & New Zealand Journal of Statistics* 57, 510–516.

Durnez, L., Eddyani, M., Mgode, G.F., Katakweba, A., Katholi, C.R., Machang'u, R.R., Kazwala, R.R., Portaeis, F. and Leirs, H. (2009) First detection of mycobacteria in African rodents and insectivores, using stratified pool screening. *Applied and Environmental Microbiology* 74, 768–773.

Fletcher, J.D., Russell, A.C. and Butler, R.C. (1999) Seed-borne cucumber mosaic virus in New Zealand lentil crops: yield effects and disease incidence. *New Zealand Journal of Crop and Horticultural Science* 27, 7–204.

Gao, H., Aban, I.B. and Katholi, C.R. (2014) Properties of a one-sided likelihood ratio test as applied to pool screening. *Communications in Statistics – Theory and Methods* 43, 4546–4565.

Gao, H., Aban, I.B. and Katholi, C.R. (2016) Pool screening: an example of independent non-identical Bernoulli trial. *Communications in Statistics – Simulation and Computation* 45, 3307–3316.

Gu, W., Lampman, R. and Novak, R.J. (2004) Assessment of arbovirus vector infection rates using variable size pooling. *Medical and Veterinary Entomology* 18, 299–204.

Hardwick, J., Page, C. and Stout, Q.F. (1998) Sequentially deciding between two experiments for estimating a common success probability. *Journal of the American Statistical Association* 93, 1502–1511.

Hepworth, G. (1996) Exact confidence intervals for proportions estimated by group testing,. *Biometrics* 52, 1134–1146.

Hepworth, G. (2004) Mid-P confidence intervals based on the likelihood ratio for proportions estimated by group testing. *Australian and New Zealand Journal of Statistics* 46, 391–405.

Hepworth, G. (2005) Confidence intervals for proportions estimated by group testing with groups of unequal size. *Journal of Agricultural, Biological, and Environmental Statistics* 10, 478–497.

Hepworth, G. and Watson, R. (2009) Debiased estimation of proportions in group testing. *JRSS-C* 58, 105–121.

Hirji, K.F. (2006) Exact Analysis of Discrete Data. New York: Chapman and Hall.

Hughes-Oliver, J.M. and Swallow, W.H. (1994) A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association* 89, 982–993.

Kendall, M.G., Stuart, A. and Ord, J.K. (1991) Kendall's Advanced Theory of Statistics (Fifth Edition), Volume 2: Classical Inference and Relationship. London: Edward Arnold.

Lehmann, E.L. (1986) Testing Statistical Hypotheses. Pacific Grove, California: Wadsworth.

Liu, A., Liu, C., Zhang, Z. and Albert, P.S. (2012) Optimality of group testing in the presence of misclassification. *Biometrika* 99, 245–251.

Mitchell S. and Pagano, M. (2012) Pooled testing for effective estimation of the prevalence of *Schistosoma mansoni*. *American Journal of Tropical Medicine and Hygiene* 87, 850–861.

Rodoni, B.C., Hepworth, G., Richardson, C. and Moran, J.R. (1994) The use of a sequential batch testing procedure and ELISA to determine the incidence of five viruses in Victorian cut-flower Sim carnations. *Australian Journal of Agricultural Research* 45, 223–230.

Sterne, T.E. (1954). Some remarks on confidence or fiducial limits. *Biometrika* 41, 275–278.

Tebbs, J.M. and Bilder, C.R. (2004) Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. *Journal of Agricultural, Biological, and Environmental Statistics* 9, 75–90.

Toribio, J.A. and Sergeant, E.S. (2007) A comparison of methods to estimate the prevalence of ovine Johnes infection from pooled faecal samples. *Australian Veterinary Journal* 85, 317–324.

Tu, X.M., Litvak, E. and Pagano, M. (1995) On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. *Biometrika* 82, 287–297.

Walter, S.D., Hildreth, S.W. and Beaty, B.J. (1980) Estimation of infection rates in populations of organisms using pools of variable size. *American Journal of Epidemiology* 112, 124–128.

Table 1: 95% mid-P confidence intervals constructed by ordering outcomes by either the MLE ($\hat{p}$), the total number of positive groups (T), or $T + \hat{p}$, for selected outcomes of a procedure testing 8 groups of 20 and 8 groups of 5

		Number of positive groups (x_1, x_2)									
		$(0,0)$	$(1,0)$	$(2,1)$	$(4,0)$	$(5,1)$	$(3,7)$	$(6,4)$	$(7,5)$	$(8,5)$	$(8,7)$
Ordering	T	0	1	3	4	6	10	10	12	13	15
method	$\hat{p}$	0	0.005	0.017	0.025	0.042	0.067	0.085	0.128	0.194	0.341
$\hat{p}$	lower	0	0.000	0.004	0.009	0.017	0.030	0.041	0.059	0.091	0.147
	upper	0.015	0.026	0.043	0.060	0.085	0.123	0.157	0.221	0.363	0.638
T	lower	0	0.000	0.004	0.007	0.016	0.044	0.044	0.068	0.086	0.146
	upper	0.015	0.026	0.046	0.057	0.085	0.185	0.185	0.288	0.362	0.638
$T + \hat{p}$	lower	0	0.000	0.004	0.009	0.017	0.040	0.042	0.067	0.090	0.147
	upper	0.015	0.026	0.043	0.060	0.085	0.161	0.163	0.252	0.362	0.638

Table 2: Summary statistics for coverage probability of 95% mid-P confidence intervals constructed by ordering outcomes by either the MLE ($\hat{p}$), the total number of positive groups (T), or $T + \hat{p}$, for a procedure testing 8 groups of 20 and 8 groups of 5, $p \leq \psi = 0.211$

Ordering method	Minimum	Lower quartile	Median	Upper quartile	Maximum	MAD	$p_{0.95}$
$\hat{p}$	0.923	0.944	0.954	0.962	0.996	0.0087	0.373
T	0.919	0.948	0.956	0.968	0.996	0.0130	0.305
$T + \hat{p}$	0.923	0.944	0.953	0.962	0.996	0.0089	0.360

Table 3: Three group testing problems involving n_i groups of size k_i

	n_1	k_1	n_2	k_2	n_3	k_3	n_4	k_4	n_5	k_5	no. outcomes
Problem 1	10	30	10	25	10	20					1331
Problem 2	20	10	15	30	10	50					3696
Problem 3	3	80	12	100	1	103	2	111	1	115	
	n_6	k_6	n_7	k_7	n_8	k_8	n_9	k_9			
	1	116	1	123	1	150	1	152			9984

Table 4: Summary statistics for coverage probability of 95% mid-P confidence intervals constructed by ordering outcomes by $T + \hat{p}$, for three group testing problems, as listed in Table 3, $p \leq \psi$

Problem	ψ	Minimum	Lower quartile	Median	Upper quartile	Maximum	MAD	$p_{0.95}$
1	0.093	0.923	0.949	0.950	0.951	0.990	0.0012	0.512
2	0.180	0.932	0.949	0.950	0.956	0.979	0.0033	0.460
3	0.019	0.926	0.947	0.950	0.955	0.993	0.0037	0.480

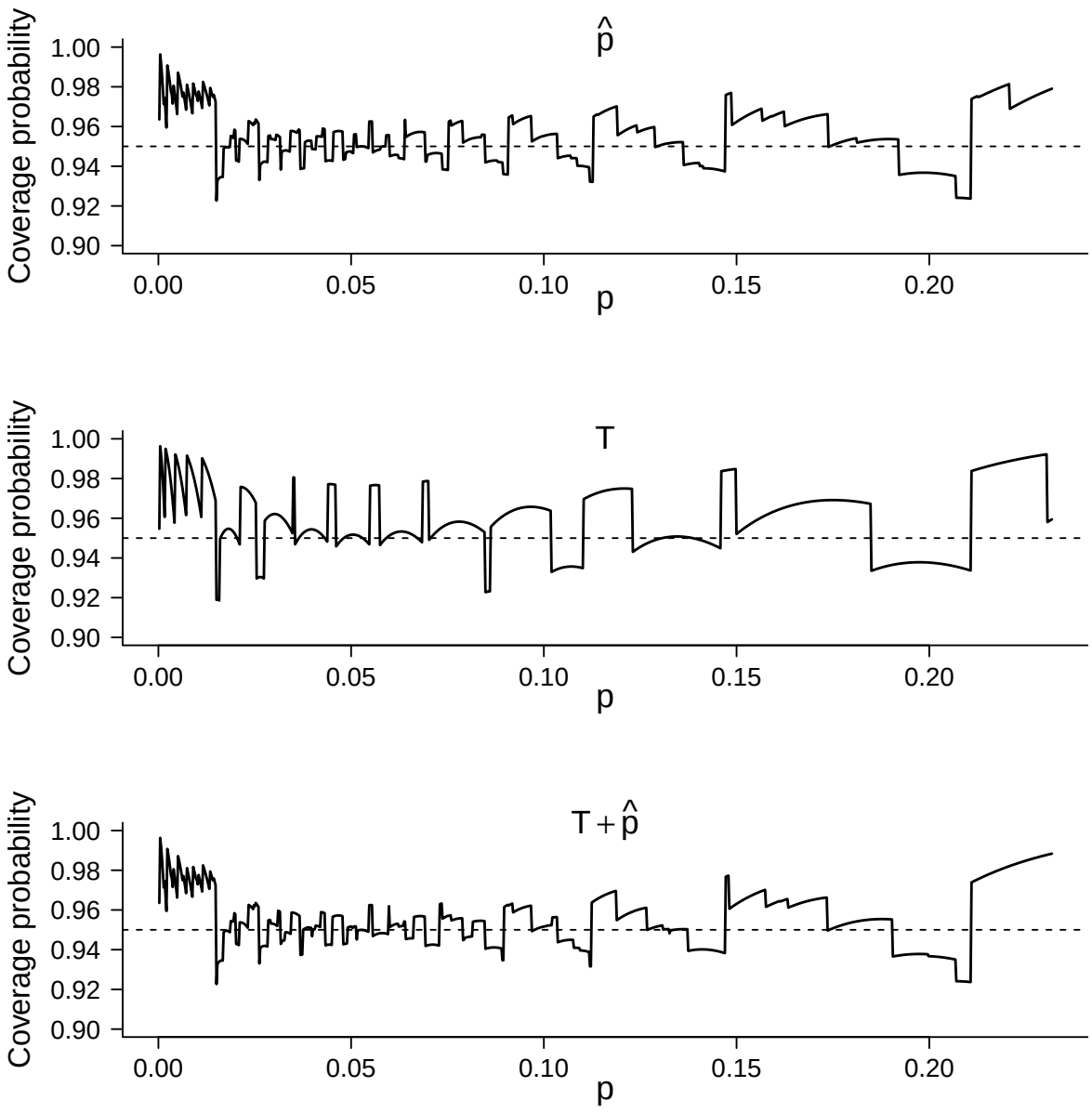

Figure 1: Coverage probability for 95% mid-P confidence intervals for three methods of ordering outcomes, for a procedure testing 8 groups of 20 and 8 groups of 5.

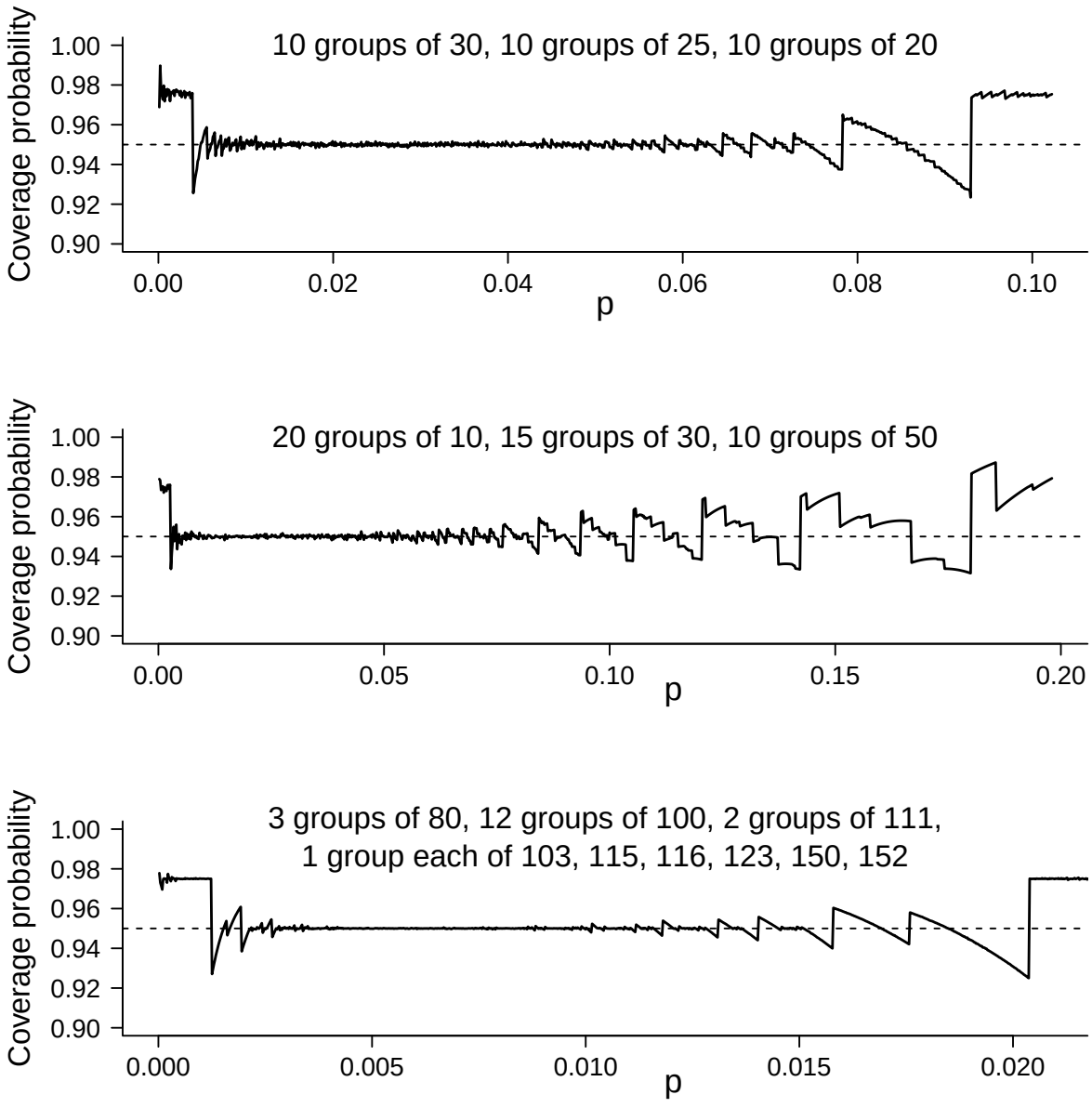

Figure 2: Coverage probability for 95% mid-P confidence intervals constructed by ordering outcomes by $T + \hat{p}$, for three group testing problems, as listed in Table 3.