Improved Estimation of Proportions Using Inverse Binomial Group Testing

Abstract

Inverse sampling for proportions is useful when there is a need to estimate the prevalence of a disease without delay. This can be combined with group (pooled) testing, in which individuals are pooled together and tested as a group for the disease. Pritchard and Tebbs (2011a) introduced this combination to the statistical literature, and we have addressed some of the key problems raised, for groups of equal size. Most point estimators of the proportion are biased, especially the MLE, but by applying a suitable correction we have developed an estimator which is almost unbiased in the region of interest. We propose two interval estimators which improve on existing methods and have excellent coverage properties. Our recommendation is a score-based method with a correction for skewness, but a good alternative is an exact method with a mid-P correction.

Keywords: Bias correction; Coverage; Estimation of proportions; Group testing; Inverse sampling; Mid-P; Negative binomial distribution

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. In most applications the attribute is a disease, whether it affects humans (e.g. Chen, Tebbs, and Bilder 2009), animals (e.g. Durnez et al. 2008) or plants and seeds (e.g. Schaarschmidt 2007). Group testing divides into two distinct fields, with a small overlap between them — *identification* (or classification) of infected individuals, and *estimation* of disease prevalence p, which is the focus of our interest. In the statistical literature, Dorfman (1943) introduced group testing as a way of assisting identification, and it grew strongly from there. Estimation took longer to develop, but has grown rapidly in recent decades, with new areas of application appearing regularly, such as estimation of the proportion of transgenic plants in a population (Montesinos-Lopez et al. 2010). The benefits of group testing are clearly greatest when p is low, but they can be substantial even for moderate prevalence. For example, Verstraeten et al. (1998) achieved a reduction of 62% in costs for a marginal loss in accuracy when the estimated prevalence was 7.3%.

In the estimation problem, a binomial model has usually been assumed, in which the outcomes for the individuals in each group follow i.i.d. Bernoulli distributions with parameter p. Hepworth (2005) gave a full description of, and justification for, this model. An extension of this is the use of the beta-binomial distribution, on which there have been a series of studies in the plant disease literature (e.g. Hughes, Gottwald, and Levy 2002). All binomial and beta-binomial models are based on the testing of a fixed total number of groups. Another scenario is where the number of positive groups is fixed, and testing continues until the required number is observed. This inverse sampling arrangement was suggested by Katholi and Unnasch (2006) in the context of disease transmission by insect vectors, an area in which group testing has a long and distinguished history. 'Inverse binomial group (pooled) testing' — inverse sampling of groups with individuals assumed to follow a Bernoulli distribution — was applied by Rodriguez-Perez et al. (2006) in an assessment of transmission of the filarial parasite *Onchocerciasis volvulus*, which causes

ocular and skin disease. Their study, based in Mexico, involved testing the disease vector (black flies) in groups of 50. In some communities they tested all available pools (direct sampling), and in others, inverse sampling was employed, with pools being tested until the first positive occurred.

Inverse sampling is appealing in situations requiring a quick response, such as a natural disaster or an outbreak of disease. Combining it with group testing creates even more powerful options for authorities responding to such events. Even if detection is of primary interest, it is often important to estimate pathogen prevalence as well. This is also the case where data collection and monitoring are performed on a continuing basis. For example, group testing of mosquitoes for West Nile Virus (WNV) is carried out regularly by the United States Centers for Disease Control and Prevention. A high prevalence may precede a significant human disease outbreak, and so prevalence estimation is an important aspect of the investigation. An inverse sampling design, combined with group testing, emerges as an excellent option.

Pritchard and Tebbs (2011a) examined inverse binomial group testing in depth, considering several point and interval estimators. They described situations where such testing would be advantageous, using as examples the transmission of WNV by mosquitoes, and the potential effect of an outbreak of foot and mouth disease on the meat industry. They addressed many of the statistical problems arising from inverse sampling, and suggested several avenues for further research. One of those was improved point estimation, necessary because of the bias associated with the standard estimators. The other concerned interval estimation, with potential improvements available for both exact and asymptotic methods. In this paper we present improvements to both point and interval estimators of p.

1.1   THE NEGATIVE BINOMIAL MODEL

In proposing inverse sampling in the context of group testing, Katholi and Unnasch (2006) stressed the importance of the statistical model matching the sampling design. They pointed out that in the case of equal group size s, the total number of groups required to

2

obtain n positive groups has a negative binomial distribution rather than a binomial, and so any inference should reflect this. Point and interval estimates will in general not be the same for the two models, given the same number of positive groups and same total number of groups. This result was reported by George and Elston (1993) for inverse sampling of individuals which stops at the first occurrence of an event, i.e. for the special case $s = 1$ and $n = 1$. In this paper we consider $s > 1$ (otherwise it is not group testing), and $n \geq 1$. The case $n = 1$, used by Rodriguez-Perez et al. (2006), gives rise to a geometric distribution. Stopping at the first positive group, as they did, is useful in signaling the presence of a disease, though it does not in general result in precise estimation of p, due to the limited information likely to be provided.

1.2   ASSUMPTIONS

We assume that the individuals' outcomes follow i.i.d. Bernoulli distributions with parameter p, and that they are randomly pooled into groups of size s. Then a random number $T = t$ tests are performed until n positive groups are observed. Pritchard and Tebbs (2011a) also considered unequal group sizes, but this makes the distribution of T intractable, so they ended up giving most of their attention to the equal group size case. Pritchard and Tebbs (2011b), in a Bayesian counterpart to their 2011a paper, considered only equal group sizes. Katholi and Unnasch (2006) commented that the 'negative binomial model will not support the use of unequal pool sizes'. We recognize that unequal group sizes can occur in practice, but give our attention to equal group sizes, thus setting a foundation which will enable some of our results to be extended to unequal group sizes.

We also assume perfect testing, i.e. that sensitivity and specificity are both 1. This is not always the best assumption, but it is reasonable in many situations, and it would not be difficult to modify our methods later to incorporate imperfect testing. The perfect testing assumption was also made by Pritchard and Tebbs (2011a), and Pritchard and Tebbs (2011b) in their work. An important work on incorporating sensitivity and specificity parameters for the binomial model was Tu, Litvak, and Pagano (1995); a more recent treatment was given by Hanson, Johnson, and Gastwirth (2006).

2   POINT ESTIMATION

2.1   MAXIMUM LIKELIHOOD AND ALTERNATIVES

The probability that a group tests positive is $\pi = 1 - (1 - p)^s$. The number of tests performed, T, has a negative binomial distribution with parameters n and π. The maximum likelihood estimator (MLE) of p is

$$\hat{p} = 1 - \left(1 - \frac{n}{T}\right)^{1/s} \tag{1}$$

which has the same form as the MLE for the binomial model, whose positive bias has been well known since the earliest work on group testing (Thompson 1962). However, the bias is even greater for the negative binomial model, because in addition to the 'group testing bias' caused by the non-linearity of (1), the MLE of the probability parameter of a negative binomial distribution is already biased. For $s = 1$ (standard inverse sampling), an unbiased estimator of p is $(n - 1)/(T - 1)$ (Haldane 1945) compared to the positively biased MLE n/T.

Pritchard and Tebbs (2011a) noted the severe positive bias of $\hat{p}$, especially for small n. Their simulations showed that bias and mean squared error converged to 0 much too slowly for the MLE to be useful in practice. This led them to look for alternative estimators, and they considered four: (i) a jackknife estimator, which was very biased negatively; (ii) a 'shrinkage' estimator, in which n/T is replaced by $\alpha(n/T)$ in (1); this provided some improvement; (iii) a 'shift' estimator, defined as

$$1 - \left(1 - \frac{n + 1}{T + \beta}\right)^{1/s}$$

and (iv) a 'combined estimator', defined as

$$1 - \left[1 - \alpha\left(\frac{n + 1}{T + \beta}\right)\right]^{1/s}.$$

Estimators (iii) and (iv), though always negatively biased, provided the most reduction in absolute bias and mean squared error, and were barely distinguishable from each other. For $p < 0.1$, the absolute bias for both (iii) and (iv) was less than 0.015 for $s = 5$ or 10, less than 0.010 for $s = 30$ and less than 0.005 for $s = 50$. These two alternative

4

estimators are in the same vein as the one proposed by Burrows (1987) for the binomial model, which removes nearly all the bias of the MLE for values of p consistent with the testing procedure (Hepworth and Watson, 2009). The effect of all of these estimators is to shrink the proportion of negative groups towards 1, which in turn adjusts the MLE downwards.

2.2   GART BIAS CORRECTION

A general bias correction to the MLE for a single unknown parameter was described by Gart (1991). Hepworth and Watson (2009) applied this correction to binomial group testing with two group sizes, and found it very effective in reducing the bias. In addition, it was simple to apply, having a closed form. Gart's correction to the MLE is now the recommended estimator in the group testing software *Pooled Infection Rate* (Biggerstaff, 2006). We now apply it to inverse binomial group testing.

Gart (1991) showed that the bias (except for terms of $O(n^{-2})$) is

$$B(p) = -\frac{2\frac{dI}{dp} + E\left[\frac{d^3 l}{dp^3}\right]}{2[I(p)]^2} \tag{2}$$

where $I(p)$ is the Fisher information and $l(p)$ is the log likelihood. $B(p)$ can be found provided the information is tractable. When group sizes are unequal, T does not have a negative binomial distribution, and the information is not tractable, but for equal group sizes it is

$$I(p) = \frac{ns^2(1-p)^{s-2}}{(1-(1-p)^s)^2} \tag{3}$$

(Pritchard and Tebbs, 2011a). It can be shown (see Appendix) that

$$\frac{dI}{dp} = -ns^2\left(\frac{(s-2)(1-p)^{s-3} + (s+2)(1-p)^{2s-3}}{(1-(1-p)^s)^3}\right) \tag{4}$$

and

$$E\left[\frac{d^3 l}{dp^3}\right] = \frac{sn}{(1-p)^3}$$
$$\times \left(\frac{s(s+1)(1-p)^s(1-(1-p)^s) + 2(s(1-p)^s + (1-p)^s - 1)^2}{(1-(1-p)^s)^3} - \frac{2}{1-(1-p)^s}\right). \tag{5}$$

5

Substituting $\hat{p}$ into (4) and (5), and then into (2) , gives an estimate of the bias, which can be subtracted from $\hat{p}$ to give a bias-corrected estimate, which we label $\hat{p}_G$. (Note that $\hat{p} > 0$ for this problem.)

Consider, for example, $n = 5$ and $s = 20$. Table 1 shows $\hat{p}$ and $\hat{p}_G$ for the first ten outcomes. The large size of the correction is evident, this being necessary because of the substantial bias of the MLE. In this example, the correction converges to around 20% of the MLE as t increases; e.g. for $t = 100$, $\frac{\hat{p}_G}{\hat{p}} = \frac{0.00205}{0.00256} = 0.800$.

Table 1: Gart bias correction applied to the MLE, $n = 5$ and $s = 20$

t	6	7	8	9	10	11	12	13	14	15
$\hat{p}$	.0857	.0607	.0479	.0397	.0341	.0299	.0266	.0240	.0218	.0201
$\hat{p}_G$	.0600	.0460	.0372	.0312	.0269	.0237	.0212	.0191	.0174	.0160

We have excluded the boundary outcome $t = 5$ (all positive groups), for which $\hat{p} = 1$, and $B(\hat{p})$ is undefined. It is a highly uninformative outcome, and one to avoid if possible in group testing. If it occurred here it would suggest that the group size $s = 20$ is too large for the underlying p. Hepworth and Watson (2009) discussed the problem of all positive groups in the binomial setting, and imposed an upper bound ψ on p for the purpose of evaluation, where ψ is the value of p at which the probability of all positive groups is 0.05, which is $1 - (1 - (0.05)^{1/n})^{1/s}$. This effectively restricts p to values consistent with the design of the testing procedure, and is the approach that we adopt in this paper. For the procedure reported in Table 1, $\psi = 0.0391$. In general, ψ increases as n increases or s decreases, since either results in a smaller probability of all positive groups. ψ is also an appropriate lower 95% confidence limit when $\hat{p} = 1$ (Hepworth, 2004).

2.3   BIAS OF CORRECTED ESTIMATOR

T has a negative binomial distribution, and so has an infinite number of outcomes. The exact bias of the corrected estimator $\hat{p}_G$ is therefore an infinite sum, which does not reduce to anything tractable, as occurred with the estimators proposed by Pritchard and Tebbs

(2011a). We approximated the bias by

$$\sum_{t=n}^{t*} \left(\hat{p}_G \binom{t-1}{n-1} (1 - (1-p)^s)^n (1-p)^{s(t-n)} \right) - p \tag{6}$$

where $\Pr(T > t*) < 0.0001$ for the smallest value of p considered, which we set to $\psi/250$.

For p larger than this, $\Pr(T > t*)$ is even smaller than 0.0001, so the approximation is very

close to the true value of the bias. The number of outcomes over which the expectation

was computed was generally around 5000. If 0.0001 is replaced by a smaller number (we

tried 0.000001), it makes virtually no difference to the results.

We calculated the bias of $\hat{p}_G$ for $n = 1, 5, 10 \times s = 5, 10, 20, 30, 50$. These values were

chosen to represent realistic scenarios and to cover the values tested by Pritchard and

Tebbs (2011a). For the all positives outcome ($t = n$, $\hat{p} = 1$) the bias corrected estimate

proposed by Burrows was used. Table 2 shows the results for the example above, for

selected values of p. It is evident that the correction proposed by Gart is extremely

effective in removing the bias for $p < \psi = 0.0391$, and almost totally removes the bias for

small p. It overcorrects more as p increases, but the maximum overcorrection for $p < \psi$

is still less than 2%.

Table 2: Expected value, bias and percentage bias for the estimator corrected for bias
using the Gart correction, for $n = 5$ and $s = 20$ ($\psi = 0.0391$)

p	.001	.003	.005	.01	.02	.03	.04	.05	.07	.1
$E(\hat{p}_G)$	.0010	.0030	.0050	.0100	.0199	.0297	.0384	.0487	.0659	.0861
$Bias(\hat{p}_G)$	.0000	.0000	.0000	$-.0000$	$-.0001$	$-.0003$	$-.0006$	$-.0013$	$-.0041$	$-.0139$
% $Bias(\hat{p}_G)$	.005	.012	.006	-0.06	-0.38	-0.88	-1.59	-2.61	-5.92	-13.9

For the other combinations of n and s (except $n = 1$), the results are similar. The

correction produces an estimator which is virtually unbiased for values of p less than

about $3\psi/4$. Close to $p = \psi$ the bias is negative and is generally between 1% and 2%.

The absolute bias is less for smaller values of s, which is generally true for group testing.

For $n = 1$, the bias is small in absolute terms, though not trivial relative to p. For

example, with $s = 20$, $E(\hat{p}_G) = 0.00106$ when $p = 0.001$, and 0.00211 when $p = 0.002$,

which is a bias of 6.3% and 5.3% respectively. Because $\psi = 0.0026$ in this case, we

need not look further at larger values of p. The larger relative bias for $n = 1$ is not of major concern—firstly, because it is already a substantial improvement over the MLE; and secondly, because a group testing procedure which stops at the first positive is unlikely to be recommended in practice due to its poor precision, unless resources are very limited or detection is of prime importance. An example of the very wide interval estimates for $n = 1$ is shown in Table 2 of Rodriguez-Perez et al. (2006), where they quote a point estimate of 2.6 infected flies/10,000 and a 95% confidence interval of $(0.06, 9.7)$.

Figure 1 (top three rows) displays the bias of $\hat{p}_G$ for $n = 5, 10 \times s = 5, 20, 50$ and $p \le 1.1\psi$. The vertical dashed line in each plot is at $p = \psi$. The plots confirm that the Gart correction to the MLE has virtually eliminated the bias for $p < \psi$. The overcorrection for p close to ψ is very small; even for the worst case of the six considered ($n = 10$ and $s = 5$), the absolute bias is less than 0.001. Comparison with the estimators proposed by Pritchard and Tebbs (2011a) is complicated by the need for their estimators to adopt values of α and β that are optimal in some way, which in practice requires setting an upper bound for p. But even with these parameters optimally chosen, our corrected estimator is a vast improvement on both the shift estimator (iii) and the combined estimator (iv), whose best result was to have an absolute bias of less than about 0.005 for $n = 5$ and $s = 50$. Even if their results are restricted to $p < \psi$ to make them more strictly comparable, the absolute bias of the Gart corrected estimator is smaller by an order of magnitude.

2.4   MEAN SQUARED ERROR OF CORRECTED ESTIMATOR

We now briefly examine the mean squared error (MSE) of $\hat{p}_G$. We approximated the MSE using the same method as for bias, with $\hat{p}_G$ replaced by $\hat{p}_G^2$ in the summation shown in (6) to calculate the expected value of the square of the estimator, and hence the variance. For the same reasons described above, this method gives a very close approximation to the true value of the MSE.

Figure 1 (bottom row) displays the MSE of $\hat{p}_G$ for $n = 5$ or 10 and $s = 20$. We have shown only these two plots because the other combinations of n and s have the same general shape. For larger n the MSE is smaller for the same p, though greater if we consider

p relative to ψ. Because the bias of the estimator is very small, it makes a minimal contribution to the MSE. For example, with $n = 5$ and $s = 20$, the contribution of the (squared) bias is only 0.07% of the MSE at $p = \psi$. Therefore it is essentially the variance of the estimator that we are examining.

The reduction in MSE by applying the Gart bias correction to the MLE is massive. For example, with $n = 5$ and $s = 20$, the MSE of the MLE at $p = \psi$ is 0.0466, which is about 100 times greater than that of the corrected estimator. At $p = \psi/2$, the ratio is about 30. The correction has therefore succeeded not only in almost eliminating the bias, but in reducing the variance as well.

Pritchard and Tebbs (2011a) plotted MSE against p for their alternative estimators, but comparisons with our corrected estimator are ambiguous. Their plot for $n = 5$ and $s = 30$ shows an MSE of around 0.0005 at $p = \psi$ for their shift and combined estimators, which is around twice the MSE for the corrected estimator. On the other hand, their plot for $n = 5$ and $s = 50$ shows an MSE of around 0.00003 at $p = \psi$, which is less than half that of the corrected estimator. This discrepancy may be due to the very different upper bounds they used for p: 0.1 for $s = 30$ (well above $\psi = 0.0261$), and 0.01 for $s = 50$ (below $\psi = 0.0158$). Note that their optimal values of α and β were defined as those which minimize the MSE for fixed n, s and p, which puts the resulting estimators at an advantage in comparisons based on MSE.

3   INTERVAL ESTIMATION

3.1   SKEW-CORRECTED SCORE INTERVALS

For inverse group testing, Pritchard and Tebbs (2011a) derived large sample Wald, score and likelihood ratio confidence intervals for p, and examined coverage probability for $n = 1, 5$ and 100, and $s = 20$. They did not make strong recommendations, but noted that all three methods performed satisfactorily for $n = 100$, had complications when $\hat{p} = 1$, and that Wald lower limits could be negative; all of these results are unsurprising. Examination of their coverage plots suggests that intervals based on the score are the

most promising of the methods they considered. In their discussion Pritchard and Tebbs suggested that a skewness correction to score intervals may provide an improvement, as it did in the binomial case (Hepworth, 2005). We now consider this.

The score function for equal group sizes s is

$$S(p) = \frac{s}{1-p}\left(t - \frac{n}{1-(1-p)^s}\right)$$

(see Appendix, equation (9)). Pritchard and Tebbs (2011a) derived the score for unequal group sizes, of which this is a special case. $S(p)$ is asymptotically normal with mean 0 and variance $I(p)$, enabling approximate confidence intervals to be found. Gart (1991) noted the problem of skewness with score-based intervals, and found the skewness to be

$$\gamma(p) = \frac{\mu_3\{S(p)\}}{\{I(p)\}^{3/2}} \tag{7}$$

where

$$\mu_3\{S(p)\} = 3\frac{dI}{dp} + 2\mathrm{E}\left(\frac{d^3l}{dp^3}\right)$$

is the third central moment of $S(p)$. Approximate confidence limits are found by solving for p the equations

$$S(p) - \frac{1}{6}\gamma(p)(z_{\alpha/2}^2 - 1) = \pm z_{\alpha/2} \tag{8}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. Hepworth (2005) derived these quantities for binomial group testing with unequal group sizes, and assessed coverage probability for this and three other asymptotic interval estimation methods. The skew-corrected score method was his final recommendation. Biggerstaff (2008) examined confidence interval methods for the difference of two proportions estimated by binomial group testing, and recommended the skew-corrected score method in that situation too.

All of the quantities given in the expressions above have already been derived while investigating the bias correction in section 2.2 (see Appendix). Using them, we found 95% confidence intervals for $n = 5, 10, 100 \times s = 5, 10, 20, 30, 50$. The case $n = 100$ was included to examine the asymptotic properties of the intervals.

We firstly examined the intervals themselves, to consider the effect of the skewness correction. Table 3 shows uncorrected and skew-corrected score 95% confidence limits for

10

the first ten outcomes for $n = 5$ and $s = 20$. The MLE $\hat{p}$ is shown as a point of reference for symmetry. We have again excluded the (uninformative) boundary outcome $t = 5$, for which $\hat{p} = 1$. For this outcome, the uncorrected score interval is $(0.0131, 1)$, and the skew-corrected interval is undefined because equation (8) has no solution for the lower limit. A better choice for the lower limit would be $\psi = 0.0391$. Equation (8) also has no solution for the lower limit when $n = 1$, in which case an exact interval or an interval based on the likelihood ratio could be used (both are described by Pritchard and Tebbs, 2011a).

Table 3: Uncorrected score, skew-corrected score, and mid-P exact 95% confidence limits, for $n = 5$ and $s = 20$

t	6	7	8	9	10	11	12	13	14	15
$\hat{p}$	.0857	.0607	.0479	.0397	.0341	.0299	.0266	.0240	.0218	.0201
Uncorrected score										
lower	.0091	.0070	.0057	.0048	.0041	.0036	.0033	.0029	.0027	.0025
upper	.1564	.1116	.0883	.0735	.0640	.0560	.0494	.0446	.0409	.0374
Skew-corrected score										
lower	.0244	.0182	.0147	.0123	.0106	.0093	.0083	.0075	.0068	.0063
upper	.2025	.1310	.1008	.0828	.0706	.0617	.0549	.0494	.0450	.0413
Mid-P exact										
lower	.0251	.0188	.0152	.0127	.0109	.0096	.0086	.0078	.0071	.0065
upper	.2060	.1322	.1010	.0827	.0704	.0614	.0546	.0491	.0447	.0410

The uncorrected intervals are almost symmetrical about $\hat{p}$, and in fact they are similar to Wald intervals (not shown), especially for larger values of t. Such symmetry is clearly undesirable with a small sample and small p. The symmetry of the interval is less about a bias-corrected estimator such as $\hat{p}_G$, but it is still a problem, particularly for estimates close to 0, for which the lower limit can be negative. For example, when $n = 3$ and $s = 20$, the lower confidence limit is negative for all outcomes except $t = 3$. For instance, $t = 4$ results in an MLE of 0.0670, an uncorrected score interval of $(-0.0077, 0.1361)$ and a corrected interval of $(0.0115, 0.1833)$. The possibility of a confidence limit outside the boundary of the parameter space is a substantial disadvantage of any interval estimation

11

method. Skew-corrected score intervals have no such disadvantage.

Another feature of all the intervals examined here is that the upper confidence limit is smaller than the corresponding limit for the binomial model with the same final outcome. The fact that the last group tested was positive gives more information than just having n positives in t tests, a point noted by George and Elston (1993) for the simple $n = 1$ and $s = 1$ case. This result is consistent with the expression for the information (3), which is the same as that for the binomial model but with the denominator squared. Because $1 - (1 - p)^s \leq 1$ the information is greater for the negative binomial case.

3.2   COVERAGE PROPERTIES

The coverage probability for skew-corrected score intervals was calculated for the combinations of n and s mentioned above. Like the bias, the exact coverage is an infinite sum, and so we approximated it in the same way, using the summation shown in (6), but with $\hat{p}_G$ replaced by an indicator function which is 1 if the interval contains p and 0 otherwise. We restricted the outcomes included in the summation in the same way as for the bias. The number of outcomes included was generally more than 10000, which produced an extremely close approximation to the true coverage.

Figure 2 (top three rows) shows the coverage probability for selected combinations of n and s and $\psi/500 \leq p \leq 1.1\psi$. There is no need to mark $p = \psi$ because it is obvious — it is the point at which the coverage jumps from generally centering around the nominal level (0.95) to a much more conservative level and a subsequent substantial departure from the nominal. This phenomenon for group testing was noted and explained by Hepworth (2004). Another point of interest is that for a fixed n, the coverage pattern for different values of s is almost identical but on a different scale depending on ψ. We have therefore not included plots for the other combinations of n and s. This extreme similarity for different s is because different combinations of s and p can produce the same probability of a positive group π, and for a fixed n, the same negative binomial distribution for T.

The plots show a very favorable coverage pattern, which converges to the nominal level as n increases. This is confirmed by the summary statistics shown for each plot in Table

4. MAD denotes the median absolute deviation about the nominal level, which is a useful overall measure of closeness. $p_{.95}$ denotes the proportion of coverage values below 0.95. The mean coverage is virtually identical to the median, and is not shown.

Table 4: Summary statistics for coverage probability of skew-corrected score and mid-P exact 95% confidence intervals, $p \leq \psi$

n	s	Minimum		Median		Maximum		MAD		$p_{.95}$	
		Score	Mid-P	Score	Mid-P	Score	Mid-P	Score	Mid-P	Score	Mid-P
5	5	.924	.924	.953	.950	.970	.967	.0057	.0044	.257	.477
5	20	.924	.923	.953	.950	.970	.967	.0057	.0045	.267	.471
10	5	.925	.925	.952	.951	.972	.972	.0041	.0037	.323	.439
10	20	.925	.925	.952	.951	.972	.972	.0042	.0041	.333	.433
50	20	.927	.933	.951	.951	.974	.974	.0033	.0034	.401	.385
100	20	.921	.932	.951	.950	.970	.976	.0029	.0031	.455	.424

Even for $n = 5$ the coverage is very good, being centered a little above the nominal level, with a small MAD. It can be seen as an advantage to average above 0.95, as this is associated with having fewer values below it. Newcombe (1998) suggested that it is best to obtain a mean coverage a little over the nominal level and a minimum a little under it, a situation that is achieved by the smaller values of n in particular. Pritchard and Tebbs (2011a) produced coverage plots for uncorrected score intervals corresponding to plots $(2, 1)$ and $(3, 2)$ in Figure 2. Neither the uncorrected nor the corrected coverage stands out as obviously better or worse, and this is confirmed by examining summary statistics, which we calculated for these two scenarios. For $n = 5$, $s = 20$ and $p \leq \psi$, the mean coverage of uncorrected score intervals is 0.956 and the MAD is 0.0063. These values are similar to the corresponding statistics for the skew-corrected intervals shown in Table 4, and so the overall coverage probability in itself does not favor the correction. However, when we consider symmetry of coverage, the correction is shown to be absolutely necessary. For example, with $n = 5$ and $s = 20$, the mean distal (left) noncoverage for the uncorrected intervals is 0.0001 and the mean mesial (right) noncoverage is 0.0437. They are therefore effectively one-sided confidence intervals, which is not the intention at all. The corresponding figures for the skew-corrected intervals are 0.0237 and 0.0236.

13

Other combinations of n and s produce similarly excellent symmetry of coverage for the skew-corrected intervals.

The MADs shown in Table 4 are small by any standards, and tiny for the larger values of n. Hepworth (2005) reported a MAD of between 0.010 and 0.013 for skew-corrected score intervals applied to binomial group testing, and similar values for other asymptotic methods. Tebbs and Bilder (2004) compared seven methods for binomial group testing, and for 40 groups of size 20 found the mean absolute deviation to be between about 0.007 and 0.020. Note that even with a set group size s, there is no single binomial group testing arrangement corresponding to a particular negative binomial arrangement, because of the unknown number of groups arising from the latter. One way of choosing a corresponding arrangement would be to use the expected number of tests for a mid-range value of p. For example, with $n = 5$ and $s = 20$, if we let $p = \psi/2 = 0.0195$, $E(T) = 10.3$. So the 40 groups used by Tebbs and Bilder can be seen as large in comparison, which further underlines the smallness of the MADs we obtained.

3.3   MID-P EXACT INTERVALS

Lui (1995) used the relationship between the negative binomial distribution and the incomplete beta function to derive an exact confidence interval for the probability parameter of the negative binomial. This is essentially a Clopper-Pearson-type interval with a convenient computational form. Pritchard and Tebbs (2011a) extended this to group testing (with equal-sized groups) by transforming the endpoints of the interval using an inversion of the relationship $\pi = 1-(1-p)^s$. They calculated the resulting exact confidence intervals for the same values of n and s that were used for the asymptotic intervals. The coverage plots shown in their paper are very conservative, as happens with two-sided $(1 - \alpha)$ Clopper-Pearson-type intervals in general, due to the requirement that the probabilities of the outcomes in each tail sum to $\alpha/2$. Pritchard and Tebbs suggested that applying a mid-P correction to these intervals be considered; this we now do. The mid-P correction is defined as the inclusion of half the probability of the observed outcome in each tail of the summation.

Unless a strictly conservative interval (i.e. one with coverage probability which is never below $1 - \alpha$) is desired, applying a mid-P correction is an obvious option. The correction generally has the effect of centering the coverage around a level close to nominal, and has the advantage of allowing a meaningful comparison with asymptotic methods. Mid-P exact intervals have been long advocated for discrete problems, beginning with Lancaster (1961). In binomial group testing they were successfully applied by Hepworth (1996) and Hepworth (2004). A minor disadvantage is that computation is not as convenient as with the strictly conservative intervals; because the probability of the observed outcome is divided between the tails, it requires an iterative solution. Having an infinite number of outcomes is not an issue in itself because the sum of probabilities can be taken over the finite number of outcomes in the other tail.

Again, we firstly examined the intervals themselves — see the last two lines of Table 3. They are very similar to the skew-corrected score intervals, with the lower limit shifted slightly to the right, and the upper limit shifted slightly to the left except for the first few outcomes. This strong similarity occurs for other values of n and s (results not shown). It is not very surprising given the symmetry of coverage observed in the skew-corrected score intervals (for which we now drop the label 'skew-corrected'). Exact intervals based on tail probabilities should have a fair degree of inherent symmetry because they effectively involve the computation of two one-sided intervals.

We then calculated the coverage probability for mid-P exact intervals, using the same combinations of n and s as used for the score intervals. The number of outcomes included in the approximation was the same (> 10000), and so again it was extremely close to the true coverage. The two bottom plots of Figure 2 show the coverage probability for $n = 5$ and 50 and $s = 20$. There is little point in showing plots for different s, as they are almost identical except for a scaling factor, as was the case for score intervals.

The two plots (and those not shown) are extremely similar to the corresponding plots for score intervals. This is not surprising, given the similarity of the intervals themselves. The only obvious difference is that for $n = 5$, the mid-P coverage centers on the nominal level even for small p, in contrast to the score coverage, which stays a little above it. For

15

$n = 50$, both center on the nominal level, because n is large enough for the score coverage to converge. The summary statistics for the mid-P coverage, shown in Table 4 for the same combinations of n and s evaluated for the score, confirm our impression from the plots. The median coverage is extremely close to 0.95 for all combinations of n and s, with a little over half the values above the nominal level. The minima, maxima, and especially the MADs, are very favorable indeed. Minimum and maximum values for the two methods hardly differ. For small n the smaller MAD for mid-P intervals is enough to be an advantage, but the larger proportion of points below 0.95 is a disadvantage. In any case, the difference in either virtually disappears for large n.

4   DISCUSSION

We have developed very good point and interval estimators of proportions obtained by inverse binomial group testing with equal group sizes. Pritchard and Tebbs (2011a) set solid foundations, and we have built on them to produce an almost unbiased point estimator, and interval estimators with excellent coverage properties. The practitioner employing group testing using inverse sampling is now well placed in regard to estimation. There is no need to (wrongly) assume a binomial model, with a fixed number of groups; the properties of the negative binomial distribution, which applies to the number of groups tested, have been used throughout this work. Even if testing begins with the simple aim of detecting a disease, estimation of prevalence often turns out to be of interest as well. Two-sided confidence intervals would be required in most situations, though a one-sided upper limit would be useful in indicating whether a threshold has been exceeded.

Gart's general bias correction to the MLE is extremely effective in addressing the bias for values of p consistent with the design of the testing procedure. We have considered a practical range of group sizes and number of positive groups, and there is no reason why the results would not extend to other combinations. The computation needed for evaluating the bias is quite complex, due to the infinite number of outcomes. However, calculation of bias-corrected estimates is easy to implement in practice because the correction has a

closed-form.

Confidence intervals based on the skewness correction to the score function give excellent coverage, with small departures from the nominal level and the majority of values above it. Exact confidence intervals based on the Clopper-Pearson method are very conservative, but applying a mid-P correction is highly effective in centering the coverage on the nominal level, even for a small number of positive groups. Computation of either type of interval requires iteration, but is quick enough to be readily implemented in practice. The code could be easily added to existing software such as the R program `invbingroup` referred to in Pritchard and Tebbs (2011a), located at the R-Forge website *http:// r-forge.r-project.org/*.

Either interval estimation method can be recommended, and the intervals do not differ much between them. The choice depends mainly on how conservative one desires the coverage to be. Our recommendation is slightly towards skew-corrected score intervals, because of the smaller proportion of points below the nominal level. The overall closeness, as measured by the MAD, is excellent, and is almost as good as that of mid-P exact intervals even for a moderate number of positive groups. A minor disadvantage is that the expected interval width is a little larger, which is not surprising given its slightly greater coverage. For example, the expected width arising from the procedure in Table 3 is up to 1.7% larger for skew-corrected score intervals (average 0.8%). But this is a small difference, both in relative and absolute terms (maximum difference = 0.00042). Another minor disadvantage of the skew-corrected score method is that the lower confidence limit cannot be found when all groups are positive. However, many interval estimation methods require modifications on the boundary of the parameter space (Hepworth 2005), and the mid-P exact method would be a reasonable replacement because it produces similar intervals. Likewise for the special case when there is only one positive, which we do not recommend in general due to the poor precision of estimates.

Choice of group size (where it is possible) is even more important in inverse binomial group testing than in the binomial case. Obtaining all positive groups is uninformative and to be avoided in either situation; it arises when the group size is too large. But a

17

group size which is too small is also problematic in the negative binomial case, because of the real possibility of obtaining no positive groups after a considerable number of tests. If the testing concludes at that point, a binomial model is needed to find an upper confidence limit for the prevalence, which violates a key modeling assumption of the testing procedure.

Extension of this work to unequal group sizes would be a useful addition to the field. Because of the intractability of the (expected) information, it is not possible to simply generalize the derivations in the Appendix, which we used for both the Gart bias correction and the skewness correction to the score. However, by using observed information and simulation, Pritchard and Tebbs (2011a) constructed (uncorrected) score and likelihood ratio intervals for unequal group sizes, and estimated the coverage probabilities by simulation. This could be extended to skew-corrected score intervals. Observed information and its derivative could also be used to find the Gart bias correction for unequal group sizes.

A greater challenge would be to extend exact confidence intervals (with or without a mid-P correction) to unequal group sizes. Pritchard and Tebbs (2011a) stated that 'to the best of our knowledge, exact intervals for p are not available when pool sizes are unequal'. There is no unambiguous ordering of the event space, which is necessary for an appropriate summation of probabilities (Hepworth 1996). This issue, combined with the infinite number of outcomes, presents a formidable problem.

APPENDIX: DERIVATION OF COMPONENTS OF THE BIAS AND THE SKEWNESS CORRECTION TO THE SCORE

DERIVATIVE OF THE INFORMATION

Let $q = 1 - p$. $I(q) = I(p) = \dfrac{ns^2 q^{s-2}}{(1 - q^s)^2}$ (see (3)).

$$
\frac{d}{dq}\left(\frac{q^{s-2}}{(1 - q^s)^2}\right) = \frac{(s-2)q^{s-3}(1 - q^s)^2 + 2sq^{2s-3}(1 - q^s)}{(1 - q^s)^4}
$$

$$
= \frac{(s-2)q^{s-3} + (s+2)q^{2s-3}}{(1 - q^s)^3} .
$$

Therefore

$$
\frac{dI}{dp} = -\frac{dI}{dq} = -ns^2 \left(\frac{(s-2)(1-p)^{s-3} + (s+2)(1-p)^{2s-3}}{(1 - (1-p)^s)^3}\right) .
$$

EXPECTATION OF THE THIRD DERIVATIVE OF THE LOG-LIKELIHOOD

Let $q = 1 - p$. The log-likelihood is

$$
l(q) = n \log(1 - q^s) + s(t - n) \log q
$$

(see Pritchard and Tebbs (2011a) who gave the log-likelihood for unequal group sizes; this is a specific case.)

$$
\frac{dl}{dq} = -\frac{nsq^{s-1}}{1 - q^s} + \frac{s(t - n)}{q}
$$

which simplifies to

$$
\frac{s}{q}\left(t - \frac{n}{1 - q^s}\right) . \tag{9}
$$

$$
\frac{d^2l}{dq^2} = -\frac{s}{q^2}\left(t - \frac{n}{1 - q^s}\right) - \frac{sn}{q}\left(-\frac{sq^{s-1}}{(1 - q^s)^2}\right)
$$

which simplifies to

$$
\frac{s}{q^2}\left(\frac{n(1 - q^s(s + 1))}{(1 - q^s)^2} - t\right) . \tag{10}
$$

19

T has a negative binomial distribution, so $E(T) = n/(1 - q^s)$. Taking the expectation of (10) gives the expression for $I(p)$ given in (3).

$$\frac{d^3l}{dq^3} = \frac{2ts}{q^3} - sn\frac{d}{dq}\left(\frac{q^s(s+1)-1}{q^2(1-q^s)^2}\right).$$

The derivative of the expression in the large bracket is

$$\frac{s(s+1)q^{s+1}(1-q^s)^2 - 2q\left((s+1)q^s - 1\right)\left((1-q^s)^2 - sq^s(1-q^s)\right)}{q^4(1-q^s)^4}$$

which simplifies to

$$\frac{s(s+1)q^s(1-q^s) + 2\left((s+1)q^s - 1\right)^2}{q^3(1-q^s)^3}.$$

Using the expression for E(T) above results in

$$E\left(\frac{d^3l}{dp^3}\right) = -E\left(\frac{d^3l}{dq^3}\right) = \frac{sn}{(1-p)^3}$$
$$\times \left(\frac{s(s+1)(1-p)^s(1-(1-p)^s) + 2(s(1-p)^s + (1-p)^s - 1)^2}{(1-(1-p)^s)^3} - \frac{2}{1-(1-p)^s}\right).$$

ACKNOWLEDGMENTS

REFERENCES

Biggerstaff, B. J. (2006) "Pooled Infection Rate: An Excel Add-In to compute point and interval estimates of infection rates and their difference using pooled samples", CDC, Fort Collins, CO, USA.

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.

Burrows, P. M. (1987), "Improved estimation of pathogen transmission by group testing," *Phytopathology* 77, 363–365.

Chen, P., Tebbs, J. M., and Bilder, C.R. (2009), "Global goodness of fit tests for group testing regression models," *Statistics in Medicine*, 28, 2912–2928.

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

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.

Gart, J. J. (1991), "An application of score methodology: Confidence intervals and tests of fit for one-hit curves," In *Handbook of Statistics*, C. R. Rao, R. Chakraborty (eds), 8, 395–406. Amsterdam: Elsevier.

George, V. T., and Elston, R. C. (1993), "Confidence limits based on the first occurrence of an event" *Statistics in Medicine*, 12, 685–690.

Haldane, J. (1945) "On a method of estimating frequencies", *Biometrika*, 33, 222–225.

Hanson, T. E., Johnson, W. O., and Gastwirth, J. L. (2006), "Bayesian inference for prevalence and diagnostic test accuracy based on dual-pooled screening," *Biostatistics*, 7, 41–57.

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. K. (2009), "Debiased estimation of proportions in group testing," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 58, 105–121.

Hughes, G., Gottwald, T. R., and Levy, L. (2002), "The use of hierarchical sampling in the surveillance program for plum pox virus incidence in the United States," *Plant Disease*, 86, 259–263.

Katholi, C.R., and Unnasch, T.R. (2006), "Important experimental parameters for determining infection rates in arthropod vectors using pool screening approaches," *American Journal of Tropical Medicine and Hygiene* 74, 779—785.

Lancaster, H. O., (1961), "Significance tests in discrete distributions", *Journal of the American Statistical Association*, 56, 223–234.

Lui, K. (1995), "Confidence limits for the population prevalence rate based on the negative binomial distribution," *Statistics in Medicine* 14, 1278–1290.

Montesinos-Lopez, O. A., Montesinos-Lopez, A., Crossa, J., Eskridge, K., and Hernandez-Suarez, C.M. (2010), "Sample size for detecting and estimating the proportion of transgenic plants with narrow confidence intervals," *Seed Science Research* 20, 123–136.

Newcombe, R. G. (1998b), "Two-sided confidence intervals for the single proportion: comparison of seven methods," *Statistics in Medicine*, 17, 857–872.

Pritchard, N. A., and Tebbs, J. M. (2011a), "Estimating disease prevalence using inverse binomial pooled testing", *Journal of Agricultural, Biological, and Environmental Statistics*, 16, 70–87.

Pritchard, N. A., and Tebbs, J. M. (2011b), "Bayesian inference for disease prevalence using negative binomial group testing," *Biometrical Journal* 53, 40–56.

Rodriguez-Perez, M. A., Katholi, C.R., Hassan, H. K., and Unnasch, T.R. (2006), "Large-scale entomologic assessment of *Onchocerca volvulus* transmission by poolscreen PCR in Mexico," *American Journal of Tropical Medicine and Hygiene* 74, 1026—1033.

Schaarschmidt, F. (2007), "Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing," *Communications in Biometry and Crop Science*, 2, 32–40.

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.

Thompson, K. (1962), "Estimation of the proportion of vectors in a natural population of insects," (*Biometrics*, 18, 568–578.

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.

Verstraeten, T., Farah, B., Duchateau, L., and Matu, R. (1998), "Pooling sera to reduce the cost of HIV surveillance: a feasibility study in a rural Kenyan district," *Tropical Medicine and International Health* 3, 747–750.
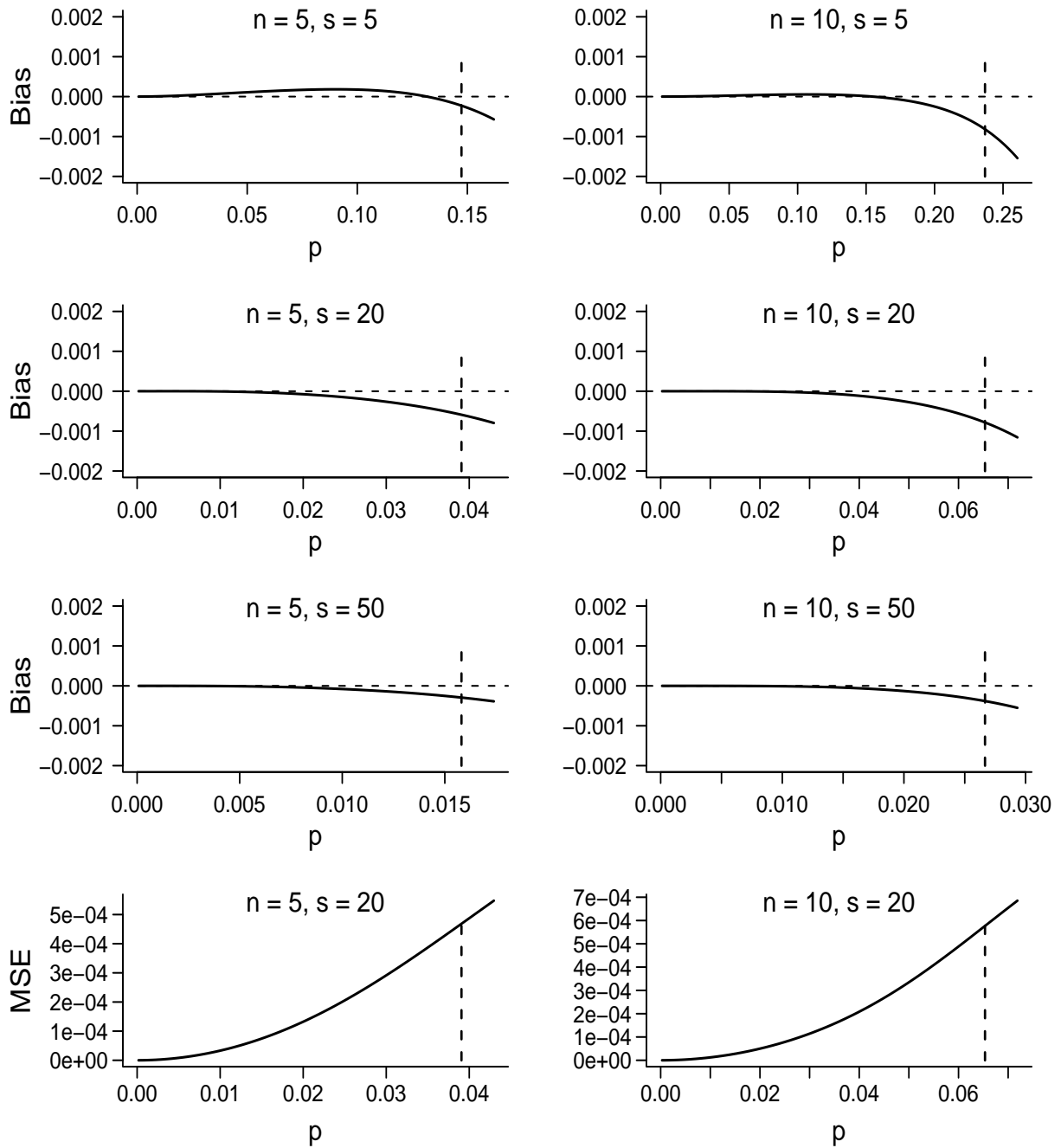
Figure 1:   Bias and MSE of corrected estimator for various combinations of n and s. Vertical dashed line is at $p = \psi$.
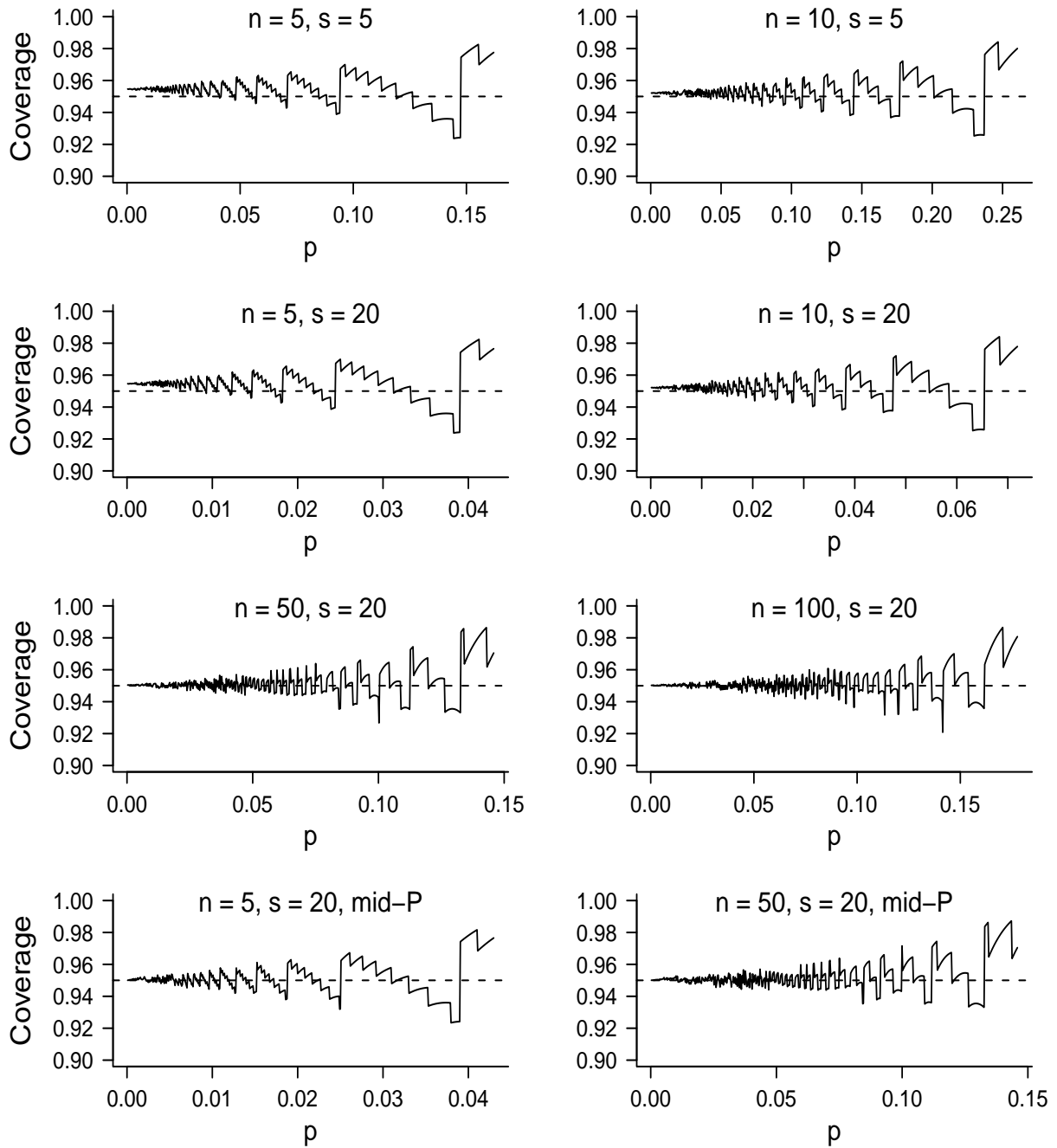
Figure 2:   Coverage probability for skew-corrected score 95% confidence intervals (top three rows) for selected combinations of n and s, and for mid-P exact intervals (bottom row).

Author/s:
Hepworth, G

Title:
Improved Estimation of Proportions Using Inverse Binomial Group Testing

Date:
2013-03-01

Citation:
Hepworth, G. (2013). Improved Estimation of Proportions Using Inverse Binomial Group
Testing. JOURNAL OF AGRICULTURAL BIOLOGICAL AND ENVIRONMENTAL STATISTICS,
18 (1), pp.102-119. https://doi.org/10.1007/s13253-012-0126-6.

Persistent Link:
http://hdl.handle.net/11343/282977