

# Confidence intervals for the difference of two proportions estimated from pooled samples

Brad J. Biggerstaff

Division of Vector-Borne Infectious Diseases

National Center for Zoonotic, Vector-Borne and Enteric Diseases

Centers for Disease Control and Prevention

3150 Rampart Road, Fort Collins, CO 80521

[BBiggerstaff@cdc.gov](mailto:BBiggerstaff@cdc.gov)

January 2, 2008

## SUMMARY

Confidence intervals for the difference of two binomial proportions estimated from pooled samples with unequal pool sizes are presented. Asymptotic methods are used to derive Wald, profile score and profile likelihood ratio intervals. Corrections for bias and skewness of the distribution of the Studentized score statistic are used to improve the profile score interval. Further, the easily computed Wilson score-based interval of Newcombe (1998) is adapted. Coverage and non-coverage probabilities and expected lengths of the confidence intervals are estimated for a range of parameter values expected in application, for both one- and two-sample cases. The skewness-corrected score interval is generally recommended. The methods are applied to a comparison of West Nile virus mosquito infection prevalences by trapping height in field collections from Louisiana in 2003.

**Keywords:** binomial, group testing, score, profile score, profile likelihood, West Nile virus

## 1. Introduction

Estimation of virus infection prevalence in field collected mosquitoes typically requires pooling or grouping of individuals for testing for the presence of virus, as individual level testing can be too expensive and time consuming to be practical. Dorfman (1943) introduced the idea of such pooling in the context of medical screening to identify infected individuals, an approach that came to be known as *group testing*. Pooling of samples for the purpose of estimation was considered by Chiang and Reeves (1962), who developed likelihood methods when pool sizes are equal, later also considered by Sobel and Elashoff (1975). Tu et al. (1995) investigated in detail properties of the maximum likelihood estimator when pool sizes are equal. Hauck (1991) gave score and exact confidence intervals (CI) in the case of equal pool sizes. Chen and Swallow (1990) further investigated appropriate pool size and diagnostic methods for evaluating the binomial assumption.

To estimate infection prevalence of yellow fever virus in mosquitoes, Walter, Hildreth and Beaty (1980) extended likelihood methods to unequal pool sizes, and Farrington (1992) showed that the standard binomial formulation with unequal pool sizes could be cast in a generalized linear models framework using the binomial model with complementary log-log link. When pool sizes differ, computation of the maximum likelihood estimate (MLE) requires iteration, and a simple formula using the Newton-Raphson algorithm is given in Walter et al. (1980). Because this requires the use of a computer, researchers often use a simpler measure given by the number of positive pools divided by the total number of individuals. This measure, called the *minimum infection rate* (MIR) in the entomology literature, assumes that only one individual is positive in a positive pool, effectively ignoring the pooling. The model under such an assumption is unclear. For the small proportions typically encountered ( $p < 1/100$ , say) the MIR does not suffer much bias for typical pool sizes. CIs constructed under the MIR assumption, however, tend to be too narrow because they do not reflect information lost in pooling (Hepworth (1999)).

Hepworth (1999) undertook a detailed investigation of point and interval estimation for

unequal pool sizes. Exact CI methods he developed were published in Hepworth (1996), however, the exact methods are infeasible with more than two or three distinct pool sizes because their computation becomes prohibitive. Field applications for mosquito infection prevalence estimation, and similar such studies, however, give rise to many distinct pool sizes. Hepworth (2005) presented asymptotic likelihood CIs, including the standard Wald, Wilson-based score and likelihood ratio test-based intervals. Additionally, he applied the methods of Bartlett (1953a) and Gart (1991) to adjust the Studentized score statistic for skewness to improve the score-based CI. Hepworth (2005) recommended the skewness-corrected score interval among the asymptotic intervals.

Comparison of proportions estimated from pooled samples has received little attention, even in the equal pools case. Walter et al. (1980) give the Wald CI on the difference when pool sizes differ. McCann and Tebbs (2007) consider multiple testing adjustments for pairwise differences of proportions, but under the restriction that each population's pools are a common size. Our motivating example concerns a comparison of West Nile virus (WNV) infection prevalences in field collected *Culex nigripalpus* mosquitoes trapped at different heights. To be able to quantify such comparisons, we derive asymptotic CIs for the difference of two proportions estimated from pooled samples where the pool sizes may differ. The difference provides a comparison on an absolute scale. One might in some applications be interested in a relative comparison, as provided for example by the ratio. The nature of the comparison in our example is not naturally a relative one, since neither population serves a referent role in the interpretation of the results, so we focus the present research on the difference. We present seven CIs, including one based on the MIR, the Wald interval, the profile score interval, and the profile likelihood interval. Following Hepworth (2005), we apply the methods of Bartlett (1953b), Gart (1991) and Gart and Nam (1990) to adjust the profile score interval for bias and skewness of the Studentized profile score statistic. Our work extends the methods developed for the difference of binomial proportions for unpooled samples (i.e., when all pool sizes are 1), a review of which can be found in Newcombe (1998),

and (Gart and Nam (1990)). In his review, Newcombe introduced an easily computed interval based on the Wilson (1927) score interval with good coverage and average length properties, and we adapt this approach for the pooled case.

The layout of the paper is as follows. The next section introduces the motivating example. One-sample intervals are reviewed in section 3, and these are extended to the two-sample case in section 4. In section 5, performance characteristics of the one-sample intervals are summarized from the literature, and results of a simulation study of coverage and non-coverage properties for both the one- and two-sample cases are presented. The methods are then applied in section 6 to the data from the field example. The paper closes with a brief discussion.

Reflecting our application of interest, we use “positive” or “infected” for a generic binomial “success,” either for individuals or for pools.

## 2. Motivating example

Godsey et al. (2007) studied WNV infection prevalences in *Culex nigripalpus* mosquitoes in Louisiana in 2003. *Cx. nigripalpus* mosquitoes feed on birds, the natural amplifying host for WNV, and mammals, and they are an important vector species in the WNV transmission cycle, which includes humans. Many aspects of the habits and WNV infection prevalences were considered in this study, and for illustration we restrict attention to the comparison of WNV infection prevalences in mosquitoes collected at two different trapping heights at a single location. For this dataset, mosquito collections were made at 1.5 meter and 6 meter heights, one night per week from July 29 through November 18, 2003. Because of the bird feeding habits of *Cx. nigripalpus* mosquitoes, it is thought that their WNV infection prevalences may be higher in those that feed higher in the tree canopy, where the birds roost.

Table 1 about here

The data are summarized in Table 1. Roughly 50% more mosquitoes were trapped at 6 m, and these were grouped into slightly smaller pools on average. Only 1 pool tested

positive for WNV from the 1.5 m height, while 7 pools tested positive at the 6 m height. Single sample estimates (see section 3) of the infection prevalences differed by over 2.5 per 1000, indicating that the prevalence may indeed be higher in the canopy.

### 3. Inference for a single proportion

Assume that the proportion of infected individuals in a given population is  $p$ . Let  $N$  individuals be sampled independently from the population, and group them independently into pools of sizes  $m_i$ , for  $i = 1, 2, \dots, M$ , where  $M$  is the number of distinct pool sizes. For  $i = 1, 2, \dots, M$ , let  $n_i$  be the number of pools of size  $m_i$ , and let  $X_i$  be the number of the  $n_i$  pools that is positive, with  $x_i$  the observed value of  $X_i$ .

The minimum infection rate (MIR)

$$\tilde{p} = \frac{\sum_{i=1}^M x_i}{\sum_{i=1}^M m_i n_i} = \frac{1}{N} \sum_{i=1}^M x_i$$

is commonly computed as a measure of infection prevalence. In using  $\tilde{p}$ , one assumes that only one individual is positive in a positive pool, so that  $\sum_i x_i$  under this assumption is just the number of positive individuals, and the problem thus appears simply to be one of a routine binomial proportion. In this case, a 95% CI for  $p$  based on  $\tilde{p}$  is often computed in practice as the simple, textbook (Wald) interval

$$\tilde{p} \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{N}}, \tag{1}$$

where  $z_{\alpha/2}$  is the  $100(1 - \alpha/2)$ th percentile point of the standard normal distribution. This method ignores the pooling of the individuals in the computation of the “standard error” of  $\tilde{p}$ , which one might expect overstates the precision. One could compute other, standard binomial CIs, e.g., the Wilson (1927) score interval, using these simplifications, but they would all ignore the loss of information through pooling.

Under the sampling described above, for each  $i = 1, 2, \dots, M$ , the distribution of  $X_i$  is

binomial( $n_i, 1 - (1 - p)^{m_i}$ ). A log-likelihood for  $p$  with data  $x = (x_1, x_2, \dots, x_M)$  is thus

$$l(p; x) = l(p) = \sum_{i=1}^M x_i \log [1 - (1 - p)^{m_i}] + \log (1 - p) \sum_{i=1}^M m_i (n_i - x_i). \quad (2)$$

The maximum likelihood estimator (MLE) of  $p$  is the maximizer of equation (2), obtained as the solution  $\hat{p}$  to the score equation

$$\frac{\partial l(p; x)}{\partial p} = S(p) = \frac{1}{1 - p} \sum_{i=1}^M \left[ \frac{m_i x_i}{1 - (1 - p)^{m_i}} - m_i n_i \right] = 0. \quad (3)$$

Unless all pool sizes  $m_i$  are equal,  $\hat{p}$  must be obtained iteratively. The Newton-Raphson method was used by Walter et al. (1980) to obtain

$$\hat{p}^{(r+1)} = \hat{p}^{(r)} + \frac{N - \sum_{i=1}^M m_i x_i / [1 - (1 - \hat{p}^{(r)})^{m_i}]}{\sum_{i=1}^M m_i^2 x_i (1 - \hat{p}^{(r)})^{m_i - 1} / [1 - (1 - \hat{p}^{(r)})^{m_i}]^2},$$

used to compute the MLE by iterating over successive values  $\hat{p}^{(r)}$ , for  $r = 0, 1, 2, \dots$ , until convergence. A convenient starting value for this iteration is the MIR,  $\hat{p}^{(0)} = \tilde{p}$ .

The case that all pools are positive, i.e.,  $x_i = n_i$ , for all  $i = 1, 2, \dots, M$ , represents a particular difficulty, and various remedies have been proposed, as summarized in Hepworth (1999). We exclude this case, acknowledging that all individual-level information is lost, possibly because too few pools have been examined or  $p$  is simply too large for pooling to be effective with the given pool sizes. Estimation procedures provide no recourse, so careful determination of pool sizes ahead of field preparation of samples for testing may be required to insure that not all pools are positive. This would involve balancing resources for testing  $N$  individuals in  $\sum_i n_i$  pools with anticipated levels for  $p$ . If possible, sequential pooling and testing, as discussed in Hepworth (1999) and references cited therein, may be the most reliable way to produce successful and efficient estimation.

Walter et al. (1980) derived the asymptotic variance of  $\hat{p}$ ,  $\text{Var}(\hat{p}) = I(p)^{-1}$ , where

$$I(p) = \sum_{i=1}^M \frac{m_i^2 n_i (1-p)^{m_i-2}}{1 - (1-p)^{m_i}} \quad (4)$$

is the Fisher information. The  $100(1 - \alpha)\%$  score CI is obtained as the solutions to

$$z(p) = \frac{S(p)}{\sqrt{I(p)}} = \pm z_{\alpha/2}. \quad (5)$$

As with all the estimating equations given, solutions to equation (5) must in general be found numerically, though simplification may be possible with equal pool sizes. Following Gart (1991), Hepworth (2005) modified the Studentized score statistic in equation (5) to adjust for skewness of the distribution of  $z(p)$ . Hepworth computed the third central moment of  $S(p)$  as

$$\mu_3[S(p)] = \sum_{i=1}^d \frac{m_i^3 n_i (1-p)^{m_i-2} [2(1-p)^{m_i} - 1]}{[1 - (1-p)^{m_i}]^2}. \quad (6)$$

The skewness-corrected score interval is then obtained as the solutions to

$$z(p) - \gamma(p) \frac{z_{\alpha/2}^2 - 1}{6} = \pm z_{\alpha/2}, \quad (7)$$

where  $\gamma(p) = \mu_3[S(p)]/I(p)^{3/2}$ . Finally, Hepworth computed the bias of  $\hat{p}$  as

$$b(p) = \frac{1}{2} [\text{Var}(p)]^2 \sum_{i=1}^d \frac{m_i^2 (m_i - 1) n_i (1-p)^{m_i-3}}{1 - (1-p)^{m_i}} \quad (8)$$

which we will use for the two-sample problem below.

#### 4. Inference for the difference of two proportions

To develop CIs for the difference of proportions from two independently sampled populations, extend the notation to include a population index  $k = 1, 2$  in the natural way:  $p_k; \tilde{p}_k; \hat{p}_k$ ;

$N_k$ ;  $m_k = (m_{ks_k})$ ;  $n_k = (n_{ks_k})$ ;  $X_k = (X_{ks_k})$ ; and  $x_k = (x_{ks_k})$ , for  $s_k = 1, 2, \dots, M_k$ . When using both sets of indices  $s_1$  and  $s_2$ , we will use simpler, single index notation using  $i$  and  $j$  when there should be no confusion. Lastly, define the parameter of interest to be  $d = p_1 - p_2$ . When a single-sample function is used for population  $k = 1, 2$  below, a subscript indicates that the corresponding values for  $x_k$ ,  $m_k$  and  $n_k$  should be used in their evaluation.

#### 4.1 Confidence intervals based on the MIR

Because of independence, an MIR-based point measure is  $\tilde{p}_1 - \tilde{p}_2$ , and the CI in equation (1) is readily adapted as

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{N_1} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{N_2}}.$$

#### 4.2 Likelihood based confidence intervals

Each  $X_{ks_k} \sim \text{binomial}(n_{ks_k}, 1 - (1 - p_k)^{m_{ks_k}})$ , for  $s_k = 1, 2, \dots, M_k$  and  $k = 1, 2$ . A log-likelihood for parameters  $p_1$  and  $p_2$  is therefore  $l(p_1, p_2; x_1, x_2) = l(p_1; x_1) + l(p_2; x_2)$ , by independence. We use  $l$  to denote all log-likelihood functions used, noting that their arguments should make clear which particular function is intended.

**4.2.1 Wald intervals** The asymptotic joint distribution of the MLEs  $(\hat{p}_1, \hat{p}_2)$  is normal with mean  $(p_1, p_2)$  and covariance matrix  $\text{diag}\{\text{Var}_1[\hat{p}_1], \text{Var}_2[\hat{p}_2]\}$ . The MLE of  $d$  is simply  $\hat{d} = \hat{p}_1 - \hat{p}_2 \stackrel{\text{asy}}{\sim} N(d, \text{Var}_1[\hat{p}_1] + \text{Var}_2[\hat{p}_2])$ . The  $100(1 - \alpha)\%$  Wald CI for  $d$  is thus

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}_1[\hat{p}_1] + \widehat{\text{Var}}_2[\hat{p}_2]},$$

where the variances have been estimated by using the MLEs for  $p_1$  and  $p_2$  in equation (4).

4.2.2 *Profile score intervals* To compute score-based CIs, it is convenient to introduce  $s = p_1 + p_2$  and reparameterized the log-likelihood in terms of  $(d, s)$  as

$$l(d, s) = l\left(\frac{s+d}{2}, \frac{s-d}{2}\right) = l_1\left(\frac{s+d}{2}\right) + l_2\left(\frac{s-d}{2}\right)$$

Inference for  $d$  based on  $l(d, s)$  treats  $s$  as a nuisance parameter, and one may use the profile log-likelihood function given by  $l_P(d) = \max_{s|d} l(d, s) = l(d, \hat{s}_d)$ , where the maximum is taken over all  $d$  consistent with the given value of  $s$  (see Barndorff-Nielsen and Cox (1994)). The maximum profile likelihood estimate of  $d$ , that is, the maximizer of  $l_P(d)$ , is simply the MLE  $\hat{d}$ . Following Gart (1991), one may compute the score functions

$$\begin{aligned} S_d(d, s) &= \frac{\partial l(d, s)}{\partial d} = \frac{1}{2} \left[ S_1\left(\frac{s+d}{2}\right) - S_2\left(\frac{s-d}{2}\right) \right] \\ S_s(d, s) &= \frac{\partial l(d, s)}{\partial s} = \frac{1}{2} \left[ S_1\left(\frac{s+d}{2}\right) + S_2\left(\frac{s-d}{2}\right) \right] \end{aligned} \tag{9}$$

Because  $d$  cannot be separated from  $s$  in equation (9), the score function  $S_d(d, s)$  depends on the nuisance parameter  $s$ . To overcome this, one fixes  $d$  and computes the solution  $\hat{s}_d$  to  $S_s(d, s) = 0$  in  $s$  and then bases inference for  $d$  on  $S_d(d, \hat{s}_d)$ .

The exact moments of  $S_d(d, \hat{s}_d)$  are not available, but Bartlett (1953b) (see also Gart (1991)) gave the asymptotic moments based on the approximation

$$S_d(d, \hat{s}_d) \approx S_d(d, s) - \frac{I_{ds}}{I_{ss}} S_s(d, s),$$

where  $I_{ss} = E\left[-\frac{\partial^2 l(d, s)}{\partial s^2}\right]$ ,  $I_{ds} = E\left[-\frac{\partial^2 l(d, s)}{\partial d \partial s}\right]$  and  $I_{dd} = E\left[-\frac{\partial^2 l(d, s)}{\partial d^2}\right]$  are the components of the Fisher information matrix. The asymptotic mean and variance of  $S_d(d, \hat{s}_d)$  are then 0 and

$$\text{Var}[S_d(d, \hat{s}_d)] = I_{dd} - \frac{I_{ds}^2}{I_{ss}} \equiv I_{d \cdot s} \tag{10}$$

Evaluating the derivatives in the present case yields

$$I_{dd} = I_{ss} = \frac{1}{4} \left[ I_1 \left( \frac{s+d}{2} \right) + I_2 \left( \frac{s-d}{2} \right) \right]; \quad I_{ds} = \frac{1}{4} \left[ I_1 \left( \frac{s+d}{2} \right) - I_2 \left( \frac{s-d}{2} \right) \right] \quad (11)$$

(See the online appendix for details.) Substituting these expressions into equation (10), the asymptotic variance of  $S_d(d, \hat{s}_d)$  is

$$\text{Var}[S_d(d, \hat{s}_d)] = \left[ \text{Var}_1 \left( \frac{\hat{s}_d + d}{2} \right) + \text{Var}_2 \left( \frac{\hat{s}_d - d}{2} \right) \right]^{-1}. \quad (12)$$

The Studentized score statistic

$$\begin{aligned} Z(d) &= \frac{S_d(d, \hat{s}_d)}{\sqrt{I_{d,s}}} \\ &= \frac{1}{2} \left[ S_1 \left( \frac{\hat{s}_d + d}{2} \right) - S_2 \left( \frac{\hat{s}_d - d}{2} \right) \right] \sqrt{\text{Var}_1 \left( \frac{\hat{s}_d + d}{2} \right) + \text{Var}_2 \left( \frac{\hat{s}_d - d}{2} \right)} \end{aligned}$$

is asymptotically standard normal, and CIs are computed by solving  $Z(d) = \pm z_{\alpha/2}$  for  $d$ . When solving this equation, for each iteration of  $d$ , a new value for  $\hat{s}_d$  must be computed by solving  $S_s(d, s) = 0$  in  $s$ .

Bartlett (1955) showed that  $Z(d)$  has non-negligible bias of first order. Following calculations detailed in the online appendix, the (approximate) bias of  $Z(d)$  is

$$B(d, \hat{s}_d) = \frac{R^{3/2}}{\sqrt{\text{Var}_2 \left( \frac{\hat{s}_d - d}{2} \right)}} b_1 \left( \frac{\hat{s}_d + d}{2} \right) - \frac{(1-R)^{3/2}}{\sqrt{\text{Var}_1 \left( \frac{\hat{s}_d + d}{2} \right)}} b_2 \left( \frac{\hat{s}_d - d}{2} \right), \quad (13)$$

where

$$R = \frac{I_1 \left( \frac{\hat{s}_d + d}{2} \right)}{I_1 \left( \frac{\hat{s}_d + d}{2} \right) + I_2 \left( \frac{\hat{s}_d - d}{2} \right)}$$

and using the one-sample bias in equation (8). Further calculation (detailed in the online

appendix) yields the third central moment of  $S_d(d, \hat{s}_d)$  as

$$M_3 [S_d(d, \hat{s}_d)] = (1 - R)^3 \mu_{3,1} \left( \frac{\hat{s}_d + d}{2} \right) - R^3 \mu_{3,2} \left( \frac{\hat{s}_d - d}{2} \right) \quad (14)$$

The approximate skewness is thus  $\gamma(d, \hat{s}_d) = M_3 [S_d(d, \hat{s}_d)] / I_{d,s}(d, \hat{s}_d)^{3/2}$ , so that the bias- and skewness-corrected Studentized score statistic is

$$\tilde{Z}(d) = Z(d) - B(d, \hat{s}_d) - \gamma(d, \hat{s}_d) \frac{z_{\alpha/2}^2 - 1}{6}.$$

Solutions to  $\tilde{Z}(d) = \pm z_{\alpha/2}$  thus yield bias- and skewness-corrected (BiasSkewScore)  $100(1 - \alpha)\%$  score confidence limits for  $d$ . We also compute CIs excluding the bias correction  $B(d, \hat{s}_d)$ , and we refer to these intervals simply as skewness-corrected score intervals (SkewScore).

*4.2.3 Profile likelihood ratio test intervals* Inversion of the profile likelihood ratio test of the hypothesis  $d = 0$  versus  $d = d_0$  provides the last, standard asymptotic CI. The profile log-likelihood ratio statistic  $r(d_0) = 2[l_P(\hat{d}) - l_P(d_0)]$  has an asymptotic chi-squared distribution with 1 degree of freedom (Barndorff-Nielsen and Cox (1994)). Because the profile log-likelihood ratio statistic equals the log-likelihood ratio statistic here, we can express  $r(d_0)$  as

$$\begin{aligned} r(d_0) &= 2[l_P(\hat{d}) - l_P(d_0)] = 2[l(d, \hat{s}_d) - l(d_0, \hat{s}_{d_0})] \\ &= 2 \left\{ \left[ l_1 \left( \frac{\hat{s}_d + d}{2} \right) - l_1 \left( \frac{\hat{s}_{d_0} + d_0}{2} \right) \right] - \left[ l_2 \left( \frac{\hat{s}_d - d}{2} \right) - l_2 \left( \frac{\hat{s}_{d_0} + d_0}{2} \right) \right] \right\} \\ &= 2 \left[ r_1 \left( \frac{\hat{s}_d + d}{2} \right) - r_2 \left( \frac{\hat{s}_d - d}{2} \right) \right] \end{aligned}$$

where  $r_k(d)$  is the one-sample log-likelihood ratio test statistic. Asymptotic  $100(1 - \alpha)\%$  confidence limits are computed by solving  $r(d_0) = \chi_{1;\alpha}^2$ , where  $\chi_{1;\alpha}^2$  is the  $100(1 - \alpha)$ th

percentile point of the chi-squared distribution with 1 degree of freedom. We denote this CI as the PLRT interval; the corresponding one-sample CI based on  $r(d)$  we shall denote LRT.

### 4.3 *Square-and-add Walter intervals*

Newcombe (1998) summarized and evaluated eleven CIs for the difference of two independent proportions in the usual, unpooled case. Among these, he introduced an interval that uses the endpoints of the individual proportions' Wilson score-based CIs (Wilson (1927)) directly in the formula for the Wald CI for the difference. Newcombe shows that these intervals perform well in the unpooled case, so we adapt this method to the pooled samples. For each  $k = 1, 2$ , let  $L_k$  and  $U_k$  be the lower and upper  $100(1 - \alpha)\%$  individual score confidence limits for  $p_k$  obtained as the solutions to equation (5). Set

$$\begin{aligned}\delta &= \sqrt{(\hat{p}_1 - L_1)^2 + (U_2 - \hat{p}_2)^2} \\ \epsilon &= \sqrt{(U_1 - \hat{p}_1)^2 + (\hat{p}_2 - L_2)^2}.\end{aligned}$$

A  $100(1 - \alpha)\%$  CI for  $d$  is then  $(\hat{d} - \delta, \hat{d} + \epsilon) = (\hat{p}_1 - \hat{p}_2 - \delta, \hat{p}_1 - \hat{p}_2 + \epsilon)$ , which we call the Square-and-Add Walter (SAW) interval.

### 4.4 *Zero positive pools*

In application, it is common for there to be no positive pools in a sample, that is, that  $x_{ks_k} = 0$ , for all  $s_k$  in either or both populations  $k = 1, 2$ . Newcombe (1998, Appendix 2) analyzes this case in detail for CIs for  $d$  without pooling, and his analysis is applicable here with little modification.

When all pools from both populations are negative,  $\hat{p}_1 = \hat{p}_2 = 0$ , and the interval endpoints for the Wald interval are degenerate; we simply set the interval to  $(0, 0)$  and acknowledge that the Wald interval is useless in this case. For the score and SAW intervals, one uses  $(-U_2, U_1)$ , where these endpoints are the single-sample score upper limits. With or without the bias correction, the statistic  $\tilde{Z}(d)$  exhibits anomalous behavior as a function of  $d$  in the case of no positive pools in both populations, and so we substitute in this case

the unadjusted score CI. For the PLRT interval, notice that the log-likelihood function in equation (2) reduces to that in the standard, unpooled case: we know that all individuals must be negative if all the pools are, so the pooling has no impact. The PLRT interval in the pooled case is thus equivalent to the unpooled case,  $(-1 + e^{-\chi_{1;\alpha}^2/(2N_1)}, 1 - e^{-\chi_{1;\alpha}^2/(2N_2)})$  (see Newcombe (1998)).

When only one population has zero positive pools, the Wald interval reduces to endpoints based only on the data from the population with a positive pool, an artifact observed for the Wald and square-and-add Wilson methods in the unpooled case by Newcombe (1998, Table II). While this is abhorrent, we compute the interval without adjustment. For the three score and the PRLT methods, the interval equations may be utilized as described; see Newcombe (1998) for discussion. The SAW interval is directly computable in this case.

## 5. Simulation

We compared all the CIs by estimating coverages, non-coverages and expected lengths using a simulation study. Coverage and directional non-coverage were evaluated following the paradigm set out in Newcombe (1998). For the one-sample CIs in the applications envisioned here, we are interested only in small values of  $p$ . Directional non-coverage may be characterized in this case as left- and right-non-coverage, indicating that the CI is wholly above or below the true value of  $p$ , respectively. Because there is symmetry in the ordering of the populations for  $d = p_1 - p_2$ , directional non-coverage is more appropriately characterized relative not only to  $d$  but to interval location with respect to  $d$  and the central parameter space point, 0. In this case, *distal* (*mesial*) non-coverage occurs when a CI misses  $d$  toward (away from) 0. Define non-coverage symmetry as the difference in proportional non-coverage, with a negative value indicating right- or distal-non-coverage; a value of 0 indicates symmetric non-coverage, while values of  $\pm 1$  represent totally asymmetric non-coverage. Symmetric non-coverage is preferable, otherwise intervals are directionally biased in location.

For the one-sample case, Hepworth (2005) evaluated exact coverage and interval length properties of the CIs for  $p$  for  $N \in \{100, 200, 400\}$ , with five scenarios of pool sizes ranging

5 to 25. Echoing standard results for the unpooled case, Hepworth concluded that the Wald interval was poor and did not warrant serious consideration. Among the other methods, he noted that the LRT interval was generally too anticonservative. The skewness-corrected score interval was recommended based on coverage and, secondarily, on length and on its property of functional invariance.

We augment the available one-sample evaluations by considering values typical in our motivating example and like applications. Specifically, we take  $p \in \{1, 1.5, 2, 5, 10\} / 1000$ . For our primary evaluation, we take  $N = 1000$ , a modest sample size in typical entomological applications; further results for  $N = 500$  and  $N = 5000$  are summarized here, with detail in the online appendix. Pool sizes take values  $m_i \in \{5, 10, 25, 50\}$ , in counts laid out in Table 2. For  $N = 500$ , the counts in Table 2 were halved, while for  $N = 5000$ , the counts were multiplied by 5. These parameter space points  $p$ ,  $N$  and  $(m_i, n_i)$  were then carried over to the two-sample evaluation of CIs for  $d$ .

Table 2 about here

The focus of this work is CIs for  $d$ , so it was necessary because of computational time to provide a somewhat limited evaluation of CIs for  $p$ . The full array of pooling combinations given in Table 2 applies to the evaluations for  $d$ , while those used for  $p$  are marked. Although more limited, these still reflect a range in the amount of pooling, and they provide results consistent with previous work and with the evaluations for  $d$ .

Each value of  $p$  was used for each pooling combination  $(m_i, n_i)$  given in Table 2, and each of these was used for each value of  $N = 500, 1000, 5000$ . This resulted in  $25 \times 3 = 75$  parameter combinations in the one-sample study, and  $4900 \times 3 = 14700$  parameter combinations in the two-sample study. For the two-sample evaluation with  $N = 1000$ , a simulation run for a given set of parameter values was composed of 25000 realizations of binomial variates with sizes  $n_{ks_k}$  and probabilities  $1 - (1 - p_k)^{m_{ks_k}}$ , for  $s_k = 1, 2, \dots, M_k$  and  $k = 1, 2$ . At nominal 95% coverage, the standard errors of the individual run coverage probability estimates are thus 0.0014, so that stated coverages are within 0.0027 with 95%

probability. One-sample realizations were generated analogously. To ease the computational burden, for the one-sample case, and for  $N = 500$  and  $N = 5000$  in both cases, the number of realizations was reduced to 1000, so that stated coverages are within 0.014 with 95% probability for these parameter settings. All computations were made in R (<http://www.r-project.org>), version 2.2.1. Whenever the term “coverage” is used, it should be understood to be the average of the individual runs’ coverages over the parameters not under consideration.

In all our interpretations, we weigh a CI’s performance by coverage primarily, non-coverage secondarily, and only then average length.

### 5.1 *Results: one-sample confidence intervals for $p$*

Table 3 records a coarse summary of the results, listing coverage, non-coverage, non-coverage symmetry and average length over all parameter settings, and Figure 1 graphs coverages and non-coverages by  $p$ . The MIR and Wald intervals had coverages far too low, though they improved with  $p$ . This was coupled with drastically asymmetric non-coverage, reflecting the inability of the intervals to handle all negative pools. The LRT, Score and SkewScore intervals had reasonably close to nominal coverage, with the LRT and SkewScore intervals exhibiting nearly the same properties. Though left-non-coverage dominated for small  $p$  for all these intervals, symmetry improved as  $p$  increased. Apparent here is the effect of the skewness correction to the score, which clearly improved non-coverage symmetry. Further, the LRT interval tended to be slightly shorter than the score intervals (see the online appendix).

Figure 3 shows a plot of coverage against the total number of pools, grouped by  $p$ . There was no strong influence of the number of pools on coverage, except for the MIR with larger values of  $p$ , when coverage was too low for few pools. This observations makes sense, for with larger pools (so fewer pools for a fixed total sample size), there is greater chance that more than one positive individual will be in a positive pool, especially as  $p$  increases. Coverage worsened slightly with the number of pools for the Wald interval, and, to a lesser degree, for the LRT interval.

To summarize results for  $N = 500$ , coverage properties noted here for  $N = 1000$  were similar in direction and greater in magnitude, and all intervals exhibited asymmetric non-coverage in the directions shown in Figure 1 until  $p$  reached 10/1000. For  $N = 5000$ , the trends were similar, but coverage improved to nearly nominal but for the smallest  $p$ , except that coverage was too anti-conservative for the MIR interval for large  $p$ , reflecting the phenomenon discussed above.

For the one-sample case, our results generally reiterate Hepworth (2005), with further detail on directional non-coverage illustrating how the SkewScore interval provides an improvement over the Score. We found, somewhat differently from Hepworth, that the LRT interval performed nearly identically to the SkewScore interval, with slightly more symmetric non-coverage and slightly shorter average length. The reasons for this discrepancy are unclear, but may be because our evaluations were for larger sample sizes, or because we focused on very small  $p$ .

Table 3 about here

Figure 1 about here

## 5.2 Results: two-sample confidence intervals for $d$

As an initial summary, the coverages, non-coverages and mean lengths for the CIs averaged over all simulations and parameter values for  $N = 1000$  are reported in Table 3. (Qualitatively similar results occur when the values are averaged over  $N$  as well; because the number of realizations in the simulations for  $N = 1000$  was higher, we report these here.) From these overall summaries we see that the MIR, Wald and PLRT intervals tended to fall short of nominal, 95% coverage, while the score intervals and the SAW intervals tended to have at least nominal coverage, with the SAW interval being most conservative. Figure 2 shows coverages and non-coverages plotted by  $|d|$ . Generally, the score and the SAW intervals tended to be conservative for smaller and larger values of  $|d|$ , with closer to nominal coverage for middle values. The SAW interval was generally more conservative than the score

intervals. The PLRT interval was anti-conservative generally, with coverage closer to nominal as  $|d|$  increased. Both the MIR and Wald intervals were generally anti-conservative, with coverages approaching 85% for  $|d| \sim 0$ , though for  $d = 0$  both of these intervals were conservative, due to the degeneracy of the intervals for zero positive pools. Coverage for the MIR interval decreased with increasing  $|d|$ , with coverage near nominal for middle values of  $|d|$ ; relationships between coverage for the MIR and  $|d|$  were not easily characterized. Coverage for the Wald interval was similar in nature to the MIR interval, though the Wald interval's coverage tended to be closer to nominal. Non-coverage was predominantly distal for the MIR and Wald intervals, except when  $d = 0$ , when it must be mesial for any method. The PRLT interval had reasonably symmetric non-coverage, except for small  $|d|$ . Finally, as expected, the SkewScore and BiasSkewScore intervals had more symmetric non-coverage than the Score, especially for middle values of  $|d|$ . The BiasSkewScore and SkewScore intervals were nearly indistinguishable.

Figure 2 about here

Figure 3 about here

Figure 3 shows that the amount of pooling had some but not a drastic effect on the coverage, except for the MIR interval. The MIR interval had, for larger values of  $|d|$ , poorer coverage for fewer total pools—that is, for more pooling—and this was seen even more drastically for sample sizes 5000, but not as much for sample sizes 500. This phenomenon was expected for the MIR, because the assumption of one positive in a positive pool is less likely to be met in these cases. Coverage also increased for the SAW interval for larger  $|d|$ , and in the score and, more clearly, in the PLRT intervals for small  $|d|$ . These effects on these intervals are more apparent for sample size 500, and less so for sample size 5000. The skewness correction to the score interval appeared to have lessened the impact of the amount of pooling on coverages for each  $N$ , as declining coverages away from nominal for the Score were flattened somewhat in the SkewScore and BiasSkewScore for each sample size.

Similar to the one-sample evaluation above, the trends in the performances of the intervals in the two-sample case were similar to those seen for  $N = 500$  and  $N = 5000$ . When both populations' total sample sizes were equal to 500, coverages were generally exaggerated in the direction away from nominal compared to  $N = 1000$ , and non-coverage was more asymmetric; and, except for the MIR, when both populations' total sample sizes equaled 5000, coverages were closer to nominal, and non-coverage is more symmetric. Again reflecting the problem with the MIR, coverage for the MIR when  $N_1 = N_2 = 5000$  dropped drastically below nominal for  $|d| > 5$  per 1000; non-coverage in this case was nearly entirely distal.

When the total sample sizes  $N_1$  and  $N_2$  were unequal, trends in coverage and non-coverage persisted and were not greatly affected, though they were of course no longer symmetric about  $d = 0$ . Because of the interplay of the  $N_k$  and  $p_k$  in the variance, trends with  $N_k$  and  $p_k$  in these cases are not easily summarized. The most obvious impacts were on the MIR and Wald intervals, where coverages differed a great deal on either side of  $d = 0$ . The PLRT interval was more affected by the asymmetry than the score intervals, while the skewness correction ameliorated the impact on the Score interval. Coverage improved with sample size in either population, but this gain depended also on the underlying value for  $p$  in that population.

### 5.3 *Summary and recommendations*

Our one-sample evaluation supports the recommendation of Hepworth (2005) for the general use of the SkewScore interval among asymptotic methods, though in our evaluation the LRT interval was a close competitor.

The results of our study of CIs for  $d$  concludes (perhaps now unsurprisingly) that the SkewScore interval is to be recommended for this case, too. The MIR and Wald intervals had poor coverage and symmetry properties and cannot be recommended, as seen repeatedly by many researchers in other contexts. The PLRT interval's coverage was slightly anticonservative, and had symmetric non-coverage, a result seen for the unpooled case in Newcombe (1998). The SAW interval was generally too conservative, if symmetric. The Score inter-

val had coverage slightly above nominal, but was too mesial, while the SkewScore interval maintained this coverage and improved symmetry. Finally, the bias correction did not offer an improvement over the skewness correction alone.

## 6. Application

Recalling the goal to compare WNV prevalences in *Culex nigripalpus* mosquitoes by two different trap heights, Table 4 records the various 95% CIs for difference in the true WNV infection prevalences. All CIs are fairly wide, reflecting the relatively few (on the scale of the prevalences) numbers of individual mosquitoes collected and pools tested. None of these CIs indicates a statistically significant difference in the WNV infection prevalences between the trapping heights. The general conclusions from the simulation study are reflected in these results. The MIR and Wald intervals are relatively narrow, with the MIR interval shifted lower because it is centered about the smaller MIR point estimate, and the SAW interval is the widest. The skewness correction has shortened the score interval, most clearly on the left endpoint, while the bias correction had only slight effect. The PLRT is narrower than the score intervals, as expected. Based on the simulation study, the SkewScore interval is preferred, and so we would report the CI  $(-0.573, 6.824)$  per 1000.

Table 4 about here

## 7. Discussion

Computation of all the likelihood-based CIs presented here requires numerical computation using specially written software and employing numerical root-finding routines. A natural alternative to avoid this is to use bootstrap methods, which would have the further benefit of not relying on asymptotic distributional approximations. We developed the following nonparametric bootstrapping method as a component of this work for the two-sample case (the one-sample version is clear from this, too): for each population, select a so-called bootstrap sample by independent, with-replacement resampling of the pools (so positive or negative for each pool) *within* each set of  $n_{ks_k}$  pools of common pool size  $m_{ks_k}$ , for each

$s_k = 1, 2, \dots, M_k$  and  $k = 1, 2$ . (Equivalently, draw independent deviates from binomial distributions with sizes  $n_{ks_k}$  and success probabilities equal to the proportions  $x_{ks_k}/n_{ks_k}$  of positive pools of sizes  $m_{ks_k}$ .) Then for each bootstrap sample  $h = 1, 2, \dots, H$ , compute the MLEs  $\hat{p}_1^{[h]}$  and  $\hat{p}_2^{[h]}$  separately and form  $\hat{d}^{[h]} = \hat{p}_1^{[h]} - \hat{p}_2^{[h]}$ . Then a point estimator of  $d$  is  $\hat{d}_{\text{boot}} = H^{-1} \sum_h \hat{d}^{[h]}$ , and a CI may be computed from the empirical distribution of the  $\{\hat{d}^{[h]}\}$  using, for example, the standard empirical quantile limits (see, among others, Davison and Hinkley (1997)). We found in preliminary evaluations that this method performed similarly to the Wald CI, owing to the propensity to produce bootstrap samples with all negative pools. We therefore excluded this approach from detailed development and evaluation, but researchers may find it useful for large sample sizes or larger prevalences, though specific recommendations await future work.

Results of the simulation study indicate that the bias correction to the score statistic is unnecessary. Because the total sample sizes and parameter settings in the simulation study were fairly comparable for the two groups, the weighted individual biases in equation (13) are on the same order and thus effectively diminish this term. A similar observation was made by Gart and Nam (1990) in the unpooled case. One may find in some applications, however, that sample sizes or variances differ enough between groups that this correction may be required, though we did not investigate here when such a correction may be necessary.

The SAW interval as implemented here used the unadjusted, one-sample score interval. A refinement to the SAW interval would be to use the one-sample SkewScore's endpoints instead of the Score's. Further, we used the MLEs of  $p$  in the SAW interval, but a bias-corrected version of the MLE was presented in Hepworth (1999) that could be used for the point estimator instead. These adjustments to the SAW interval might reasonably be expected to improve the conservative nature of the interval observed here.

Different approaches to comparing proportions from pooled samples are provided by the generalized linear models (GLM) framework. Farrington (1992) showed that the binomial model with a complementary log-log link provides a natural way to model the populations'

prevalences, but this does not provide a directly and easily interpreted comparison of the proportions, as with the difference. Although we have focused our attention on the difference  $p_1 - p_2$ , as noted in the introduction, one may profitably compare the proportions using the ratio  $\rho = p_1/p_2$ , and, with the small values considered in our application, some may consider this a more natural measure. Approximate inference for  $\rho$  may be carried out by using a Poisson GLM, also shown in Farrington (1992). The choice of which measure to use,  $d$  or  $\rho$ , in any given application is largely one of personal preference, and extensions of the methods developed here to  $\rho$  would provide researchers useful options.

To make the methods developed here more accessible to researchers, computer code for the statistical packages R ([www.r-project.org](http://www.r-project.org)) or S-Plus (Insightful Corp., Seattle, WA) is available from the author.

#### ACKNOWLEDGEMENTS

I would like to thank Graham Hepworth for sharing his dissertation and for helpful discussions, Mark Delorey for a careful reading of the manuscript, and Marvin Godsey for the data used in the example. I extend grateful appreciation to Professor Robert G. Newcombe for providing invaluable suggestions in a thorough and thoughtful review for the journal; and thanks also to the anonymous reviewer and associate editor for their detailed suggestions that helped to improve the manuscript. Finally, I thank the Department of Statistics at Colorado State University for the use of their computing facilities.

## REFERENCES

- Barndorff-Nielsen, O. and Cox, D. (1994). *Inference and asymptotics*. Chapman & Hall, London.
- Bartlett, M. S. (1953a). Approximate confidence intervals. *Biometrika* **40**, 12–19.
- Bartlett, M. S. (1953b). Approximate confidence intervals: II. more than one unknown parameter. *Biometrika* **40**, 306–317.
- Bartlett, M. S. (1955). Approximate confidence intervals: III. a bias correction. *Biometrika* **42**, 201–204.
- Chen, C. L. and Swallow, W. H. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics* **46**, 1035–1046.
- 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.
- Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, U.K.
- Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics* **14**, 436–440.
- Farrington, C. P. (1992). Estimating prevalence by group testing using generalized linear models. *Statistics in Medicine* **11**, 1591–7.
- Gart, J. J. (1991). An application of score methodology: Confidence intervals and tests of fit for one-hit curves. In *Handbook of Statistics*, volume 8, pages 395–406. Elsevier Science Publishers.
- Gart, J. J. and Nam, J.-M. (1990). Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension to multiple tables. *Biometrics* **46**, 637–643.
- Godsey, M., King, R., Burkhalter, K., Colton, L., Sutherland, G., Charnetzky, D., Ezenwa, V., Coffee, M., Milheim, L., Delorey, M., Palmisano, C., Wesson, D., Taylor, V. and

- Guptill, S. (2007). Ecology of potential vectors of West Nile virus, southeastern Louisiana. *Emerging Infectious Diseases* in preparation.
- Hauck, W. W. (1991). Confidence intervals for seroprevalence determined from pooled sera. *Annals of Epidemiology* **1**, 277–81.
- Hepworth, G. (1996). Exact confidence intervals for proportions estimated by group testing. *Biometrics* **52**, 1134–1146.
- Hepworth, G. (1999). *Estimation of proportions by group testing*. PhD thesis, The University of Melbourne.
- 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.
- McCann, M. H. and Tebbs, J. M. (2007). Pairwise comparisons for proportions estimated by pooled testing. *Journal of Statistical Planning and Inference* **137**, 1278–1290.
- Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* **17**, 873–90.
- Sobel, M. and Elashoff, R. M. (1975). Group testing with a new goal, estimation. *Biometrika* **62**, 181–193.
- 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 population of organisms using pools of variable size. *American Journal of Epidemiology* **112**, 124–8.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212.

**Table 1***Summary of Culex nigripalpus mosquitoes trapped at heights of 6 m and 1.5 m.*


---



---

Height	6 m	1.5 m
Total	2,021	1,324
No. pools	53	31
Avg. pool size	38.1	42.7
Min : max pool size	1 : 50	5 : 100
No. positive pools	7	1
$\hat{p}$ (per 1000)	3.73	0.754
95% CI (per 1000)	1.653 – 7.408	0.044 – 3.670

---



---

**Table 2**

Values of  $n_k$  for the given  $m_k$  used in the simulation study. Combinations  $m_k$  and  $n_k$  are such that in a given row  $\sum_{s_k} m_{s_k} n_{s_k} = 1000$ . The settings are grouped by the numbers of distinct pool sizes, and the number of pools that result is also reported and reflects the “amount of pooling” for the row-specified combination of  $m_k$  and  $n_k$ . Combinations used for the one-sample evaluation are indicated with an X.

---



---

Number of Pool Sizes	Pool Size $m_{ks_k}$				Number of Pools	One-Sample Evaluation	
	5	10	25	50			
1	200				200	X	
			100			100	
				20		20	X
2	100	50			150	X	
		50	20		70		
			20	10		30	
3	100	40	4		144		
		50	12	4	66		
		20	20	6	46		
		20	8	12	40	X	
4	20	40	12	4	76	X	
		10	20	14	8	52	
		10	10	22	6	48	
		10	10	10	12	42	

---



---

**Table 3**

*Confidence interval coverages and non-coverages in percent, non-coverage symmetry and  $100\times$  mean length, averaged over all simulations and parameter values for  $N = 1000$  (one-sample) and  $N_1 = N_2 = 1000$  (two-samples). Nominal coverage is 95%.*

---



---

<b>One-sample</b>					
<b>Interval</b>	<b>Coverage</b>	<b>Left- Non-cov.</b>	<b>Right- Non-cov.</b>	<b>Non-cov. Symmetry</b>	<b>Mean Length</b>
MIR	80.7	19.20	0.10	0.99	0.60
Wald	81.4	18.30	0.27	0.97	0.65
LRT	96.6	1.50	1.88	-0.11	0.76
Score	94.8	0.40	4.76	-0.84	0.80
SkewScore	96.6	1.36	2.05	-0.20	0.78

<b>Two-sample</b>					
<b>Interval</b>	<b>Coverage</b>	<b>Mesial- Non-cov.</b>	<b>Distal- Non-cov.</b>	<b>Non-cov. Symmetry</b>	<b>Mean Length</b>
MIR	93.2	0.97	5.80	-0.71	0.98
Wald	93.4	1.39	5.18	-0.58	1.06
SAW	97.3	1.49	1.26	0.08	1.29
PLRT	93.7	3.58	2.69	0.14	1.17
Score	96.3	2.45	1.26	0.32	1.54
SkewScore	96.4	2.17	1.46	0.19	1.51
BiasSkewScore	96.4	2.17	1.46	0.19	1.51

**Table 4**

*95% CIs for the difference of proportions (per 1000) of WNV-infected *Culex nigripalpus* mosquitoes trapped at heights of 6 m and 1.5 m, respectively. Point estimates for the difference are 2.708 per 1000 for the MIR and 3.008 per 1000 for the MLE.*

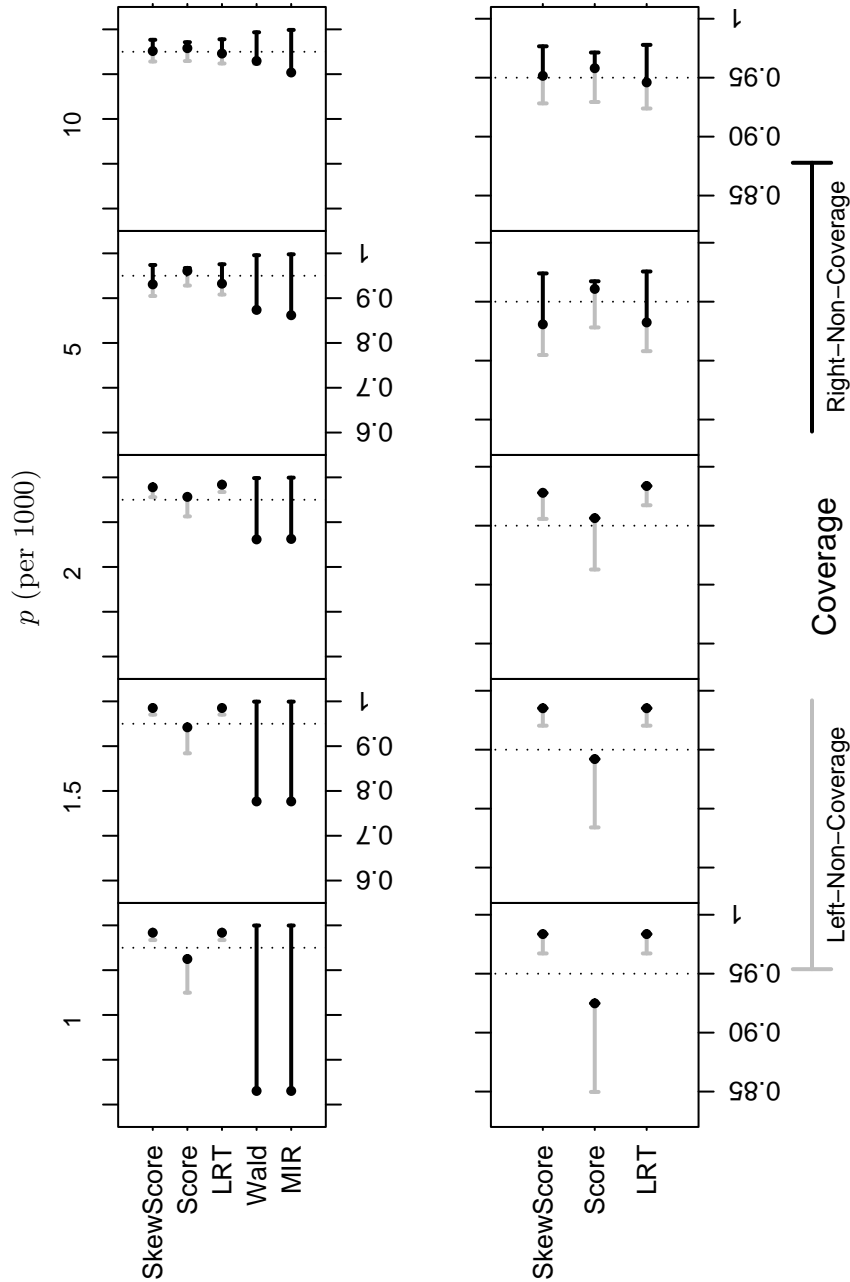
---



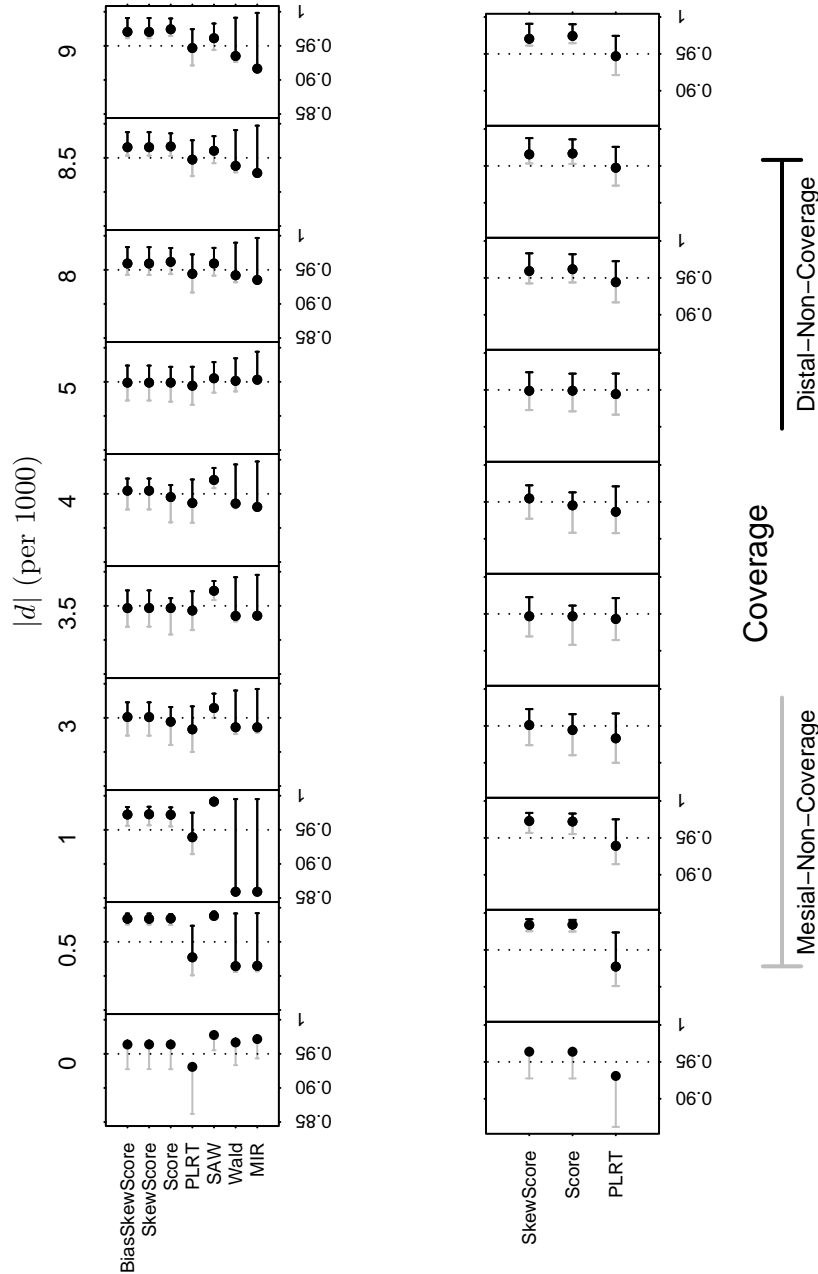
---

<b>Interval</b>	<b>95% CI</b>
MIR	(-0.250, +5.667)
Wald	(-0.165, +6.182)
Score	(-0.746, +6.935)
SkewScore	(-0.572, +6.824)
BiasSkewScore	(-0.570, +6.825)
PLRT	(-0.355, +6.729)
SAW	(-0.861, +6.852)

**Figure 1.** Coverage and non-coverage by  $p$  for each one-sample confidence interval, for  $N = 1000$ . Coverage is indicated by the dot, with nominal, 95% coverage marked by the dotted line. The amount of directional non-coverage is displayed using capped line segments (left as gray; right as black), initiating at the coverage dot. The top panels contain all intervals, while the bottom panels isolate the LRT, Score and SkewScore intervals for ease of visual interpretation of the scale. The column panels are for each values of  $p$  evaluated.



**Figure 2.** Coverage and non-coverage by  $d$  for each two-sample confidence interval, for  $N_1 = N_2 = 1000$ . Coverage is indicated by the dot, with nominal, 95% coverage marked by the dotted line. The amount of directional non-coverage is displayed using capped line segments (mesial as gray; distal as black), initiating at the coverage dot. The top panels contain all intervals, while the bottom panels isolate the PLRT, Score and SkewScore intervals for ease of visual interpretation of the scale. The column panels are for each values of  $|d| = |p_1 - p_2|$  evaluated.



**Figure 3.** Coverage by the total number of pools,  $\sum_i n_i$  for the one-sample case in the top panel, and  $\sum_i n_{1i} + \sum_j n_{2j}$  for the two-sample case in the bottom panel. Separate lines indicate grouping by values of  $p$  (one-sample) or  $|d|$  (two-sample), with darker values associated with smaller  $p$  or  $|d|$ .

