

Estimating a binomial proportion from several independent samples

C. G. QIAO

Centre for Social Research and Evaluation
Ministry of Social Development
P.O. Box 12 136
Wellington, New Zealand
email: chungui.qiao001@msd.govt.nz

G. R. WOOD

Department of Statistics
Macquarie University, Sydney
NSW 2109, Australia

C. D. LAI

Institute of Information Sciences and
Technology
Massey University
Private Bag 11 222
Palmerston North, New Zealand

Abstract This paper addresses the problem of estimating a binomial proportion from several independent samples in agricultural research, where the arithmetic average is widely used. The penalties of using a suboptimal estimator, the arithmetic estimator, relative to the preferred best estimator, the weighted average, are theoretically and empirically investigated, using numerical illustrations and simulation studies. Raw count data from a study of the proportion of inoculated transgenic hairy roots expressing resistance to cyst nematode in soybean (*Glycine max*) cultivars and a set of 10 examples of proportion estimation involving several independent samples are used for a practical evaluation of the findings. Results show that using the arithmetic average estimator can inflate variance and widen large sample confidence intervals of the estimates. The weighted average is recommended.

Keywords arithmetic average estimator; weighted average estimator; binomial proportion; confidence interval; penalty; variance ratio

THE PROBLEM: A SUBOPTIMAL ESTIMATOR IS BEING USED

Many areas of agricultural research use count data, modelled by a binomial distribution, to estimate the proportionate occurrence of a certain event. Examples include the proportion of plants affected by the incidence of a particular disease or insect pest, the survival percentage of insect pests after application of certain chemicals, the germination percentage of seeds, and the proportion of environments in which a new variety outperforms the local control. When count data from a series of such samples (generally of different sizes) have been recorded and pooled for estimating the underlying proportion, the question arises as to how this proportion should be estimated. We emphasise strongly that throughout this paper it is assumed that there is a single underlying true proportion p of interest for all the binomial samples under study.

Two methods are commonly considered for estimation of p : arithmetic averaging, which divides the sum of all these sample proportions by the total number of samples; and weighted averaging, which estimates the proportion via dividing the total count from all the samples by the total sample size. Suppose K independent samples have been taken and let X_i ($i = 1, 2, \dots, K$) represent the number of occurrences of a particular event in the i th sample, with n_i the corresponding sample size. The true underlying proportion of occurrence p for this event is estimated by $Y_i = X_i/n_i$. Evidently X_i follows a binomial distribution, or $X_i \sim B(n_i, p)$. The arithmetic average estimator of the population proportion p is defined as:

$$\bar{p}_A = \sum Y_i / K$$

The contrasting weighted average estimator of p uses the formula:

$$\bar{p}_W = \sum w_i Y_i = \left(\frac{n_1}{\sum n_i} \frac{X_1}{n_1} + \frac{n_2}{\sum n_i} \frac{X_2}{n_2} + \dots + \frac{n_K}{\sum n_i} \frac{X_K}{n_K} \right) = \frac{\sum X_i}{\sum n_i}$$

We can think of the collective outcome from all K samples as a binomial outcome from $\sum n_i$ trials; the experiment may be considered as having observed $\sum X_i$ “successes” from a binomial sample of size $n = \sum n_i$

and hence the best estimator of p is $\bar{p}_W = \sum X_i / \sum n_i$, not $\bar{p}_A = \left(\sum \frac{X_i}{n_i} \right) / K$.

Based on the concept and the definition of a “best” estimator described by Guenther (1973), \bar{p}_W is a complete and sufficient statistic, and the maximum likelihood estimator of p (Johnson et al. 1992).

In practice, \bar{p}_W has been advocated both in quality control (Pitt 1994) and in some statistical textbooks (Ott 1993; Ott & Mendenhall 1994), whereas \bar{p}_A has not been found to be recommended in any text books. Even among recent publications of the same journal, some researchers have used the weighted method (Chen et al. 2000; Choi et al. 2000; Paderson & Brink 2000). Others, however, still adopt the arithmetic average approach (Narayanan et al. 1999; Casler & van Santen 2000; Ismail et al. 2000). Although it is clear that the weighted average should be used, we have not found any report on the consequences of adopting a suboptimal estimator \bar{p}_A .

The present study examines the theoretical and practical impact of using the suboptimal estimator for estimating the binomial proportion. The findings are illustrated using data sets from both simulation experiments and agricultural research.

COMPARISON OF THE VARIANCE OF THE TWO ESTIMATORS

Variance comparison using algebra

Since $Y_i = X_i/n_i$, where $X_i \sim B(n_i, p)$, we have $E(Y_i = p)$ and

$$Var(Y_i) = \frac{1}{n_i^2} Var(X_i) = \frac{1}{n_i^2} n_i p(1-p) = \frac{1}{n_i} p(1-p)$$

Under certain conditions (e.g., $n_i \geq 25$) \bar{p}_A then follows a normal distribution with mean and variance given by:

$$E(\bar{p}_A) = E\left(\frac{\sum Y_i}{K}\right) = \frac{\sum E(Y_i)}{K} = \frac{Kp}{K} = p \quad \text{and}$$

$$Var(\bar{p}_A) = Var\left(\frac{\sum Y_i}{K}\right) = \frac{\sum Var(Y_i)}{K^2} = \frac{\sum p(1-p)/n_i}{K^2} = \frac{p(1-p)}{K^2} \sum \frac{1}{n_i} \tag{1}$$

Similarly, since $X_i \sim B(n_i, p)$, we have $\sum X_i \sim B(\sum n_i, p)$ whence $E(\sum X_i) = p \sum n_i$ and $Var(\sum X_i) = (\sum n_i) p(1-p)$. Then \bar{p}_W will follow an approximate normal distribution with mean and variance given by:

$$E(\bar{p}_W) = E\left(\frac{\sum X_i}{\sum n_i}\right) = \frac{E(\sum X_i)}{\sum n_i} = \frac{p \sum n_i}{\sum n_i} = p \quad \text{and}$$

$$Var(\bar{p}_W) = Var\left(\frac{\sum X_i}{\sum n_i}\right) = \frac{Var(\sum X_i)}{(\sum n_i)^2} = \frac{(\sum n_i) p(1-p)}{(\sum n_i)^2} = \frac{p(1-p)}{\sum n_i} \tag{2}$$

The relative merits of \bar{p}_W and \bar{p}_A can thus be measured by the ratio of these two variances:

$$R = \frac{Var(\bar{p}_W)}{Var(\bar{p}_A)} = \left\{ \frac{p(1-p)}{\sum n_i} \right\} / \left\{ \frac{p(1-p)}{K^2} \sum \frac{1}{n_i} \right\} = K^2 / \left\{ \left(\sum n_i \right) \left(\sum \frac{1}{n_i} \right) \right\} \tag{3}$$

Consider vectors \mathbf{x} and \mathbf{y} , where:

$$\mathbf{x} = \left(n_1^{\frac{1}{2}}, n_2^{\frac{1}{2}}, \dots, n_K^{\frac{1}{2}} \right) \text{ and } \mathbf{y} = \left(\frac{1}{n_1^{\frac{1}{2}}}, \frac{1}{n_2^{\frac{1}{2}}}, \frac{1}{n_K^{\frac{1}{2}}} \right)$$

Let $\langle \mathbf{x}, \mathbf{y} \rangle$ denote the inner product of \mathbf{x} and \mathbf{y} , and $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ denote the norms of \mathbf{x} and \mathbf{y} , respectively. The Cauchy-Schwarz inequality gives that the relationship between the modulus of the inner product and the norms is:

$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$, where

$$|\langle \mathbf{x}, \mathbf{y} \rangle| = \left(n_1 \frac{1}{n_1} + n_2 \frac{1}{n_2} + \dots + n_K \frac{1}{n_K} \right) = (1 + 1 + \dots + 1) = \sum 1 = K$$

$$\|\mathbf{x}\| = \left\{ \left(n_1^{\frac{1}{2}} \right)^2 + \left(n_2^{\frac{1}{2}} \right)^2 + \dots + \left(n_K^{\frac{1}{2}} \right)^2 \right\}^{\frac{1}{2}} = \left(\sum n_i \right)^{\frac{1}{2}} \text{ and}$$

$$\|\mathbf{y}\| = \left\{ \left(\frac{1}{n_1^{\frac{1}{2}}} \right)^2 + \left(\frac{1}{n_2^{\frac{1}{2}}} \right)^2 + \dots + \left(\frac{1}{n_K^{\frac{1}{2}}} \right)^2 \right\}^{\frac{1}{2}} = \left(\sum \frac{1}{n_i} \right)^{\frac{1}{2}}$$

Therefore, $K \leq \left(\sum n_i \right)^{\frac{1}{2}} \left(\sum \frac{1}{n_i} \right)^{\frac{1}{2}}$.

Since $K > 0$, this is equivalent to:

$$K^2 \leq \left(\sum n_i \right) \left(\sum \frac{1}{n_i} \right) \text{ or } K^2 / \left\{ \left(\sum n_i \right) \left(\sum \frac{1}{n_i} \right) \right\} \leq 1$$

Following Equation 3, $R = \frac{Var(\bar{p}_W)}{Var(\bar{p}_A)} = K^2 / \left\{ \left(\sum n_i \right) \left(\sum \frac{1}{n_i} \right) \right\} \leq 1$

Note that $R=1$ when there is equality in the Cauchy-Schwarz inequality. This is equivalent to having $\mathbf{x}=c\mathbf{y}$ for some real number c or $n_i^{\frac{1}{2}} = c \frac{1}{n_i^{\frac{1}{2}}}$ for all i , or $n_i=c$ for all i . Thus \bar{p}_A is as good as \bar{p}_W only when all samples have the same size. When sample sizes are different, \bar{p}_W is superior to \bar{p}_A , having smaller variance.

Variance comparison using simulation

A simulation experiment was run to compare the two averaging methods via examining their means and variances at varying population proportions. At each of nine equally-spaced population proportions p ranging from 0.1 to 0.9, a set of 10 random binomial samples (so $K = 10$) was simulated using Minitab software for three different types of sample size difference, as characterised by small, medium, and large coefficient of variation of these sample sizes CV_{SS} . The CV_{SS} was chosen as 0.10, 0.40, and 0.81, respectively to mimic realistic situations encountered in agricultural research (Narayanan et al. 1999; Verma et al. 1999; Chen 2000; Paderson & Brink 2000) for small, medium, and large sample size difference. For small sample size difference, the sizes of the 10 samples are 25, 26, 27, 28, 29, 30, 31, 32, 33, and 34 respectively (hence $n = 295$). For medium sample size difference, the sizes of the 10 samples are 25, 26, 27, 28, 29, 30, 57, 58, 59, and 60 respectively (hence $n = 399$). For large sample size difference, the sizes of the 10 samples are

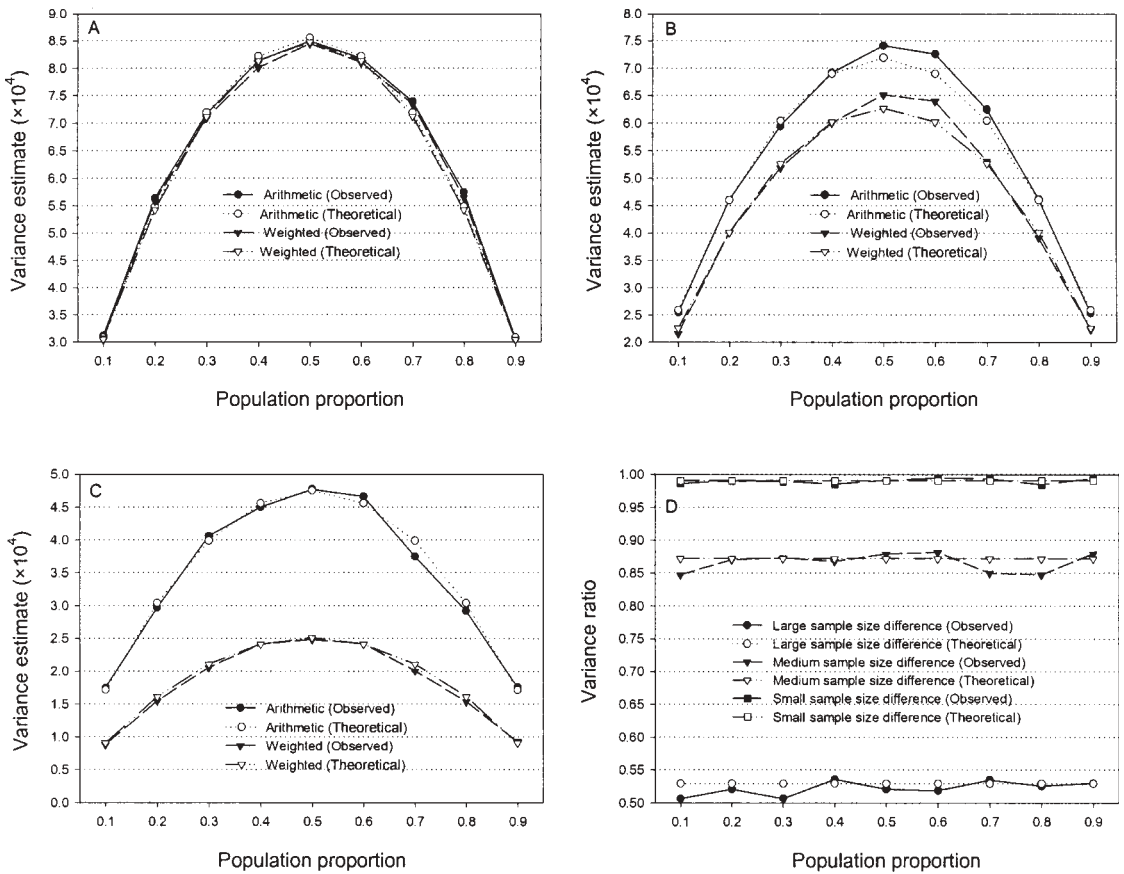


Fig. 1 Variance estimates for weighted and arithmetic averages and their ratios for: **A**, small; **B**, medium; and **C**, large sample size differences ($CV_{SS} = 0.10, 0.40, \text{ and } 0.81$) from the 2000 sets of 10 binomial samples from the simulation study. **D**, theoretical values for 10 samples with the three specified sample size differences.

25, 30, 35, 40, 45, 50, 185, 190, 195, and 200 respectively (hence $n = 995$). Estimates of \bar{p}_A and \bar{p}_W were obtained over the range of population proportions. This process was repeated 2000 times, resulting in 2000 pairs of proportion estimates (\bar{p}_A and \bar{p}_W), which then formed the data points for calculation of the observed means and variances of these two methods. Estimates of $Var(\bar{p}_A)$ and $Var(\bar{p}_W)$ as well as their ratio R are presented, together with the theoretical values based on Equations 1, 2, and 3, in Fig. 1.

The results showed that \bar{p}_A and \bar{p}_W , when averaged over the 2000 repetitions, were almost identical for each of the population proportion values for small, medium, and large sample size differences (results not listed). This agrees with the numerical illustrations in the section on Variance comparison

using algebra, that both \bar{p}_A and \bar{p}_W are unbiased estimates of the population proportion for binomial data. The variances of \bar{p}_A and \bar{p}_W , however, varied with the amount of sample size difference (Fig. 1).

When sample size difference is large, the weighted average approach consistently provided a more reliable estimate of the true underlying proportion than did the arithmetic average method, giving stable and low variance estimates over the range of population proportions (Fig. 1C). When sample size difference is decreased from large to medium, the gap between variance estimates of \bar{p}_A and \bar{p}_W closes up substantially, although the weighted average still consistently produces a lower variance estimate over the range of population proportions (Fig. 1B). For small sample size difference, there is almost no difference between variance estimates of \bar{p}_A and \bar{p}_W

over the range of population proportions (Fig. 1A). The theoretical predictions for both methods are also shown over the range of population proportions (Fig. 1A,B,C).

The variance ratios R between these methods were stable across all nine population proportions. As expected, the population proportion did not influence the magnitude of R , noted in Equation 3 (Fig. 1D). The penalty incurred by adopting the suboptimal estimator \bar{p}_A is shown for both large and medium sample size differences, with maximum penalty found for the large sample size difference (Fig. 1D). These findings illustrate the algebraic results of Equations 1, 2, and 3.

The simulation results suggest that the use of the suboptimal estimator \bar{p}_A is not penalised in variance inflation until the sample size difference is reasonably large, producing CV_{SS} significantly larger than 0.10, approaching 0.40.

CONFIDENCE INTERVAL

Confidence interval using algebraic analysis

A 95% large sample confidence interval for p , based on \bar{p}_W is:

$$\bar{p}_W \pm 1.96\sqrt{p(1-p) / \sum n_i}, \text{ with width of } D_W = 3.92\sqrt{p(1-p) / \sum n_i} \tag{4}$$

Using \bar{p}_A , in contrast, a 95% large sample confidence interval for p is:

$$\bar{p}_A \pm 1.96\sqrt{p(1-p) / [K^2(\sum \frac{1}{n_i})]}, \text{ with width of } D_A = 3.92\sqrt{p(1-p) / [K^2(\sum \frac{1}{n_i})]} \tag{5}$$

Since $\sum n_i \geq K^2 / (\sum \frac{1}{n_i})$, $D_A \leq D_W$.

Thus, the width of the confidence interval based on \bar{p}_A is always larger than that based on \bar{p}_W , as long as the sample sizes are unequal. The effect of using the suboptimal estimator \bar{p}_A is to widen the confidence interval for p .

Confidence interval using simulation analysis

The penalties for using the suboptimal estimator \bar{p}_A , relative to \bar{p}_W , were demonstrated using simulation in the section on Variance comparison using simulation. As shown in that section, \bar{p}_A and \bar{p}_W are very

close to p for each of the nine proportions ranging from 0.1 to 0.9, owing to large sample size. Hence, the difference between confidence intervals based on \bar{p}_A and \bar{p}_W is solely determined by the difference between $\sqrt{Var(\bar{p}_A)}$ and $\sqrt{Var(\bar{p}_W)}$. Results show that for both theoretical and observed estimates, the width of a confidence interval for p based on \bar{p}_W is always narrower than one based on \bar{p}_A (Table 1). The width of a 95% confidence interval for p based on \bar{p}_W is theoretically only 0.221 times of that for \bar{p}_A , and ranges from 0.205 to 0.235 for the observed simulation results (Table 1).

IMPLICATIONS AND RECOMMENDATIONS

The variance ratio $R = K^2 / \left\{ \left(\sum n_i \right) \left(\sum \frac{1}{n_i} \right) \right\}$

is an indicator of the relative merit of the weighted and arithmetic averages. It is influenced by three factors, the number of independent samples K , the total sample size $n = \sum n_i$, and the value of $\sum \frac{1}{n_i}$. Evidently $\sum \frac{1}{n_i}$ is determined by how much these n_i differ from one another and essentially it reflects the variance of N where the sample sizes $n_i, i = 1,2,\dots,K$, may be thought of as K outcomes from a random variable N .

For a given K , the ratio R will decrease if n_i are chosen such that n is kept approximately constant while allowing the variance of N to increase. This is because $\sum \frac{1}{n_i}$ increases as $Var(N)$ increases, provided the mean of N (given by $\frac{n}{K}$) remains approximately constant.

We also notice from Table 1 that the penalty is greatest when the sample size difference is large ($CV_{SS} = 0.81$), followed by the medium sample size difference ($CV_{SS} = 0.40$). There is almost no difference in confidence interval width when the sample size difference is small ($CV_{SS} = 0.10$). Population proportions have no influence on the confidence interval width. The above analysis demonstrates when \bar{p}_W is superior to \bar{p}_A in estimating p when a single underlying proportion exists.

In agricultural research or surveys, it is recommended that the weighted average approach be used when proportion data from a series of samples needs to be pooled or averaged, provided that it is reasonable that a single underlying p exists. The suboptimal estimator \bar{p}_A has to be used, however, when n_i and X_i are not known individually but only the

Table 1 Effect of using the suboptimal proportion estimator, expressed as the ratio of the width of a 95% confidence interval (CI) for \bar{p}_W to the width of a 95% CI for \bar{p}_A ($R_{CI} = \sqrt{Var(\bar{p}_W)Var(\bar{p}_A)}$) for three types of sample size difference.

p	Theoretical			Observed		
	95% CI for \bar{p}_W	95% CI for \bar{p}_A	R_{CI}	95% CI for \bar{p}_W	95% CI for \bar{p}_A	R_{CI}
Large sample size difference ($CV_{SS} = 0.81$)						
0.1	[0.0814, 0.1186]	[0.0744, 0.1256]	0.7274	[0.0817, 0.1185]	[0.0744, 0.1262]	0.7117
0.2	[0.1751, 0.2249]	[0.1658, 0.2342]	0.7274	[0.1755, 0.2242]	[0.1659, 0.2334]	0.7217
0.3	[0.2715, 0.3285]	[0.2609, 0.3391]	0.7274	[0.2719, 0.3282]	[0.2605, 0.3394]	0.7118
0.4	[0.3696, 0.4304]	[0.3582, 0.4418]	0.7274	[0.3692, 0.4301]	[0.3578, 0.4410]	0.7318
0.5	[0.4689, 0.5311]	[0.4573, 0.5427]	0.7274	[0.4690, 0.5308]	[0.4572, 0.5429]	0.7217
0.6	[0.5696, 0.6304]	[0.5582, 0.6418]	0.7274	[0.5693, 0.6302]	[0.5576, 0.6422]	0.7203
0.7	[0.6715, 0.7285]	[0.6609, 0.7391]	0.7274	[0.6728, 0.7283]	[0.6618, 0.7377]	0.7313
0.8	[0.7751, 0.8249]	[0.7658, 0.8342]	0.7274	[0.7759, 0.8245]	[0.7669, 0.8338]	0.7248
0.9	[0.8814, 0.9186]	[0.8744, 0.9256]	0.7274	[0.8813, 0.9190]	[0.8744, 0.9263]	0.7278
Medium sample size difference ($CV_{SS} = 0.40$)						
0.1	[0.0706, 0.1294]	[0.0685, 0.1315]	0.9338	[0.0713, 0.1288]	[0.0688, 0.1313]	0.9206
0.2	[0.1608, 0.2392]	[0.1580, 0.2420]	0.9338	[0.1602, 0.2387]	[0.1570, 0.2411]	0.9329
0.3	[0.2550, 0.3450]	[0.2518, 0.3482]	0.9338	[0.2547, 0.3440]	[0.2516, 0.3471]	0.9345
0.4	[0.3519, 0.4481]	[0.3485, 0.4515]	0.9338	[0.3518, 0.4478]	[0.3483, 0.4514]	0.9315
0.5	[0.4509, 0.5491]	[0.4475, 0.5525]	0.9338	[0.4498, 0.5498]	[0.4463, 0.5530]	0.9376
0.6	[0.5519, 0.6481]	[0.5485, 0.6515]	0.9338	[0.5496, 0.6487]	[0.5461, 0.6517]	0.9389
0.7	[0.6550, 0.7450]	[0.6518, 0.7482]	0.9338	[0.6540, 0.7443]	[0.6503, 0.7482]	0.9219
0.8	[0.7608, 0.8392]	[0.7580, 0.8420]	0.9338	[0.7611, 0.8386]	[0.7575, 0.8417]	0.9206
0.9	[0.8706, 0.9294]	[0.8685, 0.9315]	0.9338	[0.8710, 0.9294]	[0.8689, 0.9313]	0.9377
Small sample size difference ($CV_{SS} = 0.10$)						
0.1	[0.0658, 0.1342]	[0.0656, 0.1344]	0.9952	[0.0661, 0.1347]	[0.0658, 0.1350]	0.9932
0.2	[0.1544, 0.2456]	[0.1541, 0.2459]	0.9952	[0.1532, 0.2458]	[0.1528, 0.2459]	0.9951
0.3	[0.2477, 0.3523]	[0.2475, 0.3525]	0.9952	[0.2472, 0.3515]	[0.2468, 0.3518]	0.9945
0.4	[0.3441, 0.4559]	[0.3438, 0.4562]	0.9952	[0.3438, 0.4547]	[0.3434, 0.4552]	0.9924
0.5	[0.4429, 0.5571]	[0.4427, 0.5573]	0.9952	[0.4431, 0.5570]	[0.4430, 0.5572]	0.9971
0.6	[0.5441, 0.6559]	[0.5438, 0.6562]	0.9952	[0.5421, 0.6537]	[0.5418, 0.6539]	0.9953
0.7	[0.6477, 0.7523]	[0.6475, 0.7525]	0.9952	[0.6467, 0.7529]	[0.6465, 0.7530]	0.9966
0.8	[0.7544, 0.8456]	[0.7541, 0.8459]	0.9952	[0.7535, 0.8466]	[0.7530, 0.8469]	0.9918
0.9	[0.8658, 0.9342]	[0.8656, 0.9344]	0.9952	[0.8661, 0.9346]	[0.8660, 0.9347]	0.9969

$Y_i = X_i/n_i$ are available, for example, when the data collector only presents the percentages for different samples. In this instance we have no choice but to use the arithmetic averaging approach as a substitute for weighted averaging.

APPLICATIONS IN AGRICULTURAL RESEARCH

We used data from a study of inoculated transgenic hairy roots expressing β -glucuronidase (GUS) activity in soybean (*Glycine max* (L.) Merrill) in 1999, where a series of ratio estimates needed to be pooled over four cultivars (Narayanan et al. 1999). The GUS histochemical staining activity is measured as being regulated by three promoters: (1) the cauliflower mosaic virus (CaMV 35S); (2) the chalcone synthase-8 (CHS); and (3) the phenylalanine ammonia (PAL). The results are listed in Table 2, where the number of hairy roots expressing GUS activity and the total number of hairy roots sampled in each cultivar are shown. The aim was to estimate the average or pooled proportion of hairy roots expressing GUS activity of all four cultivars in soybean. A pooled estimate of the proportion using data from all four samples was considered appropriate, since the proportion estimates of all five locations are similar (a statistical test for difference between the four sample proportions gives $\chi^2 = 4.719$, d.f. = 3, $P = 0.194$). Thus, a single underlying proportion is considered a reasonable assumption for this data set.

The first part of Table 2 details the data collected in each cultivar (considered as a sample in this context). The second part of the table gives estimates (using the two estimators) of the population binomial proportion, the variances of the two estimators and their ratio $R = Var(\bar{p}_W)/Var(\bar{p}_A)$. Using data from all four cultivars, the average proportion of hairy roots expressing GUS is \bar{p}_A if the arithmetic average method is adopted, in comparison to a larger estimate of \bar{p}_W with the weighted average method. The difference in the overall proportion estimates is c. 2%, which is relatively large in comparison to the published data found from other available sources (Chen et al. 2000; Choi et al. 2000; Paderson & Brink 2000). This is because for most of the published data the differences between sample sizes n_i and between the individual sample proportion estimates are small, so that differences between \bar{p}_A and \bar{p}_W are also small. Data of this type can be found in Levine et al. (2000, p. 131) and Kempthorne (1957, pp. 152–154), where the \bar{p}_A and \bar{p}_W estimates were 73.3% and 72.9% for the former, and 52.0% and 51.5% for the latter. Examples of this type may explain why the issue of choice between the arithmetic and weighted average methods has not drawn sufficient attention from applied agricultural scientists, and why many keep using the arithmetic average.

In our example, the estimated variances of \bar{p}_A and \bar{p}_W are 0.00104 and 0.00093, respectively, resulting in $R = Var(\bar{p}_W)/Var(\bar{p}_A) = 0.8942$. Since the magnitudes of these variances are small, the width

Table 2 Frequency and proportion of inoculated transgenic hairy roots expressing β -glucuronidase (GUS) histochemical staining activity regulated by promoter CaMV 35S in four soybean (*Glycine max*) cultivars (data source: table 1, Crop Science 1999, 39: 1680–1686).

Cultivar	No. of hairy roots n_i	No. of hairy roots expressing GUS activity x_i	Proportion expressing GUS activity
Agassiz	44	21	0.4773
Parker	58	29	0.5000
Bell	59	28	0.4746
Faribault	105	65	0.6190
Coefficient of variation for sample sizes $CV (n_i)$			0.3995
\bar{p}_W			0.5376
\bar{p}_A			0.5177
Variance estimate for the arithmetic average $Var(\bar{p}_W)$			0.00093
Variance estimate for the weighted average $Var(\bar{p}_A)$			0.00104
Variance ratio of the two estimators $\hat{R} = Var(\bar{p}_W)/Var(\bar{p}_A)$			0.8942
95% confidence interval for p based on \bar{p}_W			0.4777~0.5975
95% confidence interval for p based on \bar{p}_A			0.4546~0.5808
Ratio of the widths of confidence intervals between \bar{p}_W and \bar{p}_A			0.9494

of a 95% confidence interval based on \bar{p}_W is only marginally smaller than that based on \bar{p}_A , with the ratio of the widths being 0.9494. As expected, the penalty for using the suboptimal estimator \bar{p}_A is relatively small in this example in the width of the confidence interval. The main reason for this result is the relatively large overall sample size $n = 266$, which causes the variances for both estimators to be very small. Nevertheless, there is gain in using \bar{p}_W because $\bar{p}_A - \bar{p}_W \approx 2\%$.

Additional examples in agricultural research

Eight real examples in agricultural research (four from Paderson & Brink 2000; one from Verma et al. 1999; one from Chen 2000; two from Narayanan et al. 1999), where raw count data are available (for details see Appendix 1), are used to illustrate the findings of the present study. For all, we assume that it is reasonable that a single underlying exists. Table 3 shows penalties for using the suboptimal estimator \bar{p}_A , relative to \bar{p}_W , in variance inflation of the estimates and in increased width of a 95% confidence interval.

It can be seen from Table 3 that the \bar{p}_A and \bar{p}_W estimates are generally close in these instances. The difference between variances of the \bar{p}_A and \bar{p}_W estimates, however, can be substantial, with the variance of \bar{p}_A always larger than that of \bar{p}_W . This is illustrated by the variance ratio of the two estimators, ranging from 0.8102 to 0.9954 (Table 3). Similarly, the width of the 95% large sample confidence interval based

on \bar{p}_W is always smaller than based on the suboptimal estimator \bar{p}_A , with the ratio of the two being smaller than 1. Thus, using \bar{p}_A for estimation of the proportion produces more uncertainty and hence is not preferred. In contrast, \bar{p}_W should be used in that its variance and confidence interval width are both small.

It should be noted that the advantage of using the weighted average approach can be fully demonstrated only when CV_{SS} is relatively large. For the examples shown in Table 3, when $CV_{SS} < 0.20$, the penalty of using the suboptimal estimator is so small that the variance ratio estimate almost reaches 1 for examples 1, 2, and 3. In comparison, the penalty is larger for other instances, where CV_{SS} is larger than 0.20, but generally with the ratio of 0.93 or below. Using the measure in this study, most of these examples belong to “medium sample size difference”, except for example 8 which demonstrates large sample size difference. This may be the reason why differentiating between the two estimators, when a single underlying p exists, has not drawn sufficient attention from agricultural research scientists.

DISCUSSION AND CONCLUSIONS

When count data from different independent samples (generally of different sizes) have been recorded and pooled for estimating the underlying proportion in agricultural research, the commonly arising

Table 3 Some real examples in agricultural research, illustrating penalties incurred through using the suboptimal estimator, where the variance ratio $\hat{R} = Var(\bar{p}_W)/Var(\bar{p}_A)$.

Example	CV_{SS}	Estimator	p estimate	Variance $\times 10^4$	\hat{R}	95% confidence interval
1	0.076	Arithmetic average	0.9188	0.0805	0.9954	[0.9133, 0.9244]
		Weighted average	0.9181	0.0801		[0.9126, 0.9237]
2	0.150	Arithmetic average	0.8516	0.0764	0.9828	[0.8462, 0.8570]
		Weighted average	0.8528	0.0751		[0.8474, 0.8581]
3	0.196	Arithmetic average	0.1012	0.2533	0.9620	[0.0913, 0.1110]
		Weighted average	0.1012	0.2437		[0.0916, 0.1109]
4	0.295	Arithmetic average	0.4606	7.5079	0.9298	[0.4069, 0.5143]
		Weighted average	0.4463	6.9808		[0.3945, 0.4981]
5	0.354	Arithmetic average	0.7728	5.0168	0.8714	[0.7289, 0.8168]
		Weighted average	0.7941	4.3715		[0.7531, 0.8351]
6	0.440	Arithmetic average	0.6541	0.4688	0.8580	[0.6407, 0.6675]
		Weighted average	0.6785	0.4022		[0.6660, 0.6909]
7	0.443	Arithmetic average	0.7502	0.4903	0.8722	[0.7365, 0.7639]
		Weighted average	0.7762	0.4276		[0.7634, 0.7890]
8	0.572	Arithmetic average	0.7667	4.1866	0.8102	[0.7266, 0.8068]
		Weighted average	0.7546	3.3918		[0.7185, 0.7907]

question is, “How should this proportion be estimated?” Two types of situation are possible: there is a single underlying true proportion p of interest for all the samples, or the population proportions differ for these samples.

We have considered the example where p itself has a distribution in the companion paper (Wood et al. 2005). This situation arises commonly in practice and requires us to estimate the underlying mean binomial proportion, with either of the two commonly used estimators, the arithmetic average and the weighted average of the observed sample proportions. The paper evaluates the variances of the two estimators and uses the difference to decide which estimator is preferred in a given context. The relative merits of the estimators depend on the distribution of the proportions and the sizes of samples under study. The findings are illustrated using both simulation and count data of feeding incidents involving fruit for nine different kaka, an endangered New Zealand parrot species.

Results of the present study show that when a single underlying proportion exists, the weighted average approach should be adopted for estimating the binomial proportion. Using the suboptimal arithmetic average estimator, which is still a popular practice in agricultural research, could lead to inflated variance of the estimator and so to a wider large sample confidence interval. The arithmetic average \bar{p}_A has to be used, however, when n_i and x_i are not known but only the y_i are available, for example, when the data collector only presents the percentages for different samples.

ACKNOWLEDGMENTS

We are indebted to Dr Nihal De Silva, the other anonymous referee and the editor for their constructive and valuable comments, which helped to significantly improve the manuscript.

REFERENCES

- Casler MD, van Santen E 2000. Patterns of variation in a collection of meadow fescue accessions. *Crop Science* 40: 248–255.
- Chen S, Lin XH, Xu CG, Zhang Q 2000. Improvement of bacterial blight resistance of ‘Minghui 63’, an elite restorer line of hybrid rice, by molecular marker-assisted selection. *Crop Science* 40: 239–244.
- Choi HW, Lemaux PG, Cho MJ 2000. Increased chromosomal variation in transgenic versus nontransgenic barley (*Hordeum vulgare* L.) plants. *Crop Science* 40: 524–533.
- Guenther WC 1973. Concepts of statistical inference. 2nd ed. Tokyo, McGraw-Hill Kogakusha Ltd. Pp. 127–131.
- Ismail AM, Hall AE, Ehlers JD 2000. Delayed-leaf-senescence and heat-tolerance trait mainly are independently expressed in cowpea. *Crop Science* 40: 1049–1055.
- Johnson NL, Kotz S, Kemp AW 1992. Univariate discrete distributions. 2nd ed. New York, Wiley Interscience. Pp. 125–127.
- Kemphorne O 1957. An introduction to genetic statistics. New York, John Wiley and Sons Inc. Pp. 95–128.
- Levine DM, Krehbiel TC, Berenson ML 2000. Business statistics: a first course. 2nd ed. Upper Saddle River, New Jersey, Prentice Hall. 645p.
- Lipschutz S 1968. Schaum’s outline of theory and problems of linear algebra. Schaum’s Outline Series, New York, McGraw Book Company. Pp. 279–306.
- Narayanan RA, Atz R, Nenny R, Young ND, Somers DA 1999. Expression of soybean cyst nematode resistance in transgenic hairy roots of soybean. *Crop Science* 39: 1680–1686.
- Ott RL 1993. An introduction to statistical methods and data analysis. 4th ed. Belmont, California, Wadsworth Inc. Pp. 380–384.
- Ott RL, Mendenhall W 1994. Understanding statistics. 6th ed. Belmont, California, Wadsworth Inc. Pp. 406–410.
- Paderson GA, Brink GE 2000. Seed production of white clover cultivars and naturalised populations when grown in a pasture. *Crop Science* 40: 1109–1114.
- Pitt H 1994. SPC for the rest of us—a personal path to statistical process control. New York, Addison-Wesley Publishing Company Inc. Pp. 259–276.
- Verma V, Bains NS, Mangat GS, Nanda GS, Gosal SS, Singh K 1999. Maize genotypes show striking differences for induction and regeneration of haploid wheat embryos in the wheatxmaize system. *Crop Science* 39: 1722–1727.
- Wood GR, Lai CD, Qiao CG 2005. Estimation of a proportion using several independent samples of binomial mixtures. *Australian and New Zealand Journal of Statistics* (in press).

