Estimation for Binomial Proportions from Pooled Samples Using an Objective Prior

Lizanne Raubenheimer[1,3] and Abrie van der Merwe[2]
[1]Department of Statistics, Rhodes University, Grahamstown, South Africa,
[2]Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa
[3]Corresponding author: Lizanne Raubenheimer, e-mail: L.Raubenheimer@ru.ac.za

Abstract

Group testing has been used in many fields of study, as individual testing can be too time consuming and pooled testing is more cost-effective. Group testing is where units are pooled together and tested as a group rather than individually. In this paper we will look into confidence intervals for linear functions of binomial proportions from pooled samples. We will investigate the performance of Bayesian confidence (credibility) intervals for a single proportion as well as the difference of two binomial proportions estimated from pooled samples. An objective (non-informative) prior, the Jeffreys prior, will be used. Results from the Bayesian method will be compared to results from some known classical methods. These intervals will be compared with each other in terms of coverage, left non-coverage, right non-coverage, symmetry and interval length.

Keywords: Bayesian inference, coverage, credibility interval, Jeffreys prior

1. Introduction

In this paper we will look into confidence intervals for linear functions of binomial rates from pooled samples. We will investigate the performance of Bayesian credibility intervals for a single proportion as well as the difference of two binomial proportions estimated from pooled samples. Where the Jeffreys prior will be used for the Bayesian method. Highest Posterior Density (HPD) intervals will be considered, where the HPD interval will have a shorter interval length than the equal tail interval. Group testing has been used in many fields of study, as individual testing can be too time consuming and pooled testing is more cost-effective. Group testing is where units are pooled together and tested as a group rather than individually. Biggerstaff (2008) used asymptotic methods to derive Wald, profile score and profile likelihood ratio intervals. Biggerstaff (2008) also adapted the Wilson score-based interval of Newcombe (1998). Tu et al. (1995) investigated the maximum likelihood estimator for equal pool sizes. Hepworth (1996) considered the sequential testing of groups of different sizes, by constructing exact confidence intervals for problems involving unequal sized groups. Hepworth (2005) also considered asymptotic interval estimation methods where groups are of different sizes. Hepworth (2005) investigated four methods, two based on the distribution of the maximum likelihood estimate (MLE), one on the score statistic and one on the likelihood ratio. Hepworth (2005) recommended the method based on the score statistic with a correction for skewness. Biggerstaff (2008) recommended the skewness-corrected profile score interval. In Section 2 the Bayesian method will be discussed, two simulation studies will be considered in Section 3. An application will be discussed in Section 4 and the conclusion will be given in Section 5. For the simulation studies and the application, the results from the Bayesian method will be compared to the results obtained by Biggerstaff (2008).

2. Prior Distribution for Binomial Proportions from Pooled Samples

Assume that the proportion of successes in a given population is p. We will refer to an infected individual as a success in a binomial trial. Using the notation from Biggerstaff (2008), let N individuals be sampled independently from the population, and then be grouped into pools. The size of a pool will be indicated by m_i, for $i = 1,2,\ldots,M$, where M is the number of distinct pool sizes, 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. Assume that $X_1, X_2, \ldots, X_M$ are independent binomial random variables with $X_i \sim Bin\left(n_i, 1-(1-p)^{m_i}\right)$.

The likelihood function is given by

$$L(p\,|x_1,x_2,\ldots,x_M) \quad \propto \quad \prod_{i=1}^{M}\left\{[1-(1-p)^{m_i}]^{x_i}\,[(1-p)^{m_i}]^{n_i-x_i}\right\}.$$

The Fisher information was derived by Walter et al. (1980), and is given by

$$F(p) \quad = \quad \sum_{i=1}^{M}\left\{\frac{m_i^2 n_i\,(1-p)^{m_i-2}}{[1-(1-p)^{m_i}]}\right\}.$$

The Jeffreys prior, from Jeffreys (1939), is proportional to the square root of the determinant of the Fisher information and is given by

$$\begin{aligned}\pi(p) \quad &\propto \quad |F(p)|^{\frac{1}{2}} \\ \therefore \pi(p) \quad &\propto \quad \left(\sum_{i=1}^{M}\left\{\frac{m_i^2 n_i\,(1-p)^{m_i-2}}{[1-(1-p)^{m_i}]}\right\}\right)^{\frac{1}{2}}. \qquad (1)\end{aligned}$$

The posterior distribution is then given by

$$\begin{aligned}\pi(p\,|data) \quad &\propto \quad \pi(p)\times L(p\,|data) \\ &\propto \quad \left(\sum_{i=1}^{M}\left\{\frac{m_i^2 n_i\,(1-p)^{m_i-2}}{[1-(1-p)^{m_i}]}\right\}\right)^{\frac{1}{2}} \\ &\quad \times\prod_{i=1}^{M}\left\{[1-(1-p)^{m_i}]^{x_i}\,[(1-p)^{m_i}]^{n_i-x_i}\right\} \qquad \text{for } 0\le p\le 1. \qquad (2)\end{aligned}$$

If $M = 1$, $m_1 = m$, $n_1 = n$ and $x_1 = x$, it follows from Equation 1 that

$$\begin{aligned}\pi(p) \quad &\propto \quad \left\{\frac{m^2 n\,(1-p)^{m-2}}{[1-(1-p)^{m}]}\right\}^{\frac{1}{2}} \\ &\propto \quad [(1-p)^{m}]^{\frac{1}{2}-\frac{1}{m}}\,[1-(1-p)^{m}]^{-\frac{1}{2}}. \qquad (3)\end{aligned}$$

The posterior distribution when using the Jeffreys prior is given by

$$\pi(p\,|data) \quad \propto \quad [(1-p)^{m}]^{n-x+\frac{1}{2}-\frac{1}{m}}\,[1-(1-p)^{m}]^{x-\frac{1}{2}} \qquad \text{for } 0\le p\le 1. \quad (4)$$

Theorem 1. When $\theta = (1-p)^m$, the posterior distribution of θ will be $Beta\left(x+\frac{1}{2}, n-x+\frac{1}{2}\right)$, i.e.

$$\pi(\theta\,|data) \quad \propto \quad (1-\theta)^{n-x-\frac{1}{2}}\,\theta^{x-\frac{1}{2}}. \qquad (5)$$

Proof. From Equation 4, the posterior distribution is given as

$$\pi(p\,|data) \propto [(1-p)^m]^{n-x+\frac{1}{2}-\frac{1}{m}}\,[1-(1-p)^m]^{x-\frac{1}{2}} \qquad \text{for } 0 \le p \le 1.$$

Let $\theta = (1-p)^m$, then $p = 1-\theta^{\frac{1}{m}}$, and

$$\left|\frac{dp}{d\theta}\right| = \frac{1}{m}\theta^{\frac{1}{m}-1}$$

$$\begin{aligned}
\pi(\theta\,|data) &\propto \left[\left(1-\left(1-\theta^{\frac{1}{m}}\right)\right)^m\right]^{n-x+\frac{1}{2}-\frac{1}{m}}\left[1-\left(1-\left(1-\theta^{\frac{1}{m}}\right)\right)^m\right]^{x-\frac{1}{2}}\frac{1}{m}\theta^{\frac{1}{m}-1} \\
&= \left[\left(\theta^{\frac{1}{m}}\right)^m\right]^{n-x+\frac{1}{2}-\frac{1}{m}}\left[1-\left(\theta^{\frac{1}{m}}\right)^m\right]^{x-\frac{1}{2}}\frac{1}{m}\theta^{\frac{1}{m}-1} \\
&= \theta^{n-x+\frac{1}{2}-\frac{1}{m}}(1-\theta)^{x-\frac{1}{2}}\frac{1}{m}\theta^{\frac{1}{m}-1} \\
&= \frac{1}{m}\theta^{n-x+\frac{1}{2}-\frac{1}{m}+\frac{1}{m}-1}(1-\theta)^{x-\frac{1}{2}} \\
\therefore \pi(\theta\,|data) &\propto (1-\theta)^{x-\frac{1}{2}}\,\theta^{n-x-\frac{1}{2}}. \qquad (6)
\end{aligned}$$

□

Transforming Equation 6, the posterior distribution for $p = 1-\theta^{\frac{1}{m}}$ can be determined, where $\left|\frac{d\theta}{dp}\right| = m(1-p)^{m-1}$.

$$\therefore \pi(p\,|data) = \frac{m}{B\left(x+\frac{1}{2}, n-x+\frac{1}{2}\right)}\,[(1-p)^m]^{n-x+\frac{1}{2}-\frac{1}{m}}\,[1-(1-p)^m]^{x-\frac{1}{2}}. \quad (7)$$

3. Simulation Studies

3.1 Simulation Study I - Single Proportion

In this section we will consider a simulation study for a single proportion from pooled samples. A single proportion will be considered where $M = 1$, $M = 2$, $M = 3$ and $M = 4$. We will look at coverage, left noncoverage, right noncoverage, symmetry and interval length. Biggerstaff (2008) defines noncoverage symmetry as the difference in proportional noncoverage, i.e.

$$\text{Symmetry} = \frac{P[\text{Left noncoverage}] - P[\text{Right noncoverage}]}{P[\text{Left noncoverage}] + P[\text{Right noncoverage}]}$$

with a negative value indicating mostly right noncoverage and a positive value indicating mostly left noncoverage. A value of zero for symmetry indicates symmetric noncoverage.

We considered the different pool size combinations which was used by Biggerstaff (2008). Table 1 gives the results from Biggerstaff (2008) and the results obtained by us using the Bayesian method. The first five intervals in Table 1 are from Biggerstaff (2008). The results in Table 1 are averages taken over the different values for p and the different pool size combinations.

Table 1: Overall averages of coverage rates, noncoverages, symmetry and average lengths. Nominal coverage is 95%.

Interval	**Coverage**	**Left non-coverage**	**Right non-coverage**	**Symmetry**	**Length** $\times 1\,000$
MIR	0.8070	0.0010	0.1920	-0.99	6.0000
Wald	0.8140	0.0027	0.1830	-0.97	6.5000
Likelihood ratio (LRT)	0.9660	0.0188	0.0150	0.11	7.6000
Profile score	0.9480	0.0476	0.0040	0.84	8.0000
Skewness corrected score	0.9660	0.0205	0.0136	0.20	7.8000
Bayesian	0.9584	0.0158	0.0258	0.34	7.0659

From Table 1 it is clear that the coverage rates obtained by the MIR and Wald intervals are far below the nominal level of 0.95, this was also stated by Biggerstaff (2008). The other four intervals give coverages close to the nominal level, with the profile score and the Bayesian intervals performing slightly better. The results obtained from the Bayesian method by us compare well with the results obtained from the other researcher. In terms of coverage we can conclude that the Bayesian interval and the Profile score interval perform the best. When having a closer look at the interval length, the Bayesian interval is shorter than the Profile score interval.

3.2 Simulation Study II - Two Proportions

In this section we will consider a simulation study for proportions from pooled samples for the difference between two proportions. Biggerstaff (2008) considered the different combinations, and listed the average of the coverage, left noncoverage, right noncoverage, noncoverage symmetry and mean length over all the different parameter values. For the Bayesian method we only considered the two cases, $M_1 = M_2 = 1$ and $M_1 = M_2 = 2$, and averaged over these values. Left noncoverage is interpretable as distal noncoverage probability and right noncoverage is interpretable as mesial noncoverage. It is desirable that these should be equal.

Table 2 gives the results from Biggerstaff (2008) and the results obtained by us using the Bayesian method. The first seven intervals in Table 2 are from Biggerstaff (2008).

Table 2: Overall averages of coverage rates, noncoverages, symmetry and average lengths for $p_1 - p_2$. Nominal coverage is 95%.

Interval	**Coverage**	**Left non-coverage**	**Right non-coverage**	**Symmetry**	**Length** $\times 1\,000$
MIR	0.9320	0.0580	0.0097	0.7100	9.8000
Wald	0.9340	0.0518	0.0139	0.5800	10.6000
Square-and-add Walter	0.9730	0.0126	0.0149	-0.0800	12.9000
Likelihood ratio (LRT)	0.9370	0.0269	0.0358	-0.1400	11.7000
Profile score	0.9630	0.0126	0.0245	-0.3200	15.4000
Skewness corrected score	0.9640	0.0146	0.0217	-0.1900	15.1000
Bias Skewness corrected score	0.9640	0.0146	0.0217	-0.1900	15.1000
Bayesian	0.9663	0.0247	0.0090	0.4653	12.2760

The coverage rate for the Bayesian method is above the nominal level of 0.95, this is the case for all the other intervals except for the MIR, Wald and likelihood ratio in-

tervals. The Profile score, Skewness corrected score, Bias Skewness corrected score and Bayesian intervals gave coverage closest to the nominal level. When looking at the intervals with coverage rates above the nominal level, it can be seen that the Bayesian interval yields the shortest interval.

4. Example - West Nile Virus

Biggerstaff (2008) considered an example where a comparison is made between West Nile virus (WNV) infection prevalences in field collected *Culex nigripalpus* mosquitoes trapped at different heights. Biggerstaff (2008) derived asymptotic confidence intervals for the difference between two proportions estimated from pooled samples, where the sizes of the pools are not equal. Biggerstaff (2008) considered seven confidence intervals: an interval based on the minimum infection rate (MIR), the Wald interval, the profile score interval, the skewness corrected score interval, the bias- and skewness-corrected score interval, square-and-add Walter (SAW) interval and the profile likelihood interval. Table 3 summarises the data from Biggerstaff (2008).

Table 3: Summary of *Culex nigripalpus* mosquitoes trapped at different heights of $6m$ and $1.5m$.

	Sample 1 **height** $= 6m$	**Sample 2** **height** $= 1.5m$
Total	2 021	1 324
Number of pools	53	31
Average pool size	38.1321	42.7097
Minimum pool size	1	5
Maximum pool size	50	100
Number of positive pools	7	1

We used the Jeffreys prior to construct a 95% Bayesian (HPD) interval for each sample. The results are shown in Table 4.

Table 4: 95% intervals and interval lengths for the proportions (per 1 000) of the two samples.

	95% HPD Interval	**Length**	**95% Confidence Interval** (Biggerstaff, 2008)	**Length**
Sample 1 height $= 6m$	$(1.444, 6.959)$	5.515	$(1.653, 7.408)$	5.755
Sample 2 height $= 1.5m$	$(0.019, 3.002)$	2.983	$(0.044, 3.670)$	3.626

From Table 4 the Bayesian intervals are shorter than those obtained by Biggerstaff (2008).

For the mosquito data we draw random samples of 100 000 from each of the two posteriors mentioned above and calculate the difference between the two proportions. We used the Jeffreys prior to construct a 95% Bayesian (HPD) interval for the difference between the two proportions. The results are shown in Table 5, the results for the first seven intervals are from Biggerstaff (2008).

Table 5: 95% intervals and interval lengths for the difference between the two proportions (per 1 000).

	95% Interval	**Length**
MIR	$(-0.250, 5.667)$	5.920
Wald	$(-0.165, 6.182)$	6.347
Profile score	$(-0.746, 6.935)$	7.681
Skewness corrected score	$(-0.572, 6.824)$	7.396
Bias- and skewness-corrected score	$(-0.570, 6.825)$	7.395
Profile likelihood	$(-0.355, 6.729)$	7.084
Square-and-add Walter	$(-0.861, 6.852)$	7.713
Bayesian	$(-0.403, 6.528)$	6.931

The Bayesian interval compares relatively well with the others, all the intervals include 0. The MIR, Wald and Bayesian intervals give shorter interval lengths than the other intervals. The MIR and Wald intervals are known for giving poor coverage. So if we compare the Bayesian interval to the other five intervals, the Bayesian interval is the shortest one.

5. Conclusion

In this paper we compared the proposed Bayesian method to results obtained by Biggerstaff (2008). The Jeffreys prior was used for the Bayesian method. Simulation studies were considered as well as an example. The Bayesian method compared well with the other results, and gave much better results than the Wald and minimum infection rate intervals. The Wald and the minimum infection rate intervals performed the poorest. When looking at a single proportion, the Bayesian interval and the Profile score interval performed the best when looking at coverage. When having a closer look at the interval length, the Bayesian interval is shorter than the Profile score interval in the case of a single proportion.

References

Biggerstaff, B. J. (2008) "Confidence intervals for the difference of two proportions estimated from pooled samples," *Journal of Agricultural, Biological, and Environmental Statistics*, 13(4), 478 – 496.

Hepworth, G. (1996) "Exact confidence intervals for proportions estimated by group testing," *Biometrics*, 52(3), 1134 – 1146.

Hepworth, G. (2005) "Confidence intervals for proportions estimated by group testing with groups of unequal size," *Journal of Agricultural, Biological, and Environmental Statistics*, 10(4), 478 – 497.

Jeffreys, H. (1939) *Theory of Probability*, Oxford University Press, Oxford, 1st edition.

Newcombe, R. G. (1998) "Interval estimation for the difference between independent proportions: Comparison of eleven methods," *Statistics in Medicine*, 17(8), 873 – 890.

Tu, X. M., Litvak, E., & Pagano, M. (1995) "On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening," *Biometrika*, 82(2), 287 – 297.

Walter, S. D., Hildreth, S.W., & Beaty, B. J. (1980) "Estimation of infection rates in populations of organisms using pools of variable size," *American Journal of Epidemiology*, 112(1), 124 – 128.