Author/s:
Hepworth, G

Title:
Approximate bias of the estimated proportion in group testing

Date:
2017-03-01

Citation:
Hepworth, G. (2017). Approximate bias of the estimated proportion in group testing. Environmental and Ecological Statistics, 24 (1), pp.1-6. https://doi.org/10.1007/s10651-016-0358-7.

Persistent Link:
https://hdl.handle.net/11343/283253

Approximate bias of the estimated proportion in group testing

Graham Hepworth

School of Mathematics and Statistics
The University of Melbourne
Victoria 3010, Australia

Abstract:

The MLE for estimating proportions by group testing is biased. An expression for the approximate bias has been previously presented, which enables the creation of a less biased estimator by removing the term of $O(n^{-1})$. However, in this previous work the term of $O(n^{-2})$ was incorrectly derived. This note gives a correct derivation, and examines the relative contribution of the two terms.

Keywords: bias correction, estimation of proportions, group testing.

1 Bias of the MLE in group testing

Estimation of proportions by group testing occurs when units from a population are pooled together and tested in groups for the presence of an attribute. Group testing is also known as pooled testing, and in the environmental sciences, as a form of composite sampling (Patil 2011). In most situations the attribute is a disease (e.g. Keys et al. 2014), but it can also be the occurrence of transgenic plants (Montesinos-Lopez et al. 2011), the presence of a chemical contaminant (Colon et al. 2001), or indeed any unwanted or rare trait (Schaarschmidt 2007). In some cases, identification of the infected or defective units is also of interest (Tebbs et al. 2013), but our concern is with estimation of the proportion of such units (p) in the population. Some studies assume equal group sizes (e.g. Liu et al. 2013). while others have involved unequal group sizes (e.g. Crockett et al. 2012). In this note, we consider equal group sizes.

The maximum likelihood estimator (MLE) of p, which we denote $\hat{p}$, is positively biased, except in the case of all groups consisting of one unit (simple binomial testing). This bias was recognized in very early group testing work (Gibbs and Gower 1960). The exact bias of $\hat{p}$ can be calculated, given the number of groups n and the group size k, using the pmf of the number of positive groups and the MLE for each outcome. Swallow (1985) did this for different n and k, and constructed tables for assisting in design of group testing studies. In searching for estimators with less bias than the MLE, authors have found approximate expressions for the bias using power series, and then endeavored to remove the term of $O(n^{-1})$. The most successful was Burrows (1987), whose adjusted estimator had bias ranging from 1% to 5% of that of $\hat{p}$, and mean squared error (MSE) uniformly less.

Lovison et al. (1994) derived expressions for the bias and MSE of $\hat{p}$ with terms up to $O(n^{-2})$. They gave the correct expression for the term of $O(n^{-1})$, but an incorrect expression for the term of $O(n^{-2})$. The purpose of this note is to give the correct expression, and to examine the relative contribution of the two terms to the bias.

2 Approximate bias derivation

Expressing a function f of a random variable Y as a Taylor series around $\mathbb{E}(Y) = \mu$, and taking expectations, we obtain

$$\begin{aligned}\mathbb{E}[f(Y)] &= f(\mu) + f'(\mu)\mathbb{E}(Y-\mu) + \frac{f''(\mu)}{2!}\mathbb{E}(Y-\mu)^2 + \frac{f'''(\mu)}{3!}\mathbb{E}(Y-\mu)^3 + \cdots \\ &= f(\mu) + \frac{f''(\mu)}{2}\text{var}(Y) + \frac{f'''(\mu)}{6}\mathbb{E}(Y-\mu)^3 + \cdots\end{aligned}$$

Let Y be the number of negative groups out of n; it is more convenient in this derivation to work with Y rather than the number of positive groups, and with $q = 1-p$, the proportion of negative units, rather than p. Under standard assumptions, Y has a binomial distribution with parameters n and q^k, and so $\mathbb{E}(Y) = nq^k$ and $\text{var}(Y) = nq^k(1-q^k)$. Let $f(Y) = (Y/n)^{1/k} = \hat{q}$, the MLE of q. Now

$$f'(Y) = \frac{1}{k}\frac{Y^{1/k-1}}{n^{1/k}}, \quad f''(Y) = \frac{1}{k}\left(\frac{1}{k}-1\right)\frac{Y^{1/k-2}}{n^{1/k}},$$

$$f'''(Y) = \frac{1}{k}\left(\frac{1}{k}-1\right)\left(\frac{1}{k}-2\right)\frac{Y^{1/k-3}}{n^{1/k}}, \quad \text{etc.}$$

If a random variable Z is binomially distributed with parameters n and θ, the third and fourth central moments of Z are

$$\begin{aligned}\mathbb{E}[(Z-n\theta)^3] &= n\theta(1-\theta)(1-2\theta) \\ \mathbb{E}[(Z-n\theta)^4] &= 3\left[n\theta(1-\theta)\right]^2 + n\theta(1-\theta)\left[1-6\theta(1-\theta)\right]\end{aligned}$$

(Johnson et al. 2005). Collecting terms of $O(n^{-1})$ and $O(n^{-2})$,

$$\begin{aligned}\mathbb{E}(\hat{q}) = q \;&+\; \frac{1-k}{2k^2}\,\frac{(nq^k)^{1/k-2}}{n^{1/k}}\,nq^k(1-q^k) \\ &+\; \frac{(k-1)(2k-1)}{6k^3}\,\frac{(nq^k)^{1/k-3}}{n^{1/k}}\,nq^k(1-q^k)(1-2q^k) \\ &-\; \frac{(k-1)(2k-1)(3k-1)}{8k^4}\,\frac{(nq^k)^{1/k-4}}{n^{1/k}}\,n^2q^{2k}(1-q^k)^2 \\ &+\; \cdots\end{aligned}$$

So the bias of $\hat{q}$ is

$$\begin{aligned}\mathbb{E}(\hat{q}) - q &= \frac{1-k}{2k^2}\,\frac{q^{1-k}(1-q^k)}{n} + \frac{(k-1)(2k-1)}{6k^3}\,\frac{q^{1-2k}(1-q^k)(1-2q^k)}{n^2} \\ &- \frac{(k-1)(2k-1)(3k-1)}{8k^4}\,\frac{q^{1-2k}(1-q^k)^2)}{n^2} + O(n^{-3}) \\ = \frac{k-1}{2k^2}\,q(1-q^k)&\left[-\frac{1}{nq^k} + \frac{(2k-1)(1-2q^k)}{3kn^2q^{2k}} - \frac{(2k-1)(3k-1)(1-q^k)^2}{4k^2n^2q^{2k}}\right] + O(n^{-3})\,.\end{aligned}$$

Therefore the bias of $\hat{p}$ is

$$(1-p)\left(1-(1-p)^k\right)\left(\frac{k-1}{2k^2}\right) \times$$
$$\left[\frac{1}{n(1-p)^k} + \frac{(2k-1)(2(1-p)^k-1)}{3kn^2(1-p)^{2k}} + \frac{(2k-1)(3k-1)(1-(1-p)^k)^2}{4k^2n^2(1-p)^{2k}}\right] + O(n^{-3})\,.$$

The formula given by Lovison et al. (1994) had a different expression for the term of $O(n^{-2})$. Their expression leads to substantially larger values of the bias than our (correct) formula, even though there is no difference in the more dominant $O(n^{-1})$ term.

For illustrate the discrepancy, consider the study of Keys et al. (2014), who estimated the prevalence of hepatitis C in North Carolina by testing 224 groups with 80 serum samples in each. They obtained 138 positive groups, which resulted in $\hat{p} = 0.0119$. At this prevalence, the bias term of $O(n^{-1})$ is 0.37%. The term of $O(n^{-2})$ is negligible (0.002%), but the formula of Lovison et al. (1994) gives 1.04%. At a higher prevalence, the difference is even greater; for example, at $p = 0.03$, the bias term of $O(n^{-1})$ is 0.93%, and the term of $O(n^{-2})$ is calculated to be 0.03% (our formula) vs 12.27% (Lovison).

Colon et al. (2001) also derived expressions for the bias and MSE of $\hat{p}$. Their expression is identical for the $O(n^{-1})$ term, and different from ours for the $O(n^{-2})$ term. In the example above, the bias term of $O(n^{-2})$ is calculated using their formula to be 0.003% at $p = 0.0119$, and 0.04% at $p = 0.03$; the differences from what we obtained are inconsequential.

The term of $O(n^{-1})$ in the bias is an increasing function of p, both in absolute bias and in percentage bias, because of the dominance of $(1-p)^{k-1}$ in the denominator.

The term of $O(n^{-2})$ is also an increasing function of p in absolute bias, but is not strictly increasing in percentage bias; for small p the percentage bias is fairly constant.

Gart (1991) found the bias of an MLE, except for a term of $O(n^{-2})$, to be

$$-\frac{2\frac{\partial I}{\partial p} + \mathbb{E}\left[\frac{\partial^3 l}{\partial p^3}\right]}{2[I(p)]^2}$$

where l is the log-likelihood and I is the Fisher information. Hepworth and Watson (2009) applied this to group testing and obtained the term of $O(n^{-1})$ listed above.

3 Bias comparisons

Table 1 shows the percentage bias contributed by the terms of $O(n^{-1})$ and $O(n^{-2})$ for $(n = 20, 50, 100) \times (k = 10, 20, 50)$, and a range of p. Also shown is the exact bias percentage, calculated using the pmf of Y and the expression above for $\hat{q}$. The values of n have been chosen as realistic numbers of groups in many situations, but large enough to approach asymptotic conditions. Similarly, the values of k have been chosen to give a range of realistic group sizes. The values of p have been chosen to be broadly consistent with the values of k; smaller k should, if possible, be used with larger p, to avoid obtaining all positive groups. Hepworth and Watson (2009) placed an upper bound ψ on p when evaluating bias, where ψ is the value of p at which the probability of all positive groups is 0.05. We adopt this approach here, and set the maximum p to be in the vicinity of ψ for $n = 100$; for example, when $k = 10$, $\psi = 0.297$, so the largest value of p used is 0.3.

These results show that group size has only a modest effect on the bias for small p, but a greater effect for larger p; larger k is associated with more bias. The overall bias is small for large n, especially for small p, where it is generally less than 1%. Large bias is evident for smaller n and larger p, though some of the results for which this is the case are for p clearly exceeding ψ; for example, when $k = 20$ and $n = 20$, $\psi = 0.094$. These general observations are totally consistent

Table 1: Percentage bias of $O(n^{-1})$ and $O(n^{-2})$ in the MLE of p, and exact percentage bias, for testing n groups of k units

	$n=20$			$n=50$			$n=100$		
p	$O(n^{-1})$	$O(n^{-2})$	exact	$O(n^{-1})$	$O(n^{-2})$	exact	$O(n^{-1})$	$O(n^{-2})$	exact
				$k=10$					
0.001	2.26	0.07	2.34	0.90	0.01	0.92	0.45	0.00	0.45
0.002	2.27	0.07	2.35	0.91	0.01	0.92	0.45	0.00	0.46
0.005	2.30	0.07	2.38	0.92	0.01	0.93	0.46	0.00	0.46
0.01	2.36	0.07	2.44	0.94	0.01	0.96	0.47	0.00	0.47
0.02	2.47	0.07	2.57	0.99	0.01	1.00	0.49	0.00	0.50
0.03	2.59	0.07	2.72	1.04	0.01	1.06	0.52	0.00	0.52
0.04	2.72	0.07	2.87	1.09	0.01	1.11	0.54	0.00	0.55
0.05	2.87	0.08	3.05	1.15	0.01	1.17	0.57	0.00	0.58
0.07	3.19	0.11	3.46	1.27	0.02	1.31	0.64	0.00	0.65
0.1	3.78	0.21	4.40	1.51	0.03	1.58	0.76	0.01	0.77
0.15	5.20	0.67	11.83	2.08	0.11	2.27	1.04	0.03	1.08
0.2	7.48	2.09	41.76	2.99	0.33	4.58	1.50	0.08	1.62
0.3	18.06	22.48	123.43	7.22	3.60	52.86	3.61	0.90	14.83
				$k=20$					
0.001	2.40	0.08	2.48	0.96	0.01	0.97	0.48	0.00	0.48
0.002	2.42	0.08	2.51	0.97	0.01	0.98	0.48	0.00	0.49
0.005	2.49	0.07	2.59	1.00	0.01	1.01	0.50	0.00	0.50
0.01	2.62	0.07	2.73	1.05	0.01	1.06	0.52	0.00	0.53
0.02	2.90	0.08	3.06	1.16	0.01	1.18	0.58	0.00	0.59
0.03	3.22	0.11	3.46	1.29	0.02	1.32	0.64	0.00	0.65
0.04	3.60	0.15	3.97	1.44	0.02	1.49	0.72	0.01	0.73
0.05	4.04	0.23	4.80	1.62	0.04	1.69	0.81	0.01	0.82
0.07	5.16	0.55	11.87	2.06	0.09	2.22	1.03	0.02	1.07
0.1	7.72	1.96	69.51	3.09	0.31	4.82	1.54	0.08	1.66
0.15	16.69	15.70	248.84	6.68	2.51	78.13	3.34	0.63	13.70
				$k=50$					
0.001	2.51	0.08	2.60	1.00	0.01	1.02	0.50	0.00	0.51
0.002	2.57	0.08	2.68	1.03	0.01	1.05	0.51	0.00	0.52
0.005	2.78	0.08	2.91	1.11	0.01	1.13	0.56	0.00	0.56
0.01	3.17	0.10	3.38	1.27	0.02	1.30	0.63	0.00	0.64
0.02	4.19	0.24	5.30	1.68	0.04	1.75	0.84	0.01	0.86
0.03	5.68	0.69	29.45	2.27	0.11	2.48	1.14	0.03	1.18
0.04	7.88	1.91	151.61	3.15	0.30	5.84	1.58	0.08	1.69
0.05	11.17	5.07	380.05	4.47	0.81	38.55	2.23	0.20	3.14
0.07	23.86	34.73	766.49	9.55	5.56	342.35	4.77	1.39	91.84

with theory; of more interest here are the relative contributions of the $O(n^{-1})$ and $O(n^{-2})$ terms, and their combined magnitude in relation to the exact bias.

The contribution of the $O(n^{-2})$ term is generally very small in percentage terms, and also in relation to the magnitude of the $O(n^{-1})$ term. This is reassuring to those seeking to adjust the MLE by removing the term of $O(n^{-1})$. This includes the estimator of Burrows (1987), as well as those applying the method of Gart (1991). The sum of the $O(n^{-1})$ and $O(n^{-2})$ terms is close to the exact bias for small p, but when n is not large (e.g. 20) the discrepancy is substantial for larger p. This sounds a note of caution regarding the use of asymptotic approximations for group testing problems unless n is large or p small.

For p close to ψ the contribution of the $O(n^{-2})$ term relative to the $O(n^{-1})$ term is non-trivial. For example, when $k = 20$ and $n = 20$, at $p = \psi = 0.094$, the contributions of the two terms are 7.1%, and 1.5% respectively. At first glance, increasing the bias adjustment from 7.1% to 8.6% might be seen as worthwhile. However, both terms are an order of magnitude less than the exact bias of 51.7%, making the choice of whether to include the $O(n^{-2})$ term of limited consequence, and highlighting the severity of the bias of the MLE in some group testing situations.

References

Burrows PM (1987) Improved estimation of pathogen transmission rates by group testing. Phytopathology 77:363–365.

Colon S, Patil GP, Taillie C (2001) Estimating prevalence using composites. Environ. Ecol. Stat. 8:213–236.

Crockett RK, Burkhalter K, Mead D, Kelly R, Brown J, Vernado W, Roy A, Horiuchi K, Biggerstaff BJ, Miller B, Nasci R (2012) Culex Flavivirus and West Nile Virus in *Culex quinquefasciatus* populations in the southeastern United States. J. Med. Entomol. 49:165–174.

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

Gibbs AJ, Gower JC (1960) The use of a multiple-transfer method in plant virus transmission studies—some statistical points arising in the analysis of results. Ann. Appl. Biol. 48: 75–83.

Hepworth G (2013) Improved estimation of proportions using inverse binomial group testing. J. Agric. Biol. Envir. S. 18:102–119.

Hepworth G, Watson R (2009) Debiased estimation of proportions in group testing. J. Roy. Stat. Soc. C-App. 58:105–121.

Johnson NL, Kemp AW, Kotz S (2005) Univariate Discrete Distributions, Third Edition. Wiley, New York.

Keys JR, Leone PA, Eron JJ, Alexander K, Brinson M, Swanstrom R (2014) Large Scale Screening of Human Sera for HCV RNA and GBV-C RNA. J. Med. Virol. 86:473–477.

Liu C, Liu A, Zhang B, Zhang Z (2013) Improved confidence intervals of a small probability from pooled testing with misclassification. Front. Pub. Heal. doi:10.3389/fpubh.2013.00039.

Lovison G, Gore SD, Patil GP (1994) Design and analysis of composite sampling procedures: a review. In: Patil GP, Rao CR (eds) Handbook of Statistics. Elsevier, Amsterdam, Vol 12, pp 103–166.

Patil GP (2011) Composite sampling: A novel method to accomplish observational economy in environmental studies: A monograph introduction. Environ. Ecol. Stat. 18:385–392.

Schaarschmidt F (2007) Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing. Commun. Biom. Crop Sci. 2:32–40.

Swallow WH (1985) Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology 75:882–889.

Tebbs JM, McMahan CS, Bilder CR (2013) Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. Biometrics 69:1064–1073.

Graham Hepworth

I am an Associate Professor of Statistics in the School of Mathematics and Statistics at the University of Melbourne. I have over 70 refereed publications, in a wide range of subject areas. My methodological research has been mainly in estimating proportions by group testing (pooled testing), and in confidence intervals for discrete data, especially in relation to the binomial parameter.

My collaborative research, much of which has arisen from consulting, has been in many fields, including animal reproduction, plant pathology, cardiac electrophysiology, dentistry, nursing, water use, compost science, ecology, entomology, veterinary science, and neurology

I have more than 30 years experience as a consulting statistician, with particular expertise and experience in the design and analysis of experiments, as well as in surveys and sampling. I have undertaken projects across a wide range of fields, including medicine, dentistry, food science, education, law, ecology, languages, psychology, and especially plant and animal sciences. I have performed work for a wide range of state and federal government agencies, small and large businesses, and individual researchers.

I regularly teach the intensive short courses “Design and "Design and Analysis of Experiments" to participants from a wide range of fields. These are courses I developed using my extensive experience in surveys and experiments, which help researchers apply sound statistical methods to their research in practical ways. I have previously taught both undergraduate and Masters level courses.

I have a Bachelor of Science (Honours), Master of Science, and a PhD, all in mathematical statistics from the University of Melbourne. I am an active member and a past president of the Australasian Region of the International Biometric Society. I am an Accredited Statistician (AStat) of the Statistical Society of Australia.