ADAPTIVE ESTIMATION OF BINOMIAL PROBABILITIES UNDER MISCLASSIFICATION

W. Albers * H.J. Veldman †

Abstract

If misclassification occurs the standard binomial estimator is usually seriously biased. It is known that an improvement can be achieved by using more than one observer in classifying the sample elements. Here it will be investigated which number of observers is optimal given the total number of judgements that can be made. An adaptive estimator for the probability of interest is introduced which uses an estimator of this optimal number of observers, obtained without additional cost. Some simulation results are presented which suggest that the adaptive procedure performs quite well.

Key words & Phrases: *adaptive estimator, binomial distribution, misclassification.*

1. Introduction

Consider a population in which a fraction p of the elements possess a certain characteristic E . To estimate p , an observer will typically classify the elements in a random sample of size n from this population and obtain the binomial estimator for p . However, if the observer is not infallible, this standard procedure is not at all satisfactory anymore. It will generally show a nonvanishing bias. Especially if p is small, which will typically be the case in many applications, this effect is relatively strong.

Several authors, some of which we will mention in the subsequent sections, have therefore devised alternative approaches to the misclassification problem. First we shall discuss the majority rule, in which for some $k = 1, 2, \dots$, each element of the sample is judged by $2k + 1$ independent observers and then classified into the category indicated by the majority. Given a fixed budget, say N , for the total number of judgements to be made, it seems interesting to determine which value of k is optimal. We shall attack this problem in the present paper. Occurring in the mean squared error (MSE) are the bias, which decreases in k , and the variance, which increases in k . Hence it seems clear that there will be a certain k at which we should stop adding observers, since from that value on the decrease of bias will be outweighed by the increase of variance.

However, such a determination of the optimal k will only mean a first step towards a procedure that can be applied in practice, since the optimal k will depend on p and the probabilities of misclassification involved. It follows that first we have to apply the majority rule with e.g. 3 observers to a preliminary sample to

* Twente University of Technology (at the time of writing this paper)

† Twente University of Technology

obtain preliminary estimates for these unknown probabilities, which in their turn lead to an estimate $\tilde{k}$ for the optimal k . To obtain the final estimate for p we use the majority rule for $\tilde{k}$ in the following way: we apply $2\tilde{k} - 2$ observers to each element of the preliminary sample and use $2\tilde{k} + 1$ observers for each of the remaining elements, thus "absorbing" the first sample in a natural way and avoiding extra cost. (If one finds it awkward to reexamine the first sample, one can apply the rule to new elements only, but this will be less efficient.)

Clearly, the adaptive estimator described above cannot be applied if $\tilde{k} = 0$. In that case we are apparently so close to the ideal model of no misclassification that the standard binomial estimator should be preferred. However, we have already been using a majority rule on the preliminary sample. Rather than simply throwing away the extra observations, we shall try to repair as much of the damage as possible. The philosophy behind this strict attitude is that as a robust competitor of the standard estimator, our adaptive estimator, besides performing well if the ideal model is incorrect, should also lose as little efficiency as possible near the ideal model. The desired effect is achieved by applying an existing method of combining inexpensive fallible and expensive infallible observers.

In Section 2 the majority rule is discussed. In Section 3 we introduce the method to be used if $\tilde{k} = 0$. An approximate solution for the optimal k , which is shown to work remarkably well, is presented in Section 4. The resulting adaptive estimator is discussed in Section 5. Some simulation results are added which indicate that the estimator indeed lives up to expectations. For convenience the precise recipe to be followed when applying the estimator is summarized in Section 6.

2. The majority rule

The majority rule (MR) is introduced as follows: for some $k \in \{0, 1, 2, \dots\}$, each element in a random sample is judged by $2k + 1$ observers. If at least $k + 1$ of them decide that it has property E , it is classified as having E . The estimator then is the fraction of elements in the sample that have been classified as having E . For sample size n we denote it by $M_{k,N}$, where $N = (2k + 1)n$ is the total number of judgments performed.

The performance of the MR with $k = 1$ has been investigated by Isaki (1974), who also remarks that such a procedure was used in several U.S. Censuses of Population and Housing. He compares $M_{1,N}$ to the average of the binomial estimators obtained by 3 observers on one and the same sample. Isaki demonstrates that typically $M_{1,N}$ is to be preferred. He conjectures that further improvement may be possible by using $k = 2$.

We shall now proceed to discuss some properties of the MR. We shall always assume that the observers are of the same quality. Hence we can define

$$\begin{aligned}\theta &= P(\text{not } E \text{ is observed} \mid E \text{ is true}), \\ \phi &= P(E \text{ is observed} \mid E \text{ is not true}).\end{aligned}\tag{2.1}$$

Moreover, we shall suppose that the observers work independently. Then it is easy to verify that the bias b_k resulting from using $M_{k,N}$ is independent of n and equals

$$b_k = -p\bar{B}(2k+1, k+1, \theta) + (1-p)\bar{B}(2k+1, k+1, \phi), \quad (2.2)$$

where $\bar{B}(n, j, \rho) = \sum_{i=j}^n \binom{n}{i} \rho^i (1-\rho)^{n-i}$. The MSE $\tau_{k,N}^2$ for a sample of size n is in terms of b_k given by

$$\tau_{k,N}^2 = b_k^2 + \frac{(p+b_k)(1-p-b_k)}{n}. \quad (2.3)$$

Note that some explanation of the nice behaviour of the MR, as was found by Isaki, is given by (2.2). From this expression it is clear that the bias of the MR only involves terms with powers of θ or ϕ of degree at least $k+1$. Moreover, suppose that all such terms are deleted from the vector of multinomial probabilities underlying the MR. For θ and ϕ sufficiently small this again leads to a vector of probabilities, which e.g. for $k=1$ equals $((1-3\phi)(1-p), 3\phi(1-p), 3\theta p, (1-3\theta)p)$. It is easy to verify that the MR precisely produces the maximum likelihood estimator (MLE) for p in this approximate model.

3. A combination of fallible and infallible observers

In this section we shall discuss a second approach to the misclassification problem, this time due to Tenenbein (1970). In Section 5 we shall demonstrate that this approach is precisely what we need in the case where $\tilde{k}=0$.

To avoid confusion later on with the parameters of interest p , θ and ϕ introduced in the previous sections, we shall introduce Tenenbein's method using instead parameters p' , θ' and ϕ' , which play a similar role.

Suppose that besides the fallible method of estimating p' , which is subject to the misclassification probabilities θ' and ϕ' , a second, infallible, but more expensive method is available. Then Tenenbein suggests to use a double sampling scheme: L elements are classified using the fallible method, after which for a sub-sample of size l of these L elements the true classifications are obtained using the infallible observer. The data of this experiment can be represented as

$$\begin{array}{cc} & \text{infallible observer} \\ & \begin{array}{cc} 0 & 1 \end{array} \\ \text{fallible} & \begin{array}{cc|c|c} 0 & l_{00} & l_{10} & l_{.0} & V \\ \text{observer} & 1 & l_{01} & l_{11} & l_{.1} & U \\ & & l_{.0} & l_{.1} & l & L-l \end{array} \end{array} \quad (3.1)$$

Tenenbein shows that the MLE for p' equals

$$T_{l,L} = \frac{l_{11}}{l_{.1}} \frac{U+l_{.1}}{L} + \frac{l_{10}}{l_{.0}} \frac{V+l_{.0}}{L}. \quad (3.2)$$

He also demonstrates that, if $L \rightarrow \infty$ and l/L tends to a positive constant, $T_{l,L}$ is asymptotically normal with mean p' , while for the asymptotic variance the following nice expression can be given

$$p'(1-p') \left\{ \frac{1-C}{l} + \frac{C}{L} \right\}, \quad \text{with } C = \frac{p'(1-p')(1-\theta'-\phi')^2}{\pi'(1-\pi')}, \quad (3.3)$$

where $\pi' = p'(1-\theta') + (1-p')\phi'$.

4. The determination of the optimal number of observers

In the present section we obtain an approximate solution for the optimal number of observers to be engaged. We shall concentrate on the situation where p , θ and ϕ are small and N , the total number of judgements to be made, is large. The idea behind this is that if θ and ϕ are not small, it becomes quite impossible to obtain reasonable estimators for p , unless extremely large samples are used. Moreover, without loss of generality we can suppose that $p \leq \frac{1}{2}$. If p is near $\frac{1}{2}$, there is not much reason to worry about some misclassification: in the bias $b_0 = -p\theta + (1-p)\phi$ (cf. (2.2)) of the standard binomial estimator $M_{0,N}$ the two parts will pretty much tend to cancel each other. Hence, for p not small, b_0 will tend to be small, whereas for p small the opposite occurs. Finally, if N is too small, under misclassification the estimator will be too unprecise to be of any use at all. To make precise what we mean by "small" for p , θ and ϕ and by "large" for N , we shall assume that

$$p \leq 0.15, \quad \theta \leq 0.20, \quad \phi \leq 0.20, \quad N \geq 50. \quad (4.1)$$

We now want to minimize with respect to k the MSE $\tau_{k,N}^2$ from (2.3), for given $N = (2k+1)n$. Since p is small and moreover

$$\bar{B}(2k+1, k+1, \phi) = \binom{2k+1}{k} \phi^{k+1} + o(\phi^{k+1}) \text{ for } \phi \rightarrow 0, \text{ we approximate } b_k \text{ in (2.2)}$$

by $\tilde{b}_k = \binom{2k+1}{k} \phi^{k+1}$ and try to minimize instead

$$\tilde{\tau}_{k,N}^2 = \tilde{b}_k^2 + \frac{(2k+1)(p + \tilde{b}_k)}{N}. \quad (4.2)$$

Let k_0 and $\tilde{k}_0$ be the values of k minimizing $\tau_{k,N}^2$ and $\tilde{\tau}_{k,N}^2$, respectively, then we have the following lower bound for $\tilde{k}_0$.

Lemma 4.1. Under (4.1) and for $p \neq 0$, $\phi \neq 0$, we have that $\tilde{k}_0 \geq \underline{\tilde{k}} = \underline{\tilde{k}}(p, \phi)$, where

$$\underline{\tilde{k}}(p, \phi) = \frac{\log(2p) - \log N - \log\{\phi^2(1-16\phi^2)\}}{2 \log(3\phi)} \quad (4.3)$$

Proof. From (4.2) it follows that $\bar{\tau}_{k,N}^2 - \bar{\tau}_{k+1,N}^2$ equals

$$\begin{aligned} & \left[\frac{2k+1}{k} \right]^2 \phi^{2k+2} \left\{ 1 - \left(\frac{4k+6}{k+2} \right)^2 \phi^2 \right\} \\ & + \frac{2k+1}{N} \left[\frac{2k+1}{k} \right] \phi^{k+1} \left\{ 1 - \frac{2k+3}{2k+1} \frac{4k+6}{k+2} \phi \right\} - \frac{2p}{N}. \end{aligned}$$

To simplify this expression we note that $\left(\frac{2k+1}{k} \right) \geq 3^k$ and $4 > (4k+6)/(k+2)$ for $k=0,1,2,\dots$. We also remark that $1 - \phi(2k+3)(4k+6)/\{(2k+1)(k+2)\} \geq 0$ for $k=0$ if $\phi \leq 1/9$, for $k=1$ if $\phi \leq 9/50$ and for $k=2,3,\dots$ if $\phi \leq 0.20$. It follows that for these values of k and ϕ

$$\bar{\tau}_{k,N}^2 - \bar{\tau}_{k+1,N}^2 \geq \phi^2(1 - 16\phi^2)(3\phi)^{2k} - \frac{2p}{N}. \quad (4.4)$$

Now it is easily verified directly that for $N \geq 50$ this result also holds for $k=0$ and $1/9 < \phi \leq 0.20$ and for $k=1$ and $9/50 < \phi \leq 0.20$. Hence we may conclude that (4.4) holds under (4.1) for all k . Since the right-hand side of (4.4) is positive if and only if $k < \bar{k}(p, \phi)$ as given in (4.3), it follows that $\bar{k}$ is indeed a lower bound for $\bar{k}_0$. $\square$

The relation between $\bar{k}$ and $\bar{k}_0$ is of secondary importance, $\bar{k}_0$ itself being a mere approximation. In view of lemma 4.1 we suggest as our approximation for k_0

$$k_a = [\bar{k}(p, \phi) + 1], \quad (4.5)$$

where $[y]$ stands for the integer part of y . In Figures 1a—1d we present both k_0 and $\bar{k}$ for $p=0.05$ and $N=300,600$ and for $p=0.15$ and $N=150,300$ as functions of ϕ for the case where ϕ and θ are equal. From these representative figures it is clear that (4.5) indeed provides an adequate approximation to k_0 , especially for small p . Since (4.5) moreover is very easy to calculate, we feel that such a simple rule is useful in practice, even though obtaining the exact value of k_0 using a computer will neither be very difficult nor expensive. We should also realize in this connection that the underlying model still merely is an approximation of the real situation, which makes using a good approximation rather than the exact value even more acceptable.

Figure 1a: $p = 0.05$, $N = 300$ Figure 1b: $p = 0.05$, $N = 600$

Figure 1c: $p = 0.15$, $N = 150$

Figure 1d: $p = 0.15$, $N = 300$

5. The adaptive estimator for p

By now we are in a position to introduce the announced adaptive procedure as a means of attacking the main problem of our paper. We begin by specifying a maximum $2k_M + 1$ for the number of observers to be used, for some $k_M = 2, 3, \dots$. We draw a preliminary sample of size

$$m = \left\lceil \frac{N}{2k_M + 1} \right\rceil, \quad (5.1)$$

and we apply the MR described in Section 2 with $k = 1$ to this sample. Let X_i be the number of elements in the sample for which i observers feel that the element has property E , for $i = 0, 1, 2, 3$. Then we have as our preliminary estimators for p and ϕ the MLE's based on the approximate multinomial distribution in Section 2,

$$\hat{p}_b = \frac{X_2 + X_3}{m}, \quad \hat{\phi}_b = \frac{X_1}{3(X_0 + X_1)}. \quad (5.2)$$

Subsequently we replace the true but unknown p and ϕ in the approximation k_a from (4.5) by these estimators $\hat{p}_b$ and $\hat{\phi}_b$, respectively, and we arrive at an estimator $2\tilde{K} + 1$ for the optimal number of observers $2k_0 + 1$, where

$$\tilde{K} = \min([\tilde{k}(\hat{p}_b, \hat{\phi}_b) + 1], k_M). \quad (5.3)$$

If the value $\tilde{k}$ attained by $\tilde{K}$ is at least 1, the remaining part of the procedure is straightforward as described in the introduction: the preliminary sample is absorbed into an overall sample of size $[N / (2\tilde{k} + 1)]$, judged by $(2\tilde{k} + 1)$ observers and the resulting estimator clearly is nothing but the MR estimator $M_{\tilde{k}, N}$ from Section 2.

It remains to deal with the situation where $\tilde{K} = 0$. Here we cannot simply use $M_{0, N}$, as we have already been spending 3 instead of 1 observers per element throughout our preliminary sample. The best MR still available merely produces the estimator $M_{0, N - 2mM}$. Since we are not willing to throw away the information contained in $2m$ of the observations, we have to look outside the class of MR's and that brings us to Tenenbein's approach as discussed in Section 3.

To adapt this method to the present situation, we begin by defining

$$l = [N / 5]. \quad (5.4)$$

On the reasons for this particular choice we shall come back later. We now let the 3 observers from our preliminary sample judge another $l - m$ elements. To exhaust the budget we then let one of the 3 observers, say the first, judge an additional $N - 3l$ objects. In Tenenbein's terminology, observer 1 is the fallible observer, who has judged a total of

$$L = N - 2l \quad (5.5)$$

observations. The role of the infallible observer is played not by a single observer

but by the majority classification, as applied to the first l elements. Consequently, although we start out with observers of equal quality, it nevertheless makes sense to invoke Tenenbein's results. As our estimator we again use the MLE given in (3.2). Since both l and L have been defined through (5.4) and (5.5) in terms of N , we shall denote this estimator in the present context by T_N rather than by $T_{l,L}$.

Note that to obtain an expression for the MSE v_N^2 of T_N we cannot simply use the asymptotic variance of T_N from (3.3). We should bear in mind that the parameters p' , θ' and ϕ' from Section 3 are almost, but not exactly, the same as the parameters p , θ and ϕ . In fact, p' is the probability that the majority of the 3 observers decide that a given element has property E , and consequently it equals

$$p' = p\{1 - 3\theta^2(1 - \theta) - \theta^3\} + (1 - p)\{3\phi^2(1 - \phi) + \phi^3\} \approx p + 3\phi^2. \quad (5.6)$$

Remember that T_N is only going to be used if $\tilde{K}$ happens to be 0, that is, if the estimated optimal number of observers equals 1. Since this will typically only happen if ϕ is quite small, (5.6) shows that the majority classification is indeed almost infallible.

To obtain the relation between θ' and θ , we note that θ' is the probability that the first observer decides that a given element does not possess E , while the majority decision is that it does have E . It follows that $\theta' = \{p\theta(1 - \theta)^2 + (1 - p)\phi^2(1 - \phi)\} / p'$. Likewise, we obtain that $\phi' = \{p\theta^2(1 - \theta) + (1 - p)\phi(1 - \phi)^2\} / (1 - p')$. We now arrive at the following asymptotic expression $\tilde{v}_N^2$ for the MSE v_N^2 of T_N :

$$\tilde{v}_N^2 = (p' - p)^2 + p'(1 - p') \left\{ \frac{1 - C}{l} + \frac{C}{L} \right\},$$

with C as in (3.3). This result can be expressed in terms of N , p , θ and ϕ by using (5.4), (5.5) and the expressions above for p' , θ' and ϕ' .

It remains to explain the choice for l given in (5.4). The objective is to choose l and $L = N - 2l$ given N in such a way that $\tilde{v}_N^2$ is minimal. This requires an estimator for C . Since for small p this estimator has a large variance, we prefer to use an "average" value of C . For $k_0 = 0$, we find that C roughly varies from 0.6 to 1, which produces $C = 0.8$ as our average. The corresponding optimal l is $0.2071N$ and hence the choice in (5.4).

To conclude this section we present some simulation results. In each simulation we generate a sample $Y_1, \dots, Y_m$, with m as in (5.1), from the distribution determined by $P(Y = j) = pb(3, j, 1 - \theta) + (1 - p)b(3, j, \phi)$ for $j = 0, 1, 2, 3$, where $b(n, k, p) = \binom{n}{k} p^k (1 - p)^{n-k}$. From this sample we evaluate $\tilde{K}$ in (5.3). For $i = 1, \dots, s$, where s is the total number of simulations, let $\tilde{k}_i$ be the value attained by $\tilde{K}$. A possible estimate of the performance of the adaptive procedure

Figure 2a: $p = 0.05$, $N = 300$, $s = 1000$

Figure 2b: $p = 0.05$, $N = 600$, $s = 1000$

Figure 2c: $p = 0.15$, $N = 150$, $s = 1000$

Figure 2d: $p = 0.15$, $N = 300$, $s = 1000$.

then is

$$\omega_N = \frac{1}{s} \sum_{i=1}^s \{r_i \tau_{\tilde{k}_i, N} + (1-r_i) \tilde{v}_N\},$$

where $r_i = 1$ if $\tilde{k}_i \geq 1$ and $r_i = 0$ if $\tilde{k}_i = 0$. In Figures 2a-2d we have sketched $\tau_{k, N}$ for $k=0,1,2,3$, $\tilde{v}_N$ and ω_N as functions of ϕ for the case where $\theta = \phi$, for $p=0.05$ and $N=300,600$, and for $p=0.15$ and $N=150,300$, respectively. In all cases $k_M=3$ and $s=1000$.

The interpretation of these figures is pretty straightforward. For ϕ (and θ) very close to zero, the standard binomial estimator $M_{0, N}$ is optimal, but as ϕ (and θ) increase it deteriorates rapidly. As k increases, $M_{k, N}$ provides better and better protection against the misclassification effects, but the price involved if θ and ϕ happen to be near 0 after all, becomes higher and higher. The Tenenbein-type estimator T_N already means some improvement, performing roughly speaking like $M_{1, N}$ for larger θ and ϕ , while sacrificing much less efficiency for θ and ϕ near zero. However, the real improvement seems to come from using our adaptive procedure. It rather nicely tends to follow $\min\{\tau_{0, N}, \tau_{1, N}, \tau_{2, N}, \tau_{3, N}, \tilde{v}_N\}$ as ϕ increases, that is, it indeed tends to select the optimal possibility for each given configuration.

6. Summary of the procedure

For convenience we here summarize the procedure in the following recipe. Select a maximal number of observers $2k_M + 1$ ($k_M \geq 2$) and a budget N . Use 3 observers to classify $m = [N / (2k_M + 1)]$ elements, where $[y]$ stands for the integer part of y . Let X_i be the number of these elements classified in E by i observers, for $i=0,1,2,3$. Use $\hat{p}_b = (X_2 + X_3) / m$ and $\hat{\phi}_b = X_1 / \{3(X_0 + X_1)\}$ as preliminary estimators for p and ϕ . Apply these to obtain an estimator $\tilde{K} = \min(\lfloor \tilde{k}(\hat{p}_b, \hat{\phi}_b) \rfloor + 1, k_M)$, where $\tilde{k}(p, \phi) = \{\log 2p - \log N - \log(\phi^2(1 - 16\phi^2))\} / \{2 \log(3\phi)\}$.

If $\tilde{K}$ attains a value $\tilde{k} \geq 1$, apply another $2\tilde{k} - 2$ observers to the first m elements and apply $2\tilde{k} + 1$ observers to a second sample of $[N / (2\tilde{k} + 1)] - m$ elements. Estimate p by the fraction of the $[N / (2\tilde{k} + 1)]$ elements classified in E by a majority of the $(2\tilde{k} + 1)$ observers.

If $\tilde{K} = 0$, let for $l = [N / 5]$ the 3 observers classify another $l - m$ elements. For the l elements now considered, let $w_1(w_0)$ be the fraction of elements classified in E by the first observer that are in E (not in E) according to the majority of the 3 observers. Then let this first observer classify another $N - 3l$ elements, thus bringing his total to $L = N - 2l$ elements. Let $V_1(V_0)$ be the fraction of these L elements classified in E (not in E) by the first observer. Estimate p by $w_1 V_1 + w_0 V_0$.

For $N \geq 50$, $p \leq 0.15$, $\phi \leq 0.2$ and $\theta \leq 0.2$, this procedure will work well.

References

- ISAKI, C.T. (1974). Bias of two classification procedures in a dichotomous population, *J. Amer. Statist. Ass.* 69, 957-960.
- TENENBEIN, A. (1970). A double sampling scheme for estimating from binomial data with misclassification, *J. Amer. Statist. Ass.* 65, 1350-1362.

Received October 1983, revised December 1983.