Ann Inst Stat Math (2012) 64:359–371
DOI 10.1007/s10463-010-0316-3

Estimators for the binomial distribution that dominate the MLE in terms of Kullback–Leibler risk

Paul Vos · Qiang Wu

Received: 23 June 2009 / Revised: 17 February 2010 / Published online: 17 September 2010

Abstract Estimators based on the mode are introduced and shown empirically to have smaller Kullback–Leibler risk than the maximum likelihood estimator. For one of these, the midpoint modal estimator (MME), we prove the Kullback–Leibler risk is below $\frac{1}{2}$ while for the MLE the risk is above $\frac{1}{2}$ for a wide range of success probabilities that approaches the unit interval as the sample size grows to infinity. The MME is related to the mean of Fisher's Fiducial estimator and to the rule of succession for Jefferey's noninformative prior.

Keywords Kullback–Leibler risk · Modal estimators · MLE

1 Introduction

Eguchi and Yanagimoto (2008) provide an asymptotic adjustment to the MLE that improves the Kullback–Leibler risk. We show that for the binomial family of distributions the Kullback–Leibler risk for the MLE is above 1/2 for a wide range of success probabilities that approaches the unit interval as the sample size grows to infinity. We introduce an estimate, the midpoint modal estimate, that has Kullback–Leibler risk below 1/2 on an interval of success probabilities that also approaches the unit interval as the sample size grows to infinity. The relationship between the MLE and the midpoint modal estimate (MME) is explored with particular attention to post-data considerations. The paper is outlined as follows. Section 2 defines modal estimators,

P. Vos (✉) · Q. Wu
Department of Biostatistics, East Carolina University,
Greenville, NC 27858, USA
e-mail: vosp@ecu.edu

Q. Wu
e-mail: wuq@ecu.edu

Sect. 3 contains the main result on Kullback–Leibler risk, Sect. 4 contains discussion and comments regarding the relationship between these estimators.

2 Modal set estimators

For fixed sample size n, we consider estimators for distributions in the binomial family

$$\mathcal{B}_n = \left\{ f_n : \mathcal{X}_n \mapsto R : B_n(x; p) = \binom{n}{x} p^x (1-p)^{n-x},\ 0 < p < 1 \right\},$$

on the sample space $\mathcal{X}_n = \{0, 1, 2, \ldots, n\}$. Sometimes the degenerate cases where $p = 0$ and $p = 1$ are considered part of the binomial family. We would not include these, but these cases can often be treated as special cases. The notation $B(n, p) \in \mathcal{B}_n$ emphasizes that points in $\mathcal{B}_n$ are distributions, but it is more common and less cumbersome to use parameter values to refer to these distributions. In particular, the estimate $B(n, \hat{p})$will be denoted by simply $\hat{p}$, $\hat{\theta}$ (log odds), or some other parameter. However, when constructing and evaluating estimators, we maintain the perspective that the object of interest is a distribution. For the estimators we consider, a better notation relating parameters and the distributions they name is $\hat{p} = p(\hat{B})$ because the estimate is obtained directly from $\mathcal{B}_n$. Our estimators, like the maximum likelihood estimator (MLE), are parameter invariant as are the evaluation criteria that we consider.

A *modal distribution for x* is any distribution in $\mathcal{B}_n$ for which x is a mode. The set of all modal distributions for x is *the modal set estimate for x*; expressed in terms of the parameter p the modal set estimate is

$$[p](x) = \left\{ p : f(x; p) \geq f(x'; p),\ \forall x' \in \mathcal{X}_n \right\}.$$

Compare this to the MLE which consists of a set with a single point:

$$\hat{p}(x) = \left\{ p : f(x; p) \geq f(x; p'),\ \forall p' \in [0, 1] \right\}.$$

The distribution that maximizes the likelihood is obtained by maximizing over the distribution space $\mathcal{B}_n$ (including the degenerate distributions with $p = 0$ and $p = 1$) while modal distributions satisfy a weaker condition, namely, that of assigning the greatest probability to the value that was observed. The median, like the mode, provides another set estimator

$$\left\{ p : F(x; p) \geq \frac{1}{2} \text{ and } F(x-1; p) \leq \frac{1}{2} \right\}$$

where F is the cumulative distribution function. The logic of the median estimator is that because data in the tails provide evidence against a value for p, an estimate for p should consist of values for which the observed data x is far from the tails.

A simple calculation shows that

$$[p](x) = \left[\frac{x}{n+1}, \frac{x+1}{n+1}\right],$$

so that

$$\bigcup_{x \in \mathcal{X}_n} [p](x) = [0, 1].$$

Except for the endpoints of the modal set estimates, this estimator partitions the parameter space.

Often it will be convenient to approximate the estimate $[p](x)$ with a single distribution p'. A simple method for approximating the distributions in $[p](x)$ is to use the midpoint as determined on the success probability scale. This gives the MME $B \in \mathcal{B}_n$ having success parameter

$$\frac{x + \frac{1}{2}}{n+1}. \tag{1}$$

Fisher (1973, pp. 62 to 68) shows that this is the mean of the posterior distribution for the success parameter when Jeffrey's prior is used. He also shows that the mean of the fiducial distribution when expanded in powers of n^{-1} is the same as (1) to at least the n^{-4} term.

3 Kullback–Leibler risk

For binomial distributions, the Kullback–Leibler divergence D is defined on $\mathcal{B}_n \times \mathcal{B}_n$ but can be expressed as a function D_n on $(0, 1) \times (0, 1)$ as

$$D(B(n, p),\ B(n, p')) = D_n(p, p') = nA(p, p') - nH(p) \tag{2}$$

where $A(p, p') = -p \log p' - (1-p) \log(1-p')$ and $H(p) = A(p, p)$ is the entropy function for a Bernoulli trial with success parameter p.

We consider the Kullback–Leibler risk for an estimator $\tilde{P}$ defined by

$$nR_{KL}(p, \tilde{P}) = E_p D(B(n, \tilde{P}), B(n, p)). \tag{3}$$

The risk can be partitioned into two non negative quantities as follows:

$$\begin{aligned} nR_{KL}(p, \tilde{P}) &= E_p D_n(\tilde{P}, p) \\ &= E_p(nA(\tilde{P}, p) - nH(\tilde{P})) \\ &= nA(E_p\tilde{P}, p) - nE_pH(\tilde{P}) \end{aligned}$$

$$= nA(E_p\tilde{P}, p) - nH(E_p\tilde{P}) + nH(E\tilde{P}) - nE_pH(\tilde{P})$$
$$= D_n(E_p\tilde{P}, p) + n\left(H(E_p\tilde{P}) - E_pH(\tilde{P})\right). \quad (4)$$

Since the MLE $\hat{P}$ is unbiased for p the first term on the right hand side of (4) is zero. We call $n^{-1}D_n(E_p\tilde{P}, p)$ the square of the Kullback–Leibler (KL) bias for $\tilde{P}$. The second term, the difference between the entropy of the expected value and the expected value of the entropy of the estimator, we call the KL variance of $\tilde{P}$. Interpreting the entropy $H(p)$ as the amount of information in the distribution named by p, the KL variance for the MLE $\hat{p}$ (or any unbiased estimator) can be understood as the loss of information due to using $\hat{p}$ rather than the true distribution p. For KL biased estimators the same interpretation holds with p replaced with the mean of the estimator.

We define the KL bias and KL variance of an estimator $\tilde{P}$ (equivalently, the distribution $B(1, \tilde{P})$) as

$$\text{Bias}_{KL}(\tilde{P}) = \text{sign}(E\tilde{P} - p) \times D_1^{1/2}(E_p\tilde{P}, p) \quad (5)$$
$$\text{Var}_{KL}(\tilde{P}) = ED_1(\tilde{P}, E_p\tilde{P})$$
$$= H(E_p\tilde{P}) - E_pH(\tilde{P}). \quad (6)$$

The equality in (6) follows from the linearity in the function A in its first argument. Note that $\tilde{P}$ stands for the *random* distribution $B(1, \tilde{P})$ but the KL bias depends only on the KL divergence from the fixed distribution $B(1, E\tilde{P})$ to $B(1, p)$; the KL variance is the difference between the entropy of this fixed distribution and the mean entropy of the random distribution $B(1, \tilde{P})$. Heskes (1998) considers bias and variance decomposition of KL divergence but defines risk as $ED(p, \tilde{P})$ which for the binomial is infinite. Eguchi and Yanagimoto (2008) define risk as we do, but do not consider the bias/variance decomposition. Using these definitions (4) becomes

$$R_{KL}(p, \tilde{P}) = \text{Bias}_{KL}^2(\tilde{P}) + \text{Var}_{KL}(\tilde{P}).$$

The risk for the MLE and the MME are shown in Fig. 1. Figure 1 suggests the following theorems.

Theorem 1 *For each sample size n, there exist* $0 < p_{lo}(n) < p_{hi}(n) < 1$ *such that when* $p_{lo}(n) \leq p \leq p_{hi}(n)$ *the Kullback–Leibler risk* $nR_{KL}(p, \hat{P}) > \frac{1}{2}$*. Furthermore,* $p_{lo}(n) = 1 - p_{hi}(n) = O(n^{-1})$.

Proof We expand $H(\hat{p})$ around $H(p)$ using a Taylor series. First,

$$\hat{p}\log(\hat{p}) - p\log(p) = (1 + \log(p))(\hat{p} - p) + \frac{1}{2!p}(\hat{p} - p)^2 + R_1$$

where

$$R_1 = n\sum_{k=2}^{\infty}\frac{(-1)^{k+1}(k-1)!}{(k+1)!p^k}E\left[(\hat{p} - p)^{k+1}\right].$$

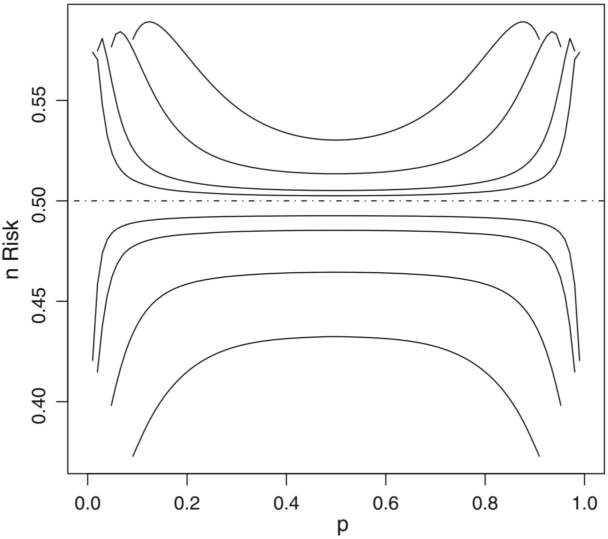

Fig. 1 Kullback–Leibler Risk risk for MLE (>1/2) and MME (<1/2) for $n = 10, 20, 50,$ and 100. Risks for these estimators converge as $n \to \infty$

Second,

$$\begin{aligned}&(1-\hat{p})\log(1-\hat{p}) - (1-p)\log(1-p)\\&\quad = (-1-\log(1-p))(\hat{p}-p) + \frac{1}{2!(1-p)}(\hat{p}-p)^2 + R_2\end{aligned}$$

where

$$R_2 = n\sum_{k=2}^{\infty} \frac{(k-1)!}{(k+1)!(1-p)^k} E\left[(\hat{p}-p)^{k+1}\right].$$

Given $E[\hat{p}] = p$ and $E[(\hat{p}-p)^2] = p(1-p)/n$, we have

$$nE[H(p) - H(\hat{p})] = \frac{1}{2} + R_1 + R_2.$$

To show $nR_{KL}(p\hat{P}) > \frac{1}{2}$, it suffices to show $R_1 + R_2 > 0$. Writing $R_1 + R_2$ in a different way gives

$$R_1 + R_2 = n\sum_{k=2}^{\infty} \frac{E[(\hat{p}-p)^{k+1}]}{(k+1)k}\left(\frac{(-1)^{k+1}}{p^k} + \frac{1}{(1-p)^k}\right).$$

It is not hard to see that the terms in $R_1 + R_2$ with $k = 3, 5, 7, \ldots$ are all positive. Or by using the mean value theorem, we have

$$R_1 + R_2 = n \sum_{k=2}^{4} \frac{E[(\hat{p} - p)^{k+1}]}{(k+1)k} \left(\frac{(-1)^{k+1}}{p^k} + \frac{1}{(1-p)^k} \right) + R,$$

where

$$R = \frac{n}{30} E\left[(\hat{p} - p)^6 \left(\frac{1}{p^{*5}} + \frac{1}{(1-p^*)^5} \right) \right] > 0$$

and p^* is in between $\hat{p}$ and p. In other words, a sufficient condition of $nE[H(p) - H(\hat{p})] > \frac{1}{2}$ is $R_1 + R_2 - R \geq 0$.

The following are the 3rd, 4th and 5th central moments of $\hat{p}$ where $q = 1 - p$.

$$\begin{aligned} E[(\hat{p} - p)^3] &= \frac{pq(q-p)}{n^2} \\ E[(\hat{p} - p)^4] &= \frac{3p^2q^2}{n^2} + \frac{pq(1-6pq)}{n^3} \\ E[(\hat{p} - p)^5] &= \frac{10p^2q^2(q-p)}{n^3} + \frac{pq(q-p)(1-12pq)}{n^4} \end{aligned}$$

Some further algebra shows that

$$\begin{aligned} E[(\hat{p} - p)^3]\left(-\frac{1}{p^2} + \frac{1}{q^2} \right) &= \frac{(-1+4pq)}{n^2 pq} \\ E[(\hat{p} - p)^4]\left(\frac{1}{p^3} + \frac{1}{q^3} \right) &= \left(\frac{3pq}{n^2} + \frac{1-6pq}{n^3} \right) \frac{(1-3pq)}{p^2q^2} \\ E[(\hat{p} - p)^5]\left(-\frac{1}{p^4} + \frac{1}{q^4} \right) &= \left(\frac{10pq}{n^3} + \frac{1-12pq}{n^4} \right) \frac{(1-2pq)(-1+4pq)}{p^3q^3} \end{aligned}$$

Let $x = \frac{1}{pq}$, then $x \geq 4$ with the equality holds if and only if $p = q = \frac{1}{2}$. Some additional algebra shows that

$$\begin{aligned} f(x) &\triangleq 60n^3(R_1 + R_2 - R) \\ &= -3x^3 - (25n - 54)x^2 + (5n^2 + 135n - 240)x - (5n^2 + 150n - 288). \end{aligned}$$

Since $f(4) = 5n(3n - 2) > 0$ and $f(0) < 0$ for $n \geq 2$, this cubic has three real roots, one of which is >4. It is enough to show that the largest root of $f(x)$ tends to ∞ as $n \to \infty$. The derivative of $f(x)$ is a quadratic whose largest root

$$-\frac{25}{9} n + 6 + \frac{1}{9} \sqrt{670\, n^2 - 1485\, n + 756}$$

goes to infinity as $n \to \infty$. Therefore, the largest root of $f(x)$ tends to infinity. □

Theorem 2 *For each sample size n, there exist* $0 < p_{lo}(n) < p_{hi}(n) < 1$ *such that when* $p_{lo}(n) \leq p \leq p_{hi}(n)$ *the Kullback–Leibler risk* $nR_{KL}(p, \tilde{P}) < \frac{1}{2}$. *Furthermore,* $p_{lo}(n) = 1 - p_{hi}(n) = O(n^{-1})$.

Proof We write the definition of the risk explicitly as

$$\begin{aligned} R(p, \tilde{p}) &= \bar{A}(E\tilde{p}, p) - EH(\tilde{p}) \\ &= -E\tilde{p}\log(p) - (1-E\tilde{p})\log(1-p) + E[\tilde{p}\log(\tilde{p})] + E[(1-\tilde{p})\log(1-\tilde{p})] \\ &= E[\tilde{p}\log(\tilde{p}) - \tilde{p}\log(p) + (1-\tilde{p})\log(1-\tilde{p}) - (1-\tilde{p})\log(1-p)] \end{aligned}$$

By using Taylor series expansion and the mean value theorem, we have

$$\begin{aligned} &\tilde{p}\log(\tilde{p}) - \tilde{p}\log(p) + (1-\tilde{p})\log(1-\tilde{p}) - (1-\tilde{p})\log(1-p) \\ &\quad = \frac{\tilde{p}(\tilde{p}-p)}{p} - \frac{(1-\tilde{p})(\tilde{p}-p)}{q} \\ &\qquad - \frac{\tilde{p}(\tilde{p}-p)^2}{2!p^2} - \frac{(1-\tilde{p})(\tilde{p}-p)^2}{2!q^2} + \frac{\tilde{p}(\tilde{p}-p)^3}{3!p^3} \\ &\qquad - \frac{(1-\tilde{p})(\tilde{p}-p)^3}{3!q^3} - \frac{\tilde{p}(\tilde{p}-p)^4}{4!p_*^4} - \frac{(1-\tilde{p})(\tilde{p}-p)^4}{4!q_*^4} \end{aligned}$$

where p_* is in between $\tilde{p}$ and p and $q_* = 1 - p_*$. It is clear that

$$E\left[-\frac{\tilde{p}(\tilde{p}-p)^4}{4!p_*^4} - \frac{(1-\tilde{p})(\tilde{p}-p)^4}{4!q_*^4}\right] < 0.$$

For $\tilde{p} = \frac{X+\frac{1}{2}}{n+1}$, we have the central moments

$$\begin{aligned} E[\tilde{p}] &= \frac{np+\frac{1}{2}}{n+1} \\ E[(\tilde{p}-E\tilde{p})^2] &= \frac{npq}{(n+1)^2} \\ E[(\tilde{p}-E\tilde{p})^3] &= \frac{npq(q-p)}{(n+1)^3} \\ E[(\tilde{p}-E\tilde{p})^4] &= \frac{3n^2p^2q^2 + npq(1-6pq)}{(n+1)^4} \end{aligned}$$

Then

$$\begin{aligned} E\big[(\tilde{p}-p)^2\big] &= E[(\tilde{p}-E\tilde{p})^2] + (E\tilde{p}-p)^2 = \frac{npq}{(n+1)^2} + \frac{1-4pq}{4(n+1)^2} \\ E\big[(\tilde{p}-p)^3\big] &= E[(\tilde{p}-E\tilde{p})^3] + 3E[(\tilde{p}-E\tilde{p})^2](E\tilde{p}-p) \\ &\quad + 3E[\tilde{p}-E\tilde{p}](E\tilde{p}-p)^2 + (E\tilde{p}-p)^3 \end{aligned}$$

$$= \frac{npq(q-p)}{(n+1)^3} + \frac{3npq(q-p)}{2(n+1)^3} + \frac{(q-p)^3}{8(n+1)^3}$$

$$\begin{aligned} E\big[(\tilde{p}-p)^4\big] &= E\left[(\tilde{p}-E\tilde{p})^4\right] + 4E\left[(\tilde{p}-E\tilde{p})^3\right](E\tilde{p}-p) \\ &\quad + 6E[(\tilde{p}-E\tilde{p})^2](E\tilde{p}-p)^2 + (E\tilde{p}-p)^4 \\ &= \frac{3n^2p^2q^2+npq(1-6pq)}{(n+1)^4} + \frac{2npq(q-p)^2}{(n+1)^4} \\ &\quad + \frac{3npq(q-p)^2}{2(n+1)^4} + \frac{(q-p)^4}{16(n+1)^4} \end{aligned}$$

As a result,

$$E\left[\frac{\tilde{p}(\tilde{p}-p)}{p} - \frac{(1-\tilde{p})(\tilde{p}-p)}{q}\right] = \frac{E\left[(\tilde{p}-p)^2\right]}{pq} = \frac{n}{(n+1)^2} + \frac{1-4pq}{4(n+1)^2pq}$$

$$\begin{aligned} E\left[-\frac{\tilde{p}(\tilde{p}-p)^2}{2!p^2} - \frac{(1-\tilde{p})(\tilde{p}-p)^2}{2!q^2}\right] &= E\left[(\tilde{p}-p)^3\right]\left(-\frac{1}{2p^2} + \frac{1}{2q^2}\right) \\ &\quad - E\left[(\tilde{p}-p)^2\right]\left(\frac{1}{2p} + \frac{1}{2q}\right) \\ &= -\frac{5n(1-4pq)}{4pq(n+1)^3} - \frac{(1-4pq)^2}{16p^2q^2(n+1)^3} \\ &\quad - \frac{n}{2(n+1)^2} - \frac{1-4pq}{8(n+1)^2pq} \end{aligned}$$

$$\begin{aligned} E\left[\frac{\tilde{p}(\tilde{p}-p)^3}{3!p^3} - \frac{(1-\tilde{p})(\tilde{p}-p)^3}{3!q^3}\right] &= E\left[(\tilde{p}-p)^4\right]\left(\frac{1}{6p^3} + \frac{1}{6q^3}\right) \\ &\quad + E\left[(\tilde{p}-p)^3\right]\left(\frac{1}{6p^2} - \frac{1}{6q^2}\right) \\ &= \left(\frac{3n^2p^2q^2+npq(1-6pq)}{(n+1)^4}\right. \\ &\quad \left. + \frac{7npq(q-p)^2}{2(n+1)^4} + \frac{(q-p)^4}{16(n+1)^4}\right)\frac{1-3pq}{6p^3q^3} \\ &\quad + \frac{5npq(1-4pq)}{12p^2q^2(n+1)^3} + \frac{(1-4pq)^2}{48p^2q^2(n+1)^3} \end{aligned}$$

Finally, let $x = \frac{1}{pq} \geq 4$ and we have

$$\begin{aligned} f(x) &\triangleq 96(n+1)^4\left(nR(p,\tilde{p}) - \frac{1}{2}\right) \\ &= nx^3 + (68n^2 - 15n)x^2 - (20n^3 + 560n^2 - 84n)x + 32n^3 + 880n^2 - 352n - 48. \end{aligned}$$

Since

$$f(0) = -48 - 352n + 880n^2 + 32n^3 > 0$$

and

$$f(4) = -48 - 192n - 48n^3 - 272n^2 < 0$$

$f(x)$ has a root for $x > 4$. It is enough to show that the largest root of $f'(x)$ tends to infinity as $n \to \infty$. This follows by noting that

$$f'(x) = 3nx^2 + 2(68n^2 - 15n)x - (20n^3 + 560n^2 - 84n)$$

whose largest root

$$-\frac{68}{3}n + 5 + \frac{1}{3}\sqrt{4684n^2 - 360n - 27}$$

tends to infinity as $n \to \infty$. □

4 Discussion and comments

We have shown how the MME dominates the MLE in terms of Kullback–Leibler risk. This estimator, and modal estimators more generally, have other important properties as well. Like the MLE, the modal estimates are defined on the space of distributions and so are parameter-invariant. In particular, the modal estimate will be the same whether the success probability or log odds are used to parametrize the binomial family of distributions.

Both the MLE and the modal estimates allow for post-data interpretations. The MLE is the distribution that assigns the largest probability to the observed value among all distributions in the family. Each modal estimate satisfies a weaker criterion: a modal estimate is a distribution that assigns the largest probability to the observed value among all values in the sample. Fisher consistency is a post-data consideration that applies to both the MLE and modal set estimators. For the binomial, the MLE is Fisher consistent because when the sample mean (i.e. $\hat{p}$) equals the true mean, the MLE is the true distribution. The modal set estimator is Fisher consistent in that it contains the MLE. We make the following comments:

(1) Fisher consistency does not apply unless the success probability is of the form x/n for some integer x. A weaker form of consistency is to require that when the observation is the most likely value under the true distribution, i.e., the mode of the true distribution, the observation is also the mode of the estimate. We call this weak Fisher consistency for the mode. By definition, all modal estimates are weakly Fisher consistent for the mode. The plus-four estimate ($(x+2)/(n+4)$) is a popular estimate that is not weakly Fisher consistent for the mode. The standard asymptotic confidence interval based on this estimate was originally suggested by Wilson in 1927 and has started to appear in popular introductory statistics text such as Moore et al. (2009).

The plus-four estimate can be a distribution that assigns much lower probability to the observed data than to other values not observed. For example, when $x_{\text{obs}} = 0$ out of $n = 10$ trials is observed, the plus-four estimate assigns greater probability to unrealized values $x = 1$ and $x = 2$. In fact, the plus-four estimate assigns greater probability to unobserved values when $x_{\text{obs}} \in \{0, 1, 2, 8, 9, 10\}$.

(2) In addition to consistency for the mean and mode, consistency can also be defined for the median. In particular, an estimator is weakly Fisher consistent for the median if the observed value is a median for the true distribution then the observed value is also a median for the estimate. Formally, the median set estimator $\widehat{[p]}_x$ is defined to be the collection of values p such that the corresponding distribution has the observed value as its median. This is an interval whose endpoints can be expressed in terms of the beta distribution:

$$\widehat{[p]}_x = \left[\dot{\beta}^x_{n+1}, \dot{\beta}^{x+1}_{n+1}\right], \tag{7}$$

where $\dot{\beta}^x_{n+1}$ is the median of the beta distribution $Beta(x, n+1-x)$. Equation (7) follows from the relationship between the binomial random variable $(B : n, p)$ and the beta random variable $(\beta : x, n+1-x)$: $\Pr[(B : n, p) \geq x] = \Pr[(\beta : x, n+1-x) \leq p]$. By definition, estimators that take values in $\widehat{[p]}_x$ for each x are weakly Fisher consistent for the median. It can be shown that the MME is Fisher consistent for the median (Wu and Vos 2009).

(3) One motivation for the set estimators is the behavior of the MLE across differing sample sizes. For a sample of size n the MLE takes values in

$$\left\{\frac{0}{n}, \frac{1}{n}, \ldots \frac{n-1}{n}, \frac{n}{n}\right\}.$$

The MLE for this sample size cannot be the MLE for a sample increased by one observation except for the case where all observations are successes or all failures. The set estimator approach is to identify all estimates (distributions) that are on an equal footing for a given value in the sample space. The point estimate, for example the MME, is obtained as a summary of the distributions in the set.

(4) The modal set estimates for $x = 0$ and $x = n$ specify the values for p that specify claims too weak to be addressed by a sample of size n. For example, the data cannot address claims of the form $p \geq p_0$ when p_0 is very small because small observations would be evidence against this claim and yet $x = 0$ is the most likely observation for small success probabilities. In particular, for a sample of size $n = 9$, the data cannot address claims of the form $p \geq 0.10$ because the smallest possible observation is the mode for success probabilities of 0.10 or less. If the claim is strengthened to $p \geq 0.11$, observing 0 out of 9 successes provides some evidence against this claim since zero is no longer the most likely observation. If claims such as $p \geq 0.10$ are of interest, a sample size of $n = 9$ is not large enough, but the claim $p \geq 0.11$ differs from $p \geq 0.10$ in that, for the latter, the extreme tail and mode are the same sample space value. Similar comments hold for success probabilities near 1 so that union of modal set estimators for sample values from $x = 1$ to $x = n - 1$, or $(\frac{1}{n+1}, \frac{n}{n+1})$, represent the *addressable claims for p for a sample of size n*. Theorem 1 claims only that the

Table 1 The first two columns list sample size n and lower endpoint of addressable values for p. The last two columns list MLE, p_{lo} the lowest value for which the Kullback–Leibler risk for the MLE is $>\frac{1}{2}$ and MME, p_{lo} the lowest value for which the Kullback–Leibler risk for the MME is $<\frac{1}{2}$. Upper endpoint of addressable values and the corresponding values for MLE, p_{hi} and MME, p_{hi} follow by symmetry

n	$0.5/(n+1)$	MLE, p_{lo}	MME, p_{lo}
1	0.2500	0.1997	0.0173
2	0.1667	0.1554	0.0219
3	0.1250	0.1226	0.0213
4	0.1000	0.0991	0.0197
5	0.0833	0.0828	0.0181
10	0.0455	0.0450	0.0124
15	0.0312	0.0308	0.0093
20	0.0238	0.0234	0.0074
25	0.0192	0.0189	0.0062
30	0.0161	0.0158	0.0053
35	0.0139	0.0136	0.0046
40	0.0122	0.0120	0.0041
45	0.0109	0.0107	0.0037
50	0.0098	0.0096	0.0034
60	0.0082	0.0080	0.0028
70	0.0070	0.0069	0.0025
80	0.0062	0.0060	0.0022
90	0.0055	0.0054	0.0019
100	0.0050	0.0048	0.0017
200	0.0025	0.0024	0.0009
300	0.0017	0.0016	0.0006
400	0.0012	0.0012	0.0005
500	0.0010	0.0010	0.0004
1000	0.0005	0.0005	0.0002
2000	0.0002	0.0002	0.0001

range of values for which the risk for the MLE exceeds $\frac{1}{2}$ approaches the unit interval. In fact, inspection of the risk shows that it exceeds $\frac{1}{2}$ for all success probabilities on an interval wider than the addressable range of values: $(.5/(n+1), (n+.5)/(n+1))$. Selected values appear in Table 1.

(5) There are other reasonable summaries measures for the modal set estimator besides the midpoint estimator we consider here. One approach to approximate the distributions in $[p](x)$ is to use the distribution that minimizes the maximum possible Kullback–Leibler error that can be incurred. The minimax approximation $\tilde{p}(x)$ to the modal estimate $[p](x)$ is defined by

$$\tilde{p}(x) = \arg \min_{p' \in [0,1]} \max_{p \in [p](x)} D_n(p, p').$$

Since $D_n(p+\varepsilon, p)$ and $D_n(p-\varepsilon, p)$ are increasing function of ε we see that $\tilde{p}$ is the point in $[p](x)$ such that

$$D_n\left(\frac{x}{n+1}, \tilde{p}\right) = D_n\left(\frac{x+1}{n+1}, \tilde{p}\right). \tag{8}$$

Using Eq. (2) and solving (8) for $\tilde{p}$ gives

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = (n+1)H\left(\frac{x}{n+1}\right) - (n+1)H\left(\frac{x+1}{n+1}\right) \tag{9}$$

which shows that the minimax Kullback–Leibler approximation to the modal set estimate is the binomial distribution in $\mathcal{B}_n$ whose log odds is the difference in the entropies for the distributions $B(n+1, \frac{x}{n+1})$ and $B(n+1, \frac{x+1}{n+1})$. The numerical values for this estimator and the MME are not too different.

(6) The post-data approach to inference is due to Fisher but has been described by others as well. See, for example, Kempthorne and Folks (1971). For data from continuous distributions the difference between the post-data approach and traditional frequentist long run inference appears to be largely a matter of interpretation with no difference in methodology. For discrete distributions, however, methodology designed for optimal, or even improved, long run properties can lead to difficulties from a post-data perspective. Problems that can arise for the binomial have been considered by Vos and Hudson (2005) and Vos and Hudson (2008).

(7) Modal estimates can be extended beyond the one sample binomial setting. Estimates for the log odds in the 2 × 2 table include the sample odds ratio and a modified version suggested by Agresti (2002)

$$\tilde{\eta} = \frac{\left(n_{11}+\frac{1}{2}\right)\left(n_{22}+\frac{1}{2}\right)}{\left(n_{12}+\frac{1}{2}\right)\left(n_{21}+\frac{1}{2}\right)}$$

where n_{ij} is the observed number of occurrences in the ij^{th} cell. It is easily shown that the modal set estimates provide the following partition of the odds ratio parameter

$$\left\{\frac{jn_{22}}{n_{12}n_{21}} : \max(0,\ n_{1\cdot} - n_{2\cdot}) \le j \le \min(n_{1\cdot},\ n_{\cdot 1})\right\}$$

where $n_{12},\ n_{21},$ and n_{22} are functions of n_{11} defined by fixing the margin totals $n_{1\cdot} = n_{11} + n_{12}$, $n_{2\cdot} = n_{21} + n_{22}$, $n_{\cdot 1} = n_{11} + n_{21}$, and $n_{\cdot 2} = n_{12} + n_{22}$. The sample odds ratio and modified version $\tilde{\eta}$ are modal estimators. Extension to logistic regression, however, is more complicated especially when the covariates are continuous. For discrete covariates where there are multiple replications on each cell, Agresti (2002, page 168) recommends plotting

$$\hat{\eta}_i = \log\frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}$$

and describes these as the least biased estimator of this form for the true logit. Writing this logit as a proportion gives the modal estimate considered in this paper.

(8) The modal estimate can be viewed as a shrinkage estimator because the sample proportion is moved closer to the value $\frac{1}{2}$. Copas (1997) explores the relationship between shrinkage and regression to the mean for both linear and logistic regression.

Whether there is a relationship between regression to the mean and the shrinkage of the modal estimate is not clear but would be interesting for further study.

Acknowledgments We are grateful to a referee for pointing out the relationship between shrinkage and regression to the mean as well as suggesting application of modal estimates to the case of two binomial samples.

References

Agresti, A. (2002). *Categorical data analysis*. New York: Wiley.

Copas, J. (1997). Using regression models for prediction: Shrinkage and regression to the mean. *Statistical Methods in Medical Research, 6*, 167–183.

Eguchi, S., Yanagimoto, T. (2008). Asymptotic improvement of maximum likelihood estimators on Kullback–Leibler loss. *Journal of Statistical Planning and Inference, 138*, 3502–3511.

Fisher, R. (1973). *Statistical methods and scientific inference* (3rd ed.). New York: Hafner Press.

Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators. *Neural Computation, 10*, 1425–1433. doi:10.1162/089976698300017232

Kempthorne, O., Folks, L. (1971). *Probability, statistics, and data analysis*. Ames, IA: The Iowa State University Press.

Moore, D., McCabe, G., Craig, B. (2009). *Introduction to the practice of statistics*. New York: Freeman.

Vos, P., Hudson, S. (2005). Evaluation criteria for discrete confidence intervals: Beyond coverage and length. *The American Statistician, 59*, 137–142.

Vos, P., Hudson, S. (2008). Problems with binomial two-sided tests and their associated intervals. *Australian & New Zealand Journal of Statistics, 50*, 81–89.

Wu, Q., Vos, P. (2009). Improving boundaries of binomial 50 percentage points with applications to the midpoint modal estimator, East Carolina University. *Biostatistics Technical Report #09.2*.