Adaptive Bayesian Analysis for Binomial Proportions

Sonali Das[1] and Sourish Das[2]

October 2008

Key words and phrases: Adaptive Bayes, Bayesian central limit theorem, Monte Carlo method, statistical power, proportion, ϕ-divergence.

Abstract

We consider the problem of statistical inference of binomial proportions for non-matched, correlated samples, under the Bayesian framework. Such inference can arise when the same group is observed a different number of times on two or more inference occasions, with the aim of testing the proportion of some trait. These scenarios can occur when we are interested to infer the proportion of *extreme* wave height per year, at a certain measuring station, where measurements are made every hour. Gaps in measurements, either due to a malfunction of the measuring instrument or another reason, can result in an unequal number of observations in different years. For such scenarios, we develop an adaptive Bayesian method, and suggest a heuristic decision procedure to conduct statistical inference. We use the ϕ-divergence measure to quantify the perturbation of the posterior distribution of the proportion in different time points. We present a simulation study for frequentist power investigation for both the adaptive Bayesian method, as well as the regular frequentist method, using the Monte Carlo technique. Based on the simulation

[1]Senior Researcher, Logistics and Quantitative Methods, CSIR BE, Pretoria, South Africa. Contact: sdas@csir.co.za

[2]Postdoctoral Fellow, SAMSI, Duke University, Durham, NC, USA.

The authors wish to acknowledge the TNE at the University of Connecticut, USA and the CSIR, South Africa for the data sets used as case studies. The research presented in this paper was supported by CSIR, Pretoria. We also take this opportunity to express our gratitude to Prof. Dipak Dey, University of Connecticut, for his valuable comments.

study of frequentist power, as well as theoretical proof, under certain design, the adaptive Bayesian method is shown to outperform the regular frequentist method. We administer the developed adaptive Bayesian method to two case studies when the total number of observation instances of the same group are unequal, at different time points of interest.

1 Introduction

Inference on the probability of 'success' of the binomial distribution is one of the most important problems in statistical literature. Popular, everyday applications involve inferring the equality of proportions of success from two samples. Such samples can be either independent groups, or the samples can be from matched pairs of the same group over the two inference points of interest. However, a third scenario can arise when for the same group, the number of measurement instances at one time point may not be equal to the number of measurement instances of the same group at a subsequent time point. In such a scenario, the samples are not independent, but neither do they satisfy the matched pair criteria. As such, usual inference procedures for the proportion cannot be conducted under the independent sample assumptions, nor under the matched pair assumption. The authors have come across a number of such real case studies, and this work arises from a need to address this inference problem.

We briefly recapitulate some popular inference methods for binomial proportions. In the simplest case of two samples, two test book scenarios for inferring the hypothesis of equality of binomial proportions occurs when the samples are independent, or when the samples are matched. When the samples are independent, in the frequentist framework, the Cochran-Mantel-Hansel type χ^2 tests for equality of proportion are commonly used (Cochran, 1954; Mantel and Haenszel, 1959; Agresti and Caffo, 2000). For matched pairs,

the McNemar (1947) χ^2 test, or the exact test by Liddell (1983) are available. To test for the difference of proportions for matched binary pairs, Lloyd (1990) proposed two methods to construct confidence intervals. In the Bayesian framework, Chen and Dey (1998) developed a model for correlated binary responses using a scale mixture of multivariate normal link functions, and implemented the model using the Markov Chain Monte Carlo (MCMC) method. Ghosh *et al* (2000) developed a Bayesian hierarchical technique for binary matched pair data, while Ghosh and Chen (2002) developed a general technique for inferring in a matched case-control setup. Berger (1985:457) provides a decision theoretic Bayesian stopping rule under a sequential analysis framework for binomial proportion.

In this paper, we develop, in stages, an adaptive Bayesian method to infer for binomial proportions, when neither assumptions of independence, nor matched-pair holds. The paper is organised as follows: In Section 2, we develop an adaptive Bayesian method for non-matched correlated binomial proportions under various non-informative conjugate Beta priors, as well as the Zellner's non-informative prior. In Section 3, we develop the ϕ-divergence measure for the adaptive Bayesian method for proportions. In Section 4, we propose a heuristic decision procedure based on the adaptive Bayesian method. In Section 5 we present two real case studies for inferring on non-matched correlated proportions using the developed adaptive Bayesian method and finally in Section 6 we conclude the paper with a short discussion.

2 Adaptive Bayesian method for binomial proportions

We first present two motivating case studies to clarify what we mean by adaptive Bayesian method for the binomial proportions. The first case study involves the inference for annual

proportion of *high* hourly wave heights measured at a station off the coast of Cape Town in South Africa between 1997-2003. The total number of observations were different in each year. The second case study involves the inference for the proportion of a particular group of teachers rated into two categories - 'proficient' and 'non-proficient'. Each teacher was observed and rated at two instances (before and after an intervention), and at each instance, was categorised as either 'proficient' or 'non-proficient', for different numbers of times. In both the cases, the sample sizes were not the same for different time points. Clearly, regular matched pair analysis of proportions is not applicable here. Also, regular analysis of binomial proportions for more than one sample cannot be applicable as it does not take care the time order dependence embedded in the data. With these case studies as background, we proceed to explain the new adaptive Bayesian method for binomial proportions from an information theory point of view.

2.1 Adaptive Bayesian method in the context of Information Processing

Suppose a process has been observed at k different time points, say $T_1, T_2, ..., T_k$. Suppose we observed the process n_i times at time point T_i, of which $s_i(\leq n_i)$ are categorised as success for a binary outcomes (where $i = 1, 2, ..., k$). If we denote S_i as the random number of successes in the first time point T_i, then we assume $S_i \sim Binomial(n_i, p_{_{T_i}}),\ i = 1, ..., k$. Hence the natural conjugate prior is $Beta(a_t, b_t)$.

Zellner's Information Processing Rule (ZIPR) (Zellner 1988, 2002) for statistical inference is based on the input and output information measure. A measure is considered 100% efficient if the output information is exactly equal to the input information. In our

set up, the information processing between two time points is

$$\begin{aligned}\Delta\big[\tilde{\pi}^t(p \mid s_t)\big] &= \text{Output information} - \text{Input information} \\ &= \int \tilde{\pi}^t(p) log \tilde{\pi}^t(p) dp \; + \; \int \tilde{\pi}^t(p) log(h(s_t)) dp \\ &\quad - \; \int \tilde{\pi}^t(p) log f(s_t \mid p) \; - \; \int \tilde{\pi}^t(p) log \tilde{\pi}^{t-1}(p \mid s_{t-1}) dp,\end{aligned}$$

where

$$\begin{aligned}h(s_t) &= \int f(s_t \mid p) \pi^{t-1}(p \mid s_t) dp \\ &= \frac{\begin{pmatrix} n_t \\ s_t \end{pmatrix}}{Beta(a_t, b_t)} \int p^{s_t} (1-p)^{n_t - s_t} p^{a_t} (1-p)^{b_t - 1} dp \\ &= \begin{pmatrix} n_t \\ s_t \end{pmatrix} \frac{Beta(s_t + a_t, n_t + b_t - s_t)}{Beta(a_t, b_t)},\end{aligned}$$

with $\tilde{\pi}^t(p)$ being proper probability density at stage/time t, such that $\int \tilde{\pi}^t(p) dp = 1$.

Observe that $h(s_t)$ is free from the parameter p, and is the likelihood at stage/time t, while $\pi^{t-1}(p)$ is the prior distribution at stage/time t.

$$\begin{aligned}\Delta\big[\tilde{\pi}^t(p \mid s_t)\big] &= \int \tilde{\pi}^t(p) log\Big(\frac{\tilde{\pi}^t(p) h(s_t)}{f(s_t \mid p) \pi^{t-1}(p)}\Big) dp \\ &= \int \tilde{\pi}^t(p) log\Big(\frac{\tilde{\pi}^t(p)}{\pi^t(p)}\Big) dp\end{aligned}$$

where $\pi^t(p) = \frac{f(s_t|p)\pi^{t-1}(p)}{h(s_t)}$ is posterior distribution of p at stage/time t where we use $\pi^t(p) = \pi^{t-1}(p \mid s_t)$ as the prior. Hence, in plain language, we are choosing the prior adaptively based on the posterior we observe in the previous stage. Now, by Zellner's criterion, a rule is 100% efficient if $\Delta\big[\tilde{\pi}^t(p \mid s_t)\big] = 0$. Clearly, $\Delta\big[\tilde{\pi}^t(p \mid s_t)\big] = 0$ if

$\tilde{\pi}^t(p) = \pi^t(p) = \frac{f(s_t|p)\pi^{t-1}(p)}{h(s_t)}$. We thus summarise the above in the following theorem:

Theorem 2.1. *For binomial proportions, $\tilde{\pi}^t(p)$ yields 100% efficient Zellner's Information Processing Rule (ZIPR) if one uses the posterior of previous stage/time point as the prior for the next stage/time point.*

In the next section we discuss the relation between adaptive Bayesian method and Markov transition model.

2.2 Relationship between the Adaptive Bayesian method and Markov transition model

Suppose $\pi^{t-1}(p \mid s_1) \sim Beta(a_{t-1} + s_{t-1}, n_{t-1} + b_{t-1} - s_{t-1})$ and is the posterior at stage $(t-1)$. Motivated from the previous section, we choose $\pi^{t-1}(p \mid s_1)$ as prior for stage t, *i.e.*, $\pi^t(p) = \pi^{t-1}(p \mid s_1)$, or simply $\pi^t(p) \sim Beta(a_t, b_t)$, where $a_t = a_{t-1} + s_{t-1}$ and $b_t = n_{t-1} + b_{t-1} - s_{t-1}$. Therefore, with $S_t \sim Binomial(n_t, p)$, the posterior of p at time point t is

$$\pi^t(p \mid s_t) \sim Beta(a_t + s_t, n_t + b_t - s_t).$$

Observe that

$$\begin{aligned}
E^{\pi^t}(p|s_t) &= \frac{a_t + s_t}{n_t + b_t + a_t} \\
&= \frac{s_{t-1} + s_t + a_{t-1}}{n_{t-1} + n_t + a_{t-1} + b_{t-1}} \\
&= \frac{s_t}{n_{t-1} + n_t + a_{t-1} + b_{t-1}} + \frac{s_{t-1} + a_{t-1}}{n_{t-1} + n_t + a_{t-1} + b_{t-1}} \\
&= \frac{n_t}{n_{t-1} + n_t + a_{t-1} + b_{t-1}} \times \frac{s_t}{n_t} + \frac{n_{t-1} + a_{t-1} + b_{t-1}}{n_{t-1} + n_t + a_{t-1} + b_{t-1}} \times \frac{s_{t-1} + a_{t-1}}{n_{t-1} + a_{t-1} + b_{t-1}} \\
&= \omega_t \frac{s_t}{n_t} + (1 - \omega_t) E^{\pi^t}(p \mid s_{t-1}), \qquad (2.1)
\end{aligned}$$

where $\omega_t = \frac{n_t}{n_{t-1}+n_t+a_{t-1}+b_{t-1}}$. Also, observe that (2.1) has an intuitive form of a weighted average of observed proportion and expected proportion from the previous stage. Thus, (2.1) can be re-written as

$$E^{\pi^t}(p \mid s_t) = \alpha_t + \beta_t E^{\pi^{t-1}}(p \mid s_{t-1}),$$

where $\alpha_t = \omega_t \frac{s_t}{n_t}$ and $\beta_t = (1 - \omega_t)$, for $t = 1, 2, ..., k$, which is a Markov transition model. Therefore, the adaptive Bayesian method developed in this paper provides solid justification for the one-step Markov transition process.

2.3 Prior at the initial stage

Now we discuss, the different choices of prior that we can use at the first stage. One must note that if we use $a_1 = b_1 = 0.5$, then we have the Jeffreys's (1961) non-informative prior, while if we use $a_1 = b_1 = 0$, then we have the Novick-Hall's (1965) non-informative prior for binomial proportions.

The Zellner's (1977) non-informative prior in the first time point $\pi^1(p)$ is proportional to $p^p \ (1-p)^{1-p}$. Note that for Zellner's prior, we do not need any hyper-parameters. The posterior for the second time point is then

$$\begin{aligned}\pi^1(p|s_1) &\propto \binom{n_1}{s_1} p^{s_1} \ (1-p)^{n_1-s_1} \times p^p \ (1-p)^{1-p} \\ &\propto p^{s_1+p} \ (1-p)^{n_1+1-(s-1+p)}.\end{aligned}$$

Following the same adaptive Bayesian idea from Section (2.1), we consider the posterior of the first time point as the prior for the second time point. That is, $\pi^2 = \pi^1(p|s_1)$, or equivalently, $\pi^2(p) \propto p^{s_1+p} \ (1-p)^{n_1+1-(s-1+p)}$. Then the posterior for the second time

point is

$$\pi^2(p|s_2) \propto p^{s_1+s_2+p}\ (1-p)^{n_1+n2+1-(s-1+s_2+p)}. \tag{2.2}$$

Because of the complicated form of the posterior distribution in (2.2) resulting from Zellner's prior, we generate 10,000 samples from the posterior distribution using the Metropolis-Hastings sampling method (Chen *et al*, 2000), and implement the adaptive Bayesian method as we have discussed in this paper.

In the next section, we discuss different measures for quantifying the amount of information that propagates from one stage to another stage through using this adaptive Bayesian method.

3 Measuring Distance Between the Posterior of Two Time Points Using the ϕ-divergence Measure

According to the Bayesian paradigm, the posterior distribution contains all the relevant information about the parameters needed for analyses. Therefore, any perturbation in the parameters between the two time points l lags apart, should depend on the discrepancy between the posteriors in the two time points with the lag l. We use the ϕ-divergence measure (Ali and Silvey, 1966; Csiszar, 1967) to capture the discrepancy between the posteriors $\pi^t(p|s_t)$ and $\pi^{t+l}(p|s_{t+l})$ at the two time points t and $t+l$. We define the ϕ-divergence in our setup as

$$D_\phi^{(t,l)} = D\Big(\pi^t(p|s_t),\ \pi^{t+l}(p|s_{t+l})\Big) = \int \phi\left(\frac{\pi^{t+l}(p|s_{t+l})}{\pi^t(p|s_t)}\right) \times \pi^t(p|s_t)\ dp,$$

where ϕ is a convex function with $\phi(1) = 0$. Several choices of ϕ are given by Dey and Birmiwal (1994). Using the Monte Carlo technique, we can easily implement the Kullback-Leibler divergence measure (Kullback and Leibler, 1951), the χ^2-divergence measure and the Hellinger distance.

In our adaptive Bayesian method, the posterior distribution, under conjugate Beta prior, in the two time points is given by

$$\pi^t(p|s_t) \sim Beta\Big(\sum_{i=1}^{t} s_i + a_1,\ \sum_{i=1}^{t} n_i + b_1 - \sum_{i=1}^{t} s_i\Big),$$

and

$$\pi^{t+l}(p|s_{t+l}) \sim Beta\Big(\sum_{i=1}^{t+l} s_i + a_1,\ \sum_{i=1}^{t+l} n_i + b_1 - \sum_{i=1}^{t+l} s_i\Big).$$

We define the Kullback-Leibler divergence measure, under the conjugate prior, for our case as

$$\begin{aligned} D_{KL}^{(t,l)} &= E\bigg[- log\ \frac{\pi^{t+l}(p|s_{t+l})}{\pi^t(p|s_t)}\bigg] \\ &= E\Big[- log\ K^{(t,l)} p^{\sum_{i=t+1}^{t+l} s_i}\ (1-p)^{\sum_{i=t+1}^{t+l} n_i - \sum_{i=t+1}^{t+l} s_i}\Big], \end{aligned} \tag{3.3}$$

where

$$K^{(t,l)} = \frac{Beta\Big(\sum_{i=1}^{t} s_i + a_1,\ \sum_{i=1}^{t} n_i + b_1 - \sum_{i=1}^{t} s_i\Big)}{Beta\Big(\sum_{i=1}^{t+l} s_i + a_1,\ \sum_{i=1}^{t+l} n_i + b_1 - \sum_{i=1}^{t+l} s_i\Big)}.$$

The χ^2-divergence, under conjugate prior in our case is

$$\begin{aligned} D_{\chi^2}^{(t,l)} &= E\left[\left(\frac{\pi^{t+l}(p|s_{t+l})}{\pi^t(p|s_t)} - 1\right)^2\right] \\ &= E\Big[\Big(K^{(t,l)}\ p^{\sum_{i=t+1}^{t+l} s_i}\ (1-p)^{\sum_{i=t+1}^{t+l} n_i - \sum_{i=t+1}^{t+l} s_i} - 1\Big)^2\Big]. \end{aligned} \tag{3.4}$$

The Hellinger distance, under conjugate prior in our case is

$$
\begin{aligned}
D_H^{(t,l)} &= 2 \times \left[1 - E\left(\sqrt{\frac{\pi^{t+l}(p|s_{t+l})}{\pi^t(p|s_t)}} \right) \right] \\
&= 2 \times \left[1 - E\left(\sqrt{K^{(t,l)}\ p^{\sum_{i=t+1}^{t+l} s_i}\ (1-p)^{\sum_{i=t+1}^{t+l} n_i - \sum_{i=t+1}^{t+l} s_i}} \right) \right]. \qquad (3.5)
\end{aligned}
$$

The forms of $D_{KL}^{(t,l)}$, $D_{\chi^2}^{(t,l)}$ and $D_H^{(t,l)}$ under conjugate prior as given in (3.3), (3.4) and (3.5) are analytically intractable, and thus we calculate these distance measures using the Monte Carlo technique. Eguchi and Copas (2006) showed that the standard non-negativity property of the Kullback-Leibler divergence is essentially a re-statement of the optimal property of the likelihood ratio as established by the Neyman-Pearson lemma. Thus, for all our analyses, we shall focus on the Kullback-Leibler divergence measure since it has a natural interpretation with the Neyman-Pearson lemma.

Observe that for $t = 1$ and $l = 1$, we have the Kullback-Leibler divergence measure between time point 1 and time point 2 as

$$
\begin{aligned}
D_{KL}^{(1,1)} &= E\left[-log\ \frac{\pi^2(p|s_2)}{\pi^1(p|s_1)} \right] \\
&= E\Big[-log\ K^{(1,1)} p^{s_2}\ (1-p)^{n_2 - s_2} \Big],
\end{aligned}
$$

where

$$
K^{(1,1)} = \frac{Beta\Big(s_1 + a_1,\ n_1 + b_1 - s_1\Big)}{Beta\Big(s_1 + s_2 + a_1,\ n_1 + n_2 + b_1 - (s_1 + s_2)\Big)}.
$$

Similarly, for $t = 1$ and $l = 2$, we have the Kullback-Leibler divergence measure between

time point 1 and time point 3 as

$$D_{KL}^{(1,2)} = E\Bigg[- log \ \frac{\pi^3(p|s_3)}{\pi^1(p|s_1)} \Bigg]$$
$$= E\Big[- log \ K^{(1,2)} p^{s_2+s_3} \ (1-p)^{(n_2+n_3)-(s_2+s_3)} \Big],$$

and for $t = 2$ and $l = 1$, we have the Kullback-Leibler divergence measure between time point 2 and time point 3 as

$$D_{KL}^{(2,1)} = E\Bigg[- log \ \frac{\pi^3(p|s_3)}{\pi^2(p|s_2)} \Bigg]$$
$$= E\Big[- log \ K^{(2,1)} p^{s_3} \ (1-p)^{n_3-s_2} \Big],$$

where $K^{(1,2)}$ and $K^{(2,1)}$ are defined accordingly.

4 Statistical Power Investigation of the Adaptive Bayesian Method

In this section, we investigate some properties of the adaptive Bayesian method, including large sample properties via the Bayesian Central Limit Theorem. We suggest a heuristic procedure to conclude whether the binomial proportion in the second time point is significantly different from the previous time point. Thus, it is important to verify the long-run behaviour of the Bayesian method, which leads us to investigate the frequentist power of the heuristic procedure based on the adaptive Bayesian method, using the Monte Carlo technique.

4.1 Inferences based on the Adaptive Bayesian Method

Bayesian inference is determined via a loss function and the prior information. Therefore, we start with the loss function as given by Geisser (2006)

$$L(p,\ C) = \begin{cases} \alpha\ \mathcal{L}(C) - A, & \text{if } p \in C, \\ \alpha\ \mathcal{L}(C), & \text{otherwise,} \end{cases}$$

where $A > 0$, and $\mathcal{L}(C)$ is the Lebesgue measure of C. Under this loss function, it can be shown that the Bayes estimator (i.e., the minimum posterior expected loss) is $100(1-\alpha)\%$ highest posterior density (HPD) interval for p, which is given by

$$C = \Big\{p : \pi\{p \mid s\} \geq \pi_\alpha\Big\},$$

where π_α is the largest constant such that $P(p \in C) \geq (1-\alpha)$. The length of C will then be an indicator of the statistical power of the method. Other than loss, statistical power in the Bayesian paradigm will also depend on the choice of the prior. Following the objective Bayesian philosophy as discussed by Berger (2006), we shall use Jeffreys non-informative prior for our case. In order to compute the HPD interval, we use the Bayesian version of the Central Limit Theorem (BCLT) of Berger (1985:224) which states:

"Suppose that $X_1, X_2, ..., X_n$ are *i.i.d.* from the density $f_0(x_i|\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^t$ being an unknown vector of parameters. (We will write $\mathbf{x} = (x_1, ..., x_n)^t$ and $f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^n f_0(x_i|\boldsymbol{\theta})$, as usual.) Suppose $\pi(\boldsymbol{\theta})$ is a prior density, and that $\pi(\boldsymbol{\theta})$ and $f(\mathbf{x}|\boldsymbol{\theta})$ are positive and twice differentiable near $\hat{\boldsymbol{\theta}}$, the (assumed to exist) maximum likelihood estimate of $\boldsymbol{\theta}$. Then, for large n and under commonly satisfied assumptions, the posterior density

$$\pi_n(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})}{m(\mathbf{x})},$$

can be approximated" in the following way: "π_n is approximately $\mathcal{N}_p(\boldsymbol{\mu}^\pi(\mathbf{x}), \mathbf{V}^\pi(\mathbf{x}))$, where $\boldsymbol{\mu}^\pi$ and $\mathbf{V}^\pi$ are the posterior mean and covariance matrix."

For the adaptive Bayesian method as described in Section(2.1),

$$\pi^1(p \mid s_1) \xrightarrow{\mathcal{L}} \mathcal{N}(m_1,\ s_1^2)$$

with $m_1 = \frac{a_2}{a_2+b_2}$ and $s_1^2 = \frac{a_2\ b_2}{(a_2+b_2)^2(a_2+b_2+1)}$, where $a_2 = a_1 + s_1$ and $b_2 = n_1 + b_1 - s_1$. Similarly, the posterior of the second time point is

$$\pi^2(p \mid s_2) \xrightarrow{\mathcal{L}} \mathcal{N}(m_2,\ s_2^2)$$

with $m_2 = \frac{a_3}{a_3+b_3}$ and $s_2^2 = \frac{a_3\ b_3}{(a_3+b_3)^2(a_3+b_3+1)}$, where $a_3 = a_2 + s_2$ and $b_3 = n_2 + b_2 - s_2$. If there is a true shift in p, then it would be expected that the $100(1-\alpha)\%$ HPD interval from the second time point would not contain the posterior mean of p from the first time point. Hence we are motivated to define a heuristic decision procedure as follows: $d = \{\textit{Shift in } p\}$ and $d^c = \{\textit{No shift in } p\}$. We reject d if

$$E^{\pi^1}\big(p \mid s_1\big) \in C_2,$$

where C_2 is the $100(1-\alpha)\%$ HPD interval at the second time point, and accept d otherwise. We can easily extend this heuristic procedure for a general setup with $k(\geq 2)$ time points.

4.2 Frequentist Power Study of the Adaptive Bayesian Method using the Monte Carlo Technique

We develop a simple Monte Carlo algorithm to compute the frequentist power of the heuristic decision procedure based on the adaptive Bayesian method as developed in the

previous sub-section. The algorithm is as follows:

Step 0: Initialise $count = 0$, and simulation size N.

Step 1: For given $(n_1,\ p_1)$ we generate a sample s_1 from $Binomial(n_1, p_1)$, where n_1 is the sample size at first time point, and p_1 is the true proportion of success.

Step 2: Given s_1 and (a, b), we compute the posterior mean at the first time point as

$$E^{\pi^1}\big(p|s_1\big) = \frac{a + s_1}{n_1 + a + b},$$

where a and b are the hyper-parameters from the conjugate prior of the first time point.

Step 3: For given $(n_2,\ p_2)$, we generate a sample s_2 from $Binomial(n_2, p_2)$, where n_2 is the sample size at the second time point, and p_2 is the true proportion of success.

Step 4: We calculate HPD interval C_2 for p for the second time point using the posterior of the second time point.

Step 5: If $E^{\pi^1}\big(p|s_1\big) \in C_2$ then $count = count + 1$. Goto Step 1.

Step 6: Repeat Steps 1-5 N times.

Step 7: Statistical power of the adaptive Bayesian method is $1 - \frac{count}{N}$.

[Insert Figure 1 and 2 here.]

We implement the above algorithm using the R-software. In Figures 1 and 2, we present the power function for some specific choices of sample sizes $n_1,\ n_2$, and p_1, for the entire range of $[0, 1]$ for p_2. We use the BCLT to compute the HPD for the adaptive Bayesian method. For equal sample sizes $n_1 = n_2 = 15, 30, 50$ and $p_1 = 0.5, 0.25$, the performance of the two methods are similar. For equal sample size of $n_1 = n_2 = 15$ and $p_1 = 0.1, 0.9$, the performance of the regular frequentist method is better than the

adaptive Bayesian method. This is becuase we used the Bayesian CLT for computing HPD, which is not valid for small sample sizes like $n_1 = n_2 = 15$. In Figure 3, we present the exact power, calculated using the exact HPD interval (Chen *et al* 2002), and observe that the performance, for equal sample size cases, is similar in both the methods. When $n_2 > n_1$, the adaptive Bayesian method uniformly outperforms the frequentist method because the frequentist method, uses the $min(n_1, n_2)$, forcing equality of sample size, thereby resulting in loss of information. When $n_2 < n_1$, the adaptive Bayesian method performs worse than the frequentist method. Hence, based on these extensive studies of frequentist power under both the methods, we strongly recommend that one uses the adaptive Bayesian method when the sample size in the second time point is larger than the sample size in the first time point.

[Insert Figure 3 here.]

The theoretical justification of the above observation is provided via the following theorem:

Theorem 4.1. *Suppose* $S_t \sim Bin(n_t, p)$ *with prior on* $p = \pi^t \sim Beta(a_t, b_t)$*, and posterior* $\pi^t(p|s_t) \sim Beta(a_t + s_t, n_t + b_t - s_t)$*. Assume* $E(S_{t-1}) = \rho_t \; E(S_t)$*, where* ρ_t *is some known constant. If* $\rho_t n_t \geq n_{t-1}$*, then* $Var(p_t|S_{t-1}) \leq Var(p_{t-1}|S_{t-1})$*, while if* $\rho_t n_t \leq n_{t-1}$*, then* $Var(p_t|S_{t-1}) > Var(p_{t-1}|S_{t-1})$*.*

Proof. Assumption $E(S_{t-1}) = \rho_t \; E(S_t)$ implies

$$n_{t-1} p_{t-1} = \rho_t n_t p_t. \tag{4.6}$$

From (4.6), the posterior variance of $n_{t-1} p_{t-1}$ is given as

$$Var(p_{t-1}|S_{t-1}) = \rho_t^2 \frac{n_t^2}{n_{t-1}^2} \; Var(p_t|S_{t-1}). \tag{4.7}$$

Thus, when $\rho_t n_t \geq n_{t-1}$, then $Var(p_t|S_{t-1}) \leq Var(p_{t-1}|S_{t-1})$, while when $\rho_t n_t \leq n_{t-1}$, then $Var(p_t|S_{t-1}) > Var(p_{t-1}|S_{t-1})$. □

In following sections, we present two case studies, where we apply the proposed adaptive Bayesian method to test for non-matched correlated binomial proportions for T time points, where the sample size in each time point is different.

5 Case Studies

5.1 Case study I: Testing the proportion of threshold exceedances of wave-height off the coast of Cape Town

Rise in sea level is a cause for concern for coastal engineers and other stake-holders, as waves above a certain threshold can affect coastal infrastructure. In this case study, we focus on data collected on PetroSA's Mossgas platform, South Africa, which is archived on the Wave database of the CSIR, South Africa. The station is off the coast of Cape Town, and hourly 'significant wave height', henceforth referred to as HMO, in metres, is measured at a depth of 113 metres. The interest is in assessing whether there have been any significant changes in the yearly proportions of HMO that are in the top 5% of all recorded HMO, henceforth referred to as *extreme* HMO, between 1997 and 2003. A summary of the number of observations and proportions, per year, is given in Table 1. Observe, the total number of observations, though intended to be hourly, are in fact, different for each year. This could be due to a number of reasons, such as instrument malfunction.

[Insert Table 1 here.]

We proceed to answer the question of whether there has been any significant change in the proportions of *extreme* HMO, using the adaptive Bayesian method. In Table 2, we present the posterior mean and the corresponding 2.5% and 97.5% quantiles, and in Figure 4 we present the corresponding 95% credible intervals (CI).

[Insert Table 2 here.]

[Insert Figure 4 here.]

From Table 2 and Figure 4, we observe that there has been significant movements of the proportion of *extreme* HMO between 1997 and 2003. In fact, in 2000 and 2001, these proportions were significantly lower than proportions in 1997, while in the 2002 and 2003, these proportions were significantly higher than that in 2000. This, in spite of the fact that for latter years, the posterior CIs are tighter.

5.2 Case study II: Evaluating teaching performance

In this study, data arose from 2 time points, where the before and after intervention observation instances of the same group were unequal. Clearly, the assumption of independence does not hold as the same group is evaluated in the two time points, while it is not strictly a matched pair problem. The Teachers for a New Era (TNE) at the University of Connecticut conducted a study between 2005 and 2006 to evaluate the effectiveness of an intervention on the teaching quality in a certain grade in a participating school. The teachers' performances are a surrogate for assessing effectiveness of the intervention. The teachers were evaluated by 4 different raters into binary categories – 'proficient' and 'non-proficient', both before and after the intervention. Each teacher was rated at a number of instances both before and after the intervention. The raters rated the teachers blinded from each other at both time points. The total number of observations per teacher at

the 2 time points were different, and consequently, the proportion of 'proficient' teachers by rater, were calculated from different total number of instances, before and after the intervention. The data is summarised in Table 3.

[Insert Table 3 and 4 here.]

We implemented the adaptive Bayesian method to the TNE data set using the conjugate Beta priors – namely, uniform, Novick-Hall, Zellner and Jeffreys. Due to space constraint, we are reporting results from only the Jeffreys prior, though the results using the other mentioned priors are very similar.

Results based on Jeffreys prior are presented in Table 4. From Table 4, we observe that for Rater 3, the posterior mean of time point 1 is not contained in the 95% HPD interval of the second time point. Also from Table 4, the 95% HPD interval of the second time point of only Rater 3 lies entirely to the left of the posterior mean of the first time point. For all others, the 95% HPD interval contains their respective posterior mean from the first time point.

[Insert Figure 5 here.]

From these results, we can conclude that by the adaptive Bayesian method, Raters 1, 2 and 4 did not observe any significant change in proficiency before and after intervention, while Rater 3, observed a significant drop in the proficiency after the intervention.

Figure 5 reveals that for Raters 2, 3 and 4, the posterior density in the second time point (solid line) shifted to the left compared to the posterior density in the first time point (dotted line), indicating a general agreement of assessment between these 3 raters. However, for Rater 1, this was the opposite. In addition, we observe from Table 4 that the Kullback-Leibler divergence measure for Rater 3 is the largest, which is in tune with the conclusive evidence that there was a significant drop in proficiency in the second time

point.

6 Conclusion

The problem of inferencing with binomial proportions is a very important problem in statistics. Inference procedures for proportions from independent samples as well as from matched-paired samples are well established in the literature. However, for atypical cases where the same sample is tested for equality of proportions, at different time points, based on an unequal number of observation instances, such inference methods do not suffice.

In this paper, we have developed an adaptive Bayesian method to address this problem. We have evaluated the proposed method by comparing its performance to the regular frequentist tests for detecting significant shifts in proportions between time points via the ϕ-divergence measure. We have also made recommendations for using the developed adaptive Bayesian method instead of the regular frequentist method in certain scenarios. We suggest a prescription, supported both by simulation studies and theoretical justification, that when the subsequent time point sample size is larger than the previous time point sample size, the adaptive Bayesian method outperforms the regular frequentist method. When the second time point sample size is smaller than the first time point sample size, the regular frequentist method outperforms the adaptive Bayesian method. For equal sample size scenarios, i.e., in the matched pair scenario, either of the methods perform equally well.

We have applied the proposed method to two real case studies — one involving the testing of annual proportion of *extreme* wave-heights between 1997 and 2003 for a station off the coast of Cape Town in South Africa; and the second involving the evaluation of a group of teachers categorised as 'proficient' or 'non-proficient', before and after an

intervention, with the number of evaluation instances being unequal at the two time points. In both the case studies, the inference problem could only be addressed via the developed method. The adaptive Bayesian method developed here does not consider any covariate information associated with the observations, which can be an immediate natural extension of this paper.

References

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician, 54*, 280-288.

Agresti, A. and Coull, B.A. (1998). Approximate is better than"Exact" for interval estimation of binomial proportions. *The American Statistician, 52*, 119-126.

Ali, S.M. and Silvey, S.D. (1966). A general class of coefficient of divergence of one distribution from another. *Journal of the Royal Statistical Society, 286*, 131142.

Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27(1)*, 3-23.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edn.). Springer Series in Statistics, New York.

Berger, J.O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3), 385-402.

Brown, L.D., Cai, T.T. and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16(2)*, 101133.

Chen, M.-H. and Dey, D. (1998). Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhyā Series A, 60*, 322-343.

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2002). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 test. *Biometrics, 10*, 417-451.

Csiszar, I. (1967). Information type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica, 2*, 105-113.

Dey, D.K. and Birmiwal, L.R. (1994). Robust Bayesian analysis using entropy and divergence measures. *Statist. Probab. Lett., 20*, 287-294.

Eguchi, S. and Copas, J. (2006). Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of Multivariate Analysis, 97(9)*, 2034-2040.

Geisser, S. (2006). *Modes of Parametric Statistical Inference*. Wiley Series in Probability and Statistics.

Ghosh, M. and Chen, M.-H. (2002). Bayesian inference for matched case-control studies. *Sankhyā, Series B, 2(64)*, 107-127.

Ghosh, M., Chen, M.-H., Ghosh, A. and Agresti, A. (2000). Hierarchical Bayesian analysis of binary matched pairs data. *Statista Sinica, 10*, 647-657.

Jeffreys, H. (1961). *Theory of Probability* (3rd. edn.). Oxford University Press, London.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

Liao, J.G. (1999). A hierarchical Bayesian model for combining multiple 2×2 tables using conditional likelihoods. *Biometrics, 55(1)*, 268-272.

Liddell, F.D. (1983). Simplified exact analysis of case-referent studies: matched pairs; dichotomous exposure. *Journal of Epidemiology and Community Health, 37*, 82-84.

Lloyd, C. J. (1990). Confidence intervals from the difference between two correlated proportions. *Journal of the American Statistical Association, 85*, 1154-1158.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22,* 719-748.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions of percentages. *Psychometrika, 12*, 153-157.

Newcombe, R.G. (1998). Improved confidence interval for the difference between binomial proportions based in paired data. *Statistics in Medicine, 17*, 2635-2650.

Novick, M.R. and Hall, W.J. (1965). A Bayesian indifference procedure. *Journal of the American Statistical Association, 60*, 1104-1117.

Zellner, A. (1977). Maximal data information prior distributions. In *Methods in the Applications of Bayesian Methods*, A. Aykac and C. Brumat (Eds.), 211-232. North-Holland, Amsterdam.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician, 42(4)*, 278-280.

Zellner, A. (2002). Information processing and Bayesian analysis. *Journal of Econometrics, 107*, 41-50.

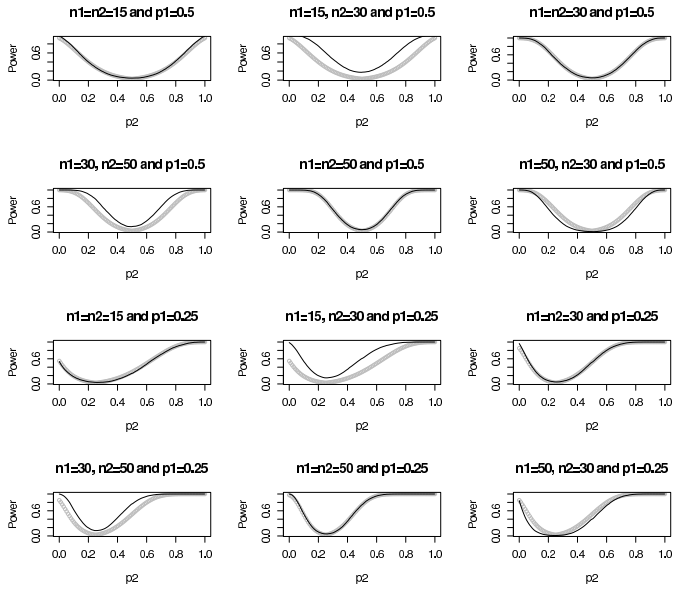

Figure 1: Simulation study of power function ($p_1 = 0.5, 0.25$).
Note, Grey circles: Frequentist method, Black solid: Adaptive Bayesian method.

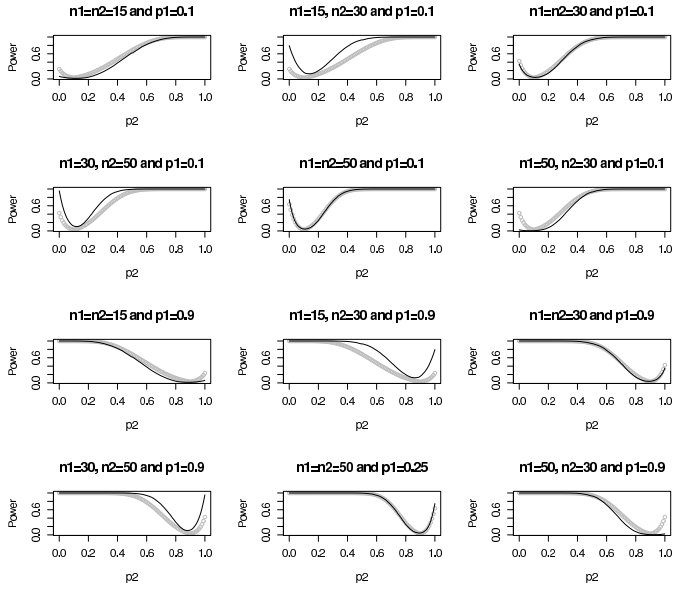

Figure 2: Simulation study of power function ($p_1 = 0.1, 0.9$).
Grey circles: Frequentist method, Black solid: Adaptive Bayesian method.

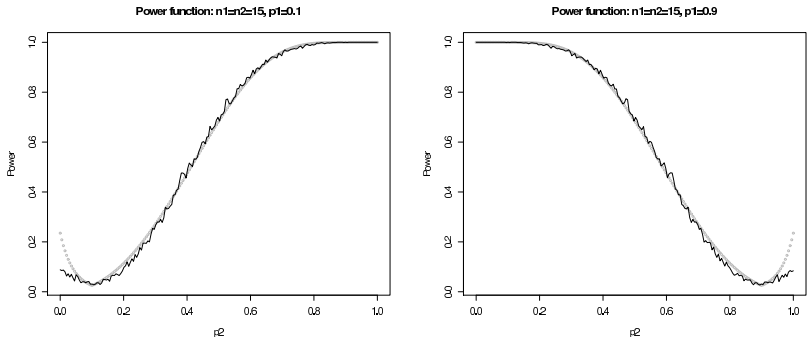

Figure 3: Exact power calculation for small sample size.
Note, Grey circles: Frequentist, Solid line: Adaptive Bayesian.

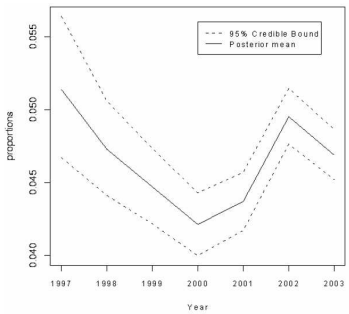

Figure 4: 95% CI of extreme HMO

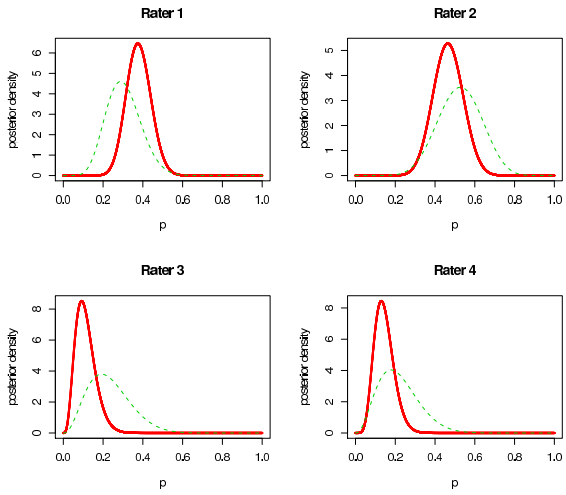

Figure 5: Posterior density of TNE data.
Note: Dotted lines: first time point; solid line: second time point.

Table 1: Summary of HMO (in metres) recorded at a station off Cape Town coast

Year	Total observed HMO	Number of HMO in top 5%	Proportion of HMO in top 5%
1997	8020	411	0.051
1998	8608	374	0.044
1999	8432	334	0.040
2000	7377	247	0.034
2001	8742	433	0.050
2002	8729	671	0.077
2003	8470	265	0.031

Table 2: Posterior mean and the corresponding 2.5% and 97.5% quantiles of HMO data

Year	2.5% posterior quantile	Posterior mean	97.5% posterior quantile
1997	0.047	0.051	0.056
1998	0.044	0.047	0.051
1999	0.042	0.045	0.047
2000	0.040	0.042	0.044
2001	0.042	0.044	0.046
2002	0.048	0.050	0.051
2003	0.045	0.047	0.049

Table 3: Summary of 4 raters evaluation at time points T_1 and T_2

	Rater 1		Rater 2		Rater 3		Rater 4	
	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
Proficient	8	15	10	10	3	1	3	4
Not-Proficient	19	19	9	14	11	24	12	32
Total	27	34	19	24	14	25	15	36

Table 4: Posterior estimates of mean, standard error (s.e.), 95% HPD and the Kullback-Leibler (KL) divergence for the 4 raters at two time points.

		Post. Mean	Post. s.e.	95% HPD	KL divergence (+/-)
Rater 1	Time 1	0.304	0.085	(0.136, 0.471)	1.025 (+)
	Time 2	0.379	0.061	(0.259, 0.499)	
Rater 2	Time 1	0.525	0.109	(0.311, 0.739)	0.524 (-)
	Time 2	0.466	0.074	(0.322, 0.611)	
Rater 3	Time 1	0.250	0.105	(0.044, 0.456)	**2.238** (-)
	Time 2	0.122	0.050	**(0.023, 0.221)**	
Rater 4	Time 1	0.235	0.100	(0.039, 0.431)	1.452 (-)
	Time 2	0.151	0.049	(0.055, 0.246)	

Note: '+/-' indicates a positive/negative shift of proficiency after intervention.