Journal of Data Science **6**(2008), 261-268

A Bayesian Approach to Zero-Numerator Problems Using Hierarchical Models

Zhongxue Chen and Monnie McGee
Southern Methodist University

Abstract: The rule of three gives $3/n$ as the upper 95% bound for the success rate of the zero-numerator problems. However, this bound is usually conservative although it is useful in practice. Some Bayesian methods with beta distributions as priors have been studied. However, choosing the parameters for the priors is subjective and can severely impact the corresponding posterior distributions. In this paper, some hierarchical models are proposed, which provide practitioners other options for those zero-numerator problems.

Key words: Bayesian hierarchical model, binomial model, Markov chain Monte Carlo, rare events.

1. Introduction

1.1 Introduction

Suppose we want to know the probability of the occurrence of a certain kind of event when in n independent trials, the event never occurs. This situation is referred as the zero-numerator problem. A probability model for this issue can be built by a binomial distribution with sample size n and the probability p, which is usually very small. Based on this binomial model, the point estimate of p by the maximum likelihood estimator is $p = x/n = 0$ since here $x = 0$. This estimate is not accurate and may not useful in practice. Although this event is rare, it may occur on occasion based on our previous experience.

1.2 Frequentist method and the rule of three

Louis (1981) gave a $(1 - \alpha) \times 100$ percent confidence interval for p:

$$[0, p_n], \quad p_n = 1 - \alpha^{1/n} = S_b/n, \tag{1.1}$$

where S_n can be considered as a number of successes in a future experiment of the same size. By taking limit as $n \to \infty$, then we have

$$\lim_{n\to\infty} S_n = -\ln(\alpha). \tag{1.2}$$

When $\alpha = 0.05, -\ln(\alpha) \approx 3$. So the rule of three states that the upper 95% bound for p is about $3/n$.

Jovanovic and Levy (1997) obtained the same results from a different way. Suppose random variable X has a Binomial distribution with parameters n and p, then

$$P(X = 0 \,|\, n, p) = (1-p)^n \tag{1.3}$$

A $(1-\alpha)\times 100\%$ bound can be obtained by solving $(1-p)^n > \alpha$. It gives the upper bound $p_u = 1 - \alpha^{1/n}$. By using a Taylor expansion, we then have

$$p_u = 1 - \alpha^{1/n} \approx -\ln(\alpha)/n \tag{1.4}$$

Both Louis (1981) and Jovanovic and Levy (1997) gave numerical examples to show that when n is large enough, the rule of three gives a good approximation of the upper 95% bound. However, sometimes we are more interested in predicting the occurrence rate rather than obtaining upper bounds. Obviously, if this is the case, the upper 95% bound does not help us much. Louis (1981) cited Bickel and Doksum (1980) and pointed out that this bound corresponds to the Bayesian 95% credibility bound for a uniform prior on p.

1.3 An example and Bayesian models

One of the applications for the zero-numerator problems is for the false-positive rate for a medical test with no previous record of positive results. Another example, given by Hanley and Lippman-Hand (1983) and cited by Winkler *et al.* (2002), involved two different contrast agents used by radiologists over a long time. The standard one has been shown to cause a serious reaction in about 15 of every 10,000 patients exposed to it. The new contrast agent was applied to 167 patients and none of them reported having the reaction. By the rule of three, the upper 95% confidence bound for the probability of a serious reaction with the new contrast agent is about $3/167 = 0.018$, while the standard contrast agent has the probability of 0.0015. What can we say about the probability of a serious reaction for the new contrast agent?

Bayesian models may shed some light on this problem. For a Bayesian model, the commonly used prior for Binomial distributions is the class of beta distributions $Beta(a, b)$. Geisser (1984) has discussed several different prior distributions that were used in binary trials, for example, the noninformative distributions $Beta(0.5, 0.5)$ (a Jeffreys prior) and $Beta(1, 1)$ (uniform).

Jovanovic and Levy (1997) suggested using $Beta(1, b), b \geq 1$ as the prior because when $b > 1$ the prior favors values of p close to zero. In addition, values of a other than 1 provide the prior with a local maximum away from zero that cannot be justified without additional information. It is well known that $Beta(a, b)$ is a conjugate prior for the binomial distribution $bin(n, p)$ and the corresponding posterior distribution is $Beta(y + a, n + b)$, where y is the observed value in n trials. In zero-numerator problems $y = 0$ and therefore the posterior is $Beta(n, n + b)$. Winkler *et al.* (2002) discussed this problem for both noninformative and informative beta distributions. They stated that $a = 1$ in Jovanovic's prior $Beta(a, b)$ was unduly restrictive and suggested trying any $Beta(a, b)$ with $a > 0$ and $b > 0$ when assessing priors.

2. Hierarchical Models

2.1 Hierarchical models

In Bayesian models for the zero-numerator problems, the prior has a huge impact on the posterior distributions due to the limited information available in the data (i.e., no event has been observed). Following Jovanovic and Levy (1997), we use the prior $Beta(1, b)$ where b is a random variable taking values greater than or equal to 1, which has its own distribution, the hyperprior. In the previous Bayesian models, b is a constant number that may vary according to the person assigning the values. Therefore, it is reasonable for us to treat b as a random variable with a given distribution. In the zero-numerator problem, we know that the probability of p is small. To capture this information, as Jovanovic and Levy (1997) have stated, b is usually greater than 1. So the random variable b may take values on $(1, \infty)$. Based on this, a reasonable hyperprior can be assigned. For example, we can assume $1/(b - 1)$ is distributed as a $Beta(c, d)$ where c and d are constants. Therefore we have the following hierarchical model:

$$\begin{aligned} y \,|\, n, p &\sim Bin(n, p) \\ p \,|\, b &\sim Beta(1, b) \\ 1/(b-1) &\sim Beta(c, d). \end{aligned} \tag{2.1}$$

In this model, the hyperparameter c should not be too small; otherwise the mass of the distribution will concentrate around zero and it will be very likely to obtain a large value of b. In other words, the posterior distribution will concentrate around zero and underestimate the probability of p.

Another reasonable choice is an exponential hyperprior:

$$\begin{aligned} y \,|\, n, p &\sim Bin(n, p) \\ p \,|\, b &\sim Beta(1, b) \\ (b-1) &\sim \exp(\lambda), \end{aligned} \tag{2.2}$$

where $\exp(x \,|\, \lambda) = \lambda e^{-x\lambda}$.

We can also choose gamma hyperprior on $b - 1$

$$\begin{aligned} y \,|\, n, p &\sim Bin(n, p) \\ p \,|\, b &\sim Beta(1, b) \\ (b-1) &\sim Gamma(\alpha), \end{aligned} \tag{2.3}$$

where $Gamma(x \,|\, \alpha) = x^{\alpha-1}e^{-x}/\Gamma(\alpha)$.

In the next subsection, we will give the numerical results obtained from different methods, including Bayesian models with noninformative and formative prior as well as hierarchical models (2.1)-(2.3).

2.2 Numerical results of the example

The popular priors of noninformative Beta distributions for this problem are $Beta(0.5, 0.5)$ (a Jefferys prior) and $Beta(1, 1)$ (uniform distribution). If we apply these two distributions to the previous example, the corresponding posterior distributions are $Beta(0.5, 167.5)$ and $Beta(1, 168)$, respectively. The means are $0.5/(0.5 + 167.5) = 0.00298$ and $1/169 = 0.00592$. Although these two numbers are very different, both of them are greater than 0.0015.

In this example, the risk proportion for the standard agent is known and this may provide us some information about the proportion of the new agent. Winkler *et al.* (2002) chose a prior $Beta(a, b)$ such that the prior mean equals to 0.0015 and has 95% chance that p is less than 0.75%, five times the risk of the old agent. They obtained $\alpha = 0.042$ and $b = 27.96$. The posterior is $Beta(0.042, 194.96)$ and its mean is 0.00022,which is much less than 0.0015. If we set $a = 1$ and want the mean of $Beta(1, b) = 0.0015$, then $b = 665.71$. The resulting posterior is $Beta(1, 832.71)$ and its mean is 0.0012.

If we use posterior mean to estimate the new risk, the method based on Jovanovic's suggestion seems give a value closer to the previous information than the one obtained by Winkler. Furthermore, if we believe that under the true risk p the probability that no serious reaction occurs in 167 observations is between 5% and 95%, then p should be between 0.00031 and 0.01778. Therefore the risk value given by Winkler seems too small. In addition Winkler's posterior median is

2×10^{-10}, which is too small. That means the posterior distribution concentrates too much mass at zero. These four posteriors are plotted in Figure 1 and some statistics (5% quantile, median, mean and 95% quantile) for these posteriors are shown in Table 1. From Figure 1 and Table 1, we can see that with different priors, the densities and the statistics may be very different. Choosing a "good" prior is a very important and difficult task.

Table 1: Summary statistics for the posteriors of Bayesian model with different a and b's

	$a = b = 0.5$	$a = b = 1$	$a = 0.042, b = 27.96$	$a = 1, b = 665.71$
5% quantile	1.12×10^{-5}	0.00029	3.14×10^{-34}	6.16×10^{-5}
Median	0.00140	0.00387	2.03×10^{-10}	0.00083
Mean	0.00298	0.00559	0.00022	0.00120
95% quantile	0.0144	0.0167	0.00106	0.00359

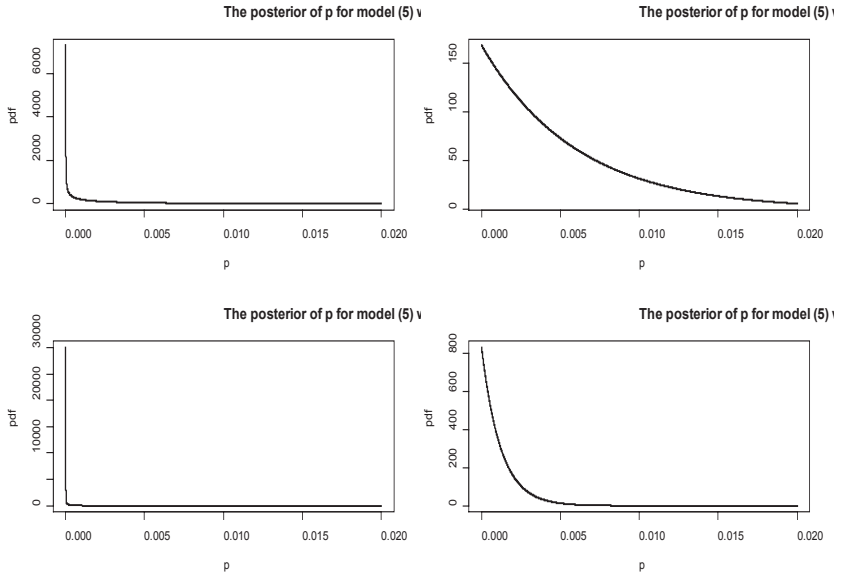

Figure 1: Posteriors for Bayesian models with different parameters

Hierarchical models (2.1), (2.2) and (2.3), each with different hyperparameters in the hyperprior distributions, are also developed for this example. Markov chain Monte Carlo (MCMC) method is used to approximate the corresponding posterior distributions. The calculation and plots are done by using software Winbugs[1].

[1]see http://www.mrc-bsu.cam.ac.uk/bugs/

Table 2: Summary statistics for the posteriors of models (2.1), (2.2) and (2.3)

Medel (2.1)				
	$c = 0.5, d = 0.5$	$c = 1, d = 1$	$c = 1, d = 10$	$c = 1, d = 100$
5%	8.1×10^{-6}	1.5×10^{-4}	1.1×10^{-4}	4.9×10^{-5}
Median	0.0013	0.0030	0.0024	0.0013
Mean	0.0028	0.0047	0.0039	0.0024
95%	0.0110	0.0151	0.0128	0.0085

Model (2.2)				
	$c = 0.01$	$c = 1$	$c = 10$	$c = 1000$
5%	1.6×10^{-4}	3.0×10^{-4}	3.2×10^{-4}	3.0×10^{-4}
Median	0.0022	0.0041	0.0041	0.0041
Mean	0.0034	0.0059	0.0058	0.0059
95%	0.0105	0.0174	0.0175	0.0138

Model (2.3)				
	$c = 0.1$	$c = 1$	$c = 100$	$c = 1000$
5%	3.2×10^{-4}	3.0×10^{-4}	2.0×10^{-4}	4.4×10^{-5}
Median	0.0041	0.0041	0.0026	6.0×10^{-4}
Mean	0.0058	0.0059	0.0037	8.6×10^{-4}
95%	0.0175	0.0138	0.0110	0.0026

Some statistics (5% quantile, median, mean and 95% quantile) are summarized in Table 2. Unlike the Bayesian models, the summary statistics from our Bayesian hierarchical models have very close values even for different hyperpriors with different hyperparameters. Figure 2 shows the densities of the posteriors from model (2.1) with different parameters. It is clear that those posteriors have very similar densities. For models (2.2) and (2.3), we obtained similar plots (not shown) as Figure 2.

3. Conclusion

For the zero-numerator problem, the rule of three gives conservative results of the upper 95% bounds. Bayesian models typically use beta priors, and the posterior distribution depends heavily on the values of the parameters used in the priors. Choosing the parameters is a hard task that needs more attention. The numerical results of an example shows that the traditional way of assessing

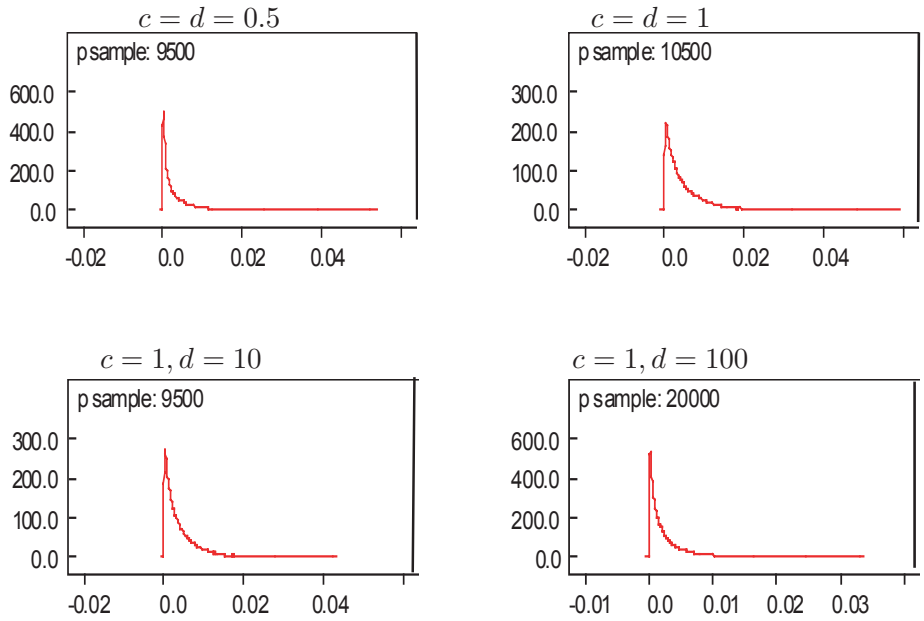

Figure 2: Posteriors for model (2.1)

a prior is not suitable for zero-numerator problems because of the sensitivity to choice of priors. To solve this problem, we proposed several hierarchical models. The hierarchical models give very consistent results, regardless of the choice of prior parameters.

References

Louis, T. A. (1983). Confidence intervals for a binomial parameter after observing no successes. *The American Statistician* **35**, 154-154.

Jovanovic, B. D. and Levy, P. S. (1997). A look at the rule of three. *The American Statistician* **51**, 137-139.

Bickel, P. J. and Doksum, K. A. (1980). *Mathematical Statistics*. Holden-Day.

Hanley, J. A. and Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? interpreting zero numerators. *Journal of the American Medical Association* **249**, 1743-1745.

Winkler, R. L., Smith, J. E. and Fryback D. G. (2002). The role of informative priors in zero-numerator problems: being conservative versus being candid. *The American Statistician* **56**, 1-4.

Geisser, S. (1984). On prior distributions for binary trials (with comments). *The American Statistician* **38**, 244-251.

Received January 29, 2006; accepted March 7, 2007.

Zhongxue Chen
Department of Statistical Science
Southern Methodist University
3225 Daniel Avenue
P.O. Box 750332
Dallas, Texas 75275, USA
zhongxue@mail.smu.edu

Monnie McGee
Department of Statistical Science
Southern Methodist University
3225 Daniel Avenue
P.O. Box 750332
Dallas, Texas 75275, USA
mmcgee@mail.smu.edu