Bayes Factors for Grouped Data

Lizanne Raubenheimer[1] **and Abrie J. van der Merwe**[2]

[1] Department of Statistics, Rhodes University, Grahamstown, South Africa,

L.Raubenheimer@ru.ac.za

[2] Department of Mathematical Statistics and Actuarial Science, University of the Free State,

Bloemfontein, South Africa, matheeme@ufs.ac.za

Abstract

In this paper we apply Bayes factors to grouped data. Group testing is where units are pooled together and tested as a group rather than individually. The Bayes factor is the ratio of the posterior probabilities of the null and the alternative hypotheses divided by the ratio of the prior probabilities for the null and the alternative hypotheses. A beta prior will be used, also known as a conjugate prior for the binomial distribution. An application to mosquito data will be considered, where a comparison is made between West Nile virus (WNV) infection prevalences in field collected *Culex nigripalpus* mosquitoes trapped at different heights.

Key words: Bayes factors, conjugate prior, group testing

1 Introduction

In Bayesian terminology we are not testing, but doing model comparison. Jeffreys (1961) introduced and developed the Bayesian approach to hypothesis testing. See Kass & Raftery (1995) and Robert et al. (2009) for a detailed discussion and explanation of Bayes factors, where they emphasize different points on Bayes factors. In this paper we will focus on Bayes factors for grouped data, where model comparison will be made for two proportions from grouped data. Group testing is where units are pooled together and tested as a group rather than individually. Group testing is also known as pooled testing, where pooled testing was introduced by Dorfman (1943). Dorfman (1943) used group testing for medical screening purposes to identify infected individuals. Bayes factors will be applied to an example by Biggerstaff (2008), where a comparison was made between West Nile virus (WNV) infection prevalences in field collected *Culex nigripalpus* mosquitoes trapped at different heights.

Not much has been done in literature from a Bayesian point of view on group testing. Hanson et al. (2006) used a two-stage sampling procedure and developed a Bayesian method that allows for sampling multiple sites in a specific region. Gastwirth & Johnson (1994) used independent beta priors. Chick (1996) used the $beta(\alpha,\beta)$ prior for obtaining posterior distributions of the unknown proportion p. The methods were applied to grouped test data for gene transfer experiments and limiting dilution assay data for immunocompetency studies.

Notation, the likelihood function and some theoretical aspects will be considered in Section 2. Bayes factors will be discussed and shown in Section 3, simulation studies will be considered in Section 4 and the application will be considered in Section 5. The discussion and conclusion will be given in Section 6.

2 Notation and Likelihood Function for Binomial Proportions from Pooled Samples

Assume that the proportion of successes in a given population is p. We will refer to an infected individual as a success in a binomial trial.

The following notation will be used in this paper:

N - number of individuals to be sampled independently from the population

m_i - the size of a pool where $i = 1, 2, \ldots, M$

M - the number of distinct pool sizes

n_i - the number of pools of size m_i

X_i - the number of the n_i pools that is positive.

In the case of grouped data assume that $X_1, X_2, \ldots, X_M$ are independent binomial random variables with parameters n_i and $1-(1-p)^{m_i}$, i.e. $X_i \sim Bin\left(n_i, 1-(1-p)^{m_i}\right)$.

The likelihood function is given by

$$\begin{aligned} L(p|data) &= \prod_{i=1}^{M} \begin{pmatrix} n_i \\ x_i \end{pmatrix} \left\{ \left[1-(1-p)^{m_i}\right]^{x_i} \left[(1-p)^{m_i}\right]^{n_i - x_i} \right\} \\ &\propto \prod_{i=1}^{M} \left\{ \left[1-(1-p)^{m_i}\right]^{x_i} \left[(1-p)^{m_i}\right]^{n_i - x_i} \right\}. \end{aligned} \tag{1}$$

In this paper we are interested in comparing two proportions, say p_1 and p_2. The likelihood function will then be

$$\begin{aligned}
L(p_1, p_2 | \underset{\sim}{x}_1, \underset{\sim}{x}_2) &= \prod_{i=1}^{2} \prod_{j=1}^{M_i} \begin{pmatrix} n_{ij} \\ x_{ij} \end{pmatrix} \left\{ [1-(1-p_i)^{m_{ij}}]^{x_{ij}} [(1-p_i)^{m_{ij}}]^{n_{ij}-x_{ij}} \right\} \\
&\propto \prod_{i=1}^{2} \prod_{j=1}^{M_i} \left\{ [1-(1-p_i)^{m_{ij}}]^{x_{ij}} [(1-p_i)^{m_{ij}}]^{n_{ij}-x_{ij}} \right\}. \qquad (2)
\end{aligned}$$

A conjugate prior to the binomial distribution is used. Conjugacy may be defined as a joint property of the prior and the likelihood function that provides a posterior from the same distribution family as the prior, (Robert, 2001). Statisticians make use of conjugate priors to be certain that the posterior is predictable in its form. Consider a beta prior, i.e. $p_i \sim Beta(\alpha, \beta)$ for the p's

$$\pi(p_1, p_2) = \prod_{i=1}^{2} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1} (1-p_i)^{\beta-1}. \qquad (3)$$

3 Bayes Factors

The Bayes factor is the ratio of the posterior probabilities of the null and the alternative hypotheses divided by the ratio of the prior probabilities for the null and the alternative hypotheses (Robert, 2001). The classical approach to hypothesis testing is not probability based; one could not place a probability on a hypothesis because a hypothesis is not a random variable in the frequentist sense. Using a frequentist approach, one has to make do with quantities like the p - value where this is conditional on H_0 being true. We do not know if H_0 is true, the real question is actually $P(H_0 \text{ is true} | \text{data})$. The Bayesian wants to find a probability that H_0 is true. The Bayes factor is a summary of the evidence provided by the data in favour of a scientific theory, represented by a statistical model, as opposed to another (Kass & Raftery, 1995).

In Bayesian terminology we are not testing in the classical sense, but we are comparing two possible models. This is also known as model comparison or Bayes factor analysis. For example, comparing model $f(x|\theta_0, \gamma)$ with model $f(x|\theta, \gamma)$, where θ is the unspecified parameter and γ is a nuisance parameter. In this instance we are interested in testing $H_0 : \theta = \theta_0$

against $H_1 : \theta \neq \theta_0$, where H_0 is the null hypothesis and H_1 the alternative hypothesis. Instead of calling the two options hypotheses, we shall call them models $\mathscr{M}_0$ and $\mathscr{M}_1$, respectively. The probability that $\mathscr{M}_0$ is the 'correct' model will then be calculated.

3.1 Two Samples with $M_1 = M_2 = 1$

We first consider the simplest case where $n_1 = n_2 = n$ and $m_1 = m_2 = m$. The n's can be different, as long as the m's are the same. Then the equality of the p's is equivalent to the model $\mathscr{M}_0 : \theta_1 = \theta_2 = \theta$, which will be compared to the model $\mathscr{M}_1 : \theta_1 \neq \theta_2$. Here we have $\theta = (1-p)^m$. Under $\mathscr{M}_0$ the prior on θ is $beta(\alpha,\beta)$, while under $\mathscr{M}_1$ we have two independent $beta(\alpha,\beta)$ priors.

The Bayes factor in favour of $\mathscr{M}_0$ is given by

$$\begin{aligned}
B_{01} &= \frac{\int_0^1 L(\theta\,|x,\mathscr{M}_0)\,\pi(\theta)\,d\theta}{\int_0^1\int_0^1 L(\theta_1,\theta_2\,|x_1,x_2,\mathscr{M}_1)\,\pi(\theta_1)\,\pi(\theta_1)\,d\theta_1 d\theta_2} = \frac{f(x\,|\mathscr{M}_0)}{f(x_1,x_2\,|\mathscr{M}_1)} \\
&= \frac{B(x+\alpha,2n-x+\beta)}{B(\alpha,\beta)} \div \frac{B(x_1+\alpha,n-x_1+\beta)\,B(x_2+\alpha,n-x_2+\beta)}{B(\alpha,\beta)^2} \\
&= \frac{B(x+\alpha,2n-x+\beta)\,B(\alpha,\beta)}{B(x_1+\alpha,n-x_1+\beta)\,B(x_2+\alpha,n-x_2+\beta)},
\end{aligned}$$

where $x = x_1 + x_2$.

Another approach to calculate Bayes factors, is to use fractional Bayes factors. This was proposed by O'Hagan (1995). Here one uses part of the information from the data to create proper priors from improper priors. It uses a fraction of the likelihood to obtain proper priors. If we let $\alpha = \beta = 1/2$ i.e. considering a $beta(1/2,1/2)$ prior for the p's, we actually make use of the Jeffreys prior. In this case the Jeffreys prior is proper, and there is no need to make use of the fractional Bayes factor. If we let $\alpha = \beta = 0$ i.e. considering a $beta(0,0)$ prior for the p's, we actually make use of the Haldane prior. This prior was introduced by Haldane (1932). According to Zellner (1977) the Haldane prior is popular due to the posterior mean being equal to the maximum likelihood estimator. In this case the Haldane prior is improper, and we can't use the Bayes factor and therefore have to make use of partial Bayes factors, to be more specific the fractional Bayes factor. To create a proper prior for the parameters under the models, a fraction b of the likelihood should be used.

For illustration and comparison purposes we will consider the fractional Bayes factor when using the Jeffreys and Haldane priors.

When using the Jeffreys prior, the fractional Bayes factor in favour of model $\mathscr{M}_0$ is given by

$$\begin{aligned} B_{01} &= \frac{f^F(x|\mathscr{M}_0)}{f^F(x_1,x_2|\mathscr{M}_1)} = \frac{f(x|\mathscr{M}_0)}{f_b(x|\mathscr{M}_0)} \div \frac{f(x_1,x_2|\mathscr{M}_1)}{f_b(x_1,x_2|\mathscr{M}_1)} \\ &= \frac{B(2n-x+1/2,x+1/2)}{B(b(2n-x)+1/2,bx+1/2)} \div \frac{B(n-x_1+1/2,x_1+1/2)B(n-x_2+1/2,x_2+1/2)}{B(b(n-x_1)+1/2,bx_1+1/2)B(b(n-x_2)+1/2,bx_2+1/2)} \\ &= \frac{B(2n-x+1/2,x+1/2)B(b(n-x_1)+1/2,bx_1+1/2)B(b(n-x_2)+1/2,bx_2+1/2)}{B(b(2n-x)+1/2,bx+1/2)B(n-x_1+1/2,x_1+1/2)B(n-x_2+1/2,x_2+1/2)}. \end{aligned}$$

When using the Haldane prior, the fractional Bayes factor in favour of model $\mathscr{M}_0$ is given by

$$\begin{aligned} B_{01} &= \frac{f^F(x|\mathscr{M}_0)}{f^F(x_1,x_2|\mathscr{M}_1)} = \frac{f(x|\mathscr{M}_0)}{f_b(x|\mathscr{M}_0)} \div \frac{f(x_1,x_2|\mathscr{M}_1)}{f_b(x_1,x_2|\mathscr{M}_1)} \\ &= \frac{B(2n-x,x)}{B(b(2n-x),bx)} \div \frac{B(n-x_1,x_1)B(n-x_2,x_2)}{B(b(n-x_1),bx_1)B(b(n-x_2),bx_2)} \\ &= \frac{B(2n-x,x)B(b(n-x_1),bx_1)B(b(n-x_2),bx_2)}{B(b(2n-x),bx)B(n-x_1,x_1)B(n-x_2,x_2)}. \end{aligned}$$

3.2 General Case for Two Samples

For the choice of prior given in the previous section, let $\alpha = \beta = 1/2$ i.e. considering a $beta(1/2,1/2)$ prior for the p's. Consider two models, $\mathscr{M}_0 : p_1 = p_2 = p$ and $\mathscr{M}_1 : p_1 \neq p_2$.

Under model $\mathscr{M}_0$, the likelihood will be

$$L(p|\underline{x},\mathscr{M}_0) \propto \prod_{i=1}^{2}\prod_{j=1}^{M_i}\left\{[1-(1-p)^{m_{ij}}]^{x_{ij}}[(1-p)^{m_{ij}}]^{n_{ij}-x_{ij}}\right\},$$

and the marginal likelihood is then

$$f(\underline{x}|\mathscr{M}_0) = \frac{1}{\pi}\int_0^1 p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}L(p|\underline{x},\mathscr{M}_0)\,dp.$$

Under model $\mathscr{M}_1$, the likelihood will be

$$L(p_1, p_2 | \underline{x}_1, \underline{x}_2, \mathscr{M}_1) \propto \prod_{i=1}^{2} \prod_{j=1}^{M_i} [1-(1-p_i)^{m_{ij}}]^{x_{ij}} [(1-p_i)^{m_{ij}}]^{n_{ij}-x_{ij}},$$

and the marginal likelihood is then

$$f(\underline{x}_1, \underline{x}_2 | \mathscr{M}_1) = \frac{1}{\pi^2} \int_0^1 \int_0^1 p_i^{-\frac{1}{2}} (1-p_i)^{-\frac{1}{2}} L(p_1, p_2 | \underline{x}_1, \underline{x}_2, \mathscr{M}_1) \, dp_1 dp_2.$$

The Bayes factor in favour of $\mathscr{M}_0$ is given by

$$\begin{aligned} B_{01} &= \frac{f(\underline{x} | \mathscr{M}_0)}{f(\underline{x}_1, \underline{x}_2 | \mathscr{M}_1)} \\ &= \frac{\frac{1}{\pi} \int_0^1 p^{-\frac{1}{2}} (1-p)^{-\frac{1}{2}} L(p | \underline{x}, \mathscr{M}_0) \, dp}{\frac{1}{\pi^2} \int_0^1 \int_0^1 p_i^{-\frac{1}{2}} (1-p_i)^{-\frac{1}{2}} L(p_1, p_2 | \underline{x}_1, \underline{x}_2, \mathscr{M}_1) \, dp_1 dp_2}. \end{aligned}$$

If one assumes that the two models are equally likely before hand, i.e. $P(\mathscr{M}_0) = P(\mathscr{M}_1)$, the posterior probability of model $\mathscr{M}_0$ is

$$P(\mathscr{M}_0 | \underline{x}) = \left(1 + \frac{1}{B_{01}}\right)^{-1}.$$

4 Simulation Results for Two Samples with $M_1 = M_2 = 1$

Here we consider the simplest case where $n_1 = n_2 = n$ and $m_1 = m_2 = m$. Then the equality of the p's is equivalent to the model $\mathscr{M}_0 : \theta_1 = \theta_2 = \theta$, which will be compared to the model $\mathscr{M}_1 : \theta_1 \neq \theta_2$. Here we have $\theta = (1-p)^m$. Under $\mathscr{M}_0$ the prior on θ is $beta(\alpha, \beta)$, while under $\mathscr{M}_1$ we have two independent $beta(\alpha, \beta)$ priors. We consider two different prior here, one where $\alpha = \beta = 1/2$, the Jeffreys prior, and one where $\alpha = \beta = 1$, the uniform prior. Figures 1 and 2 show the posterior probabilities for $\mathscr{M}_0$ when $\alpha = \beta = 1/2$ as well as when $\alpha = \beta = 1$. This is for the selected value of x_1 and a range of outcomes for x_2 when $n = 20, 50, 100, 200$. In general the results look reasonable, with probabilities usually lower with the smaller values of α and β, except when n is small.

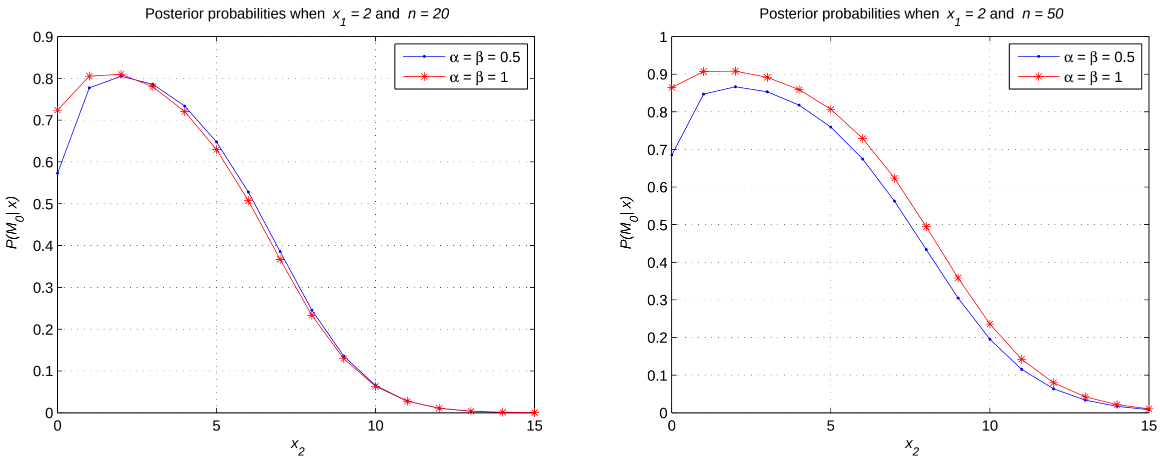

Figure 1: Posterior probabilities, $P(\mathscr{M}_0 | \underline{x})$ given that $x_1 = 2$ for $n = 20$ and 50.

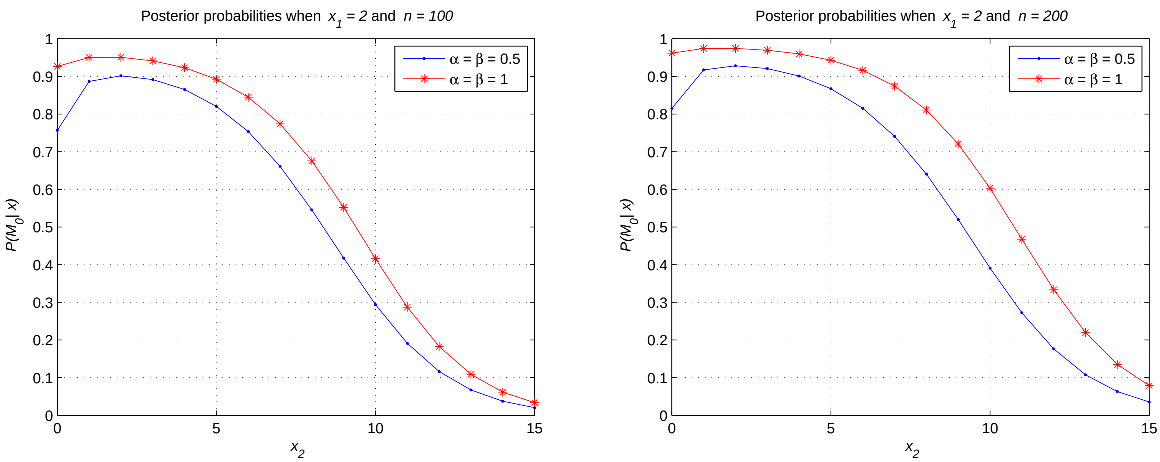

Figure 2: Posterior probabilities, $P(\mathscr{M}_0 | \underline{x})$ given that $x_1 = 2$ for $n = 100$ and 200.

We will now apply the fractional Bayes factor using the Jeffreys and Haldane priors when $b = 0.01$, $x_1 = 2$ and $n = 20$, 50, 100 and 200. The results are displayed in Table 1. In this case the Jeffreys prior is proper, and there is no need to make use of the fractional Bayes factor. It is used here just for comparison purposes.

Table 1: Posterior probabilities, $P(\mathscr{M}_0 | \underline{x})$ given that $b = 0.01$, $x_1 = 2$ for $n = 20, 50, 100$ and 200.

	Jeffreys	**Haldane**	**Jeffreys**	**Haldane**	**Jeffreys**	**Haldane**	**Jeffreys**	**Haldane**
x_2	$n = 20$		$n = 50$		$n = 100$		$n = 200$	
0	0.5592		0.6414		0.6774		0.6971	
1	0.7679	0.9748	0.8202	0.9746	0.8418	0.9745	0.8531	0.9745
2	0.7971	0.9755	0.8432	0.9749	0.8622	0.9748	0.8722	0.9747
3	0.7777	0.9708	0.8288	0.9701	0.8496	0.9700	0.8606	0.9700
4	0.7251	0.9604	0.7900	0.9603	0.8159	0.9604	0.8294	0.6905
5	0.6390	0.9412	0.7267	0.9433	0.7606	0.9440	0.7782	0.9445
6	0.5195	0.9068	0.6369	0.9150	0.6815	0.9175	0.7047	0.9188
7	0.3783	0.8457	0.5231	0.8692	0.5789	0.8756	0.6081	0.8787
8	0.2408	0.7422	0.3962	0.7975	0.4597	0.8115	0.4936	0.8180
9	0.1330	0.5851	0.2744	0.6928	0.3381	0.7190	0.3734	0.7312
10	0.0643	0.3914	0.1740	0.5558	0.2299	0.5977	0.2625	0.6171
11	0.0276	0.2139	0.1023	0.4029	0.1459	0.4579	0.1725	0.4841
12	0.0106	0.0959	0.0565	0.2616	0.0874	0.3204	0.1073	0.3496
13	0.0037	0.0363	0.0298	0.1535	0.0501	0.2055	0.0639	0.2331
14	0.0011	0.0120	0.0151	0.0830	0.0278	0.1226	0.0368	0.1450
15	0.0003	0.0034	0.0074	0.0423	0.0150	0.0692	0.0208	0.0857

The probabilities when using the Haldane prior are considerably higher than those from the Jeffreys prior. In the case of the Haldane prior all x's must be larger than zero, and $b > 0$. One of the main questions is: What should the value of b be? We know that $P(\mathscr{M}_0 | \underline{x}) \to 1$ when $b \to 0$ and $P(\mathscr{M}_0 | \underline{x}) \to 0.5$ when $b \to 1$, so the posterior probability can be manipulated by the choice of b. The usual practice is to choose $b \propto n^{-1}$, and O'Hagan (1995) suggested $b = {}^{q}/_{n}$ where q is the minimal sample size.

5 Application

The Bayes factors discussed in the previous section will be applied to an example considered by Biggerstaff (2008). Biggerstaff (2008) considered an example where a comparison is made between West Nile virus (WNV) infection prevalences in field collected *Culex nigripalpus* mosquitoes trapped at different heights. Table 2 summarises the data used by Biggerstaff (2008). The general case for two samples will be considered here.

Table 2: Summary of *Culex nigripalpus* mosquitoes trapped at different heights of $6m$ and $1.5m$.

	Sample 1 **height** $= 6m$	**Sample 2** **height** $= 1.5m$
Total	2 021	1 324
Number of pools	53	31
Average pool size	38.1321	42.7097
Minimum pool size	1	5
Maximum pool size	50	100
Number of positive pools	7	1

Using numerical integration, the Biggerstaff data yielded $B_{01} = 2.3331$ with corresponding posterior probability of $P(\mathscr{M}_0 | \underline{x}) = 0.7000$. This is moderate evidence in favour of model $\mathscr{M}_0$.

Using the sample and pool sizes as given in Biggerstaff (2008) where $M_1 = 19$, with $p_1 = 0.004$, we simulated 10 000 outcomes of the 19×1 vector $\underline{x}_1$. This was done by simulating 19 binomial observations, each with a sample size and a different probability, since the pool sizes differ. The same was done with the second sample where $M_2 = 16$, with $p_2 = 0.001$. Using numerical integration, the Bayes factors and posterior probabilities were calculated and the histograms are shown in Figure 3.

The mean of B_{01} is 3.6241 and the mean posterior probability is 0.6202, still favouring a single p slightly.

It is interesting to note that 626 of the 10 000 simulations gave the same result as the Biggerstaff (2008) data, 7 positives from the samples with p_1 and one positive from the samples with p_2, although not necessarily from samples with the same pool sizes. The range of posterior probabilities for the 626 simulations is $(0.6925, 0.7208)$, with mean of 0.7030. So the pools from which the positive observations come do not have a large affect on the posterior.

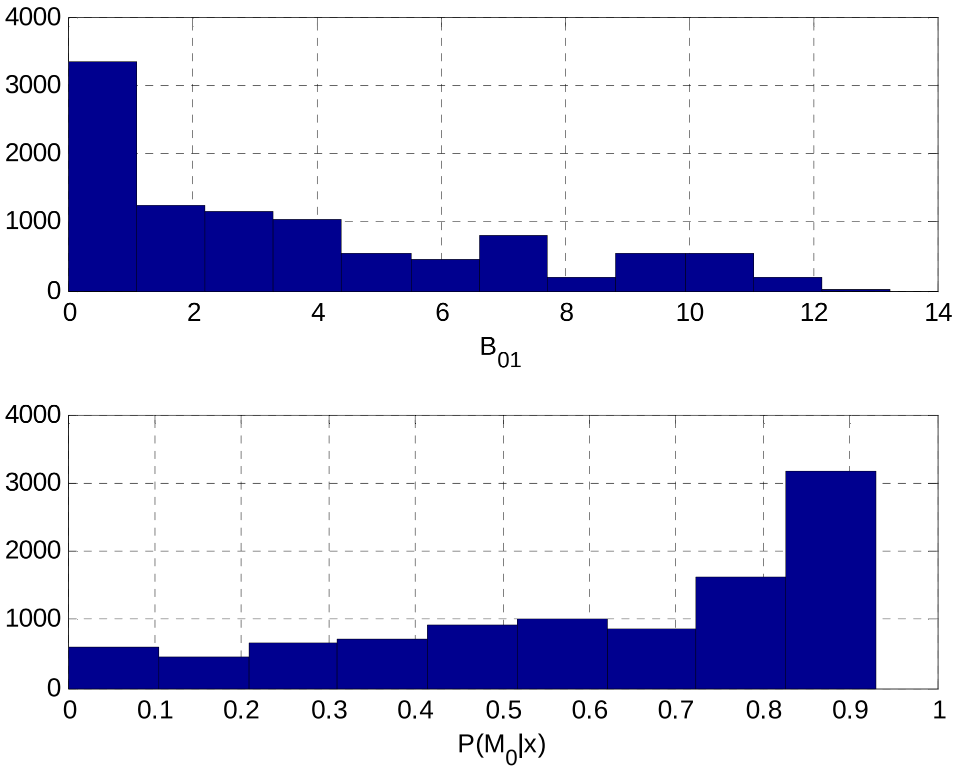

Figure 3: Histograms of the Bayes factor and posterior probabilities.

Kass & Raftery (1995) gave the following categories for interpreting the Bayes factor, B_{10}:

$\log_{10}(B_{10})$	B_{10}	Evidence against H_0
0 to 0.5	1 to 3.2	Not worth more than a bare mention
0.5 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Using these scales and categories to judge the evidence against $\mathscr{M}_0$ for B_{01}, we obtain the following results:

- 85.12% of the time, the evidence was poor;
- 9.06% of the time, it was substantial;
- 5.49% of the time, it was strong;
- 0.33% of the time, it was decisive.

6 Conclusion

In this paper we looked at the Bayes factor to grouped data. We also considered factional Bayes factors. The Bayes factor was applied to an example considered in Biggerstaff (2008), where a comparison was made between West Nile virus (WNV) infection prevalences in field collected *Culex nigripalpus* mosquitoes trapped at different heights. The two sample case with $M_1 = M_2 = 1$ was first considered, where two priors were used $beta\,(\alpha = 1/2, \beta = 1/2)$ and $beta\,(\alpha = 1, \beta = 1)$. The posterior probabilities were usually lower with the smaller values of α and β, except for small n. For the fractional Bayes factor two priors were considered a $beta\,(\alpha = 1/2, \beta = 1/2)$, Jeffreys prior, and a $beta\,(\alpha = 0, \beta = 0)$, Haldane prior. The probabilities when using the Haldane prior are considerably higher than those from the Jeffreys prior.

For the general case a $beta\,(1/2, 1/2)$ prior was used for the Bayes factor. Using numerical integration, the Bayes factors and posterior probabilities were calculated. The mean of B_{01} is 3.6241 and the mean posterior probability is 0.6202, favouring a single p slightly.

References

Biggerstaff, B. J. (2008). Confidence intervals for the difference of two proportions estimated from pooled samples. *Journal of Agricultural, Biological, and Environmental Statistics*, 13, 478 – 496.

Chick, S. E. (1996). Bayesian models for limiting dilution assay and group test data. *Biometrics*, 52, 1055 – 1062.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14, 436 – 440.

Gastwirth, J. L. & Johnson, W. O. (1994). Screening with cost-effective quality control: Potential applications to HIV and drug testing. *Journal of the American Statistical Association*, 89, 972 – 981.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28, 55 – 61.

Hanson, T. E., Johnson, W. O., & Gastwirth, J. L. (2006). Bayesian inference for prevalence and diagnostic test accuracy based on dual-pooled screening. *Biostatistics*, 7, 41 – 57.

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

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773 – 795.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society Series B*, 57, 99 – 138.

Robert, C. P. (2001). *The Bayesian Choice*. Springer, New York, 2nd edition.

Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's Theory of Probability revisited. *Statistical Science*, 24, 141 – 172.

Zellner, A. (1977). Maximal data information priors. In *New Developments in the Applications of Bayesian Methods*, Aykac, A. & Brumat, C. (eds). North-Holland-Publishing Company, Amsterdam, 211 - 232.