

CONDITIONAL BAYESIAN HYPOTHESIS TESTING FOR 2×2 TABLES

M. Ganjali

Department of Statistics, Faculty of Mathematical Sciences
Shahid Beheshti University, Tehran, Iran.
Email: m-ganjali@sbu.ac.ir

and

D. Berridge

Centre for Applied Statistics, Fylde College,
Lancaster University, Lancaster LA1 4YF, U.K.
Email: d.berridge@lancaster.ac.uk

ABSTRACT

In this paper, after discussing the conservative nature of some tests with small sample sizes, a conditional Bayesian approach for Fisher's exact test is presented. In this approach the non-null conditional distribution of Fisher's exact test statistic is reparametrized to simplify the choice of a natural prior distribution. The acceptance-rejection algorithm is used to generate random samples from the posterior distribution and then the Bayes factor is used as a summary of evidence. The robustness of results to different prior distributions is also discussed. In an application, we find that the Bayesian approach is less conservative than Fisher's exact and Yates' continuity tests.

KEYWORDS

Fisher's exact test; Small sample; Mid P-value; Randomized test; Odds ratio; Acceptance-rejection algorithm.

1. INTRODUCTION

Hypothesis Tests, using classical methods, have high power when the sample size is large. However, in some applications, where the sample size is small, classical methods are so conservative. There are examples in which researchers intuitively see some evidence against the null hypothesis, but for any possible value of the test statistic that a classical method suggests, they have to say "there is no or weak evidence to reject H_0 " (two examples, in those this may happen, are, Fisher's exact test, Fisher, 1935a, pp. 11-25, and Wilcoxon's rank-sum test, Wilcoxon, 1945).

There are two classical approaches for exact test inference of 2×2 contingency tables. These are unconditional and conditional approaches. The Bayesian approach may also be used in two different ways, unconditionally or conditionally. In the unconditional approach, one starts with independent binomial sampling in the two rows and then assumes a joint prior density for the success probabilities in the two rows. Thus, finding a joint posterior distribution (or some transformation of it) and integrating out the nuisance

parameter from this joint posterior distribution leads to the marginal posterior density of the parameter of interest (Johnson and Albert, 1999, Ch. 2).

In this paper, we shall use a conditional Bayesian approach. This approach may be used for two different sampling frameworks of which the first one is conditioning on all row and column totals, i.e. the experimental design will be the same as that of Fisher in his famous example. In the second sampling framework, sampling is initially unconditional, but for analyzing data, conditioning on sufficient statistics for nuisance parameters eliminates these parameters. For example, for Poisson or full multinomial sampling, conditioning on row totals gives binomial sampling and statistical independence is equivalent to homogeneity. Under independence, conditioning as well on column totals to eliminate the nuisance parameter yields the hypergeometric distribution (used in Fisher's exact test). This approach is also used by Boschloo (1970) who showed that Fisher's conditional test, when maximized with respect to the nuisance parameter performs well compared to other exact unconditional test. Thus we no longer, need to specify a prior distribution for the nuisance parameters which are not of main interest.

The details of our conditional Bayesian approach are given in the next Section where we use the acceptance-rejection algorithm to generate random samples from the posterior distribution and then explain how to use the Bayes factor as a summary of evidence. In Section 3 sensitivity of the results with respect to different prior distributions is investigated. In Section 4, we have an applied example. In Section 5 conclusion is given.

2. THE BAYESIAN APPROACH FOR FISHER EXACT TEST AND BAYES FACTOR

In this section the Bayesian approach for fisher exact test is presented and Bayes factor as a summary of evidence for testing independence in 2×2 tables is given.

2.1 The Bayesian Approach for Fisher Exact Test

Let us test independence (H_0) against positive correlation in a two by two contingency table with given column and row margins. Suppose N_{11} is the number of events in the cell at the intersection of the first row and the first column. Under H_0 , N_{11} has the hypergeometric distribution. Independence can be rejected against positive correlation if N_{11} is large. For this test, the P-value is $P_{H_0}(N_{11} \geq n_{11})$ where n_{11} is the observed value of N_{11} .

Suppose in Fisher's tea taster example, Fisher's colleague is given 4 cups of tea (two cups had milk added first, and the other two had tea added first) to distinguish whether milk or tea was added to the cup first. Suppose also that she distinguished all cups correctly ($n_{11} = 2$). For this example the P-value is $\frac{1}{6} = 0.167$. It can be seen that $n_{11} = 2$ is the best value in favor of H_1 . Since the colleague was able to distinguish all cups correctly, it seems, intuitively, there is evidence against H_0 . However, the classical method with regard to the P-value will say "there is no evidence to reject H_0 ".

A Bayesian approach in this example may be chosen as a better option. Let us reparametrize the non-null distribution of Fisher's exact test statistic in order to simplify the choice of a prior distribution. The non-null conditional distribution is:

$$f(n_{11} | n, n_{1+}, n_{+1}; \theta) = \frac{\binom{n_{1+}}{n_{11}} \binom{n - n_{1+}}{n_{+1} - n_{11}} \theta^{n_{11}}}{\sum_u \binom{n_{1+}}{u} \binom{n - n_{1+}}{n_{+1} - u} \theta^u}; \quad \theta > 0$$

where n_{1+} and n_{2+} are row margins, n_{+1} is the first column margin, and n is the sample size. The index of summation ranges from $\text{Max}(0, n_{1+} + n_{+1} - n)$ to $\text{Min}(n_{1+}, n_{+1})$, the possible values for the given marginal totals, and θ is the odds ratio, which takes value 1 under the null hypothesis of independence. This is the noncentral hypergeometric distribution (Fisher, 1935b, Agresti, 2002).

Let reparametrize this distribution using $\varphi = \ln \theta$ in order to provide a natural prior distribution for φ . The log odds ratio (φ) is 0 when two variables are independent and it is symmetric about 0. It is also known that the empirical estimate of the log odds ratio based on the observed data is approximately normally distributed in studies of even moderate sample sizes. This reparametrization helps because we can use the normal distribution as a natural prior distribution for the log odds ratio. Suppose $\pi(\varphi)$ is the chosen prior distribution. The posterior distribution is

$$\pi(\varphi | n, n_{1+}, n_{+1}, n_{11}) = \frac{f(n_{11} | n, n_{1+}, n_{+1}; \varphi) \pi(\varphi)}{\int_{-\infty}^{\infty} f(n_{11} | n, n_{1+}, n_{+1}; t) \pi(t) dt}$$

Now, the test of independence against positive correlation is the test of $H_0 : \varphi = 0$ against $H_1 : \varphi > 0$.

We can use a direct simulation to obtain random samples from the posterior distribution. Suppose $\pi(\varphi)$ is chosen to be a proper prior and c is chosen to be a positive constant such that $f(n_{11} | n, n_{1+}, n_{+1}; \varphi) \pi(\varphi) < c \pi(\varphi)$. If generating random values from the prior distribution is computationally tractable, one may use the following acceptance-rejection algorithm:

1. Simulate φ from the prior distribution, and U uniformly on $(0,1)$.
2. If $U < \frac{f(n_{11} | n, n_{1+}, n_{+1}; \varphi)}{c}$, then accept φ as a draw from the posterior distribution. If not, reject φ and try again.

The algorithm is repeated until the desired sample size is obtained.

2.2 Bayes Factor

The Bayes factor in favor of the null hypothesis ($H_0 : \varphi = 0$) versus $H_1 : \varphi > 0$ for Fisher's exact test is:

$$B_{\varphi=0, \varphi>0} = \frac{f(n_{11} | n, n_{1+}, n_{+1}; \varphi = 0)}{\int_0^{\infty} f(n_{11} | n, n_{1+}, n_{+1}; t) g(t) dt}$$

where $g(\varphi)$, is a proper prior defined on $\varphi > 0$ (for definition of Bayes factor see Good, 1950 who names Bayes factor as "the Factor"; Jeffreys, 1961, Section 5 who does not use the phrase Bayes factor but simply denotes the Bayes factor by K and always refers to it as such, and Lee, 2004, Section 4.1). For Fisher's exact test let us assume a normal prior distribution for the log odds ratio with empirical mean

$$\mu = \ln \frac{(n_{11} + 1/2)(n_{22} + 1/2)}{(n_{12} + 1/2)(n_{21} + 1/2)}$$

and variance

$$\sigma^2 = [1/(n_{11} + 1/2) + 1/(n_{12} + 1/2) + 1/(n_{21} + 1/2) + 1/(n_{22} + 1/2)]$$

for a two-sided test ($\varphi = 0$ versus $\varphi \neq 0$). A choice for $g(\varphi)$ in one-sided test ($\varphi = 0$ versus $\varphi > 0$) is $N(\mu, \sigma^2)$ truncated at 0 which gives

$$B_{\varphi=0, \varphi>0} = \frac{[1 - \Phi(-\mu/\sigma)] f(n_{11} | n, n_{1+}, n_{+1}; \varphi = 0)}{\int_{-\infty}^{\infty} I(\varphi > 0) f(n_{11} | n, n_{1+}, n_{+1}; \varphi) \frac{1}{\sigma} \phi\left(\frac{\varphi - \mu}{\sigma}\right) d\varphi}$$

where $I(\varphi > 0)$ is an indicator function which is 1 if $\varphi > 0$ and 0 otherwise, $\Phi(\cdot)$ and $\phi(\cdot)$ are distribution and density functions of the standard normal distribution, respectively. This form helps us to approximate the mean of $I(\varphi > 0) f(n_{11} | n, n_{1+}, n_{+1}; \varphi)$ using sample means of the simulated values of the normal distribution with mean μ and variance σ^2 .

This empirical prior can be chosen if we believe that Fisher's colleague would have guessed for many other sets of 4 cups in the same manner (the same distribution for φ) as she did for the current 4 cups. This method can be regarded as an empirical Bayes approach where one chooses the prior after observing the data.

Small values of Bayes factor indicate evidence against H_0 . Jeffreys (1961, Appendix B) states that values of $B_{\varphi=0, H_1}$ between 0.1 and $10^{-\frac{1}{2}} = 0.316$ indicate moderate or substantial evidence against H_0 , values of $B_{\varphi=0, H_1}$ between 0.01 and $10^{-\frac{3}{2}} = 0.0316$

indicate strong evidence against H_0 , and values of $B_{\varphi=0,H_1}$ less than 0.0316 indicate decisive (positive) evidence against H_0 .

2.3 Other Choices of Prior

Although the mean estimate of $\hat{\mu} = \ln \frac{(n_{11}+1/2)(n_{22}+1/2)}{(n_{12}+1/2)(n_{21}+1/2)}$ and the variance estimate

$\hat{\sigma}^2 = [1/(n_{11}+1/2) + 1/(n_{12}+1/2) + 1/(n_{21}+1/2) + 1/(n_{22}+1/2)]$ for the log odds ratio are used empirically in categorical data analysis, they are not what Bayesians call empirical Bayes estimates of μ and σ^2 . To obtain the empirical Bayes estimates of these parameters, assuming a normal distribution for φ with mean μ and variance σ^2 we should maximize

$$m(n_{11} | \mu, \sigma^2) = \int_{-\infty}^{\infty} f(n_{11} | n, n_{1+}, n_{+1}; \varphi) \pi(\varphi | \mu, \sigma^2) d\varphi,$$

with respect to μ and σ^2 which can be done using, for example, function `optim` in R. We preferred to use the corrected estimators because of their simpler form and the possibility of observing zero cell counts. Gart and Zweifel (1967) showed also that, in terms of bias and mean-squared error, the corrected estimators behave well.

The other choice of prior is the Jeffreys' prior:

$$\pi(\varphi) \propto \left[\frac{1}{\text{Var}(n_{11} | \varphi)} \right]^{1/2}.$$

However, the $\text{Var}(n_{11} | \varphi)$ in Jeffreys' prior has a complicated form and the prior may not be a proper prior.

3. SENSITIVITY TO DIFFERENT PRIOR DISTRIBUTION

Let us check the sensitivity of the results with respect to different prior distributions. Suppose $n_{11} = 2$ is observed in Fisher's exact test with $n=4$, $n_{+1}=2$ and $n_{1+}=2$. Table 1 shows the posterior mean, posterior standard error and posterior probability of $\varphi > 0$ for prior distributions which are normal with mean 0 and variances equal to 1, 2, 3, 4, 5, 10, and 50 respectively. This table also shows the Bayes factors for the one-sided test, when $g(\varphi)$ is $N(0, \sigma^2)$ for $\sigma^2 = 1, 2, 3, 4, 5, 10$ and 50 truncated at 0.

For small variances, the prior is more informative and is in favor of the null hypothesis. For higher variances, the prior is less informative as it tends to flatten out. The less informative the prior, the more likely it is that the null hypothesis of independence will be rejected.

Table 1:
Posterior mean, standard error (SE) and probability of $\varphi > 0$;

when $n_{11} = 2$ and prior distribution is $N(0, \sigma^2)$; the Bayes factors are given for one-sided test of independence versus positive association using prior distribution to be $N(0, \sigma^2)$ truncated at 0.

σ^2	Posterior mean	Posterior S.E.	$pr(\varphi > 0 n_{11} = 2)$	Bayes Factor
1	0.758	0.872	0.808	0.489
2	1.245	1.133	0.866	0.399
3	1.615	1.309	0.896	0.355
4	1.905	1.449	0.913	0.327
5	2.157	1.573	0.924	0.308
10	3.072	2.044	0.952	0.261
50	6.450	4.209	0.982	0.204

It can be inferred from our example that if we generate from the non-null distribution, say 1000 tables, with $\varphi > 0$ in this example ($n = 4, n_{+1} = 2, n_{1+} = 2$), Fisher's exact test would never reject the null hypothesis of independence, but the Bayesian approach, which is of course dependent on our choice of prior, would reject the null hypothesis in some cases. If one uses the empirical prior, the power function of the Bayesian approach, $pr(n_{11} = 2 | n = 4, n_{+1} = 2, n_{1+} = 2; \varphi)$, is an increasing function of φ (for $\varphi = 2$ it is 0.641 and for $\varphi = 3$ it is 0.832), in contrast to the insensitivity of Fisher's exact test to values of φ .

4. APPLICATION

Essenberg (1952) conducted a clinical trial to investigate the effect of tobacco on tumor risk. To this end, 72 albino mice were randomly divided into two groups. One group was kept in a chamber filled with a certain dose of tobacco smoke, one cigarette per hour. The other (control) group was kept in a chamber without smoke. After a year, the 55 surviving mice were sacrificed and autopsied for tumors. The data are presented in table 2. We are interested in the null hypothesis that treatment with smoke has no effect on tumor rate. For this example, Fisher's exact test gives a P-value=0.013 for a two-sided test, Yates' continuity correction test gives a P-value=0.021 and a chi-squared test gives a P-value=0.009. So, Yates' corrected chi-squared and Fisher's exact test are more conservative than the chi-squared test.

Table 2:
Tumor prevalence among mice by exposure to tobacco smoke [Source: C.J. Lloyd (1999), Statistical analysis of categorical data. John Wiley & Sons, New York].

	Presence or absence of tumor		Total
	Tumor	No tumor	
Smoke treatment	21	2	23
Control	19	13	32
Total	40	15	55

Using a normal distribution for φ with empirical mean and variance of the estimate of φ (1.784 and 0.572 respectively), we found $B_{\varphi=0, \varphi \neq 0} = 0.029$. All tests show that

there is strong evidence against the null hypothesis. So, treatment with smoke has a strong effect on tumor rate. However, for this application, Fisher's exact test and Yates' corrected chi-squared test are more conservative than our Bayesian approach.

5. CONCLUSION

A Bayesian approach for Fisher's exact test has been described. As an empirical and natural prior, a normal distribution has been used for the logarithm of the odds ratio. Sensitivity analyses have demonstrated that results are robust to changes in the prior close to the empirical prior. In our application the conditional Bayesian approach was less conservative than Fisher's exact test and Yates' corrected chi-squared test. A comparison of the unconditional and conditional Bayesian approaches is an ongoing research on our part.

ACKNOWLEDGMENT

The first author would like to thank the Research Council of Shahid Beheshti University of Tehran for supporting the research grant.

REFERENCES

1. Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.
2. Boschloo R.D. (1970). Raised conditional level of significance for the 2×2 table when testing the equalities of probabilities. *Statistica Neerlandica*, 24, 1-35.
3. Essenberg, J.M. (1952). Cigarette smoke and the incidence of primary neoplasm of the lung in albino mice. *Science*, 116, 561-562.
4. Fisher, R.A. (1935a). *The Design of Experiments*. Edinburgh, London: Oliver and Boyd.
5. Fisher, R.A. (1935b). The logic of inductive inference (with discussion). *J. Roy. Statist. Soc. Ser. A*, 98, 39-82.
6. Gart, J.J. and Zweifil, J.R., (1967). On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biometrika*, 54, 181-187.
7. Good, I.J. (1950). *Probability and the weighting of evidence*. London: Griffin.
8. Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.
9. Johnson, V.E. and Albert, J.H. (1999). *Ordinal Data Modeling*. Springer Verlage, USA.
10. Lee, P.M. (2004). *Bayesian Statistics: An Introduction*. Third Ed., Arnold.
11. Lloyd, C.J. (1999). *Statistical Analysis of Categorical Data*. John Wiley & Sons, New York.
12. Wilcoxon, F. (1945). Individual comparison by ranking methods. *Biometrics*, 1, 80-83.