Journal of Modern Applied Statistical Methods

Volume 14 | Issue 2 Article 16

11-1-2015

The Bayes Factor for Case-Control Studies with Misclassified Data

Tzesan Lee
Centers for Disease Control & Prevention, leetzesan@gmail.com

The Bayes Factor for Case-Control Studies with Misclassified Data

Cover Page Footnote
All the calculations were done on the EXCEL spreadsheet.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 201-218.

ISSN 1538 − 9472

The Bayes Factor for Case-Control Studies with Misclassified Data

Tzesan Lee
Centers for Disease Control & Prevention
Atlanta, GA

The question of how to test if collected data for a case-control study are misclassified was investigated. A mixed approach was employed to calculate the Bayes factor to assess the validity of the null hypothesis of no-misclassification. A real-world data set on the association between lung cancer and smoking status was used as an example to illustrate the proposed method.

Keywords: Bayes factor, Misclassification, *p*-value.

Introduction

Misclassification is a ubiquitous problem in epidemiologic studies. Particularly, it often occurs if the data are obtained from the proxy or surrogate (Nelson, Longstreth, Koesell, and van Belle 1990). Methods for dealing with misclassified data from case-control studies have been widely studied. See, for example, Kleinbaum, Kupper & Morgenstern (1982), Fleiss, Levin & Paik (2003), and Rothman, Greenland & Lash (2008). Almost all studies make an assumption in the beginning that the collected data are misclassified. Yet how to test the validity of this assumption has not been addressed.

These issues can also be considered from a Bayesian perspective. First, the misclassification probabilities are included in both the null and alternative hypothesis. Second, bias-adjusted estimators for the proportion of exposure in cases or controls are presented. Third, the uniform and the Beta distributions are adopted respectively as the prior distribution for the misclassification probability and population proportion parameter in cases or controls. Finally, the lower-bound for the Bayes factor is calculated. A real-world data set was used as an example to illustrate the proposed method. A comparison between the *p*-value and the Bayes factor is made.

Tzesan Lee was retired from the Centers for Disease Control and Prevention and is currently working as President, Applied Math Press, LLC. Email at leetzesan@gmail.com.

Methodology

Consider the data for case-control studies given in Table 1. The random variable E^* denotes the classified surrogate for the true exposure variable E, while the variable D indicates the disease status of the subjects with $D = 1$ and $D = 0$ representing cases and controls respectively. Suppose that E^* is misclassified, but D is not misclassified.

Table 1. Case-control studies with misclassified data

Classified exposure status	Group of subjects	
	***D* = 1 (cases)**	***D* = 0 (controls)**
E^* = 1 (exposed)	n_{11}	n_{10}
E^* = 0 (unexposed)	n_{01}	n_{00}
Sample size	$n_{[1]}$	$n_{[0]}$

It is well known that the traditional sample proportion estimator of the exposed group given by

$$\hat{p}_i = n_{ji} / n_{[i]}, \hat{q}_i = 1 - \hat{p}_i \tag{1}$$

In terms of the sensitivity and specificity defined by

$$\varphi_i = \Pr\left(E^* = 1 \middle| E = 1, D = i\right), \bar{\varphi}_i = 1 - \varphi_i \tag{2}$$

$$\psi_i = \Pr\left(E^* = 1 \middle| E = 0, D = i\right), \bar{\psi}_i = 1 - \psi_i \tag{3}$$

it was shown (Lee, 2009) that

$$E\left(\hat{p}_i\right) = \varphi_i p_i + \left(1 - \psi_i\right) q_i = p_i \cdot \Delta_i + 1 - \psi_i \tag{4}$$

$$E\left(\hat{q}_i\right) = \left(1 - \varphi_i\right) p_i + \psi_i q_i = q_i \cdot \Delta_i + 1 - \varphi_i \tag{5}$$

From Equations 4 and 5 it is seen that the traditional sample proportion estimators, $\hat{p}_i$ and $\hat{q}_i$, are no longer unbiased. By solving Equations 4 and 5 with the left-side $E(\hat{p}_i)$ or $E(\hat{q}_i)$ being replaced by $\hat{p}_i$ or $\hat{q}_i$, it follows

$$\breve{p}_i = (\psi_i - \hat{q}_i)/\Delta_i, \tag{6}$$

$$\breve{q}_i = (\varphi_i - \hat{p}_i)/\Delta_i, \tag{7}$$

where

$$\Delta_i = \varphi_i + \psi_i - 1, \; i = 0, 1. \tag{8}$$

Equations 6 and 7 are called the bias-adjusted proportion (BAP) estimators of p_i and q_i. The BAP estimators are said to be admissible if they are greater than zero but less than one plus their sum equals to one. Evidently, the following constraints are required to be imposed on the sensitivity and specificity in order for Equations 6 and 7 to be admissible (Lee, 2009):

$$\begin{aligned} \varphi_i &> \hat{p}_i, \\ \psi_i &> \hat{q}_i, \\ \varphi_i + \psi_i &> 1. \end{aligned} \tag{9}$$

A concern is aimed at testing whether the given data in Table 1 are misclassified - whether the exposure rates for cases and control are the same. This can be tested through the hypothesis testing which is formulated as follows:

$$H_0 : \varepsilon_{RD} = 0 \quad \text{versus} \quad H_1 : \varepsilon_{RD} \neq 0, \tag{10}$$

where $\varepsilon_{RD} = p_1 - p_0$, the subscript "*RD*" means the rate difference. However, Equation 10 can't be used to test whether the observed data of Table 1 are misclassified. In order to test if the data are misclassified, the hypotheses of Equation 10 has to be enlarged by including the misclassification probabilities associated with both cases and controls given as follows:

$$H_0 : \varepsilon_{RD} = 0, \bar{\varphi}_i = \bar{\psi}_i = 0 \text{ versus } H_1 : \varepsilon_{RD} \neq 0, \bar{\varphi}_i \neq 0, \bar{\psi}_i \neq 0, \; i = 0, 1, \tag{11}$$

To test the hypotheses of Equation 11, a mixed Bayesian approach is taken to tackle this problem (Kass & Raftery, 1995).

Let

$$\breve{\varepsilon}_{RD} = \breve{p}_1 - \breve{p}_0 - \varepsilon_{RD} \tag{12}$$

It can be shown

$$E\left(\breve{\varepsilon}_{RD}\right) = 0, \tag{13}$$

$$\begin{aligned} Var\left(\breve{\varepsilon}_{RD}\right) &= Var\left(\breve{p}_1\right) + Var\left(\breve{p}_0 + \varepsilon_{RD}\right) \\ &= \sum_{i=0}^{1}\left(p_i \cdot \Delta_i + 1 - \psi_i\right)\left(q_i \cdot \Delta_i + 1 - \varphi_i\right) \cdot n_{[i]}^{-1} \end{aligned} \tag{14}$$

Define

$$\breve{x}_{RD} = \breve{\varepsilon}_{RD}^2 \big/ Var\left(\breve{\varepsilon}_{RD}\right) \tag{15}$$

To assess the evidence in favor of supporting the null against the alternative hypothesis of Equation 11, the Bayes factor for favoring H_0 relative H_1 from using Equation 15 can be calculated as follows:

$$B^g\left(\breve{x}_{RD}\right) = \frac{f\left(\breve{x}_{RD} \middle| H_0\right)}{m_g\left(\breve{x}_{RD}\right)} \tag{16}$$

where

$$m_g\left(\breve{x}_{RD}\right) = \iint_{R\times\Omega} f\left(\breve{x}_{RD} \middle| H_1\right) \prod_{i=0}^{1} h_0\left(\varphi_i, \psi_i\right) g\left(p_i, q_i\right) d\varphi_i d\psi_i dp_i dq_i \tag{17}$$

$f\left(\breve{x}_{RD} \mid H_1\right)$ is the central chi-square distribution with one degree of freedom, $g\left(p_i, q_i\right) = \Gamma\left(\eta+\tau\right) p_i^{\eta-1} q_i^{\tau-1} \big/ \left[\Gamma\left(\eta\right)\Gamma\left(\tau\right)\right]$, the beta distribution with the parameters η and τ over [0, 1], and $h_0\left(\varphi_i, \psi_i\right) = \left[\bar{\varphi}_i \bar{\psi}_i\right]^{-1}$ is the uniform distribution over $\Omega_i = [a_i ,1] \times [b_i,1]$, where a_i and b_i are specified in the Appendix. Although the posterior marginal probability density function of m_g (Equation 17) depends on two hyper-parameters η and τ, a Bayes/non-Bayes compromise rather than a type III hyper-distribution for η and τ is adopted to estimate η and τ (Good & Crook, 1974). As a result, the parameters η and τ are estimated by employing the likelihood method. The maximum likelihood estimators for η and τ and the relative maximum value of m_g of Equation 17 are denoted respectively by $\left(\eta_{\max}, \tau_{\max}\right)$ and $m_g^{\max} = m_g\left(\eta_{\max}, \tau_{\max}\right)$. Thus, define the lower bound of the Bayes factor (Equation 16) as follows:

$$\underline{B}^g = f\left(\breve{x}_{RD} \middle| H_0\right) \big/ m_g^{\max} \tag{18}$$

The details of calculating Equation 18 are given in the Appendix.

Example

Although there is some evidence of a greater than average risk in some occupations to have the lung cancer, these occupations could not account for the general increase in pulmonary cancer. It is thought of interest to select a particular population group, homogeneous economically, with little occupational exposure to respiratory irritants and with equal access to diagnostic facilities. Physicians are believed to represent such a group. Wynder and Cornfield (1953) reported a study on the exposure to tobacco and other possible respiratory irritants of 63 physicians with lung cancer and 133 physicians with cancers in areas where respiratory irritants are not believed to play a part. Among these 133 physicians, 43 cases were cancer of stomach and kidney, 45 cases cancer of colon and lymphoma, and 45 cases cancer of bladder, leukemia and sarcoma. The data in Table 2 is taken from Cornfield (1956) who only used 43 cases from cancer of stomach and kidney as a control group. The non-smoker is defined to be those who smoked the equivalent of less than 1 cigarette a day. Here it is of interest to test whether the data concerning the smoking status in Table 2 for both cases and controls are misclassified.

Table 2. The data of physicians with and without lung cancer by smoking status

Smoking status	Lung cancer patients	Controls
Smoker	60	32
Nonsmoker	3	11
Total	63	43

Before calculating the Bayes factor, the data in Table 2 are first to be checked if the two required conditions are satisfied before using the formula derived in the Appendix. Because $\hat{p}_1 = 0.952381 > \hat{p}_0 = 0.744186$ and $\hat{\sigma}_{\hat{p}_1} = \sqrt{n_{[1]}^{-1}\hat{p}_1\hat{q}_1} = 0.027 > \hat{\sigma}_{\hat{p}_0} = \sqrt{n_{[0]}^{-1}\hat{p}_0\hat{q}_0} = 0.067$, where $n_{[1]} = 63$, $n_{[0]} = 43$, the two required conditions are indeed being satisfied; hence it was free to use the formula in the Appendix. Let $a_i = \hat{p}_i + 0.005$ and $b_i = \hat{q}_i + 0.005$, $i = 0,1$, be substituted into Equations A17 to A11, it follows that $\dot{M}_{[1,1,0,0]} = 1.1011$, $M_{[1,0,1,0]} = 0.0828$, $M_{[1,0,0,1]} = -0.0037$, $\dot{M}_{[1,1,0,1]} = 0.0513$, $M_{[1,0,1,1]} = 1.2369$, $\dot{M}_{[0,1,0,0]} = 1.1169$, $M_{[0,0,1,0]} = 0.6287$, $M_{[0,0,0,1]} = -0.0567$, $\dot{M}_{[0,1,0,1]} = 0.4819$, and $M_{[0,0,1,1]} = 4.8652$. Then, substituting the above information into Equations A12 and A14, this leads to that $N_0 = 0.1957$, $N_1 = 5.4652$, $N_2 = -31.4597$, $R_0 = 0.0016$, $R_1 = 0.1967$, $R_2 = -0.0041$, $R_3 = 0.0704$, $R_4 = 0.234$, $R_5 = -0.0252$, $R_6 = -0.1988$, and $a = 133.5876$. Again, by substituting the above information into Equations A13 and A16, it follows that

$$m_g^{(1)}(\eta,\tau) \equiv \frac{-400.8(\eta+\tau)\left\{\begin{array}{l}\eta\tau(\eta+\tau)\left[0.003\eta(\eta+\tau)-0.002\right] \\ +0.017\eta\tau(\eta+\tau)+0.002\eta\big]+0.009\tau\end{array}\right\}+5.97\tau}{2\sqrt{\eta\tau}\eta^2(\eta+\tau)^3} \tag{19}$$

and

$$m_g^{(2)}(\eta,\tau) \equiv \frac{2.33\eta\tau(\eta+\tau)^2 + 2.23\tau(\eta+\tau) - 3.82}{2\left[\sqrt{\eta\tau}(\eta+\tau)\right]^3} \tag{20}$$

Consequently, $m_g(\eta, \tau)$ was readily obtained from substituting Equations 19 and 20 into Equation A17.

To find the relative maximum of $m_g(\eta, \tau)$, the 2-dimensional unit square $[0,1] \times [0,1]$ was partitioned into 100 lattice points $(0.1, 0.1), (0.1, 0.2), \ldots, (1.0, 0.9), (1.0, 1.0)$ and then evaluated the function value of $m_g(\eta, \tau)$ at these lattice points. After identifying the proximity of the relative maximum a finer neighborhood was then searched to locate it. Equation A17 was found to have a unique relative maxima: $m_g^{\max}(0.14, 1.0) = 2.15$. The value of $f(\breve{x}_{RD} \mid H_0)$ was evaluated directly from the probability density function of the central chi-square distribution with one degree of freedom; hence we have $f(\breve{x}_{RD} \mid H_0) = 6.4 \times 10^{-6}$. After dividing the value of $f(\breve{x}_{RD} \mid H_0) = 6.4 \times 10^{-6}$ by $m_g^{\max} = 2.15$, we thus obtained the lower bound of the Bayes factor given by $\underline{B}^g(\breve{x}_{RD}) = 3.0 \times 10^{-6}$.

Since $\breve{x}_{RD} \big| H_0 = \hat{x}_{RD} = \hat{p}_D^2 \big/ Var(\hat{p}_D) = 19.1$ (p-value = 1.2×10^{-5}), where $\hat{p}_D \equiv \hat{p}_1 - \hat{p}_0$, the null hypothesis H_0 was rejected for Table 2. Yet, the evidence from the lower bound of the Bayes factor ($\underline{B}^g(\breve{x}_{RD})$ = 3.0×10^{-6}) was in favor of supporting H_1 (Equation 11) by at most a factor of "3.3×10^5 to 1". Hence the data in Table 2 are likely to be misclassified.

Discussion

Although both the p-value and the Bayes factor rejected the null hypothesis H_0 with respect to the data in Table 2, the p-value seemed much inclined to reject the null hypothesis H_0 in Equation 10 rather than that in Equation 11. In other words, the p-value is inadequate to reject the null hypothesis in Equation 11. This study provides another example to corroborate the p-value fallacy (Goodman 1999a, Goodman 1999b).

Because the Beta distribution which is the conjugate family of the binomial distribution was used as the prior distributions, the Bayes factor could of course change accordingly if other family of distributions is used as the prior distribution (Delampady & Berger, 1990).

The derivation of the formula provided in the Appendix was based on the two assumptions: (i) $p_1 > p_0$, and (ii) $\sigma_{\hat{p}_1} = \sqrt{n_{[1]}^{-1} p_1 q_1} < \sigma_{\hat{p}_0} = \sqrt{n_{[0]}^{-1} p_0 q_0}$. These two assumptions can be verified if it is valid by substituting the crude prevalence estimator ($\hat{p}_i$, $i = 0, 1$) into the inequality. Should the both of the two assumptions fail to be satisfied, all we need to do is to switch the index accordingly for cases and controls before using the formula provided in the Appendix. However, if only one of the assumptions is violated, Equation A4 has to be revised accordingly.

References

Askey, R. A. & Roy, R. (2010). Gamma function, *NIST handbook of mathematical functions* (pp.135). F. W. Olver, D. W. Lozier, R. F. Boisvert, C. W. Clark (Eds.), United Kingdom: National Institute of Standards and Technology, Cambridge University Press.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 135-148). Berkeley: University of California Press.

Delampady, M. & Berger, J. O. (1990). Lower bounds on Bayes factors for multinomial distributions with application to chi-squared tests of fit. *The Annals of Statistics, 18*(3), 1295-1316. doi:10.1214/aos/1176347750

Fleiss, J., Levin, B. & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York: John Wiley & Sons.

Good, I. J. & Crook, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of American Statistical Association*, *69*(347), 711-720. doi:10.2307/2286006

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The P-value fallacy, *Annals of Internal Medicine, 130*(12), 995-1004. doi:10.7326/0003-4819-130-12-199906150-00008

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor, *Annals of Internal Medicine*, *130*(12), 1005-1013. doi:10.7326/0003-4819-130-12-199906150-00019

Kass, R. E. and Raftery, A. E. (1995). Bayes factor, *Journal of American Statistical Association, 90*(430), 773-795. doi:10.1080/01621459.1995.10476572

Kleinbaum, D. G., Kupper, L. L. & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods*. Belmont, CA: Lifetime Learning.

Lee, T-S. (2009). Bias-adjusted exposure odds ratio for misclassified data, *The Internet Journal of Epidemiology*, *6*(2), 1-19.

Mietinen, O. & Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine, 4*(2), 213-226. doi:10.1002/sim.4780040211

Nelson, L. M., Longstreth, W. T., Koesell, T. D., & van Belle, G. (1990). Proxy respondents in epidemiologic research. *Epidemiologic Reviews*, *12*(1), 71-86.

Rothman, K. J., Greenland, S. & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Wynder, E. L. & Cornfield, J. (1953). Cancer of the lung in physicians, *New England Journal of Medicine*, *248*(11), 441-444. doi:10.1056/NEJM195303122481101

Appendix

By applying the quadratic approximation to the probability density function of the central chi-square distribution with one degree of freedom in Equation 17, we have

$$\begin{aligned} f\left(\breve{x}_{RD}\middle|\varepsilon_{RD},\varphi_0,\psi_0,\varphi_1,\psi_1\right) &= \frac{1}{\sqrt{2\pi}}\breve{x}_{RD}^{-\frac{1}{2}}e^{-\frac{1}{2}\breve{x}_{RD}} \\ &\approx \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{\breve{x}_{RD}}}\left(1-\tfrac{1}{2}\breve{x}_{RD}+\tfrac{1}{8}\breve{x}_{RD}^2\right) \\ &= \frac{1}{\sqrt{2\pi}}\left[\frac{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}}{\breve{\varepsilon}_{RD}}-\tfrac{1}{2}\frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}}+\tfrac{1}{8}\left(\frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}}\right)^3\right], \end{aligned} \tag{A1}$$

where $\breve{\varepsilon}_{RD}$ and $Var(\breve{\varepsilon}_{RD})$ are given by Equations 12 and 14, respectively.

By using the linear approximation:

$$\left[\left(1-\varepsilon_{RD}^{-1}\left(\breve{p}_1-\breve{p}_0\right)\right]^{-1}\approx 1+\varepsilon_{Rd}^{-1}\left(\breve{p}_1-\breve{p}_0\right),$$

it follows that

$$
\begin{aligned}
\frac{\sqrt{Var\left(\breve{\varepsilon}_{Rd}\right)}}{\breve{\varepsilon}_{Rd}} &= \frac{\sqrt{\begin{array}{l}\Delta_1^{-2} n_{[1]}^{-1}\left(p_1\Delta_1 + 1 - \psi_1\right)\left(q_1\Delta_1 + 1 - \varphi_1\right)\\ +\Delta_0^{-2} n_{[0]}^{-1}\left(p_0\Delta_0 + 1 - \psi_0\right)\left(q_0\Delta_0 + 1 - \varphi_0\right)\end{array}}}{\breve{p}_1 - \breve{p}_0 - \varepsilon_{RD}} \\
&= -\frac{\sqrt{A}}{\varepsilon_{RD}} \cdot \frac{\sqrt{1 + A^{-1}\left\{\sum_{i=0}^{1} n_{[i]}^{-1}\Delta_i^{-1}\left[\left(1 - p_i\varphi_i - q_i\psi_i\right) + \Delta_i^{-1}\bar{\varphi}_i\bar{\psi}_i\right]\right\}}}{1 - \varepsilon_{RD}^{-1}\left(\breve{p}_1 - \breve{p}_0\right)} \\
&= -I^{-1} \cdot \varepsilon_{RD}^{-1} \cdot \sqrt{1 + I^2 J} \cdot \left[1 - \varepsilon_{RD}^{-1}\left(\breve{p}_1 - \breve{p}_0\right)\right]^{-1} \\
&\approx -I^{-1} \cdot \varepsilon_{RD}^{-1}\left(1 + \tfrac{1}{2} I^2 J\right)\left[1 + \varepsilon_{RD}^{-1}\left(\breve{p}_1 - \breve{p}_0\right)\right] \\
&= -I^{-1} \cdot \varepsilon_{RD}^{-1}\left[1 + \varepsilon_{RD}^{-1}\left(\breve{p}_1 - \breve{p}_0\right) + \tfrac{1}{2} I^2 J + \tfrac{1}{2}\varepsilon_{RD}^{-1} I^2 J\left(\breve{p}_1 - \breve{p}_0\right)\right] \\
&= -I^{-1} \cdot \varepsilon_{RD}^{-1}\left\{\begin{array}{l}1 + \varepsilon_{RD}^{-1}\left[\Delta_1^{-1} u\left(\varphi_1\right) - \Delta_0^{-1} u\left(\varphi_0\right)\right] + \tfrac{1}{2} I^2 J \\ + \tfrac{1}{2}\varepsilon_{RD}^{-1} I^2 J\left[\Delta_1^{-1} u\left(\varphi_1\right) - \Delta_0^{-1} u\left(\varphi_0\right)\right]\end{array}\right\}
\end{aligned}
\tag{A2}
$$

where

$$
\begin{aligned}
A &= n_{[1]}^{-1} p_1 q_1 + n_{[0]}^{-1} p_0 q_0 \\
I &= A^{-\frac{1}{2}} \\
J &= \sum_{i=0}^{1} K_i \\
K_i &= n_{[1]}^{-1}\left[\begin{array}{l}\Delta_1^{-1}\left(1 - p_i\varphi_i - q_i\psi_i\right) \\ +\Delta_i^{-2}\bar{\varphi}_i\bar{\psi}_i\big)\end{array}\right] = n_{[1]}^{-1}\left[\begin{array}{l}-q_i + \Delta_i^{-1} s\left(\varphi_i\right) \\ +\Delta_i^{-2} t\left(\varphi_i\right)\end{array}\right] \\
s\left(\varphi_i\right) &= q_i\left(2\varphi_i - 1\right) \\
t\left(\varphi_i\right) &= \varphi_i\left(1 - \varphi_i\right) \\
u\left(\varphi_i\right) &= \hat{p}_i - \varphi_i
\end{aligned}
\tag{A3}
$$

By using the quadratic approximation on ε_{RD}^{-1}, I^{-1} and I, we have by assuming that $p_1 > p_0$ and $n_{[1]}^{-1} p_1 q_1 < n_{[0]}^{-1} p_0 q_0$

$$\begin{aligned}
&\varepsilon_{RD}^{-1} \approx p_1^{-1} + p_0 p_1^{-2} + p_0^2 p_1^{-3} \\
&I^{-1} \approx \frac{1}{\sqrt{n_{[0]}}}\left[\sqrt{p_0 q_0} + \tfrac{1}{2}\frac{n_{[0]}}{n_{[1]}}\cdot\frac{p_1 q_1}{\sqrt{p_0 q_0}} - \tfrac{1}{8}\frac{n_{[0]}^2}{n_{[1]}^2}\cdot\frac{\left(p_1 q_1\right)^2}{\left(p_0 q_0\right)^{\frac{3}{2}}}\right] \\
&I \equiv \frac{1}{\sqrt{A}} \approx \sqrt{\frac{n_{[0]}}{p_0 q_0}}\left[1 - \tfrac{1}{2}\cdot\frac{n_{[0]} p_1 q_1}{n_{[1]} p_0 q_0} + \tfrac{3}{8}\left(\frac{n_{[0]} p_1 q_1}{n_{[1]} p_0 q_0}\right)^2\right]
\end{aligned} \tag{A4}$$

For fixed i = 0, 1 let

$$M_{[i,j,k,l]} \equiv \int_{a_i}^{1}\int_{b_i}^{1}\left[s^j\left(\varphi_i\right)t^k\left(\varphi_i\right)u^l\left(\varphi_i\right)\right]\Big/\Delta_i^{j+2k+l}\,d\psi_i\,d\varphi_i \tag{A5}$$

where $a_i = \hat{p}_i + 0.005$, $b_i = \hat{q}_i + 0.005$, $s(\varphi_i)$, $t(\varphi_i)$ and $u(\varphi_i)$ are all defined in Equation A3. Let us calculate some of Equation A5 which will be needed later. For j = 1, k = l = 0 we have

$$\begin{aligned}
M_{[i,1,0,0]} &\equiv \int_{a_i}^{1}\int_{b_i}^{1}\left[s\left(\varphi_i\right)/\Delta_i\right]d\psi_i\,d\varphi_i = \int_{a_i}^{1} s\left(\varphi_i\right)\left[\ln\varphi_i - \ln\left(\varphi_i - \bar{b}_i\right)\right]d\varphi_i \\
&= \int_{a_i}^{1}\left\{\left[\left(s'(0)\varphi_i + s(0)\right]\ln\varphi_i - \begin{bmatrix} s'\left(\bar{b}_i\right)\left(\varphi_i - \bar{b}_i\right) \\ + s\left(\bar{b}_i\right)\end{bmatrix}\ln\left(\varphi_i - \bar{b}_i\right)\right\}d\varphi_i \\
&= q_i \dot{M}_{[i,1,0,0]}
\end{aligned} \tag{A6}$$

where $\delta_i = a_i + b_i - 1$, $\bar{b}_i = 1 - b_i$, and

$$\begin{aligned}
\dot{M}_{[i,1,0,0]} \equiv\; &\delta_i^2 \ln\delta_i - a_i^2 \ln a_i - b_i^2 \ln b_i + a_i \ln a_i \\
&+ \left(2\bar{b}_i - 1\right)\left(\delta_i \ln\delta_i - b_i \ln b_i + b_i - \delta_i\right) + \tfrac{1}{2}\left(1 + a_i^2 + b_i^2 - \delta_i^2\right)
\end{aligned}$$

For j = l = 0, k = 1 we have

$$
\begin{aligned}
M_{[i,0,1,0]} &= \int_{a_i}^{1}\int_{b_i}^{1}\left[t\left(\varphi_i\right)/\Delta_i^2 d\psi_i d\varphi_i\right. \\
&= \int_{a_i}^{1}\left[\begin{array}{c}\dfrac{\frac{1}{2}t''\left(\bar{b}_i\right)\left(\varphi_i-\bar{b}_i\right)^2+t'\left(\bar{b}_i\right)\left(\varphi_i-\bar{b}_i\right)+t\left(\bar{b}_i\right)}{\varphi_i-\bar{b}_i} \\ -\dfrac{\frac{1}{2}t''\left(0\right)\varphi_i^2+t'\left(0\right)\varphi_i+t\left(0\right)}{\varphi_i}\end{array}\right]d\varphi_i \\
&= \sum_{m=1}^{3}\left(d_{m[i,0,1,0]}+e_{m[i,0,1,0]}\right)=b_i\bar{b}_i\ln\left(b_i/\delta_i\right)
\end{aligned}
\tag{A7}
$$

where $\bar{a}_i = 1-a_i$, and

$$
d_{1[i,0,1,0]} = -\tfrac{1}{2}\bar{a}_i\left(b_i+\delta_i\right), d_{2[i,0,1,0]} = \bar{a}_i\left(1-2\bar{b}_i\right), d_{3[i,0,1,0]} = b_i\bar{b}_i\ln\left(b_i/\delta_i\right),
$$
$$
e_{1[i,0,1,0]} = \tfrac{1}{2}\bar{a}_i\left(1+a_i\right), e_{2[i,0,1,0]} = -\bar{a}_i, e_{3[i,0,1,0]} = 0.
$$

For $j = k = 0$, $l = 1$ we have

$$
\begin{aligned}
M_{[i,0,0,1]} &\equiv \int_{a_i}^{1}\int_{b_i}^{1}\frac{u\left(\varphi_i\right)}{\Delta_i}d\psi_i d\varphi_i \\
&= -\hat{p}_i a_i\ln a_i+\left(\hat{q}_i-b_i\right)\left(b_i\ln b_i-\delta_i\ln\delta_i\right) \\
&\quad +\tfrac{1}{2}\left(a_i^2\ln a_i+b_i^2\ln b_i-\delta_i^2\ln\delta_i\right) \\
&\quad -\tfrac{1}{4}\left(a_i^2+b_i^2-\delta_i^2-1\right)-\bar{a}_i\bar{b}_i
\end{aligned}
\tag{A8}
$$

For $j = l = 1$, $k = 0$ we have

$$
\begin{aligned}
M_{[i,1,0,1]} &= \int_{a_i}^{1}\int_{b_i}^{1} \frac{s(\varphi_i)}{\Delta_i}\cdot\frac{u(\varphi_i)}{\Delta_i} d\psi_i d\varphi_i \\
&\quad -\int_{a_i}^{1}\left[\begin{array}{l} \dfrac{\frac{1}{2}v_1''(\bar{b}_i)(\varphi_i-\bar{b}_i)^2 + v_i'(\bar{b}_i)(\varphi_i-\bar{b}_i)+v(\bar{b}_i)}{\varphi_i-\bar{b}_i} \\ -\dfrac{\frac{1}{2}v''(0)\varphi_i^2 + v'(0)\varphi_i + v(0)}{\varphi_i} \end{array}\right] d\varphi_i \qquad \text{(A9)} \\
&= \sum_{m=1}^{3}\left(d_{m[i,1,0,1]} + e_{m[i,1,0,1]}\right) = q_i\dot{M}_{[i,1,0,1]}
\end{aligned}
$$

where

$$
\begin{aligned}
&v_i(\varphi_i) = s(\varphi_i)u(\varphi_i) = q_i(2\varphi_i - 1)(\hat{p}_i - \varphi_i), \\
&d_{1[i,1,0,1]} = -q_i\bar{a}_i(b_i + \delta_i), d_{2[i,1,0,1]} = q_i\bar{a}_i\left[1 + 2(\hat{p}_i - \bar{b}_i)\right], \\
&d_{3[i,1,0,1]} = q_i\bar{b}_i\left[1 + 2(\hat{p}_i - \bar{b}_i) - \hat{p}_i\right]\ln(b_i/\delta_i), \\
&e_{1[i,1,0,1]} = q_i\bar{a}_i(1 + a_i), e_{2[i,1,0,1]} = -q_i(1 + 2\hat{p}_i)\bar{a}_i, e_{3[i,1,0,1]} = \hat{p}_i q_i \ln a_i, \\
&\dot{M}_{[i,1,0,1]} \equiv \bar{b}_i\left[(1 + \hat{p}_i - 2\bar{b}_i)\ln(b_i/\delta_i) + \hat{p}_i \ln a_i\right]
\end{aligned}
$$

For $j = 0$, $k = l = 1$ we have

$$
\begin{aligned}
M_{[i,0,1,1]} &= \int_{a_i}^{1}\int_{b_i}^{1} \frac{t(\varphi_i)}{\Delta_i^2}\cdot\frac{u(\varphi_i)}{\Delta_i}\,d\psi_i\,d\varphi_i \\
&= \tfrac{1}{2}\int_{a_i}^{1}\left[\begin{array}{c}\dfrac{\begin{array}{l}\tfrac{1}{6}v_2'''(\bar{b}_i)(\varphi_i-\bar{b}_i)^3+\tfrac{1}{2}v_2''(\bar{b}_i)(\varphi_i-\bar{b}_i)^2\\ +v_2'(\bar{b}_i)(\varphi_i-\bar{b}_i)+v_2(\bar{b})_i\end{array}}{(\varphi_i+\bar{b}_i)^2}\\ -\dfrac{\tfrac{1}{6}v_2'''(0)\varphi_1^3+\tfrac{1}{2}v_2''(0)\varphi_1^2+v_2'(0)\varphi_1+v_2(0)}{\varphi_1^2}\end{array}\right]d\varphi_i \\
&= \sum_{m=1}^{4}\left(d_{m[i,0,1,1]}+e_{m[i,0,1,1]}\right) \\
&= 2\bar{a}_i\bar{b}_i+\tfrac{1}{2}\left\{\begin{array}{l}\left[3\bar{b}_i^2-2(1+\hat{p}_i)\bar{b}_i+\hat{p}_i\right]\ln(b_i/\delta_i)\\ +b_i\bar{a}_i\bar{b}_i(\hat{p}_i-\bar{b}_i)(b_i+\delta_i)/(b_i\delta_i)^2\end{array}\right\}
\end{aligned}
\tag{A10}
$$

where

$$
\begin{aligned}
&v_2(\varphi_i)=t(\varphi_i)u(\varphi_i)=\varphi(1-\varphi_i)(\hat{p}_i-\varphi_i),\\
&d_{1[i,0,1,1]}=\tfrac{1}{4}\bar{a}_i(b_i+\delta_i),\, d_{1[i,0,1,1]}=\tfrac{1}{2}\bar{a}_i\left[3\bar{b}_i-(1+\hat{p}_i)\right],\\
&d_{3[i,0,1,1]}=\tfrac{1}{2}\left[3\bar{b}_i^2-2(1+\hat{p}_i)\bar{b}_i+\hat{p}_i\right]\ln(b_i/\delta_i),\\
&d_{4[i,0,1,1]}=\tfrac{1}{2}b_i\bar{b}_i(\hat{p}_i-\bar{b}_i)\bar{a}_i(b_i+\delta_i)/(b_i\delta_i)^2,\\
&e_{1[i,0,1,1]}=-\tfrac{1}{4}\bar{a}_i(1+a_i),\, e_{2[i,0,1,1]}=\tfrac{1}{2}(1+\hat{p}_i)\bar{a}_i,\, e_{3[i,0,1,1]}=\tfrac{1}{2}\hat{p}_i\ln a_i,\, e_{4[i,0,1,1]}=0.
\end{aligned}
$$

Note that in all of the above calculations I first integrate with respect to ψ_i and then integrate with respect to φ_i by employing the Taylor's series expansion to expand the function about $\varphi_i=\bar{b}_i$ or 0.

Now we are ready to calculate the marginal probability density function of Equation A1 one by one

$$\int_{a_0}^{1}\int_{b_0}^{1}\int_{a_1}^{1}\int_{b_1}^{1}\frac{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}}{\breve{\varepsilon}_{RD}}\prod_{i=0}^{1}d\psi_i d\varphi_i = -\left\{\begin{array}{l} I^{-1}\varepsilon_{RD}^{-1}\varsigma^{-1}+I^{-1}\varepsilon_{RD}^{-2}\begin{pmatrix}\bar{a}_0\bar{b}_0 M_{[1,0,0,1]}\\ -\bar{a}_1\bar{b}_1 M_{[0,0,0,1]}\end{pmatrix}\\ +\frac{1}{2}I\varepsilon_{RD}^{-1}\sum_{i=0}^{1}n_{[i]}^{-1}\begin{bmatrix}-\varsigma^{-1}q_i\\ +\bar{a}_{1-i}\bar{b}_{1-i}\begin{pmatrix}q_i\dot{M}_{[i,1,0,0]}\\ +M_{[i,0,1,0]}\end{pmatrix}\end{bmatrix}\\ +\frac{1}{2}I\varepsilon_{Rd}^{-2}n_{[1]}^{-1}\bigl[-q_1\,\bar{a}_0\bar{b}_0 M_{[1,0,0,1]}\\ +\bar{a}_0\bar{b}_0\begin{pmatrix}q_1\dot{M}_{[1,1,0,1]}\\ +M_{[1,0,1,1]}\end{pmatrix}-M_{[0,0,0,1]}\Biggl(q_1\begin{pmatrix}\dot{M}_{[1,1,0,0]}\\ -\bar{a}_1\bar{b}_1\end{pmatrix}\end{array}\right.$$

$$= -\left\{\begin{array}{l} I^{-1}\varepsilon_{RD}^{-1}\varsigma^{-1}+I^{-1}\varepsilon_{RD}^{-2}R_0\\ +\frac{1}{2}I\varepsilon_{Rd}^{-1}\begin{bmatrix}n_{[1]}^{-1}\bigl(q_1R_1+\bar{a}_0\bar{b}_0 M_{[1,0,1,0]}\\ +n_{[0]}^{-1}\bigl(q_0R_2+\bar{a}_1\bar{b}_1 M_{[0,0,1,0]}\end{bmatrix}\\ +\frac{1}{2}I\varepsilon_{RD}^{-2}\bigl[n_{[1]}^{-1}\left(q_1R_3+R_4\right)\end{array}\right. \qquad \text{(A11)}$$

where

$$\begin{aligned}
&R_0=\bar{a}_0\bar{b}_0 M_{[1,0,0,1]}-\bar{a}_1\bar{b}_1 M_{[0,0,0,1]},R_1\equiv\bar{a}_0\bar{b}_0\dot{M}_{[1,1,0,0]}-\varsigma^{-1},R_2\equiv\bar{a}_1\bar{b}_1 M_{[1,0,1,0]}-\varsigma^{-1},\\
&R_3\equiv M_{[0,0,0,1]}\left(\bar{a}_1\bar{b}_1-\dot{M}_{[1,1,0,0]}\right)+\bar{a}_0\bar{b}_0\left(\dot{M}_{[1,1,0,1]}-M_{[1,0,0,1]}\right),\\
&R_4\equiv\bar{a}_0\bar{b}_0 M_{[1,0,1,1]}-M_{[0,0,0,1]}M_{[1,0,1,0]}, \qquad \text{(A12)}\\
&R_5\equiv M_{[1,0,0,1]}\left(\dot{M}_{[[0,1,0,0]]}-\bar{a}_0\bar{b}_0\right)+\bar{a}_1\bar{b}_1\left(M_{[0,0,0,1]}-\dot{M}_{[0,1,0,1]}\right),\\
&R_6\equiv M_{[1,0,0,1]}M_{[0,0,1,0]}-\bar{a}_1\bar{b}_1 M_{[0,0,1,1]}
\end{aligned}$$

$\Rightarrow$

$$
\begin{aligned}
m_g^{(1)} &\equiv \omega\varsigma \iint_{R\times\Omega} \frac{\sqrt{Var\left(\breve{\varepsilon}_{Rd}\right)}}{\breve{\varepsilon}_{Rd}} \prod_{i=0}^{1} p_i^{\eta-1} q_i^{\tau-1} dp_i dq_i d\psi_i d\varphi_i \\
&= -\varsigma \left\{
\begin{array}{l}
\frac{3N_0\varsigma^{-1}\tau}{\sqrt{\eta\tau}} + \left(9N_0R_0\tau + \frac{3}{2}N_1\left(n_{[1]}^{-1}R_1\tau + \bar{a}_0\bar{b}_0M_{[1,0,1,0]}\right)\frac{1}{\eta(\eta+\tau)\sqrt{\eta\tau}}\right. \\
+\frac{3}{2}N_1\left(n_{[0]}^{-1}R_2\tau + \bar{a}_1\bar{b}_1M_{[1,0,1,0]}\right)\frac{1}{\eta(\eta+\tau)^2\sqrt{\eta\tau}} \\
+\frac{9}{2}N_1\tau\left[\eta^2(\eta+\tau)^3\sqrt{\eta\tau}\right]^{-1}\left[n_{[1]}^{-1}\left(R_3(\eta+\tau)+R_4\right)+n_{[0]}^{-1}\left(R_5(\eta+\tau)+R_6\right)\right]
\end{array}
\right\} \\
&= \frac{-3\varsigma(\eta+\tau)\left\{\begin{array}{l}2N_0\eta\tau(\eta+\tau)\cdot \\ \left(\varsigma^{-1}\eta(\eta+\tau)+3R_0\right)\end{array}\right. + N_1\left[\begin{array}{l}\eta\tau(\eta+\tau)\left(n_{[1]}^{-1}R_1+n_{[0]}^{-1}R_2\right) \\ +\eta\left(n_{[1]}^{-1}\bar{a}_0\bar{b}_0+n_{[0]}^{-1}\bar{a}_1\bar{b}_1\right)M_{[1,0,1,0]}+3\cdot\end{array}\right.}{2\sqrt{\eta\tau}\eta^2(\eta+\tau)^3} \\
&\quad \left.\left.\tau\left(n_{[1]}^{-1}R_3+n_{[0]}^{-1}R_5\right)\right]\right\} - 9N_1\varsigma\tau\left(n_{[1]}^{-1}R_4+n_{[0]}^{-1}R_6\right)
\end{aligned}
\tag{A13}
$$

where for $i, j, k, l = 0, 1$ $M_{[i,j,k,\ell]}$ and $\dot{M}_{[i,j,k,\ell]}$ are given respectively by Equations A6-A10,

$$
\begin{aligned}
&\omega = \left\{\Gamma(\eta+\tau)/\left[\Gamma(\eta)\Gamma(\tau)\right]\right\}^2, \\
&\varsigma = \left(\bar{a}_0\bar{b}_0\bar{a}_1\bar{b}_1\right)^{-1}, \bar{a}_i = 1-a_i, \bar{b}_i = 1-b_i, \\
&N_0 = n_{[0]}^{-\frac{1}{2}}\left(1+\tfrac{1}{2}n_{[1]}^{-1}n_{[0]} - \tfrac{1}{8}n_{[1]}^{-2}n_{[0]}^2\right), \\
&N_1 = n_{[0]}^{\frac{1}{2}}\left(1-\tfrac{1}{2}n_{[0]}n_{[1]}^{-1} + \tfrac{3}{8}n_{[0]}^2n_{[1]}^{-2}\right)
\end{aligned}
\tag{A14}
$$

On the other hand, by integrating the following equation with respect to φ_i, ψ_i, $i = 0, 1$

$$\frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}} = \frac{\breve{p}_1 - \breve{p}_0 - \varepsilon_{RD}}{\sqrt{A}}\left(1 - \frac{J}{2A}\right)$$
$$= I\left(\frac{u(\varphi_1)}{\Delta_1} - \frac{u(\varphi_0)}{\Delta_0} - \varepsilon_{RD}\right) - \tfrac{1}{2}I^3\left(\frac{u(\varphi_1)}{\Delta_1} - \frac{u(\varphi_0)}{\Delta_0} - \varepsilon_{RD}\right)J$$

This leads to

$$\int_{a_0}^{1}\int_{b_0}^{1}\int_{a_1}^{1}\int_{b_1}^{1}\frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}}\prod_{i=0}^{1}d\psi_i d\varphi_i = I\left(R_0 - \varsigma^{-1}\varepsilon_{RD}\right)$$
$$-\tfrac{1}{2}I^3\left\{\begin{array}{l} n_{[1]}^{-1}\left[\begin{array}{l} q_1R_3 + R_4 \\ -\varepsilon_{RD}\left(q_1R_1 + \bar{a}_0\bar{b}_0M_{[1,0,1,0]}\right)\end{array}\right] \\ +n_{[0]}^{-1}\left[\begin{array}{l} q_0R_5 + R_6 \\ -\varepsilon_{RD}\cdot\left(q_0R_2 + \bar{a}_1\bar{b}_1M_{[0,0,1,0]}\right)\end{array}\right]\end{array}\right\} \quad \text{(A15)}$$

Further, we obtain by integrating Equation A15 with respect to p_i, q_i, i = 0, 1

$$m_g^{(2)} \equiv \omega\varsigma\iint_{\Omega\times R}\frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}}\prod_{i=0}^{1}p_i^{\eta-1}q_i^{\tau-1}dp_i dq_i d\psi_i d\varphi_i$$
$$= \varsigma\left\{\frac{N_1R_0}{(\eta+\tau)\sqrt{\eta\tau}} - \tfrac{1}{2}\left[\frac{N_2\left(n_{[1]}^{-1}R_3 + n_{[0]}^{-1}R_5\right)}{\eta(\eta+\tau)^2\sqrt{\eta\tau}} + \frac{N_2\left(n_{[1]}^{-1}R_4 + n_{[0]}^{-1}R_6\right)}{\eta\tau(\eta+\tau)^3\sqrt{\eta\tau}}\right]\right\} \quad \text{(A16)}$$
$$= \frac{\varsigma\left\{2N_1R_0\eta\tau(\eta+\tau)^2 - N_2\left[\begin{array}{l}\tau(\eta+\tau)\left(n_{[1]}^{-1}R_3 + n_{[0]}^{-1}R_5\right) \\ +n_{[1]}^{-1}R_4 + n_{[0]}^{-1}R_6\end{array}\right]\right\}}{2\left[\sqrt{\eta\tau}(\eta+\tau)\right]^3}$$

where ς, N_1, R_0, and R_j, j = 3, 4, 5, 6 are given respectively by Equations A12 and A14, and

$$N_2 \equiv \sqrt{n_{[0]}^3}\left(1 - \frac{3n_{[0]}}{2n_{[1]}} - \frac{3n_{[0]}^2}{8n_{[1]}^2} - \frac{n_{[0]}^3}{8n_{[1]}^3} + \frac{45n_{[0]}^4}{64n_{[1]}^4} - \frac{27n_{[0]}^5}{128n_{[1]}^5} + \frac{27n_{[0]}^6}{512n_{[1]}^6}\right)$$

Note that in calculating Equations A13 and A16 I used an approximation on the Gamma function: $\Gamma(z+a)/\Gamma(z+b) \approx z^{a-b}$ (Askey & Roy, 2010).

By integrating Equation 12 with respect to (φ_i, ψ_i) first and then (p_i, q_i) for i = 0, 1 we obtain m_g (η, τ) by substituting Equations A13 and A16 into Equation A17:

$$m_g(\eta,\tau) = (2\pi)^{-\frac{1}{2}} \left\{ m_g^{(1)}(\eta,\tau) - \tfrac{1}{2} m_g^{(2)}(\eta,\tau) + \tfrac{1}{8}\left[m_g^{(2)}(\eta,\tau) \right]^3 \right\} \tag{A17}$$