Empirical Bayesian analysis of dichotomic data with errors and repeated classifications

Magda Carvalho Pires Roberto da Costa Quinino

Emilio Suyama

Departamento de Estatística - ICEx - UFMG

31270-901 - Belo Horizonte - MG - Brazil

Anderson Laécio Galindo Trindade

Departamento de Engenharia de Produção - USP - São Paulo - Brazil

November 27, 2006

Abstract

This paper discusses the problem of Bayesian estimation of a proportion p of interest when the classification of conforming and non-conforming items is subject to diagnosis errors. The use of non-informative prior distributions for the errors generates a symmetrical posterior distribution for p with great variability. It is therefore necessary that prior distribution be informative, although that is not always possible. In this paper, the authors classify items repeatedly so as to generate empirical prior distributions of errors and thus present a solution to the problem. Results from simulation experiments reveal that the methodology proposed here leads to a reasonable estimation of the proportion of interest when one uses the posterior mode or the posterior median and makes at least three repeated classifications.

Keywords: Quality Control, Empirical Bayes method, Binomial distribution, classification errors, repeated classifications

1 Introduction

When implementing quality control for attributes, one needs to take into account the efficacy of the system used to classify manufactured items as conforming or non-conforming. Two types of errors may occur during inspection: the first, known as type I, occurs when a conforming item is classified as non-conforming; the second, called type II, is when an item is considered conforming when it is actually non-conforming.

In a pioneering paper, Bross (1954) has shown that, in the presence of classification errors, the estimators obtained by means of a classical statistics approach are biased. Other authors, such as Johnson and Kotz (1988), Johnson et al. (1991), Evans et al. (1996), Viana (1994), Gustafson (2003) have emphasized that, if ignored, classification errors may jeopardize the entire process of inference and, consequently, quality control.

Let us suppose that in a random sample of n units, there are X conforming items. The random variable X has binomial distribution with parameters (n, p), that is, $X \sim Bi(n, p)$. However, the occurrence of classification errors in the system implies a modification in this probability function. Let e_1 be the probability that a conforming item be wrongly classified as non-conforming, and e_2 be the probability that a non-conforming item be classified as conforming. So, the probability that an item be classified as conforming is $q = p(1 - e_1) + (1 - p)e_2$, which yields an random variable X whose binomial distribution has parameter q instead of p.

The difficulty found in such analysis can be better grasped through the establishment of the estimator of maximum likelihood. The likelihood function for the case that presents classification errors may be written as $L(x|n, q) = q^x(1 - q)^{n-x}$. This is maximized to all points (p, e_1, e_2) so that $p(1-e_1)+(1-p)e_2 = x/n$ (Gaba and Winkler, 1992). Therefore, the estimator of maximum likelihood is not unique.

In order to solve this problem, various classical methods have been suggested, and a review can be found in Johnson et al. (1991). In general, the methods proposed rely on alternative sampling plans for a preliminary estimation of classification errors. From a Bayesian point of view, Gaba and Winkler (1992) considered an approach that requires the use of a informative prior distribution. This may be a considerable restriction because

in many cases this information is not available. They have also verified that the use of independent non-informative prior and uniform distributions between zero and one for parameters (p, e_1, e_2) yields a posterior mean for p equals $1/2$, regardless of the sample result and, besides that, every point (p, e_1, e_2) so that $p(1 - e_1) + (1 - p)e_2 = x/n$ were posterior modes.

In research papers on Bayesian sampling size for dichotomic data in the presence of classification errors, Dendukuri et al. (2004) and Rahme et al. (2000) have also noted the primordial need for informative prior distribution.

The present article proposes a model in which the process of Bayesian inference for proportion in the presence of classification errors includes making repeated classifications, both in order to elicit empirical prior distribution as well as to minimize the impact of such errors. The final classification of an item is the one that most appears after repeated classifications. In practical terms, we believe that the methodology developed here will be useful when making repeated classifications become easier and more operational than informative prior distributions.

Section 2 presents a method to include repeated classifications along with the respective establishment of the likelihood function. Section 3 presents an empirical Bayesian analysis for the proportion of interest, and numerical examples are described in Section 4. Section 5 presents the conclusions.

2 Likelihood Function

Suppose each item in a random sample of size n be independently classified m times as conforming or non-conforming, with m being an odd number. Let $C_{ij} (i = 1, 2, \ldots, n; j = 1, 2, \ldots, m)$ be a Bernoulli random variable corresponding to the *j-th* classification of the *i-th* item. So, $C_{2,3} = 1$ means that the second item was classified as conforming in the third classification. Let F_i be a random Bernoulli variable that denotes the final classification of the *i-th* item after m classifications. Consider that $F_i = 1$ if, and only if, $\sum_{j=1}^{m} C_{ij} > m/2$. The choice of an odd number for m avoids a tie and consequently avoids difficulty in reaching a final classification for an item. Table 1 describes this classification

procedure.

Table 1: Repeated classifications of n items m times each

	Classifications (C_{ij})					Final
Item	1	2	3	$\cdots$	m	Classification
1	C_{11}	C_{12}	C_{13}	$\cdots$	C_{1m}	F_1
2	C_{21}	C_{22}	C_{23}	$\cdots$	C_{2m}	F_2
3	C_{31}	C_{32}	C_{33}	$\cdots$	C_{3m}	F_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	C_{n1}	C_{n2}	C_{n3}	$\cdots$	C_{nm}	F_n

Let E_i be also another Bernoulli random variable that denotes the real state of the *i-th* item, so that the interest is to estimate $P(E_i = 1) = p$. So, we have $e_1 = P(C_{ij} = 0 \mid E_i = 1)$ e $e_2 = P(C_{ij} = 1 \mid E_i = 0)$. The probability that an item be classified as conforming results from

$$P(F_i = 1) = pBi\left(\tfrac{m}{2}; m, e_1\right) + (1-p)\overline{Bi}\left(\tfrac{m}{2}; m, e_2\right) \tag{1}$$

where $Bi\left(\frac{m}{2}; m, e_k\right)$ denotes the cumulative binomial distribution function defined at $\frac{m}{2}$ and $\overline{Bi}\left(\frac{m}{2}; m, e_k\right) = 1 - Bi\left(\frac{m}{2}; m, e_k\right)$. Note that if $m \to \infty$ and the probabilities associated to classification errors are smaller than 0.5 then (1) converges to p, thus corroborating the benefit of using repeated classifications.

Now let us suppose a random sample of n items where r items are considered conforming; the likelihood function can be written as

$$\begin{aligned} L(r|n, m, p, e_1, e_2) = &\left[pBi\left(\tfrac{m}{2}; m, e_1\right) + (1-p)\overline{Bi}\left(\tfrac{m}{2}; m, e_2\right)\right]^r \times \\ &\left[1 - pBi\left(\tfrac{m}{2}; m, e_1\right) - (1-p)\overline{Bi}\left(\tfrac{m}{2}; m, e_2\right)\right]^{n-r} \end{aligned} \tag{2}$$

Note that if $m = 1$, then (2) equals

$$L\left[r|n, p, e_1, e_2\right] = \left[p(1-e_1) + (1-p)e_2\right]^r \left[pe_1 + (1-p)(1-e_2)\right]^{n-r} \tag{3}$$

Expression (3) is precisely the likelihood function used by Gaba and Winkler (1992)

and Viana et al. (1993), which indicates that expression (3) is a generalization of these models obtained through the introduction of repeated classifications.

3 Empirical Bayesian analysis

Consider a joint prior distribution of (p, e_1, e_2) given by:

$$f(p, e_1, e_2) = f_\beta(p|\alpha, \beta) f_\beta(e_1|\alpha_1, \beta_1) f_\beta(e_2|\alpha_2, \beta_2) \tag{4}$$

where $f_\beta(a|b, c)$ denotes a Beta density function for the random variable a with parameters b and c. Beta distributions are widely used in Bayesian models to describe information concerning proportions (Gupta and Nadarajah, 2004). In this article, we consider the random variables (p, e_1, e_2) to be prior independent. As in Rahme et al. (2000) and Stamey et al. (2004) a natural way to obtain the posterior distribution of p could be the use of MCMC methods. Another way would be to use the Sampling/Importance Resampling (SIR) technique or Bayesian weighted bootstrap (Rubin, 1988). However, we have chosen an approach based on numerical integration since the posterior distribution of p could be made explicit despite not having a closed form. For this, the equation (2) must be rewritten as

$$\begin{aligned} L(r|n, m, p, e_1, e_2) = \sum_{j=0}^{r} \sum_{t=0}^{n-r} \tbinom{r}{j} \tbinom{n-r}{t} p^{n-j-t} (1-p)^{j+t} \times \\ Bi(\tfrac{m}{2}; m, e_1)^{r-j} \overline{Bi}(\tfrac{m}{2}; m, e_1)^{n-r-t} Bi(\tfrac{m}{2}; m, e_2)^{t} \overline{Bi}(\tfrac{m}{2}; m, e_2)^{j} \end{aligned} \tag{5}$$

The posterior joint density of (p, e_1, e_2) is obtained by multiplying the prior distribution (4) by the likelihood function (5) and normalizing as required by Bayes' Theorem (Winkler, 2003). Integrating with respect to e_1 and e_2, one finds the marginal posterior density function for p, which can be written as:

$$f(p|r, n, m) = \sum_{j=0}^{r} \sum_{t=0}^{n-r} w^*_{jt} f_\beta\left(p|\alpha^*, \beta^*\right) \tag{6}$$

where $w_{jt}^{*} = \frac{a_{jt}^{*}}{\sum_{j=0}^{r}\sum_{t=0}^{n-r} a_{jt}^{*}}$, with $a_{jt}^{*} = \binom{r}{j}\binom{n-r}{t} B\left(\alpha^{*}, \beta^{*}\right) k_1\left(j,t\right) k_2\left(j,t\right)$ and

$$k_1(j,t) = \int_0^1 e_1^{\alpha_1 - 1}\left(1-e_1\right)^{\beta_1 - 1} Bi\left(\tfrac{m}{2}; m, e_1\right)^{r-j} \overline{Bi}\left(\tfrac{m}{2}; m, e_1\right)^{n-r-t} de_1;$$

$$k_2(j,t) = \int_0^1 e_2^{\alpha_2 - 1}\left(1-e_2\right)^{\beta_2 - 1} Bi\left(\tfrac{m}{2}; m, e_2\right)^{t} \overline{Bi}\left(\tfrac{m}{2}; m, e_2\right)^{j} de_2;$$

and $B\left(\alpha^{*}, \beta^{*}\right)$ the value of the Beta function being calculated at (α^{*}, β^{*}) with $\alpha^{*} = \alpha + n - j - t$ and $\beta^{*} = \beta + j + t$.

The fact that the information needed to define informative prior distributions for classification errors and interest proportion is insufficient implies, for instance, the use of $U(0,1)$ distributions, which is particular to Beta distribution when parameters are equal to unit $[B(1,1)]$. The employment of these distributions yielded a posterior multimodal distribution for p with great variability. Repeated classification extenuates the problem, but its results are still unsatisfactory. The situation is even worse when a single classification is made. In this case, one finds a posterior mean of p that equals 0.5 regardless of the sample result (Gaba and Winkler, 1992). The use of prior distribution $B(0.5, 0.5)$ as an alternative to the representation of the non-information of parameters generated results similar to the use of $B(1,1)$.

Thus, the posterior distribution obtained may indeed be of little use to provide necessary information on the proportion of interest, which evidences the need to obtain additional information on classification errors.

An alternative that could minimize this problem is the use of repeated classification results ($m > 1$) in order to estimate the hiperparameters (α_1, β_1) and (α_2, β_2) of the Beta prior distribution for classification errors together with the use of $U(0,1)$ distribution for p. This procedure may be understood as a process of parametric empirical Bayes inference, as discussed by Carlin and Louis (2000), Gupta and Nadarajah (2004), Morris (1983) and Gelman et al. (2004).

Hiperparameters (α_1, β_1) and (α_2, β_2) were estimated according to the Method of Moments. First, the random sample of size n was divided into two sub-

samples: one made up of items whose final classification was conforming ($F_i = 1$) and the other with items whose final classification was non-conforming ($F_i = 0$). For each item in the first subsample we calculated the proportion of non-conforming repeated classifications, bearing in mind that the mean and the variance of these proportions estimate, respectively, the mean and variance of the Beta prior distribution for e_1. In the second subsample, we calculated the proportion of conforming classifications for each item. The mean and the variance of these proportions estimate, respectively, the mean and variance of the Beta prior distribution for e_2. Finally, through the closed expressions for the mean and variance of the Beta distribution we were able to estimate (α_1, β_1) and (α_2, β_2) thus solving systems from two equations and two unknowns. The estimates for (α_1, β_1) and (α_2, β_2) can be written, respectively, as:

$$\hat{\alpha}_1 = k_3(k_4^2 + k_3^2 - k_3)/k_4^2 \tag{7}$$

$$\hat{\beta}_1 = (k_4^2 + k_3^2 - k_3)(k_3 - 1)/k_4^2 \tag{8}$$

$$\hat{\alpha}_2 = k_5(k_6^2 + k_5^2 - k_5)/k_6^2 \tag{9}$$

$$\hat{\beta}_2 = (k_6^2 + k_5^2 - k_5)(k_5 - 1)/k_6^2 \tag{10}$$

where

$$k_3 = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(1 - C_{ij})\, I_{\{F_i=1\}}}{m \sum_{s=1}^{n} I_{\{F_s=1\}}}; \qquad k_4 = k_3\,(1 - k_3) \sum_{i=1}^{n} I_{\{F_i=1\}};$$

$$k_5 = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(1 - C_{ij})\, I_{\{F_i=0\}}}{m \sum_{s=1}^{n} I_{\{F_s=0\}}}; \qquad k_6 = k_5\,(1 - k_5) \sum_{i=1}^{n} I_{\{F_i=0\}}.$$

The use of empirical prior distribution for $m = 1$ is not viable in the method proposed here, given the impossibility to estimate (α_1, β_1) and (α_2, β_2) from proportions of mistaken

classifications. In the event that, in a particular subsample, all repeated classifications generate identical results, one needs to increase n or m so as to capture the effects of classification errors, and allow for the estimation of (α_1, β_1) and (α_2, β_2) according to the Method of Moments.

4 Numerical examples and discussions

The numerical performance of the methodology proposed in this article was evaluated through the analysis of 48 cases randomly chosen with all possible combinations of the following parameter values: $p = 0.55$; 0.75 or 0.9; $e_1 = 0.05$ or 0.15; $e_2 = 0.05$ or 0.15; $n = 250$. We have also used 1, 3, 5, and 7 for repeated classifications. Moreover, all combinations were analyzed using empirical prior distributions and $U(0, 1)$ for errors. The prior distribution for p was $U(0, 1)$ in all cases studied. We used software Matlab to create a program that calculates and graphically generates the posterior distribution for p with its respective mean, mode and median (available to download from corresponding author homepage - `www.est.ufmg.br/~roberto`). Figures 1 to 7 were simulated using parameters $p = 0.75$; $e_1 = 0.15$; $e_2 = 0.15$; $n = 250$, corresponding to 1, 3, 5 and 7 repeated classifications; these figures illustrate all cases studied.

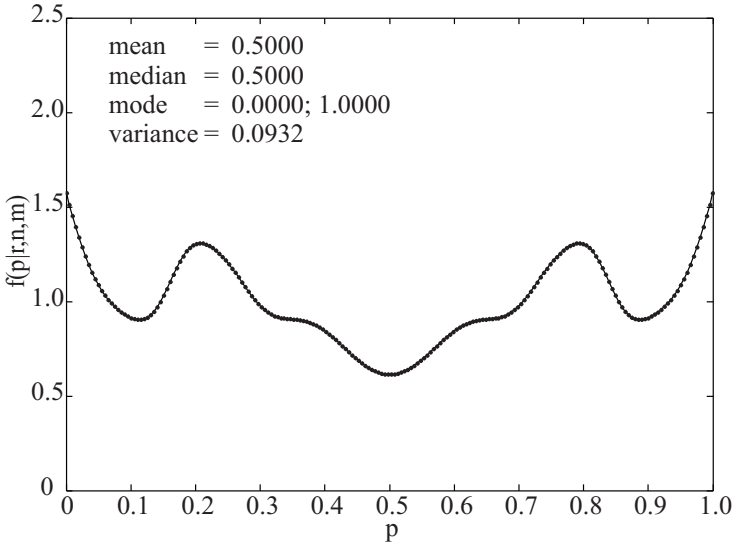

Figure 1: Posterior distribution of p with n=250, m=1 and prior distribution $U(0, 1)$ for e_1 e e_2.

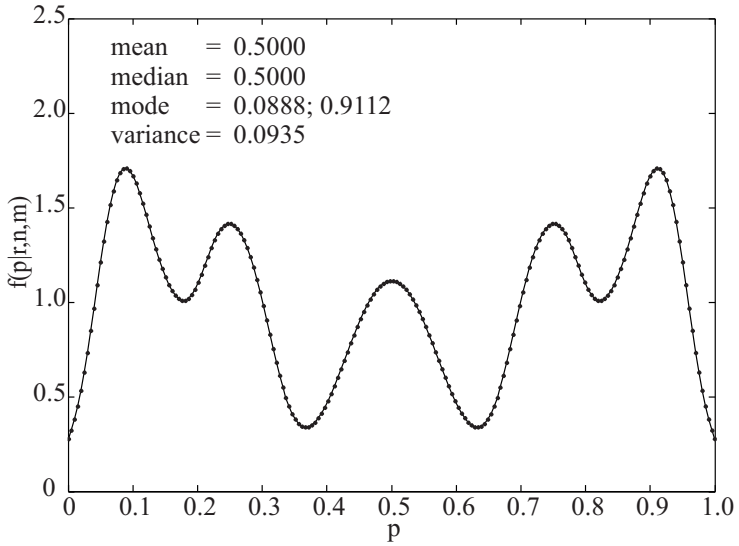

Figure 2: Posterior distribution of p with n=250, m=3 and prior distribution $U(0,1)$ for e_1 e e_2.

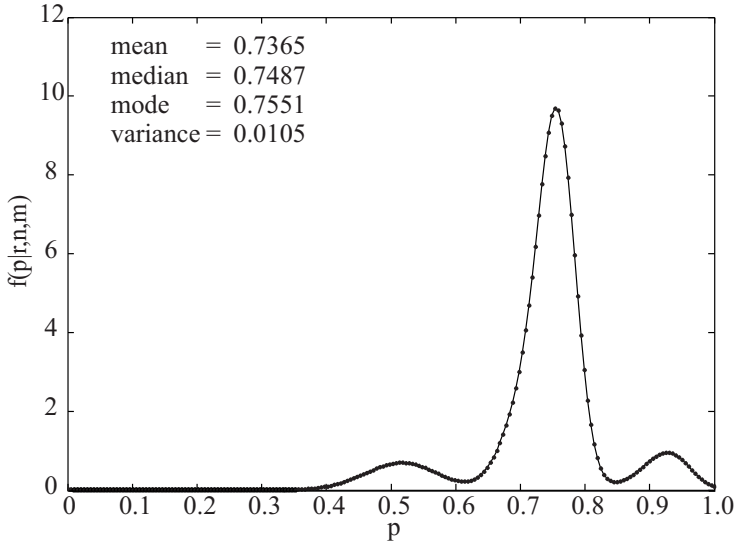

Figure 3: Posterior distribution of p with n=250, m=3 and empirical prior distribution for e_1 e e_2.

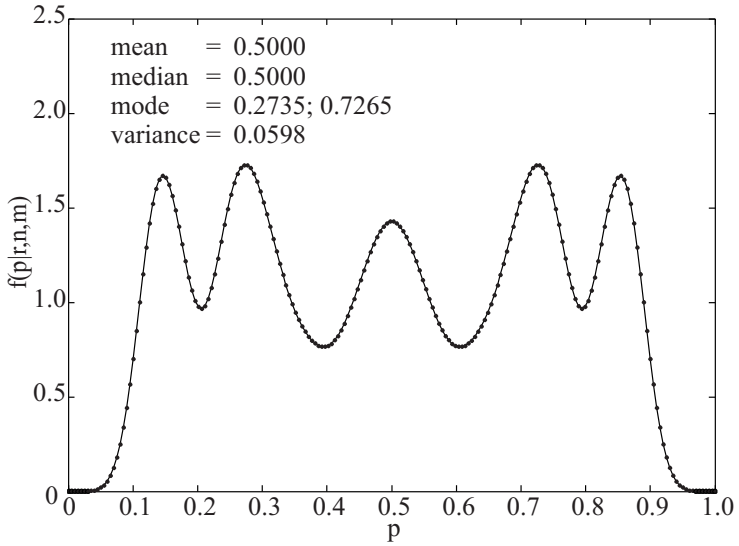

Figure 4: Posterior distribution of p with n=250, m=5 and prior distribution $U(0,1)$ for e_1 e e_2.

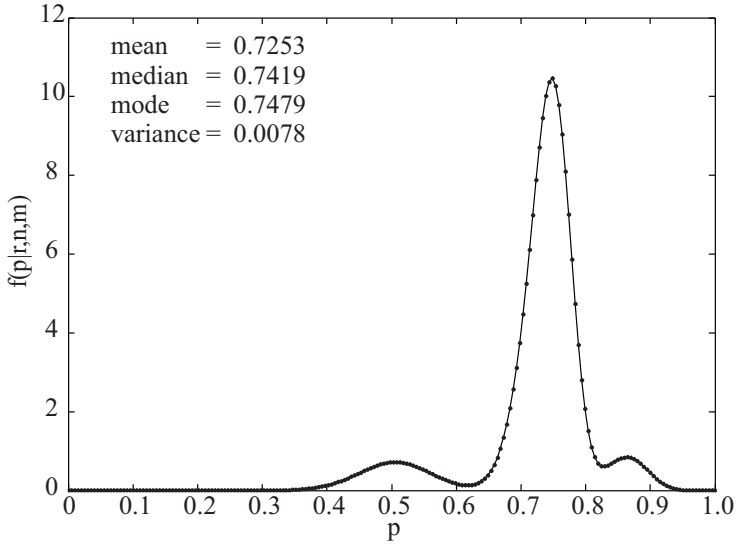

Figure 5: Posterior distribution of p with n=250, m=5 and empirical prior distribution for e_1 e e_2.

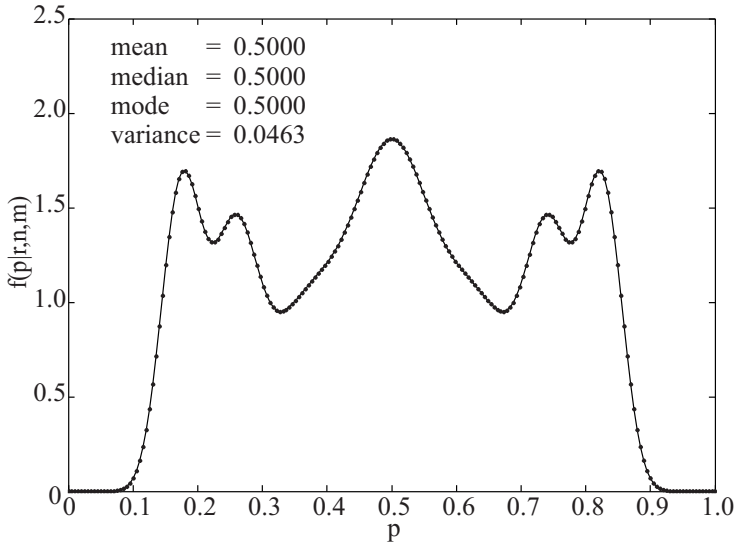

Figure 6: Posterior distribution of p with n=250, m=7 and prior distribution $U(0,1)$ for e_1 e e_2.

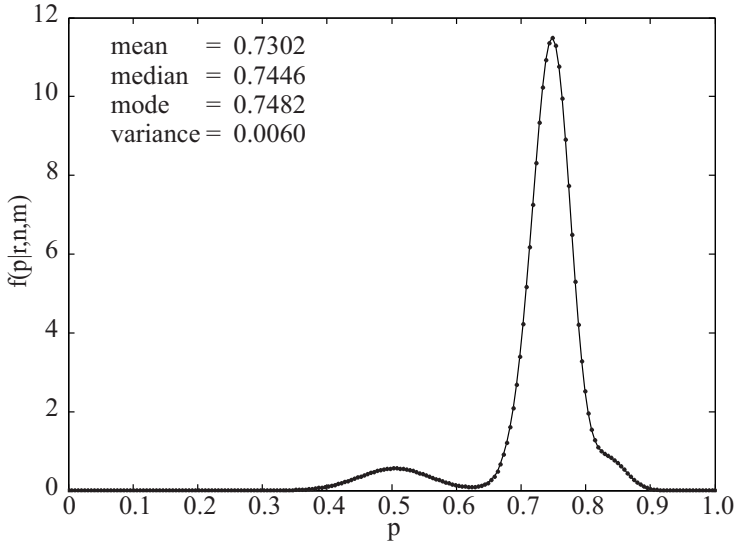

Figure 7: Posterior distribution of p with n=250, m=7 and empirical prior distribution for e_1 e e_2.

We noted that, considering the same number of repeated classifications, the posterior distributions for p that were obtained using empirical prior distribution for classification errors present less variability and smaller number of regions with significant probabilistic mass in comparison to the posterior distributions for p obtained from prior distribution $U(0,1)$ for classification errors.

Tables 2 presents the average absolute bias for simulated combinations. The use of empirical prior distribution reveals a better performance, since it generates less absolute bias than the use of uniform prior distribution. The posterior mean presents the greatest average absolute bias, and thus it is not a good choice when estimating p. The median and mode, though, present values for absolute average bias smaller than 4% for 3, 5, or 7 repeated classifications, while the mode has even a slightly better performance.

Table 2: Average absolute bias of p with n=250

		Empirical prior distribution			Uniform prior distribution		
m	p	Mean	Median	Mode	Mean	Median	Mode
	0.55	-	-	-	0.05	0.05	0.25
1	0.75	-	-	-	0.25	0.25	0.33
	0.90	-	-	-	0.40	0.40	0.54
	0.55	0.04	0.02	0.02	0.05	0.05	0.21
3	0.75	0.03	0.02	0.02	0.25	0.25	0.41
	0.90	0.07	0.03	0.03	0.40	0.40	0.58
	0.55	0.03	0.01	0.01	0.05	0.05	0.04
5	0.75	0.02	0.01	0.01	0.25	0.25	0.22
	0.90	0.06	0.02	0.01	0.40	0.40	0.85
	0.55	0.02	0.01	0.00	0.05	0.05	0.04
7	0.75	0.01	0.01	0.00	0.25	0.25	0.25
	0.90	0.02	0.01	0.01	0.40	0.40	0.09

Table 3 presents the maximal absolute bias obtained for simulated error combinations. Considering 3, 5, or 7 repeated classifications, one notices that the mode presents better performance, with values for maximal absolute bias smaller than 6%, whereas the median presents values smaller than 7%.

When using empirical or $U(0,1)$ prior distribution for errors, we noticed a tendency toward the occurrence of negative bias, that is, the p proportion of interest is underestimated. This may occur due to the criterion used when making the final classification

Table 3: Maximal absolute bias of p with n=250

		Empirical prior Distribution			Uniform *prior* Distribution		
m	p	Mean	Median	Mode	Mean	Median	Mode
	0.55	-	-	-	0.05	0.05	0.45
1	0.75	-	-	-	0.25	0.25	0.66
	0.90	-	-	-	0.40	0.40	0.86
	0.55	0.05	0.04	0.04	0.05	0.05	0.45
3	0.75	0.05	0.04	0.04	0.25	0.25	0.66
	0.90	0.12	0.06	0.05	0.40	0.40	0.80
	0.55	0.05	0.01	0.01	0.05	0.05	0.05
5	0.75	0.02	0.01	0.01	0.25	0.25	0.59
	0.9	0.10	0.04	0.03	0.40	0.40	0.90
	0.55	0.05	0.01	0.01	0.05	0.05	0.05
7	0.75	0.02	0.01	0.01	0.25	0.25	0.25
	0.90	0.04	0.02	0.02	0.40	0.40	0.10

as conforming ($F_i = 1$) or non-conforming ($F_i = 0$). Since in the simulated examples $p > 0.5$, $e_1 \leq 0.15$ and $e_2 \leq 0.15$ then on the average the number of items that were really conforming tends to be greater, which implies that the number of occurrences when a conforming item is classified as non-conforming may be greater than the number of cases when a non-conforming item is classified as conforming. Consequently, the proportion tends to be underestimated. In case $p < 0.5$ there will be a tendency to overestimate the proportion.

Note also that bias is asymptotically null as repeated classifications increase. Figure 8 illustrates this situation through the establishment of the posterior distribution for p with mean, median and mode obtained, respectively, from a simulation in which $n = 500$, $p = 0.75$, $m = 99$ and with empirical prior distribution for errors. In this scenario, the maximal bias is about -0,1% resulting from the posterior mean. The graph also indicates that, as m increases, the posterior mode tends to present less bias.

This article presents an empirical Bayesian methodology to estimate a proportion when evaluations are subject to classification errors and when prior information on such errors is not available. We propose the use of repeated classifications and, through them, one can elicit empirical prior distributions for classification errors.

A simulation study revealed that the methodology presents satisfactory performance,

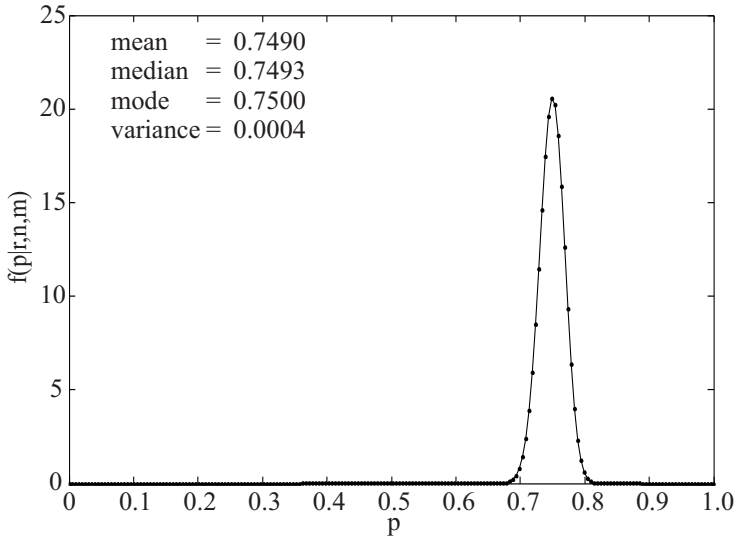

Figure 8: Posterior distribution of p with n=500, m=99 and empirical prior distribution for e_1 e e_2.

since, when compared to prior distribution $U(0,1)$, empirical prior distribution generates posterior estimates with less absolute deviation and posterior distributions with less variability.

For posterior estimation of proportion of interest we recommend as best alternative the posterior mode with at least three repeated classifications.

References

Bross, I. (1954). Misclassification in 2×2 tables. *Biometrics 10*, 478–486.

Carlin, B. P. and T. A. Louis (2000). *Bayes and empirical bayes methods for data analysis*. London: Chapman & Hall.

Dendukuri, N., E. Rahme, P. Bélisle, and L. Joseph (2004). Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics 60*, 388–397.

Evans, M., I. Guttman, Y. Haitovsky, and T. Swartz (1996). Bayesian analysis of binary data subject to misclassification. In D. Berry, K. Chaloner, and J. Geweke (Eds.),

Bayesian Analysis In Statistics and Econometrics: Essays In Honor Of Arnold Zellner, pp. 66–77. New York: North Holland.

Gaba, A. and R. L. Winkler (1992). Implications of errors in survey data: A bayesian model. *Management Science 38*(7), 913–925.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian data analysis* (2 ed.). London: Chapman & Hall.

Gupta, A. K. and S. Nadarajah (2004). *Handbook of beta distribution and its applications*. New York: Marcel Dekker.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and bayesian adjustments*. New York: Chapman & Hall.

Johnson, N. L. and S. Kotz (1988). Estimation from binomial data with classifiers of known and unknown imperfections. *Naval Research Logistics 35*, 147–156.

Johnson, N. L., S. Kotz, and X. Wu (1991). *Inspection errors for attributes in quality control*. London: Chapman & Hall.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association 78*, 47–65.

Rahme, E., L. Joseph, and T. W. Gyorkos (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics 49*(1), 119–128.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*, Cambridge, pp. 395–402. Oxford University Press.

Stamey, J. D., J. W. Seaman, and D. M. Young (2004). Bayesian analysis of complementary poisson rate parameters with data subject to misclassification. *Journal of Statistical Planning and Inference 134*, 36–48.

Viana, M. A. G. (1994). Bayesian small-sample estimation of misclassification multinomial data. *Biometrics 50*, 237–243.

Viana, M. A. G., V. Ramakrishnan, and P. S. Levy (1993). Bayesian analysis of prevalence from the results of small screening samples. *Communications in statistics - V theory and methods 22*(2), 575–85.

Winkler, R. L. (2003). *Bayesian inference and decisions* (2 ed.). London: Probabilistic Publishing.