

Bayesian Approach to Ranking and Selection for a Binary Measurement System

Mark Eschmann¹, James D. Stamey², Phil D. Young³, Dean M. Young⁴

¹Department of Statistical Science, Waco, TX, USA

²Department of Statistical Science, Baylor University, Waco, TX, USA

³Department of Information Systems, Baylor University, Waco, TX, USA

⁴Department of Statistical Science, Baylor University, Waco, TX, USA

Email: Mark_Eschmann@baylor.edu, James_Stamey@baylor.edu, Philip_Young@baylor.edu, Dean_Young@baylor.edu

How to cite this paper: Eschmann, M., Stamey, J.D., Young, P.D. and Young, D.M. (2019) Bayesian Approach to Ranking and Selection for a Binary Measurement System. *Open Journal of Statistics*, 9, 436-444. <https://doi.org/10.4236/ojs.2019.94029>

Received: June 25, 2019

Accepted: August 10, 2019

Published: August 13, 2019

Copyright © 2019 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International

License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Binary measurement systems that classify parts as either pass or fail are widely used. Inspectors or inspection systems are often subject to error. The error rates are unlikely to be identical across inspectors. We propose a random effects Bayesian approach to model the error probabilities and overall conforming rate. We also introduce a feature-subset selection procedure to determine the best inspector in terms of overall classification accuracy. We provide simulation studies that demonstrate the viability of our proposed estimation ranking and subset-selection methods and apply the methods to a real data set.

Keywords

Bayesian Statistics, Quality Control, Binary Measurement Systems, Misclassification

1. Introduction

Repeated binary testing, often referred to as a binary measurement system (BMS), is regularly used in quality control studies as a means of assessing the quality of the units produced. However, these inspection methods are highly dependent on the quality of the individual inspectors, thus making the inspection itself an integral part of the quality control process. Two aspects of evaluating the inspection process are repeatability and reproducibility. A process's repeatability refers to how frequently a single inspector inspecting a single item will obtain the same result, while reproducibility refers to how often different inspectors inspecting the same item will reach the same conclusion. Estimating classification rates of a system has been considered by several authors. [1] considered

various sampling plans to assess the qualities of a BMS. [2] found maximum likelihood estimators and method of moment estimators for the case of multiple raters, assuming fixed effects. When there are multiple inspectors, it may be of interest to determine which of several inspectors or inspection systems is performing best.

The model we consider here is a Bayesian version of [3]. Specifically, we consider a random effects model for multiple testers and multiple inspections per inspector from a Bayesian perspective. There are multiple advantages to using a Bayesian approach. For example, prior knowledge can be incorporated into the study with the use of informative prior distributions. This knowledge can be obtained either from previous data or expert opinion. Also, even in the absence of prior knowledge where the asymptotic dominance of the prior by the likelihood is present, interval estimates generated from the Bayesian paradigm are based largely on the likelihood which has been shown to be superior to other interval estimation methods [4]. Another advantage of the Bayesian paradigm is that if the prior is sufficiently informative, then, assumptions required for identifiability can be relaxed. Thus, our Bayesian approach can be used in situations when the parameters of a likelihood function are not identifiable. The Bayesian estimators considered here have no known closed form and, thus, must be found approximately. We use Markov Chain Monte Carlo (MCMC) simulations to sample from the model's posterior distribution and obtain parameter estimates.

The remainder of the paper is outlined as follows. In Section 2, we present the model and give identifiability assumptions. In Section 3, we describe a simulation study and present the simulation results for the Bayesian estimation. In Section 4, we apply our model to two parameter-ranking applications and two subset selection problems for multiple sites, and in Section 5, we perform an additional simulation to determine the effectiveness of our subset selection procedure. Finally, in Section 6, we provide several comments summarizing our results.

2. The Model

Assume that N randomly selected items to be inspected are sampled from the general population of items. Let the true quality state of an item be denoted by T , where $T = 1$ indicates a good item and $T = 0$ denotes an item that fails to meet the quality specifications. The symbol τ denotes the overall conforming rate. Because we assume that no gold standard is used and because T is a latent variable, we also assume $T \sim \text{Bernoulli}(\tau)$.

Repeated independent, fallible observations are then derived by m different inspectors on the i^{th} unit to indirectly assess the true state of the i^{th} unit, where $i \in \{1, \dots, N\}$. Let Y_{ijk} denote the result of the k^{th} inspection on the i^{th} item by the j^{th} inspector, where $Y_{ijk} = 1$ denotes a passed inspection, $Y_{ijk} = 0$ denotes a failed inspection, and $k \in \{1, \dots, n_{ij}\}$. For each Y_{ijk} and inspector j , we further define the conditional probabilities $\theta_{j,+} = P(Y_{ijk} = 1 | T_i = 0)$ (false positive rate) and $\theta_{j,-} = P(Y_{ijk} = 0 | T_i = 1)$ (false negative rate) with respect to the true state of

the item, T_i . Further, assume

$$(Y_{ijk} | T_i = 0) \sim \text{Bernoulli}(\theta_{j,+}) \tag{1}$$

and

$$(Y_{ijk} | T_i = 1) \sim \text{Bernoulli}(1 - \theta_{j,-}). \tag{2}$$

here, we initially assume that inspections are independent, given the true latent state of the i^{th} part. This conditional independence assumption yields

$$\left(\sum_{k=1}^l Y_{ijk} | T_i = 0 \right) \sim \text{Binomial}(l, \theta_{j,+}). \tag{3}$$

To relax assumptions that the inspectors all have the same probability of classifying correctly and allow for other random heterogeneity, we consider the random effects model where

$$\begin{aligned} \theta_{j,+} &\sim \text{Beta}(\mu_+, \gamma_+), \\ \theta_{j,-} &\sim \text{Beta}(\mu_-, \gamma_-), \end{aligned} \tag{4}$$

where the Beta distribution has been reparameterized such that $\mu = \alpha/(\alpha + \beta)$ and $\gamma = \alpha + \beta$. Thus, the reparameterized Beta probability density function (PDF) is

$$f(x) = \frac{x^{\mu\gamma-1} (1-x)^{\gamma-\mu\gamma-1}}{B(\mu\gamma, \gamma - \mu\gamma)}. \tag{5}$$

To complete the hierarchical model we require priors for μ_+ , μ_- , γ_+ and γ_- . Specifically we assume $\text{Beta}(\alpha_+, \beta_+)$ and $\text{Beta}(\alpha_-, \beta_-)$ priors for μ_+ and μ_- , respectively. Finally, $\text{Gamma}(c_+, d_+)$ and $\text{Gamma}(c_-, d_-)$ priors are used for γ_+ and γ_- , respectively. In the absence of prior information, $\text{Beta}(1,1)$ priors can be used for μ_+ and μ_- and diffuse Gamma priors are used for γ_+ and γ_- .

We have chosen a Beta distribution to model the random effects because of its interpretability under this reparameterization. An often used alternative model structure is

$$\text{logit}(\theta) \sim N(\mu, \sigma), \tag{6}$$

where μ is generally given a normal prior and σ is often given a half-t or half-Cauchy prior.

For the parameters $\Theta = \{\theta_-, \theta_+\}$ and $\Psi := \{\mu_+, \gamma_+, \mu_-, \gamma_-, \tau\}$, the likelihood of the latent vector $\mathbf{t} = [t_1, \dots, t_N]^T$, the observed data matrix is $\mathbf{X} = [x'_1, \dots, x'_N]^T$, where $\mathbf{x} = [x_{i,1}, \dots, x_{i,m}]^T$ and

$$\begin{aligned} f(\mathbf{x}, \mathbf{t}, \Theta | \Psi) &= f(\mathbf{x} | \mathbf{t}, \Theta, \mathbf{Y}) f(\mathbf{t}, \Theta | \Psi) \\ &= \tau^{\sum_{i=1}^N t_i} (1-\tau)^{N-\sum_{i=1}^N t_i} \left(\prod_{j=1}^m (1-\theta_{j-})^{\sum_{i=1}^N t_i x_{ij} + \mu_-\gamma_- - 1} \right. \\ &\quad \times (\theta_{j-})^{\sum_{i=1}^N t_i (n_{ij} - x_{ij}) + \gamma_- - \mu_-\gamma_- - 1} \left. \right) \left(\prod_{j=1}^m (\theta_{j+})^{\sum_{i=1}^N (1-t_i) x_{ij} + \mu_+\gamma_+ - 1} \right. \\ &\quad \times (1-\theta_{j+})^{\sum_{i=1}^N (1-t_i) (n_{ij} - x_{ij}) + \gamma_+ - \mu_+\gamma_+ - 1} \left. \right). \end{aligned} \tag{7}$$

For the random effects model, the first assumption necessary for identifiability [3] is

$$\mu_+ + \mu_- < 1. \quad (8)$$

The interpretation of (8) is that the overall expected probability of correctly classifying an item is greater than the chance of misclassifying it. This assumption is required due to the bimodal nature of the likelihood [4].

The second identifiability assumption assures that there are enough degrees of freedom to estimate all model parameters. This assumption requires two things: that enough inspectors and inspections per inspector are available to estimate the status of each item, and that enough inspectors are available to estimate the inspectors' random effects parameters. The second condition requires at least two inspectors while letting $l_j = \min(n_{ij}, \dots, n_{Nj})$. A sufficient condition to meet the first requirement is that

$$-1 + \prod_{j=1}^m (l_j + 1) \geq 2m + 1. \quad (9)$$

In the present model, (9) is sufficient because additional inspections do not harm the model identifiability.

The third identifiability assumption is that both true negatives ($T_i = 0$) and true positives ($T_i = 1$) exist in the sample. This assumption is necessary because the absence of true negatives indicates one cannot estimate false negative rates. [3] have demonstrated that the absence of either true negatives or true positives essentially implies that there is enough data to estimate only half of the variables, namely θ_+ , μ_+ , and γ_+ or θ_- , μ_- , and γ_- . We remark that the last two identifiability requirements can be omitted if one uses sufficiently informative priors on at least some parameters.

3. Ranking and Selecting Inspectors

Suppose we are interested in determining which inspector has the lowest overall error rate. Here, we have chosen to combine the false positive and false negative rates into a single positive likelihood ratio (LR), $\eta_j = (1 - \theta_{j,+}) / \theta_{j,-}$. Whichever inspector has the highest likelihood ratio would be determined to be the best. The positive likelihood ratio may not always be the most appropriate combination of the error rates, however, it is simply the one we use here as an example. In some cases, the negative likelihood ratio, $\eta_j = \theta_{j,+} / (1 - \theta_{j,-})$ or even a weighted sum of θ_+ and θ_- may be more appropriate. This approach can be decided on a case by case basis. We follow the method of [5] who have derived a decision-theoretic approach to partition parameters into two sets based on an ordering of the parameters of interest. Also, [6] extended their work to determine the largest Poisson rate when counts are subject to misclassification. Here we apply the method to subset a group of inspectors into a superior set, S , and an inferior set, S^C .

In the creation of a best subset, there are m separate two-state decision prob-

lems. Each decision involves whether or not to place an inspector’s likelihood ratio in the superior set, $d_+^k : \eta_k \in S$. We assign following constant loss functions:

$$L_+^k(\eta) = \begin{cases} 0 & \text{if } \eta_k = \eta_{best} \\ c_1 & \text{if } \eta_k \neq \eta_{best} \end{cases} \text{ and } L_-^k(\eta) = \begin{cases} c_2 & \text{if } \eta_k = \eta_{best} \\ 0 & \text{if } \eta_k \neq \eta_{best} \end{cases} \quad (10)$$

where L_+^k and L_-^k are the loss functions for d_+^k and d_-^k , respectively. To make a decision, only $c = c_2/c_1$ is required. These loss functions determine the decision criteria: take action d_+^k and include η_k as a candidate for the largest parameter if $P(\eta_k = \eta_{best} | \mathbf{x}) \geq 1/(c+1)$. Here, generally, $c_2 > c_1$ because failing to place the best η_k in S is the more serious error. Thus, c should be chosen larger than 1.

The probability that η_k is the best of the likelihood ratios is

$$P(\eta_i = \eta_{best} | \mathbf{x}) = \int_0^1 \int_0^{\eta_i} \dots \int_0^{\eta_i} p(\eta | \mathbf{x}) d\eta_1 \dots d\eta_{i-1} d\eta_{i+1} \dots d\eta_m d\eta_i, \quad (11)$$

where $p(\eta | \mathbf{x})$ is the marginal posterior of the likelihood ratios. MCMC methods are used to approximate (11) numerically. To accomplish this task, we generate a sample $(\eta_{k1}, \eta_{k2}, \dots, \eta_{kB})'$, for $k = 1, 2, \dots, m$ of size B from the posterior distribution, and then approximate the posterior probability that η_k is the best parameter by

$$\hat{P}(\eta_k = best(\eta_1, \dots, \eta_m) | \mathbf{x}) = \frac{\#(\eta_{ki} = best(\eta_{1i}, \dots, \eta_{mi}))}{B}, \quad (12)$$

where $k = 1, \dots, m$ and $i = 1, 2, \dots, B$, and B is the Monte Carlo repetition size.

The parameter η_k is included in the superior set S if

$$\hat{P}(\eta_k = best(\eta_1, \dots, \eta_m) | \mathbf{x}) \geq 1/(c+1), k = 1, \dots, m. \quad (13)$$

4. Example

As an example we consider data from [4] on a sample of 38 prints produced by inkjet cartridges. Three inspectors analyzed each print 3 times. Only the total number of passes out of the 9 inspections was provided, so for illustrative purposes, for those parts that did not have 0 or 9 passes, we distributed the number of passes across the three inspectors in a way to best match the frequentist parameter estimates provided in [4]. We assign beta (1, 9) priors to both μ_+ and μ_- since both of these quantities are expected to be considerably below 0.50. Our expert was 95% certain that both misclassification rates were less than 0.40, and a beta (1, 9) prior appropriately modeled the uncertainty. These distributions have prior 95% intervals of (0.003, 0.336) and have an equivalent sample size of 10 observations, and, therefore, would be considered mildly informative. A beta (1, 1) prior is used for τ , and Gamma (0.1, 0.1) priors are used for both γ_+ and γ_- . A burn-in of 10,000 iterations was used and inferences were based on the 20,000 subsequent iterations. The posterior summaries for each model parameter are provided in **Table 1**.

Table 1. Posterior summary for [4] example.

	Post.Median	Post.SD	val2.5pc	val97.5pc
μ_-	0.07818	0.06041	0.01913	0.251
μ_+	0.126	0.06176	0.04415	0.2841
γ_-	4.423	6.154	0.5477	22.93
γ_+	7.836	8.252	1.183	31.93
θ_{1-}	0.08885	0.03579	0.03632	0.1776
θ_{2-}	0.001537	0.01049	2.20E-06	0.03754
θ_{3-}	0.03101	0.02175	0.005326	0.08751
θ_{1+}	0.04653	0.03027	0.009797	0.1253
θ_{2+}	0.1283	0.04255	0.06432	0.2303
θ_{3+}	0.148	0.04656	0.07318	0.2543
η_1	19.52	32.73	7.21	92.31
η_2	7.746	2.995	4.31	15.45
η_3	6.525	2.471	3.798	13.19

From **Figure 1**, we see that when combined into the positive likelihood ratio, where a higher number is better, Inspector 1 has the overall highest LR. To apply the decision theoretic procedure to determine if any inspector is “best,” we compute the posterior probabilities of each likelihood ratio parameter being the largest. Here, a value of $c = 10$, implies that it is 10 times worse to leave the best inspector out of the superior set than to put an inferior inspector in the superior set, the critical probability would then be $1/(10 + 1) = 0.091$. The probabilities that Inspectors 1, 2, and 3 are each in the superior set are 0.891, 0.083, and 0.026, respectively. Thus, here, only Inspector 1 exceeds the 0.091 probability threshold. Thus, inspector 1 would be the only inspector placed in the superior set.

5. A Simulation Study

We conducted a simulation study to determine the effectiveness of the subset selection procedure. We set the number of inspectors to be $m = 7$ and the number of repeats to be $l = 3$. For $\tau = 0.5$, $\mu_+ = 0.15$, $\mu_- = 0.1$, $\gamma_+ = 20$, and $\gamma_- = 40$ we generated a single set of $\theta_{+,j}$'s and $\theta_{-,j}$'s. The values for $\theta_{+,j}$, $\theta_{-,j}$ and the corresponding likelihood ratios are presented in **Table 2**.

The prior distributions used were

$$\mu_+ \sim \text{beta}(1,1), \quad (14)$$

$$\mu_- \sim \text{beta}(1,1), \quad (15)$$

$$\gamma_+ \sim \text{Gamma}(0.1,0.1), \quad (16)$$

$$\gamma_- \sim \text{Gamma}(0.1,0.1), \quad (17)$$

and

$$\tau_+ \sim \text{beta}(1,1). \quad (18)$$

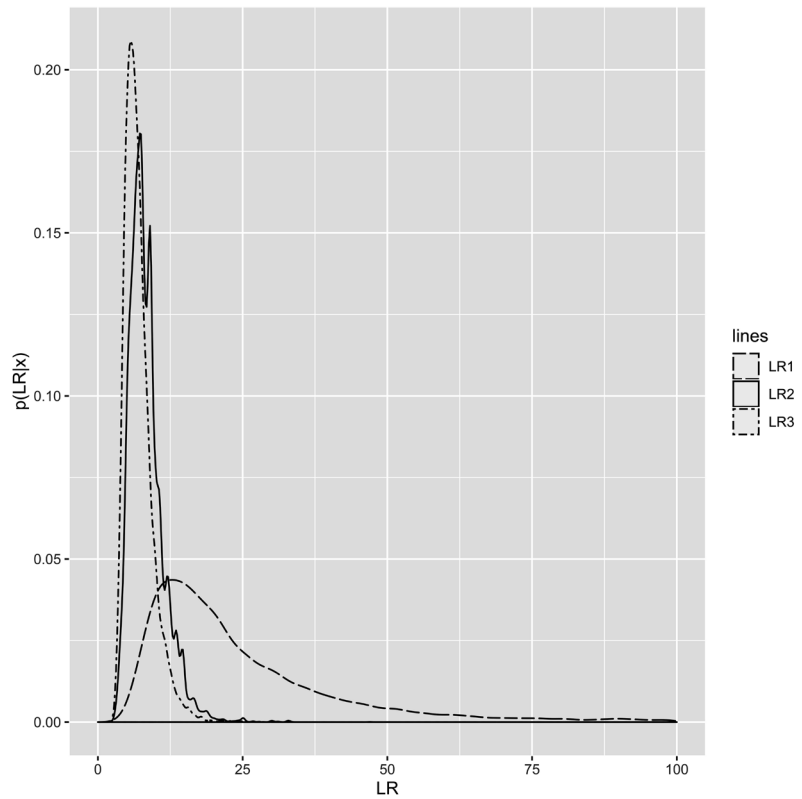


Figure 1. Posterior distributions of likelihood ratios.

Table 2. Misclassification parameters for simulation study.

Inspector	θ_-	θ_+	LR
1	0.014	0.09	10.96
2	0.084	0.218	4.20
3	0.094	0.119	7.60
4	0.224	0.168	4.62
5	0.175	0.153	5.39
6	0.033	0.076	12.72
7	0.02	0.105	9.33

Table 3. Simulation results for $N = 50$.

Inspector	$P(\eta_k = \eta_{best})$	2.5%	Median	97.5%
1	0.29	1.05	2.59	5.16
2	0.01	3.34	5.77	6.98
3	0.09	1.43	3.8	6.3
4	0.01	2.91	5.55	6.97
5	0.03	2.39	5.06	6.87
6	0.37	1.01	2.29	4.89
7	0.20	1.13	2.95	5.58

Table 4. Simulation results for $N = 100$.

Inspector	$P(\eta_k = \eta_{best})$	2.5%	Median	97.5%
1	0.31	1.06	2.22	4.08
2	0	4.66	6.23	6.98
3	0.05	1.84	3.73	5.55
4	0	4.15	5.9	6.94
5	0	3.39	5.24	6.72
6	0.47	1.01	1.88	3.72
7	0.16	1.24	2.79	4.65

Table 5. Simulation results for $N = 200$.

Inspector	$P(\eta_k = \eta_{best})$	2.5%	Median	97.5%
1	0.3	1.07	2.11	3.62
2	0	5.25	6.43	6.99
3	0.03	2.13	3.69	4.98
4	0	4.81	6.06	6.94
5	0	4.01	5.25	6.52
6	0.55	1.01	1.66	3.15
7	0.11	1.37	2.81	4.19

Thus, relatively non-informative priors were employed for all parameters. We considered sample sizes of $N = 50$, 100, and 200 and generated 1000 data sets for each sample size. We monitored the probability that each likelihood ratio is the largest and the 95% credible set of the rank for each η_i . These results are provided in **Tables 3-5**. For the decision theory problem we used $c = 10$ and, thus, also monitored whether the true “best” inspector was included in the superior set as well as the average size of the superior set. In this paper we are focusing on the ranking and selection methods, so those are the simulation results we report here. We also monitored posterior means and found they were close to the true values with small bias and coverage of intervals close to nominal for all parameters. The bias and coverage results are available upon request.

For all simulations, Inspector 6, who was the “best” inspector, yielded the highest probability of having the largest likelihood ratio, and, therefore, was correctly estimated to be the best inspector the most times. Also, the credible intervals on the ranks for Inspector 6 were consistently closest to the top rank. Conversely, Inspector 2, who was the “worst” inspector, produced the lowest probability of having the largest likelihood ratio and, was correctly considered the worst inspector the most times. Inspector 2 also yielded credible intervals for the rank with the largest values, implying this inspector was generally ranked last. Thus both the ranking and selection procedures performed well.

For all three considered sample sizes, the probability of the “best” inspector

being included in the superior set was greater than 0.9. The average size of the superior set was 2.8 for a sample size of 50, 2.4 for a sample size of 100 and 2.2 for a sample size of 200.

6. Conclusions

In this paper we have proposed a Bayesian random effects model for a binary measurement system. As shown in our real data example, combining the data with mildly informative priors yields an identifiable model where inferences can be made on the overall classification rates along with comparisons of individual inspectors. Our simulation study shows that for moderate sample sizes, even when information is not available for priors, the procedure works well with the best inspector being included in the superior set a large percentage of the time.

The methods we have proposed could be extended to comparing overall defective rates and classification probabilities of manufacturing plants instead of inspectors, as we have done here. Expanding to continuous measurements from binary is also potentially of interest.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Danila, O., Steiner, S.H. and MacKay, R.J. (2008) Assessing a Binary Measurement System. *Journal of Quality Technology*, **40**, 310-318. <https://doi.org/10.1080/00224065.2008.11917736>
- [2] van Wieringen, W.N. (2008) Measurement System Analysis for Binary Data. *Technometrics*, **50**, 468-478. <https://doi.org/10.1198/004017008000000415>
- [3] Danila, O., Steiner, S.H. and MacKay, R.J. (2012) Assessing a Binary Measurement System with Varying Misclassification Rates Using a Latent Class Random Effects Model. *Journal of Quality Technology*, **44**, 179-191. <https://doi.org/10.1080/00224065.2012.11917894>
- [4] Boyles, R.A. (2001) Gauge Capability for Pass-Fail Inspection. *Technometrics*, **43**, 223-229. <https://doi.org/10.1198/004017001750386332>
- [5] Bratcher, T.L. and Bhalla, P. (1974) On the Properties of an Optimal Selection Procedure. *Communications in Statistics: Simulation and Computation*, **3**, 191-196. <https://doi.org/10.1080/03610927408827118>
- [6] Stamey, J.A., Bratcher, T.L. and Young, D.M. (2004) Parameter Subset Selection and Multiple Comparisons of Poisson Rate Parameters with Misclassification. *Computational Statistics & Data Analysis*, **45**, 467-479. [https://doi.org/10.1016/S0167-9473\(03\)00068-9](https://doi.org/10.1016/S0167-9473(03)00068-9)