Contents lists available at GrowingScience

International Journal of Industrial Engineering Computations

homepage: www.GrowingScience.com/ijiec

Modeling quality control data using mixture of parametrical distributions

Jorge Alberto Achcar[a], Claudio Luis Piratelli[b,c] and Roberto Molina de Souza[d*]

[a] *Department of Social Medicine, University of São Paulo. Av. Bandeirantes, 3900 - Monte Alegre - CEP: 14049-900. Ribeirão Preto - SP - Brazil*
[b] *Master Program in Production Engineering, University Center of Araraquara. Rua Carlos Gomes, 1338 - Centro - CEP 14801-340. Araraquara - SP – Brazil*
[c] *Aeronautics Institute of Technology (ITA), Brazil*
[d] *Coordination of Mathematics, Federal Technological University of Paraná. Av. Alberto Carazzai, 1640 - Centro - CEP: 86300-000. Cornélio Procópio - PR - Brazil*

CHRONICLE

Article history:
Received March 1 2013
Received in revised format March 8 2013
Accepted March 8 2013
Available online March 11 2013

Keywords:
Regression
Quality control times
Mixture models
Bayesian methods
MCMC methods

ABSTRACT

In this paper, we present a Bayesian analysis of a data set selected from a Brazilian food company. This data set represents the times taken for different quality control analysts to test manufactured products arriving at the company's quality control department. The samples selected from each batch contain mixtures of different products, which may be submitted to quality testing taking different times. From preliminary analysis of the data, it was observed that the histograms presented two clusters, indicating a mixture of distributions. A mixture of parametrical distributions was thus assumed in the presence of a covariate in order to analyze the data set and to establish standards to be used by the company for the times taken by the analysts. Inferences and predictions are obtained using a Bayesian approach with standard existing Markov Chain Monte Carlo (MCMC) methods.

1. Introduction

The times taken in carrying out quality control tests can often vary greatly, influenced by a range of factors, including the experience and skill of the quality control analysts, and the presence of different products being analyzed. It is, then, of interest to industrial managers to model these data sets, from which they can make inferences and predictions and identify important factors that could affect these times. In the study herein, we consider a data set from a food company in São Paulo state, Brazil. This data set comprises quality control times for two different analysts observed on different days. This data set comprises random samples selected from all the batches of manufactured products. These different products are assessed by quality control tests lasting for different timeframes. The batches arrive in random order at the quality control department. Fig. 1 shows the histograms for the test times for the two analysts.

* Corresponding author.
E-mail: rmolinasouza@utfpr.edu.br (R. Molina de Souza)

doi: 10.5267/j.ijiec.2013.03.003

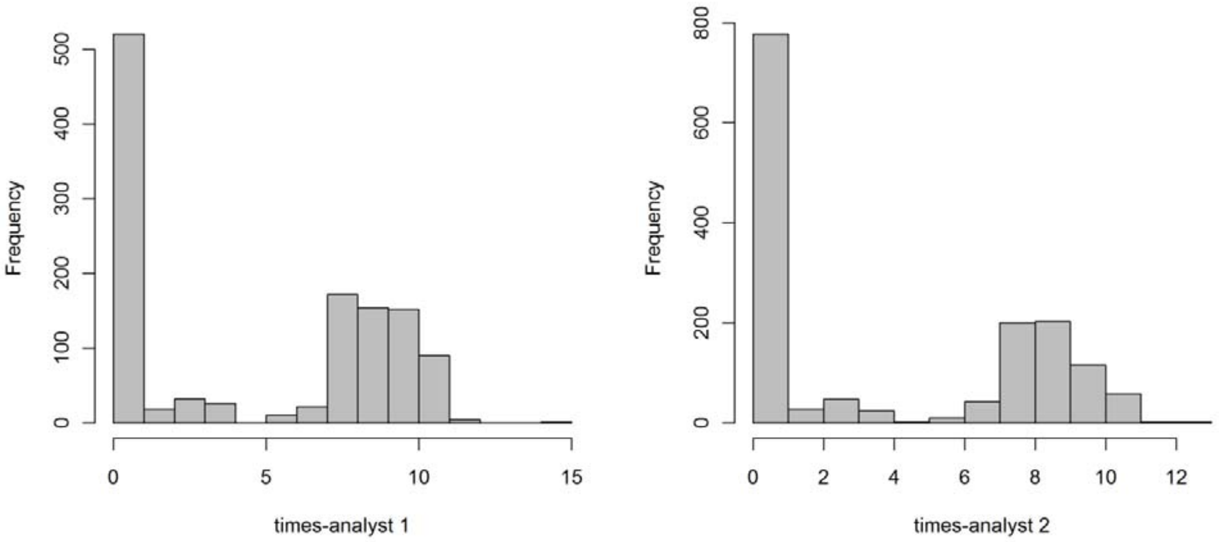


Fig. 1. Histograms for the test times considering the two analysts.

From the histograms shown in Fig. 1, it can be observed that the two analysts perform the control tests differently (analyst 1 with 1,200 samples, and analyst 2 with 1,504 samples). A discordant observation (greater than 64 minutes) for analyst 1 was discarded. It was observed that analyst 2 took less time than analyst 1. Fig. 2 shows the histogram for all the combined data for both analysts (n=2704 observations). From the histograms in Fig. 1 and Fig. 2, the mixture of two distributions for the times taken for the quality control tests is observed, where a proportion of units has short times and a second proportion of the data has long times. This made it possible to use a mixture of parametrical distributions to analyze the data. A mixture of parametrical distributions has been considered by many authors in the literature to analyze non-homogeneous data sets (see, for example, Titterington et al. 1985; Stephens, 2000a; Stephens, 2000b; Richardson & Green, 1997; Diebolt & Robert, 1994; Dey et al., 1995; Finkelstein & Esaulova, 2001).

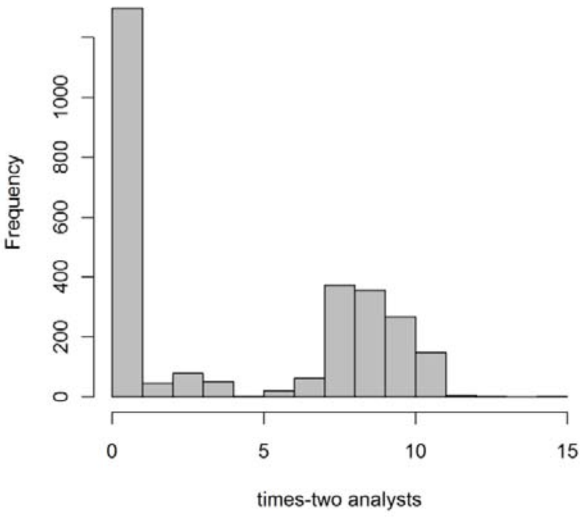


Fig. 2. Histogram of times for two analysts

In the parametric mixture model, the component distributions are from a parametric family with known parameters θ_j with the probability density function given by

$$f(t) = \sum_{j=1}^{k} p_j f_j (t \mid \theta_j) \tag{1}$$

for some mixture proportions $0 \leq p_j \leq 1$, where $p_1 + p_2 + \ldots + p_k = 1$. If $k=2$, we have a mixture of two distributions.

Inferences for finite mixture models could be obtained using Bayesian methods (see, for example, Mengersen & Robert, 1996; Carroll et al., 1999) where the posterior summaries of interest are obtained using simulation methods, especially standard Markov Chain Monte Carlo (MCMC) methods, such as the popular Gibbs sampling algorithm (see for example, Gelfand & Smith, 1990) and the Metropolis-Hastings algorithm (see, for example, Chib & Greenberg (1995)).

Recently, Achcar et al. (2012) published a paper with the same data analyzed in this article where the focus of analysis was the use of a Weibull distribution in the presence of a changing point. In this paper, the authors compare the results obtained from the use of the change point model with the results obtained from a model considering the mixture of two Weibull distributions. As the two competing methodologies showed to be appropriated to analyze this data set, this article aims to explore in more detail the use of mixtures of Weibull distributions, as a good alternative for quality engineers, also introducing a comparative study with the use of other mixture models like the mixture of normal distributions.

This paper is organized as follows: in section 2, the models and inference are presented considering a mixture of two normal distributions and a mixture of two Weibull distributions; in section 3, a Bayesian analysis for the data of the food company is presented. Finally, in section 4, some concluding remarks are presented.

2. Models and inference

In this section, we introduce two mixture models for the times taken for the quality control tests at the food company: a mixture of two normal distributions and a mixture of two Weibull distributions.

2.1. Mixture of two normal distributions

Since we have two analysts receiving samples for quality control tests in the food company, we first assume a mixture of two normal distributions considering a covariate X (an indicator variable for each analyst), where $X = 0$ for analyst 1 and $X = 1$ for analyst 2. Let T_i be a random variable denoting the quality control test time for the i^{th} sample $(i = 1,2,\ldots,n)$ where $n = 2704$, assuming a mixture of two normal distributions $N(\mu_{ji}; \sigma_j^2)$, $j = 1,2$ given (from Eq. (1)) by the density,

$$f(t_i) = p f_1(t_i \mid \mu_{1i}; \sigma_1^2) + (1-p) f_2(t_i \mid \mu_{2i}; \sigma_2^2) \tag{2}$$

where $\mu_{ji} = \lambda_j + \beta_j X_i$; $j = 1,2$; $i = 1,2,\ldots,2704$; $\lambda_2 = \lambda_1 + \theta$; $X_i = 0$ (analyst 1); $X_i = 1$ (analyst 2) and $f_j(t_i \mid \mu_{ji}; \sigma_j^2)$ is a normal density given by,

$$f_j(t_i \mid \mu_{ji}; \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2\sigma_j^2} (t_i - \mu_{ji})^2 \right] \tag{3}$$

For a Bayesian analysis of the regression mixture model defined by Eq. (2) and Eq. (3), we assume the following prior distributions for the parameters θ, β_j, λ_1 and $\tau_j = \frac{1}{\sigma_j^2}$, $j = 1,2$:

$$\begin{aligned} \theta &\sim N\left(a, 10^3\right); truncated(\theta > 0) \\ \beta_j &\sim N\left(0, 10^2\right) \\ \lambda_1 &\sim N\left(0, 10^6\right); truncated(\lambda_1 > 0) \\ p &\sim Beta(1,1) \\ \tau_j &\sim Gamma(0.1, 0.1) \end{aligned} \tag{4}$$

where $N\left(\mu, \sigma^2\right)$ denotes a normal distribution with mean μ and variance σ^2; $Gamma(b,c)$ denotes a gamma distribution with mean $\frac{b}{c}$ and variance $\frac{b}{c^2}$; $Beta(d,e)$ denotes a beta distribution with mean $\frac{d}{(d+e)}$ and variance $\frac{de}{\left[(d+e)^2(d+e+1)\right]}$. The hyperparameter a for the normal prior of θ is assumed known from a preliminary data analysis. This choice of a also implies in the identifiability of the mixture model. Note that we are assuming large variances for the prior distributions, that is, approximately non-informative priors (see for example, Paulino et al. (2003)). We further assume prior independence among the parameters.

2.2 Mixture of two Weibull distributions

Another possibility is to assume a mixture of two Weibull distributions for the times of the two analysts. In this case, we assume in Eq. (1), a mixture of two Weibull distributions (see for example, Lawless, 1982) given by the density,

$$f\left(t_i\right) = p f_1\left(t_i \mid \mu_{1i}; v_1\right) + (1-p) f_2\left(t_i \mid \mu_{2i}; v_2\right), \tag{5}$$

where $f_j\left(t_i \mid \mu_{ji}; v_j\right) = v_j \mu_{ji} t_i^{v_j - 1} \exp\left(-\mu_{ji} t_i^{v_j}\right)$ and the scale parameter of the Weibull distributions is given by $\mu_{ji} = \lambda_j \exp\left(\beta_j X_i\right)$, $j = 1,2;$ $i = 1,2,\ldots,2704;$ $\lambda_2 = \theta\lambda_1;$ v_j is the shape parameter of the Weibull distribution; $X_i = 0$ (analyst 1) and $X_i = 1$ (analyst 2). Note that the mean time for each component distribution in the mixture model (5) is given by,

$$mean_{ji} = \frac{1}{\mu_{ji}^{\frac{1}{v_j}}} \Gamma\left(1 + \frac{1}{v_j}\right) \tag{6}$$

where $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$ is a gamma function; $j = 1,2;$ $i = 1,2,\ldots,2704$. For a Bayesian analysis of the model we assume the following prior distributions for the parameters:

$$p \sim Beta(1,1) \quad \theta \sim Gamma(0.1, 0.1) \quad \lambda_1 \sim Gamma(1,1) \quad v_j \sim Gamma(0.1, 0.1); j = 1,2 \quad \beta_j \sim N(0,10) \tag{7}$$

3. A Bayesian analysis for the data of the food industry

To analyze the times taken for quality control tests by the two analysts at the food company using a Bayesian approach, we first assume a mixture of two normal distributions defined by Eq. (2) and Eq. (3) with priors Eq. (4) with $a = 8$. This model is denoted model 1. The value $a = 8$ was chosen from a preliminary data analysis (see Fig. 1 and Fig. 2).

In the simulation procedure of samples for the joint posterior distribution of $p, \beta_1, \beta_2, \lambda_1, \lambda_2, \tau_1, \tau_2$ and θ , we used OpenBUGS, an open source software available from http://openbugs.info/w/Downloads (see, for example, Lunn et al. (2009)).

OpenBUGS requires only the distribution of the data and the prior distributions for the parameters of the model, and the conditional posterior distributions used for the Gibbs sampling algorithm do not have to be specified; that is to say, the samples for the joint posterior distribution of interest is greatly simplified.

In the simulation procedure, a sample size of $5,000$ was initially simulated from the joint posterior distribution discarded to eliminate the effect of the initial values used in the iterative routine (burn-in sample). Following this burn-in sample another $20,000$ Gibbs samples were generated, taking every 20^{th} sample to have approximately uncorrelated samples, from which a final simulation sample of size $1,000$ was used to get the posterior summaries of interest. Convergence of the Gibbs sampling algorithm was monitored from the usual traceplots for each parameter sample. Table shows the posterior summaries obtained assuming the mixture of two normal distributions.

Table 1
Posterior summaries ("model 1")

parameter	mean	S. D.	95% credible interval
p	0.4812	0.0096	(0.4622;0.4998)
$1-p$	0.5188	0.0096	(0.5003;0.5380)
β_1	−0.0247	0.0068	(−0.0376;−0.0116)
β_2	−0.4060	0.1221	(−0.6458;−0.1623)
λ_1	0.7521	0.0053	(0.7421;0.7632)
λ_2	8.011	0.0845	(7.852;8.177)
τ_1	67.39	2.75	(62.16;72.99)
τ_2	0.2005	0.0076	(0.1857;0.2166)
σ_1	0.1219	0.0025	(0.1171;0.1269)
σ_2	2.234	0.043	(2.149;2.320)
θ	7.258	0.085	(7.100;7.424)

Fig. 3 shows times observed for quality control versus samples, and the fitted means (Bayesian estimates for the means) versus samples.

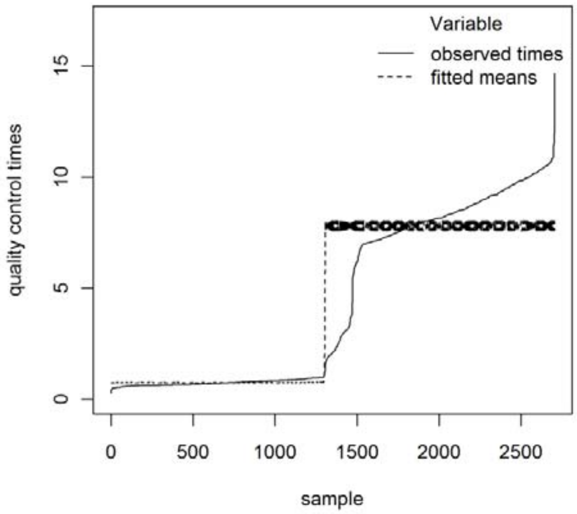


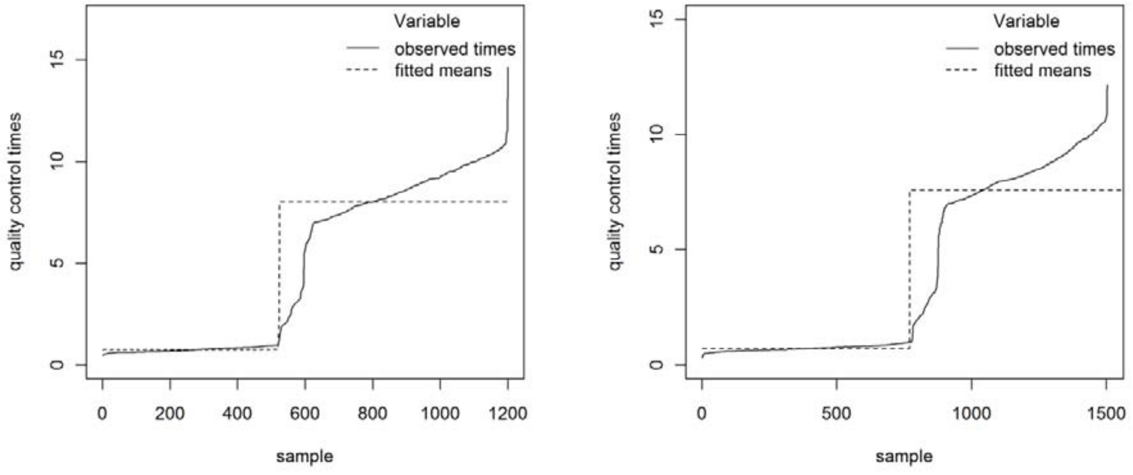
Fig. 3. Quality control times and fitted means versus samples (all data set)

A good fit was observed for the data for the proposed model. The posterior mean for the first normal cluster for analyst 1 had a Monte Carlo estimate based on the $1,000$ simulated Gibbs samples given by $\hat{\lambda}_1 = 0.7521$, and a 95% credible interval for λ_1 given by $(0.7421;0.7632)$. For the second normal cluster, the posterior mean for λ_2 is estimated by $\hat{\lambda}_2 = 8.011$ with a 95% credible interval given by $(7.852;8.177)$.

From $\mu_{ji} = \lambda_j + \beta_j X_i$, for analyst 2 $(X_i = 1)$, there is a Bayesian estimate for the mean for the first normal cluster given by $\hat{\lambda}_1 + \hat{\beta}_1 = 0.7521 - 0.0247 = 0.7273$, and for the mean of the second normal cluster given by $\hat{\lambda}_2 + \hat{\beta}_2 = 8.011 - 0.406 = 7.6050$. That is, analyst 2 has less time to perform the quality tests than analyst 1. It is also observed that the regression parameters β_1 and β_2 have significant effects on the quality control times for the analysts, since zero is not included in the 95% credible interval for β_1 and β_2 (see Table 1). Similar proportions of samples in the two clusters of data are observed. To check the quality of fit for the data for the mixture of normal distributions, we could calculate the differences of observed and fitted means given by

$$fit(l) = \sum_{i=1}^{2704} |t_i - \hat{\mu}_i| \tag{8}$$

where $\hat{\mu}_i$ are the fitted means, $i = 1,2,\ldots,2704$ and l indexes model (here, $l = 1$). We observe $fit(1) = 2338.59$. From (8), it is observed that $fit(1)$ for analyst 1 is given by 1115.80 and for analyst 2 is given by 1222.79. In Fig. 4, we have the observed values and fitted means for each analyst.



(a) (b)

Fig. 4. Quality control times and fitted means versus samples: (a) Analyst 1 and (b) Analyst 2

In the Bayesian analysis of the data from the food company, a mixture of two Weibull distributions is also assumed as defined by Eq. (5) and the prior distributions Eq. (7). Let us denote this model as model 2. Considering the same simulation steps used for model 1, Table 2 shows the posterior summaries of interest based on $1,000$ simulated Gibbs samples using the OpenBUGS software.

From the results in Table 2, the covariate X (analyst) shows a significant effect in the means of the two cluster Weibull distributions since zero is not included in the 95% credible intervals for β_1 and β_2. Note that the regression parameter β_j, $j = 1,2$ has a multiplicative effect on the scale parameter for the Weibull distributions $\mu_{ji} = \lambda_j \exp(\beta_j X_i)$, $j = 1,2$; $i = 1,2,\ldots,2704$.

From the results in Table 2, we observe that since $\mu_{ji} = \lambda_j \exp(\beta_j X_i)$, $j = 1,2$; $i = 1,2,\ldots,2704$, we get Bayesian estimates for the scale parameter of the Weibull distributions (5), given, respectively by $\hat{\mu}_1^{(1)} = \hat{\lambda}_1 = 4.265$ and $\hat{\mu}_2^{(1)} = \hat{\lambda}_2 = 0.000076$ for analyst 1 $(X_i = 0)$ and $\hat{\mu}_1^{(2)} = \hat{\lambda}_1 \exp(\hat{\beta}_1) = 4.265\exp(0.1733) = 5.0720$ and $\hat{\mu}_2^{(2)} = \hat{\lambda}_2 \exp(\hat{\beta}_2) = 0.000076\exp(0.1839) = 0.000091$ for analyst 2 $(X_i = 1)$.

Table 2
Posterior summaries ("model 2")

parameter	Mean	S. D.	95% credible interval
p	0.4807	0.0099	(0.4615;0.5015)
$1-p$	0.5193	0.0099	(0.4991;0.5386)
β_1	0.1733	0.0594	(0.0592;0.2868)
β_2	0.1839	0.0554	(0.0760;0.2945)
λ_1	4.265	0.228	(3.842;4.699)
λ_2	0.000076	0.00002	(0.000046;0.00016)
v_1	6.5520	0.1460	(6.2630;6.8580)
v_2	4.3970	0.1170	(4.1950;4.5970)
θ	0.000018	0.000005	(0.00001;0.00003)

From Eq. (6), we get Bayesian estimates for the means of both analysts, given by:

(i) Analyst 1:

$$\widehat{mean}_1^{(1)} = \frac{1}{\left[\hat{\mu}_1^{(1)}\right]^{\frac{1}{\hat{v}_1}}}\Gamma\left(1+\frac{1}{\hat{v}_1}\right) = \frac{1}{4.265^{\frac{1}{6.552}}}\Gamma\left(1+\frac{1}{6.552}\right) = 0.7471$$

for cluster 1, and

$$\widehat{mean}_2^{(1)} = \frac{1}{\left[\hat{\mu}_2^{(1)}\right]^{\frac{1}{\hat{v}_2}}}\Gamma\left(1+\frac{1}{\hat{v}_2}\right) = \frac{1}{0.000076^{\frac{1}{4.397}}}\Gamma\left(1+\frac{1}{4.397}\right) = 7.8796$$

for cluster 2 of the mixture of two Weibull distributions.

(ii) Analyst 2:

$$\widehat{mean}_2^{(1)} = \frac{1}{\left[\hat{\mu}_1^{(2)}\right]^{\frac{1}{\hat{v}_1}}}\Gamma\left(1+\frac{1}{\hat{v}_1}\right) = \frac{1}{5.07205^{\frac{1}{6.552}}}\Gamma\left(1+\frac{1}{6.552}\right) = 0.7276$$

for cluster 1, and

$$\widehat{mean}_2^{(2)} = \frac{1}{\left[\hat{\mu}_2^{(2)}\right]^{\frac{1}{\hat{v}_2}}}\Gamma\left(1+\frac{1}{\hat{v}_2}\right) = \frac{1}{0.000091^{\frac{1}{4.397}}}\Gamma\left(1+\frac{1}{4.397}\right) = 7.5633$$

for cluster 2 of the mixture of two Weibull distributions.

Similar results are observed for the means of the two distributions in the mixture model assuming normal or Weibull distributions (see Table 1). Fig. 5 and Fig. 6 show the timings of quality control versus samples and also of the fitted means (see Eq. (6)) versus samples considering respectively, the combined data of the two analysts; the data of analyst 1 and the data of analyst 2.

From Eq. (8), we get $fit_{(2)}^{(all)} = 2357$ (all combined data); $fit_{(2)}^{(1)} = 1131.98$ (data of analyst 1) and $fit_{(2)}^{(2)} = 1225.42$ (data of analyst 2). Overall, both models give similar inference results, but model 1 (mixture of two normal distributions) shows a small improvement in the fit for the data.

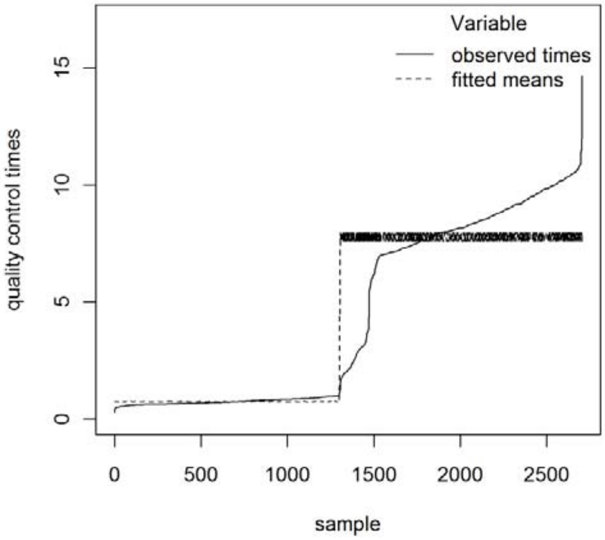
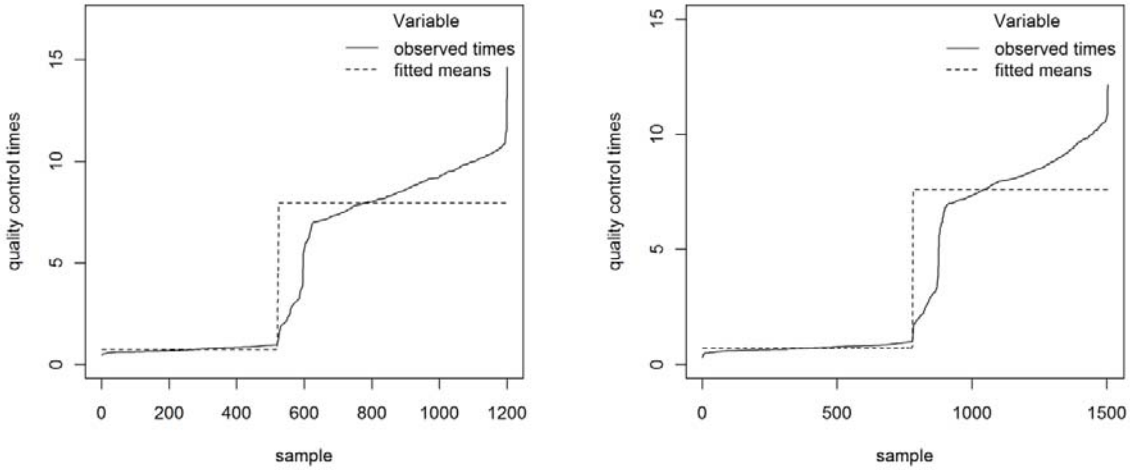


Fig. 5. Quality control times and fitted means versus samples (all data set)



(a) (b)

Fig. 6. Quality control times and fitted means versus samples: (a) Analyst 1 and (b) Analyst 2

4. Concluding remarks

In industrial applications, managers and industrial engineers are usually interested in modeling the time taken in tasks carried out by different operators, especially in order to get performance indicators for their systems (by Discrete Event simulation, for example). This paper analyzed the times taken in quality control carried out by two different analysts in a Brazilian food company. The main goal at every company is to standardize and optimize these times, as a reference that should be followed by all analysts in the company. In many cases, as was considered in this paper based on the data set from the food company, the batches of manufactured products arrive in a random order at the company's quality control department, with a mixture of different products, and the quality control tests usually take different times. Hence, the use of a mixture of parametrical distributions in the presence of a covariate, considered as it was to be of great interest in industrial applications. In this case, we were able to consider Bayesian confidence intervals, or classical confidence intervals, for the means of the two component distributions for the best analyst (short times) as standard reference intervals to be followed by all the operators in the company's quality control department.

It is important to point out that these results could be generalized for other mixed data sets consisting of more than two clusters, and in the presence of other covariates that could affect the performance of analysts, such as skill in performing the quality control tests, experience, calibration of the test equipment, day of the week, temperature, and many other factors. The use of Bayesian methods for a mixture of parametrical distributions especially considering existing simulation MCMC methods, such as the Gibbs sampling algorithm and OpenBUGS software, could be of great interest, since the computational cost to get the posterior summaries required is not high.

References

Carroll, R.J., Roeder, K., & Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics,* 55, pp. 44–54.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49, 327–335.

Dey, D.K., Kuo, L., & Sahu, S.K. (1995). A bayesian predictive approach to determining the number of components in a mixture distribution. *Statistics and Computing*, 5, 297–305.

Diebolt, J., & Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological),* 56, pp. 363–375.

Finkelstein, M.S., & Esaulova, V. (2001). Modeling a failure rate for a mixture of distribution functions. *Probab. Eng. Inf. Sci.*, 15, 383–400.

Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, 398–409.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data* (Wiley Series in Probability & Mathematical Statistics). John Wiley & Sons.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, 28, 3049–3067.

Mengersen, K., & Robert, C. (1996). *Bayesian Statistics,* 5. Oxford University Press, Oxford. chapter Testing for mixtures: a Bayesian entropy approach.

Paulino, D.C., Turkman, M.A.A., & Murteira, B. (2003). *Estatstica Bayesiana*. Fundação Calouste, Lisboa.

Richardson, S., & Green, P.J. (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological),* 59, pp. 731–792.

Stephens, M., 2000a. *Bayesian methods for mixtures of normal distributions*. Master's thesis.

Stephens, M., 2000b. Dealing with label switching in mixture models. *Journal Of The Royal Statistical Society Series B*, 62, 795–809.

Titterington, D.M., Smith, A.F.M., & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons Ltd. SERIES: Wiley Series in probability and mathematical statistics.