Modeling the Distribution of the Patients' Stay at PACU

by Jingyi Yang
Mentor: Dr. Yang Li
Department of Mathematics and Statistics, University of Minnesota Duluth

Seeing a doctor has become an indispensable part of everyone's life. With the increasing demand of patients' medical experience, the patient service level should also be improved. In a recent study to improve the patient flows at a local hospital, data of the patients' length of stay at the Emergency Department and the roomed time in the post-anesthesia care unit (PACU) were collected. Those data need to be analyzed to fit into different theoretical distributions to be used in mathematical and simulation models. However, due to the irregularity of the data, normal software that comes with the simulation package cannot find a theoretical distribution for the data. Since the length of stay in the hospital greatly affects the patients' feelings about the medical services and the overall satisfaction rate, we are suggesting a statistical approach in this UROP proposal.

The data set we are considering in this project includes $N = 30{,}534$ patients' stay in the PACU. The distribution of these data set is shown in Figure 1. The unit is in days. Most of the data points are below one but some large values do exist. The mean and median are 0.193 and 0.11, respectively. The maximum value is 29.96 (not shown in the figure).

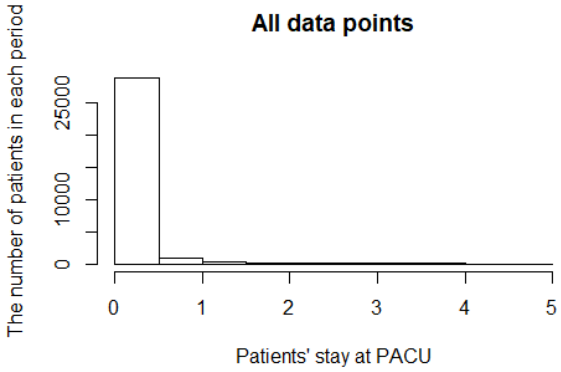

Figure 1 Histogram of patients' stay at PACU.

Since all values are positive, a gamma distribution was first fit to the whole data set. The probability density function of a gamma distribution is [1]

$$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

where $k > 0$ is the shape parameter and $\theta > 0$ is the scale parameter. I use maximum likelihood estimation (MLE) to find the parameter estimates. The histograms and the fitted MLE line for data points less than 0.5 is shown in the left panel of Figure 2. A large discrepancy between the fit and the real data is clearly displayed, implying the inadequacy of fitting a single

gamma distribution. The reason for the inconsistency could stem from the fact that there are more than one underlying groups of patients. Hence, we try a mixture of two gamma distributions with the argument that one of them represents those patients with mild conditions who need to stay at PACU for a short period of time, while the second distribution for patients with more serious conditions and longer stays. In statistics, the traditional way of fitting a mixture model is using the Expectation-Maximization (EM) algorithm which can maximize the likelihood function and give us the optimal estimates for the parameters [2]. The key point of the EM algorithm is the assumption of missing values and incomplete data. In this mixture model, the missing information is which group a given patient belongs to. The distribution function for observation x is

$$p(x|\Theta) = \sum_{i=1}^{M} \alpha_i \, p_i(x|\boldsymbol{\theta}_i)$$

where $p_i(x|\boldsymbol{\theta}_i)$ is the ith gamma density of the mixture with $\boldsymbol{\theta}_i = (k_i, \theta_i)$ being the set of its parameters, and α_i is the weight of the ith density (with the constraints $0 \leq \alpha_i \leq 1$ and $\sum_{i=1}^{M} \alpha_i = 1$). For our data set, $M = 2$ while α_1 and α_2 represent the weights of two gamma distributions with pdf $p_1(x|\boldsymbol{\theta}_1)$ and $p_2(x|\boldsymbol{\theta}_2)$. There are totally six parameters, that is, $\alpha_1, \alpha_2, k_1, k_2, \theta_1, \theta_2$.

The incomplete-data log-likelihood function

$$\log(\mathcal{L}(\Theta|\mathrm{X})) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{2} \alpha_j \, p_j(x_i|\boldsymbol{\theta}_j)\right)$$

can be maximized by the EM algorithm. The E-step gives

$$Q(\Theta, \Theta^g) = \sum_{l=1}^{2}\sum_{i=1}^{N} \log(\alpha_l)\, p(l|x_i, \Theta^g) + \sum_{l=1}^{2}\sum_{i=1}^{N} \log\big(p_l(x_i|\boldsymbol{\theta}_l)\big) p(l|x_i, \Theta^g)$$

where Θ^g is the estimate from the previous step of iteration. We need to maximize it over Θ in the M-step. This optimization with linear constraints can be done by the `constrOptim` function in R. The results of the optimal estimates are $\alpha_1 = 0.892, \alpha_2 = 0.108$, $k_1 = 2.642$, $k_2 = 1.076$, $\theta_1 = 0.0445$, $\theta_2 = 0.776$. The curve in the right panel of Figure 2 shows the fitted line from these values which matches the real data much better than using a single a gamma distribution.

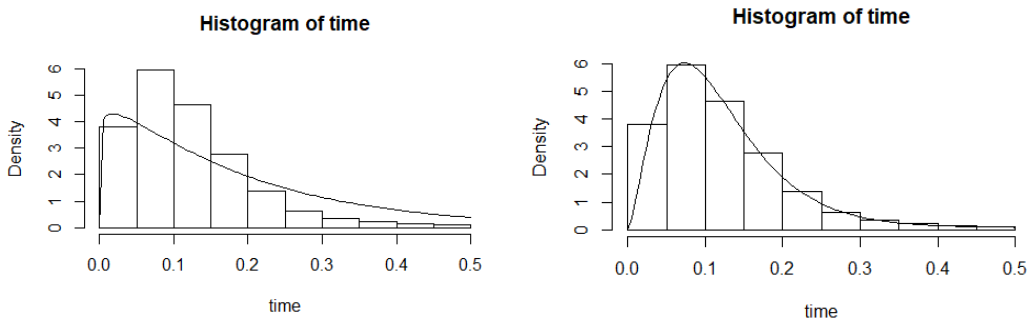

Figure 2 Histograms and the fitted MLE lines. The left figure fits a single gamma distribution to the data, while the right figure fits two gamma distributions using the EM algorithm.

The next step is to get the confidence intervals for the parameters. Assuming a normal distribution, the formula of the confidence interval for parameter θ is [1]

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{Var(\hat{\theta})}$$

where $\hat{\theta}$ is the MLE estimate and $z_{\alpha/2}$ is the standard normal quantile. The variance can be numerically evaluated by using the bootstrap method. Bootstrap is a resampling method where samples are selected from the observed data through sampling with replacement. After getting a sequence of bootstrap samples, we can approximate the variance of the original estimator by the variance of these bootstrap samples. In this research, I resampled 50 times and obtained the 95% confidence intervals for all six parameters shown in Table 1. The table shows that the first component (with an estimated mean $\widehat{k_1\theta_1} = 0.117$) contributes about 90% of the data while the second component has an estimated mean 0.835. As we mentioned earlier, the first gamma distribution represents patients with a shorter stay, and the second gamma distribution stands for those patients with more serious conditions and longer stay in the PACU.

Parameter	MLE Estimate	95% Bootstrap CI
α_1	0.892	(0.887, 0.897)
α_2	0.108	(0.103, 0.113)
k_1	2.642	(2.602, 2.683)
k_2	1.076	(0.984, 1.169)
θ_1	0.0445	(0.0436, 0.0455)
θ_2	0.776	(0.664, 0.8890)

Table 1 MLE estimates and 95% confidence interval using bootstrap.

In this research project, the most important things I have learned are the EM algorithm and resampling method such as bootstrap. I have also gained experience in writing R code to solve real world problems. I believe that the results of our study will lead to the improvement of hospital departments in medical facility allocation, patient appointment system, doctor diagnosis time planning and many other aspects.

References

[1] L. J. Bain and M. Engelhardt. Introduction to Probability and Mathematical Statistics (2nd edition). Cengage, (2000).

[2] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, 4, 126 (1998).

[3] S. F. Chong and R. Choo. Introduction to Bootstrap. Proceedings of Singapore Healthcare, 20, 236, (2011).