Bayesian Updating of the Gamma Distribution for the Analysis of Stay Times in a System

K. Aboura and Johnson I. Agbinya

College of Business Administration, University of Dammam, Dammam, Saudi Arabia, kaboura@ud.edu.sa

*Electronic Engineering, Latrobe University, Melbourne, Australia
J.Agbinya@latrobe.edu.au*

Received 4 July 2013; Accepted 15 November 2013

Publication 23 January 2014

Abstract

The evaluation of traffic in a system is an important measurement in many studies. Counting the number of items in a system has applications in all processing operations. Electronic messages circulating in a network, clients shopping in a supermarket and students attending programs in a school are examples of entities entering, staying and exiting a system. We introduce a Bayesian updating methodology for the gamma distribution for the analysis of stay times in a system. The methodology was first developed for areas monitored by surveillance cameras. The number of people in the covered area was determined and the average stay time was estimated using a gamma probability distribution. We extend the application to the generic case and present a simple updating methodology for the estimation of the model parameters.

Keywords: Traffic estimation, gamma distribution, Bayesian statistics.

1 Introduction

Counting the number of entities in a system is essential for a multitude of management and monitoring functions. For example, forecasting the number of students in a school is essential to the planning of all functions in the

Journal of Machine to Machine Communications, Vol. 1, 1–16.

doi: 10.13052/jmmc2246-137X.114

© 2014 River Publishers. All rights reserved.

school. Students register, follow programs and at times drop from the school before completing their degrees. When the school size is large, it becomes essential to estimate the flow of students and regress it on capacity, staffing and funding variables to forecast needed resources. In general, it is often the case that management desires to estimate stay times in a system. We introduce a Bayesian updating methodology for the gamma distribution of stay times in a system. The methodology was initially developed for counting people in areas monitored by surveillance cameras [1]. The method was useful in the development of technology to estimate traffic through real time video information processing. The number of people in the covered area was determined and the average stay time was estimated using a gamma probability distribution model. A Bayesian updating methodology was used. We extend the application to the generic case and present a simple updating methodology for the estimation of the probability model parameters.

2 The Number of Entities in the System

Although there is often an error associated with the assessment of the number of entities in a system, one can achieve a high reliability in many situations. In the case of a school, the exact number of students can be obtained through registration records. If the entities are students in a school or people in some geographic area, the count results in $N(t)$, the number of entities in the system at time t (Figure 1). The units of time in Figure 1 are study dependent. They can

Figure 1 Nt - the Number of Entities in the System at Time t

Figure 2 $N(t)$ is the compounding of all processes

be frames per second, months or semesters. $N(t)$ is a stochastic process, the compounding of all processes representing the arrival, stay and departure of entities in the system. For simplicity, we will use the example of people being counted in a geographic area. Starting a cycle at time $t = 0$, where $N(t) = 0$, at time T_1 the first person arrives to the system and stays a time S_1 . The second person arrives at time $T_1 + T_2$, $T_2 > 0$ and stays a time S_2 etc. $T_1, T_2, T_3 \dots$ are the inter-arrival times. $S_1, S_2, S_3 \dots$ are the times spent in the system (Figure 2).

In [1], depending on the location of system, we found that the probabilistic nature of the T_i 's and S_i 's differed according to the time of day, day of the week and season. This observation required the analysis of all data and their separation into classes of homogeneity. This data stratification is necessary in most studies and leads to results for the system in different states. Systems are often complex and can show chaotic behaviour as opposed to a steady state behaviour. An analysis of the system must take into account different periods of homogeneity in the data.

3 Incoming Flow and the Arrival Process

In [1], we derived probability models for the arrival process and the time spent by each person in the system. We used homogeneous data. The probability models were applied within periods of time where the data showed the same probabilistic behaviour. Using prior knowledge and the results of statistical analyses, we classified the data into periods of homogeneity. Considering

the incoming flow of the arrival process, we applied probability models and estimated the parameters. Based on preliminary studies, we observed that the T_i 's had exponential distributions with mean θ^{-1} (Figure 3). This, in addition to other assumptions, implied that the arrival stochastic process was a Poisson process. The more general candidate for the arrival process was the non-homogeneous Poisson process (NHPP) [2] where θ varies with time and is not a constant. However, we restricted ourselves to homogeneity periods of time where the inter-arrival times are considered independent and identically distributed (IID). Many standard techniques can be used to estimate θ . Using the inter-arrival times and their distribution $T_i \sim \text{Exp}(\theta)$, one can obtain an accurate estimate of θ that gets refined with time. To conduct such an analysis, we used a Bayesian approach with a conjugate gamma prior distribution for θ . The use of a conjugate prior leads to a posterior gamma distribution. We found the approach to be effective in the estimation of the parameter θ . While a number of techniques can be used to estimate the mean of an exponential distribution, we preferred the Bayesian approach for its probabilistic estimation of the parameter θ . The posterior gamma distribution of θ offers probability intervals surrounding the posterior mode as an estimate of the inverse of the mean inter-arrival. These probability intervals can be used effectively in studies that conduct sensitivity analyses when the system shows important variability. Often such studies involve simulation for the prediction

Figure 3 Exponential Distribution Fit for the Inter-arrival Times

of complex situations that cannot be handled analytically. The input to such simulations includes the probability models of inter-arrival times. The use of a point estimate for θ without adequate probability bounds may limit the simulation results. In our approach, we use the assumption of exponential inter-arrival times with mean θ^{-1} , with a prior gamma distribution for θ .

4 Time Spent in the System

$S_1, S_2, S_3 \dots$ are the times spent in the system. At times, we cannot observe the S_i 's directly unless we track each person. In other cases, we can use existing records. In the case of video surveillance, the image processing task provides only $N(t)$, the number of people in the system at time t . Using $N(t)$ we determine exact estimates of the average time spent in the system and probability models for the S_i 's. It can be shown that for a cycle, where $N(t)$ start at $N(t) = 0$ and returns to $N(t) = 0$, that

$$\int_{t_1}^{t_2} N(t)dt = \sum_{i=1}^n S_i \quad (1)$$

where n is the number of people who arrived to the system between t_1 and t_2 . By collecting the statistic $\sum_{i=1}^n S_i$ for all cycles of the same homogeneous time period and taking the average over the total number of people, we obtain the exact sample average time spent in the system by a person and an excellent estimator of the time spent in the system. Often, this estimate is enough for simple purposes. To conduct an inference and predict, one must derive probability models. $\sum_{i=1}^n S_i$ is observed for all cycles. Consider the variable $Y = \sum_{i=1}^n S_i$. Using the collected data $(Y_1, Y_2, Y_3 \dots)$ for the same n , and assuming a period of homogeneity, that is the data are IID, we can fit a probability distribution to the Y_i 's. From such distribution, one would derive the distribution of the S_i 's. For example, if the S_i 's are gamma distributed then the sum $\sum_{i=1}^n S_i$ will be gamma distributed. As it turns out, our study has shown that the data fit a gamma distribution. However, this approach requires collecting considerable amount of data, as one needs large samples $(Y_1, Y_2, \dots)$ having all the same n number of people that have entered and stayed in the system. Instead, we prefer to use a Bayesian approach that makes use of all collected cycle data. Let $(Y_1, Y_2, \dots, Y_m)$ be the collected statistics for m cycles, where

$$Y_i = \sum_{j=1}^{n_i} S_{i,j}, i = 1, 2, \dots, m \quad (2)$$

Using $(Y_1, Y_2, \dots, Y_m)$, we want to assess the probability distribution of $S_{i,j}$, the time spent in the system by a person. Based on probability model assumptions, we assume that the $\{S_{i,j}, i, j = 1, 2, \dots\}$ are independent random variables that are identically distributed. $S_{i,j} \sim \text{Gamma}(\alpha, \lambda)$, $i, j = 1, 2, \dots$. That is

$$f_{S_{i,j}}(x) = \frac{x^{\alpha-1} \lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)} \quad (3)$$

We want to calculate

$$p(S_{i,j} | y_1, y_2, \dots, y_m) \quad (4)$$

where $(y_1, y_2, \dots, y_m)$ are the realizations of $(Y_1, Y_2, \dots, Y_m)$. Using the Chapman-Kolmogorov equation or the Law of Total Probability, we condition and average over all possible values of α and λ . To do so, we use probability models for these two parameters and make a few model simplifying assumptions. Let $(\alpha_1, \alpha_2, \dots, \alpha_K)$ be the most likely values for the shape parameter α of the Gamma (α, λ) distribution of $S_{i,j}$. Let Gamma (a, b) be the distribution of the scale parameter λ . It is a natural conjugate prior distribution. Having discretized α , let $p(\alpha)$ be the ensuing discrete prior distribution. If prior information is available on α , it can be used to construct the discrete distribution $p(\alpha)$, either directly or through an Expert Opinion procedure [3], [4]. Otherwise, a flat discrete prior can be used, that is the Uniform distribution over the set $(\alpha_1, \alpha_2, \dots, \alpha_K)$. Further assume that, to start the procedure, α and λ are independent. This assumption is not a strong one, as the two parameters α and λ do not remain independent long once the data is used. We then have $p(S_{i,j} | y_1, y_2, \dots, y_m)$ equal to

$$\sum_{\alpha} \int_{\lambda} p(S_{i,j} | \alpha, \lambda, y_1, \dots, y_m) p(\alpha, \lambda | y_1, \dots, y_m) d\lambda \quad (5)$$

Given (α, λ) , $S_{i,j}$ is conditionally independent of $(y_1, y_2, \dots, y_m)$ and is the Gamma (α, λ) distribution. It remains to evaluate $p(\alpha, \lambda | y_1, y_2, \dots, y_m)$, the posterior distribution of (α, λ) . Using Bayes theorem,

$$p(\alpha, \lambda | y_1, \dots, y_m) = \frac{1}{\delta} p(y_1, \dots, y_m | \alpha, \lambda) p(\alpha, \lambda) \quad (6)$$

$$p(\alpha, \lambda | y_1, \dots, y_m) = \frac{1}{\delta} \prod_{i=1}^m p(y_i | \alpha, \lambda) p(\alpha) p(\lambda) \quad (7)$$

where δ is the normalizing factor,

$$\delta = \sum_{\alpha} \int_{\lambda} \prod_{i=1}^m p(y_i | \alpha, \lambda) p(\alpha) p(\lambda) d\lambda \quad (8)$$

$$\delta = \sum_{\alpha} \int_{\lambda} \prod_{i=1}^m \frac{y_i^{\alpha n_i - 1} \lambda^{\alpha n_i} e^{-\lambda y_i}}{\Gamma(\alpha n_i)} p(\alpha) \frac{\lambda^{a-1} b^a e^{-b\lambda}}{\Gamma(a)} d\lambda \quad (9)$$

$$\delta = \sum_{\alpha} \int_{\lambda} \frac{\Gamma(a + \alpha \sum_{i=1}^m n_i)}{\Gamma(a) \prod_{i=1}^m \Gamma(\alpha n_i)} \frac{\prod_{i=1}^m y_i^{\alpha n_i - 1} b^a}{(b + \sum_{i=1}^m y_i)^{a + \alpha \sum_{i=1}^m n_i}} \times \quad (10)$$

$$\frac{\lambda^{(a + \alpha \sum_{i=1}^m n_i) - 1} (b + \sum_{i=1}^m y_i)^{a + \alpha \sum_{i=1}^m n_i} e^{-\lambda(b + \sum_{i=1}^m y_i)}}{\Gamma(a + \alpha \sum_{i=1}^m n_i)} p(\alpha) d\lambda \quad (11)$$

$$\delta = \sum_{\alpha} \frac{\Gamma(a + \alpha \sum_{i=1}^m n_i)}{\Gamma(a) \prod_{i=1}^m \Gamma(\alpha n_i)} \frac{\prod_{i=1}^m y_i^{\alpha n_i - 1} b^a}{(b + \sum_{i=1}^m y_i)^{a + \alpha \sum_{i=1}^m n_i}} p(\alpha) \quad (12)$$

Since Y_i is the sum of the independent gamma random variables $S_{i,j}$ as defined in Equation 4.2, then $Y_i \sim \text{Gamma}(\alpha n_i, \lambda)$. δ is computed in a non-expensive K large summation.

5 Example

The inter-arrival times $T_1, T_2, T_3 \dots$ are obtained from the process $N(t)$. We used several data sets separately for the estimation of θ , the inverse of the mean of the inter-arrival times. The prior distribution on θ is a Gamma (a, b). We varied the prior input and observed the behaviour of the posterior distribution as a function of the choice of the parameters (a, b). The posterior gamma distribution responded robustly to the prior input. In Table 1, using data set 1, we observe the prior mean and corresponding standard deviation followed by the posterior mode and posterior mean and the actual sample mean or empirical mean of θ , for different prior distributions. This is standard Bayesian conjugate analysis of the parameter of an exponential distribution, namely θ , which is well known and well documented. It is an efficient procedure that provides probability bounds rather than just a point estimate. Such probability intervals can be used effectively as input to simulation studies for example where sensitivity analysis is a must. In our case, we provide the methodology for the arrival process in addition to the

Table 1 Bayesian Estimation of The prior distribution on θ Using Data Set 1

a	b	Prior Mean	Prior Standard Deviation	Posterior Mode	Posterior Mean	Empirical Mean
10	900	.0111	.003	.0062	.0064	.0057
5	900	.0055	.0024	.0054	.0057	.0057
3	900	.0033	.0019	.0051	.0054	.0057
2	900	.002	.0016	.0055	.0052	.0057
1	900	.0011	.0011	.0049	.0051	.0057
5	90	.05	.024	.0062	.0064	.0057
3	90	.0333	.019	.0059	.0061	.0057
2	90	.022	.0157	.0057	.0059	.0057
1	90	.0111	.0111	.00557	.0058	.0057

Table 2 Bayesian Estimation of The prior distribution on θ Using Data Set 2

a	b	Prior Mean	Prior Standard Deviation	Posterior Mode	Posterior Mean	Empirical Mean
10	900	.0111	.003	.0053	.0056	.0042
5	900	.0055	.0024	.0042	.0045	.0042
3	900	.0033	.0019	.0037	.0040	.0042
2	900	.002	.0016	.0035	.0038	.0042
1	900	.0011	.0011	.0033	.0036	.0042
10	1200	.008	.0026	.0050	.0052	.0042
5	1200	.00416	.0018	.0039	.0042	.0042
3	1200	.0025	.0014	.0035	.0037	.0042
1	1200	.0008	.0008	.0031	.0033	.0042

analysis of stay times in the system. Table 1 uses one example of a set of homogeneous data taken from an actual situation where people arriving to the system were counted using a video surveillance system [1]. Table 2 is a different set of data with homogeneous probabilistic behaviour. The data were studied separately then combined to provide a measure for the inter-arrival process.

For the stay times, we report an example for the purpose of illustration. The data is shown in Table 3. The number of people is counted for each cycle along with the total time of stay. In this example, we looked at 33 cycles.

Using only the data where the number of people in a cycle is 1, we conducted the estimation for the actual stay times. Figure 4 shows the statistical analysis of that data. Using a typical classical fit program, we found the mean to be 71.84 and the variance 989.16 for a Gamma distribution. The Gamma distribution had estimated shape parameter of 5.21 (1.64 standard error) and scale parameter 13.76 (4.54 standard error). In another analysis, an

Table 3 Total Time of Stay per Cycle

Cycle No.	Number of People	Total Time of Stay	Cycle No.	Number of People	Total Time of Stay
1	5	456	18	1	74
2	3	652	19	4	534
3	1	86	20	1	82
4	6	928	21	1	146
5	2	188	22	1	132
6	1	40	23	11	1096
7	2	212	24	8	476
8	6	2264	25	1	48
9	1	40	26	1	40
10	5	1035	27	5	394
11	1	63	28	11	1468
12	1	62	29	1	100
13	1	52	30	1	40
14	1	130	31	1	40
15	1	44	32	5	552
16	1	92	33	1	54
17	2	484

Figure 4 Classical Statistical Analysis of 1 person per Cycle Data

approximation to the actual stay times is obtained by dividing the total stay time in a cycle by the number of people in that cycle Y_i/n_i . Figure 5 shows the statistical analysis of that data. The distribution was found to be a Gamma with

Figure 5 Classical Statistical Analysis of Averaged Stay Times Y_i/n_i .

Figure 6 Uniform Prior Distribution and Posterior Distribution of α

estimated shape parameter of 2.98 (0.69 standard error) and scale parameter 35.3 (8.9 standard error). The distribution had a mean of 105.14 and variance 3712.

Using the full Bayesian analysis described in this report, Figure 6 shows the posterior distribution of α using all the data. It is customary in such a Bayesian methodology to use a uniform distribution if no prior knowledge exists about the parameter of interest, namely α in this case. This means that

a-priori, all values of α are discrete and considered equally likely before the observation of the data. In our example, once the data has been accumulated and included in the Bayesian analysis, the first two values of α emerged as the most likely values, as shown in Figure 6

Figure 7 shows the prior distribution of λ and the two first posterior distributions obtained using 1 data point and two data points (more peaked distribution). As more and more data is introduced into the analysis, the posterior distribution becomes more concentrated around the posterior estimate

Figure 7 Prior and Posterior Distributions of λ with 1 and 2 Points

Figure 8 Posterior Distributions of λ with 8 data Points

Figure 9 Posterior Distributions of λ at All Stages **Figure 10** Distribution of Stay Times

of λ . This is typical of the behaviour of a posterior distribution. The Bayesian probabilistic updating algorithm provides a mean to refine an estimate within a probability interval as more and more data is gathered. At times, the data will also point to the possibility of bi-modality. This is often due to the mixing of two populations in the data, as was observed in our example when we introduced 8 data points. Figure 8 illustrates the posterior distribution with the first 8 data points, that is $p(\lambda|y_1), \dots, P(\lambda|y_1, \dots, y_8)$.

Finally, Figure 9 shows the posterior distributions at all the stages.

The interesting observation is the clear bimodality of the posterior distribution. This reflects the two types of people who enter the system, those who come in for a short period of time and those who stay longer. This corroborated our prior knowledge of the location we observed. Figure 10 shows the prior and posterior distribution of the stay time, after averaging over the parameters α and λ . The dashed plot is the prior distribution. The straight line is the distribution of $S_{i,j}$ averaged over α and λ . The x marked plot is a numerical approximation to the posterior distribution.

6 Conclusion

We present a Bayesian updating methodology for the parameters of the gamma distribution. We apply it to the study of the data of stay times in a system. Statistics for the stay times are deducted from the stochastic process of the number of incoming people. A gamma probability model is assumed for the stay times with unknown parameters. The Bayesian methodology is applied to update knowledge on these parameters probabilistically using the existing statistical information. Based on collected information on the number in the system, stay times are estimated in a probability model. The method is useful in the development of technology to estimate traffic in a system.

References

- [1] S. Challa, K. Aboura, K. Ravikanth, S. Deshpande, 'Estimating the number of people in buildings using visual information', *Information, Decision and Control*, pp. 124–129 (2007).
- [2] D. L. Snyder, M. I. Miller, 'Random Point Processes in Time and Space', Springer-Verlag, NJ (1991).
- [3] D. V. Lindley, 'Reconciliation of probability distributions', *Operations Research*, pp. 866–880 (1983).
- [4] K. Aboura, J. I. Agbinya, 'Adaptive maintenance optimization using initial reliability estimates', *Journal of Green Engineering*, pp. 121 (2013).

Biographies

Khalid Aboura teaches quantitative methods at the College of Business Administration, University of Dammam, Saudi Arabia. Khalid Aboura spent several years involved in academic research at the George Washington University, Washington D.C, U.S.A, where he completed the Master of Science and the Doctor of Science degrees in Operations Research. Dr Aboura has extensive experience in Stochastic Modelling, Operations Research, Simulation, Maintenance Optimization and Mathematical Optimization. He worked as a Research Scientist at the Commonwealth Scientific and Industrial Research Organization of Australia and conducted research at the School of Civil and Environmental Engineering and the School of Computing and Communication, University of Technology Sydney, Australia. Khalid Aboura was a Scientist at the Kuang-Chi Institute of Advanced Technology of Shenzhen, China.

Johnson I. Agbinya is an Associate Professor in the department of electronic engineering at La Trobe University, Melbourne, Australia. He is Honorary Professor at the University of Witwatersrand, South Africa, Extraordinary Professor at the University of the Western Cape, Cape Town and the Tshwane University of Technology, Pretoria, South Africa. Prior to joining La Trobe, he was Senior Research Scientist at CSIRO Telecommunications and Industrial Physics, Principal Research Engineering at Vodafone Australia and Senior Lecturer at University of Technology Sydney, Australia. His research activities

cover remote sensing, Internet of things, bio-monitoring systems, wireless power transfer, mobile communications and biometrics systems. He has authored several technical books in telecommunications. He published more than 250 peer-reviewed research publications in international Journals and conference papers. He served as expert on several international grants reviews and was a rated researcher by the South African National Research Fund.