JMT (Jurnal Matematika dan Terapan) **4** (2) (2022), 20-26
https://doi.org/10.21009/jmt.4.2.3
e-ISSN: 2615-6792

Metode Bayesian untuk Estimasi Parameter Distribusi Eksponensial pada Data Tersensor

Reza Anjab Ramadhan[1, a)], Widyanti Rahayu[1, b)], Ibnu Hadi[1, c)]

[1] *Program Studi Matematika,Fakultas Matematika dan Ilmu Pengetahuan Alam,Universitas Negeri Jakarta*

Email: [a)]rezaanjabramadhan1305617017@mhs.unj.ac.id, [b)]wrahayu@unj.ac.id, [c)]ibnu_hadi@unj.ac.id

Abstract

Parameter is a value that describe the characteristics of a population. But the parameter's value of a real data is unknown. To estimate the value of the parameter, there are several methods, which are maximum likelihood estimation method (MLE) and Bayesian parameter estimation method. In Bayesian method, the prior information is applied to update the current data. The prior is determined based on the information in the data. This mini thesis is using censored data with exponential distribution, and using the conjugate prior. Followed by squared error loss function (SELF), the estimated value function of the λ parameter, that is $\hat{\lambda}$. When the function was applied on Stanford heart transplant data, the value of $\hat{\lambda} = 0.00089$, which means the patient's failure (death) probability is low and the patient's probability to survive is high.

Keywords: Censored Data, Bayesian Parameter Estimation, Conjugate Prior, SELF, Survival Analysis

Abstrak

Parameter merupakan nilai yang menjelaskan karakteristik dari suatu populasi. Namun parameter pada data yang sebenarnya, tidak diketahui nilainya. Untuk menduga nilai parameter dari data tersebut, terdapat beberapa metode estimasi parameter, dua diantaranya adalah metode estimasi kemungkinan maksimum (MLE) dan metode estimasi parameter Bayesian. Pada metode estimasi Bayesian, digunakan informasi awal (*prior*) untuk memperbarui data saat ini. *Prior* ditentukan berdasarkan informasi pada data. Pada skripsi ini digunakan data tersensor yang berdistribusi eksponensial, dengan *prior* yang digunakan adalah *prior* konjugat. Selanjutnya dengan fungsi kerugian error kuadrat (SELF), didapatkan fungsi nilai parameter estimasi dari parameter λ, yakni $\hat{\lambda}$. Ketika diaplikasikan pada data transplantasi jantung Stanford, didapatkan nilai $\hat{\lambda} = 0.00089$, yang berarti kemungkinan pasien mengalami kegagalan (mati) adalah rendah dan kemungkinan pasien bertahan adalah tinggi.

Kata-kata kunci: Data Tersensor, Estimasi Parameter Bayesian, Prior Konjugat, SELF, Analisis Survival

PENDAHULUAN

Analisis survival merupakan kumpulan dari prosedur statistika dalam menganalisis data, dimana hasil yang diteliti adalah waktu hingga *event* terjadi, dengan dua komponen utama yaitu fungsi survival dan fungsi *hazard* [1]. Dalam analisis survival, terdapat istilah penyensoran data, yang dilakukan karena waktu bertahan objek tidak diketahui pasti oleh peneliti.

Data tersensor, data yang sudah dilakukan penyensoran, bisa saja mengikuti suatu distribusi yang sudah diketahui, dengan nilai parameter yang tidak diketahui. Parameter dari data tersebut, dapat digunakan untuk mendapatkan fungsi survival dan fungsi *hazard*. Untuk menduga nilai dari parameter tersebut, dapat digunakan metode pendekatan seperti MLE atau metode estimasi Bayesian. Pada

Diterima: 24 Mei 2022, Direvisi: 23 Agustus 2022, Disetujui: 23 Agustus 2022

metode MLE, penduga yang dihasilkan cenderung merupakan penduga tak bias, sedangkan pada metode estimasi Bayesian pendugaan parameternya sulit dilakukan secara analitik, akan tetapi metode estimasi Bayesian, membebaskan peneliti untuk menentukan *prior* yang akan digunakan. Selain itu, karena metode estimasi Bayesian hasilnya masih berupa distribusi atau peubah acak, maka diperlukan suatu fungsi kerugian untuk mengestimasi nilai parameter yang sedang diestimasi, dan peneliti diberi kebebasan untuk menentukan fungsi kerugian yang akan digunakan.

Ningrum *et al* (2020) [2], telah melakukan penelitian untuk mengestimasi parameter dari distribusi Rayleigh dengan menggunakan metode estimasi Bayesian *squared error loss function* (SELF), dengan menggunakan *prior Vague*. Dengan studi kasus penderita kanker ovarium, yang diambil dari program R, mereka menyimpulkan bahwa jika hasil estimasi yang didapatkan, dibandingkan dengan hasil estimasi parameter dengan metode dan data yang sama, didapatkan bahwa dengan menggunakan *prior uniform* akan dihasilkan estimasi parameter yang lebih baik daripada menggunakan *prior Vague*. Hal ini didapatkan dengan membandingkan grafik fungsi survival dan fungsi *hazard* milik mereka dengan grafik dari tulisan sebelumnya dengan data yang sama.

Berdasarkan hal tersebut, penulis ingin membahas mengenai metode Bayesian SELF untuk mengestimasi parameter distribusi eksponensial pada data tersensor. Distribusi eksponensial, dipilih karena merupakan distribusi yang berkaitan dengan waktu. Selain itu, distribusi ini juga merupakan bagian dari keluarga eksponensial, yang salah satu keuntungannya adalah distribusi konjugat juga terdapat pada keluarga eksponensial [3]. Adapun data pada penelitian ini adalah data sekunder yang bersumber dari *software* R, yaitu data pasien transplantasi jantung Stanford [4], yang dapat diakses dari library "survival". Tujuan utama dari penelitian ini adalah mendapatkan fungsi nilai parameter estimasi, lalu diaplikasikan pada analisis survival untuk mengetahui performa dari parameter estimasi tersebut.

METODE

Metode penelitian yang digunakan dalam penelitian ini adalah metode penelitian kuantitatif, yaitu data yang digunakan dapat dicatat dalam bentuk angka. Metode penelitian ini dilakukan dengan cara membentuk dan menganalisis data sesuai dengan tujuan akhir peneliti.

Teknik pengumpulan data yang digunakan dalam penelitian ini, yaitu:

1. Menggunakan data sekunder berupa data survival pasien transplantasi jantung Stanford yang dimulai dari tanggal 1 November 1967 hingga 1 April 1974. Data tersebut diperoleh dari jurnal milik J. Crowley dan M. Hu (1977) yang berjudul Covariance Analysis of Heart Transplant Survival Data.
2. Studi literatur, yaitu dengan mengumpulkan materi-materi yang berkaitan dengan permasalahan yang dibahas dalam penelitian ini.

HASIL DAN PEMBAHASAN

Distribusi Eksponensial

Pada penelitian ini, parameter yang akan diestimasi adalah parameter *rate* dari distribusi eksponensial yang dinotasikan dengan λ. Karena data yang digunakan adalah data tersensor, maka untuk membentuk fungsi *likelihood*, dibutuhkan fungsi survival dari distribusi eksponensial, yaitu:

$$S(t) = 1 - F(t) = e^{-\lambda t} \tag{1}$$

dan fungsi *hazard* dari distribusi eksponensial adalah:

$$h(t) = \frac{f(t)}{S(t)} = \lambda \tag{2}$$

Setelah didapatkan fungsi survival, fungsi *likelihood* dapat dibentuk, yaitu:

$$\begin{aligned} f(t_i|\lambda, \delta_i) &= \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \\ &= \lambda^{\Sigma_{i=1}^{n} \delta_i} exp(-\lambda \Sigma_{i=1}^{n} t_i) \end{aligned} \tag{3}$$

Estimasi Parameter

1. *Prior* Chi-Square

Karena distribusi Chi-Square dan distribusi eksponensial, merupakan distribusi spesial dari distribusi Gamma, maka peneliti berasumsi bahwa distribusi Chi-Square merupakan distribusi konjugat dari distribusi eksponensial. Untuk membuktikan hal tersebut, perlu didapatkan distribusi *posterior* terlebih dahulu, diawali dengan menentukan fungsi peluang gabungan (*joint probability*) antara fungsi *likelihood* pada persamaan 17 dan distribusi *prior*, yaitu:

$$f(\lambda)_{Chi-sq} f(t|\lambda) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \lambda^{\frac{k}{2}+\Sigma_{i=1}^{n}\delta_i-1} exp\left(-\lambda\left(\frac{1}{2}+\Sigma_{i=1}^{n}t_i\right)\right) \tag{4}$$

Dilanjutkan dengan menentukan fungsi peluang marginalnya, yaitu sebagai berikut:

$$\int_{-\infty}^{\infty} f(\lambda) f(t_i|\lambda) d\lambda = \left(2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)\right)^{-1} \Gamma\left(\frac{k}{2}+\Sigma_{i=1}^{n}\delta_i\right)\left(\frac{1}{\frac{1}{2}+\Sigma_{i=1}^{n}t_i}\right)^{\left(\frac{k}{2}+\Sigma_{i=1}^{n}\delta_i\right)} \tag{5}$$

posterior, yang dinotasikan dengan $f(\lambda|t)$ atau $p(\lambda|t)$, bisa didapatkan dengan menyubstitusikan fungsi peluang gabungan dan fungsi peluang marginal pada persamaan 13, yaitu:

$$\begin{aligned} f(\lambda|t) &= \frac{f(t|\lambda)f(\lambda)_{Chi-sq}}{f(t)} \\ &= \frac{\left(\frac{1}{2}+\Sigma_{i=1}^{n}t_i\right)^{\left(\frac{k}{2}+\Sigma_{i=1}^{n}\delta_i\right)}}{\left(\Gamma\left(\frac{k}{2}+\Sigma_{i=1}^{n}\delta_i\right)\right)} \lambda^{\left(\frac{k}{2}+\Sigma_{i=1}^{n}\delta_i-1\right)} exp\left(-\lambda\left(\frac{1}{2}+\Sigma_{i=1}^{n}t_i\right)\right) \end{aligned} \tag{6}$$

Berdasarkan persamaan di atas, dapat dilihat bahwa posterior dapat dinyatakan berdistribusi Gamma dengan parameter *shape*, $\alpha = \frac{k}{2} + \Sigma_{i=1}^{n}\delta_i$ dan parameter *rate*, $\beta = \frac{1}{2} + \Sigma_{i=1}^{n}t_i$. Artinya distribusi *posterior* tidak sama dengan distribusi *prior*, artinya distribusi Chi-Square bukan merupakan distribusi konjugat dari distribusi eksponensial sehingga perlu ditentukan distribusi *prior* konjugat yang baru.

2. *Prior* Gamma

Prior yang berdistribusi Gamma dipilih karena, ketika menggunakan *prior* yang berdistribusi Chi-Square, *posterior* yang dihasilkan berdistribusi Gamma. Karena distribusi Chi-Square merupakan distribusi spesial dari distribusi Gamma, seharusnya *posterior* yang dihasilkan dengan *prior* Gamma, akan berdistribusi Gamma pula. Oleh karena itu, akan ditentukan distribusi gabungan antara fungsi *likelihood* pada persamaan 17 dan distribusi *prior*, yaitu sebagai berikut:

$$f(\lambda)_{Gam} f(t|\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha+\Sigma_{i=1}^{n}\delta_i-1} e^{-\lambda\left(\beta+\Sigma_{i=1}^{n}t_i\right)} \tag{7}$$

dan fungsi peluang marginalnya adalah:

$$\int_{-\infty}^{\infty} f(\lambda)_{Gam} f(t|\lambda) d\lambda = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\left(\frac{1}{\beta+\Sigma_{i=1}^{n}t_i}\right)^{\alpha+\Sigma_{i=1}^{n}\delta_i} \Gamma(\alpha+\Sigma_{i=1}^{n}\delta_i) \tag{8}$$

sehingga, *posterior*-nya adalah sebagai berikut:

$$\begin{aligned} f(\lambda|t) &= \frac{f(t|\lambda)f(\lambda)_{Gam}}{f(t)} \\ &= \frac{\left(\beta+\Sigma_{i=1}^{n}t_i\right)^{\alpha+\Sigma_{i=1}^{n}\delta_i}}{\Gamma(\alpha+\Sigma_{i=1}^{n}\delta_i)} \lambda^{(\alpha+\Sigma_{i=1}^{n}\delta_i)-1} e^{-\lambda\left(\beta+\Sigma_{i=1}^{n}t_i\right)} \end{aligned} \tag{9}$$

Berdasarkan persamaan di atas, *posterior* yang dihasilkan berdistribusi Gamma, dengan parameter *shape*, $\alpha + \Sigma_{i=1}^{n}\delta_i$, dan parameter *rate*, $\beta + \Sigma_{i=1}^{n}t_i$. Artinya distribusi Gamma merupakan distribusi konjugat dari distribusi eksponensial.

3. *Squared Error Loss Function* (SELF)

Setelah didapatkan distribusi *posterior*, akan digunakan SELF untuk menentukan nilai parameter estimasi. Hal ini disebabkan karena hasil dari *posterior* masih berupa variabel acak. adapun fungsi kerugian galat kuadrat ketika diaplikasikan pada metode estimasi Bayesian, sama dengan fungsi ekspektasi dari *posterior*, yaitu sebagai berikut:

$$\begin{aligned} \hat{\lambda} &= E(\lambda) \\ &= \frac{\alpha}{\beta} \\ &= \frac{\alpha+\Sigma_{i=1}^{n}\delta_i}{\beta+\Sigma_{i=1}^{n}t_i} \end{aligned} \tag{10}$$

Karena fungsi nilai parameter estimasi telah didapatkan, maka estimasi fungsi survival dan fungsi *hazard* dari data tersensor yang berdistribusi eksponensial adalah:

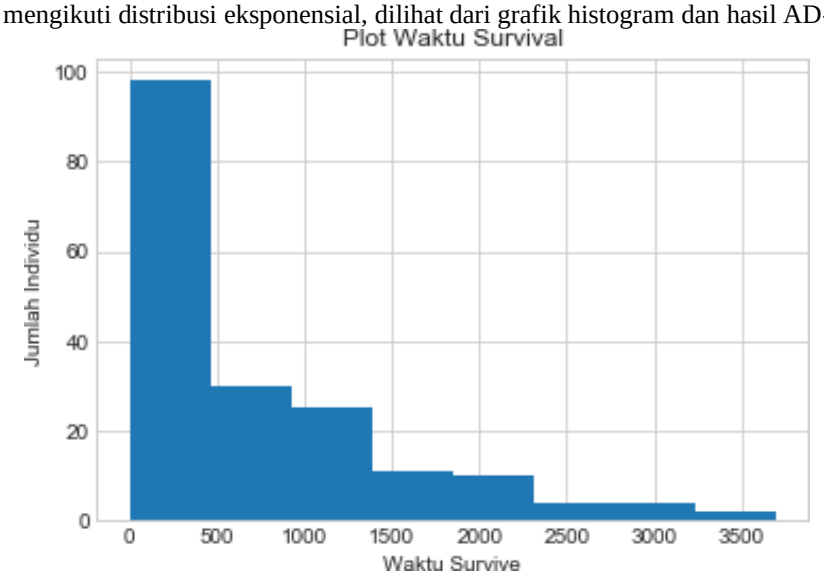
$$\begin{aligned}\hat{S}(t) &= exp(-\hat{\lambda}t) \\ &= exp\left(-\frac{\alpha+\Sigma_{i=1}^{n}\delta_i}{\beta+\Sigma_{i=1}^{n}t_i}t\right)\end{aligned} \tag{11}$$

$$\begin{aligned}\hat{h}(t) &= \hat{\lambda} \\ &= \frac{\alpha+\Sigma_{i=1}^{n}\delta_i}{\beta+\Sigma_{i=1}^{n}t_i}\end{aligned} \tag{12}$$

Simulasi pada Data

1. Deskripsi Data

Pada penelitian ini, digunakan data sekunder yang bersumber dari aplikasi R, *library* "survival", yaitu data mengenai pasien transplantasi jantung Stanford. Pada data tersebut, variabel yang akan digunakan adalah variabel status dan waktu bertahan. Variabel yang akan diestimasi adalah variabel waktu bertahan, dengan asumsi bahwa variabel tersebut mengikuti distribusi eksponensial, dilihat dari grafik histogram dan hasil AD-test.



GAMBAR 1. Plot Histogram Variabel *Time*

Grafik pada gambar 1 merupakan grafik histogram dari variabel waktu bertahan (*time*). Dari grafik tersebut dapat dilihat bahwa variabel waktu bertahan dapat dinyatakan berdistribusi eksponensial. Selain itu dapat ditentukan pula bahwa dari 184 pasien, terdapat hampir 100 pasien memiliki waktu bertahan mendekati 500 hari, dan waktu bertahan paling lama mencapai lebih dari 3500.

Ketika dilakukan test AD (Anderson-Darling), dengan taraf signifikansi sebesar 5% dan hipotesis nol, data berdistribusi normal, dan hipotesis satu, data tidak berdistribusi normal didapatkan *p-value* $< 2.2 \times 10^{-16}$. Artinya hipotesis nol ditolak, sehingga data tidak mengikuti distribusi normal, dan data dinyatakan berdistribusi eksponensial.

2. Simulasi

Berdasarkan data yang digunakan, didapatkan informasi bahwa $\Sigma_{i=1}^{184}\delta_i = 113$ dan $\Sigma_{i=1}^{184}t_i = 128237.5$, sedangkan untuk menentukan nilai α dan β, didapatkan dari mencocokkan fungsi ekspektasi dan varian antara distribusi eksponensial dan distribusi Gamma, hasilnya yaitu:

$$\alpha = 1$$
$$\beta = \bar{t} \approx 696.943$$

didapatkan nilai parameter estimasi dari data tersebut adalah:

$$\begin{aligned}\hat{\lambda} &= \frac{\alpha+\Sigma_{i=1}^{n}\delta_i}{\beta+\Sigma_{i=1}^{n}t_i} \\ &\approx 0.00089\end{aligned}$$

Berdasarkan nilai parameter estimasi yang didapat, peluang terjadinya *event* adalah rendah, yaitu sebesar 0.089%. Selain itu, hal tersebut dapat dilihat pula melalui fungsi survival, yaitu:

$$\hat{S}(t) = exp(-0.00089t)$$

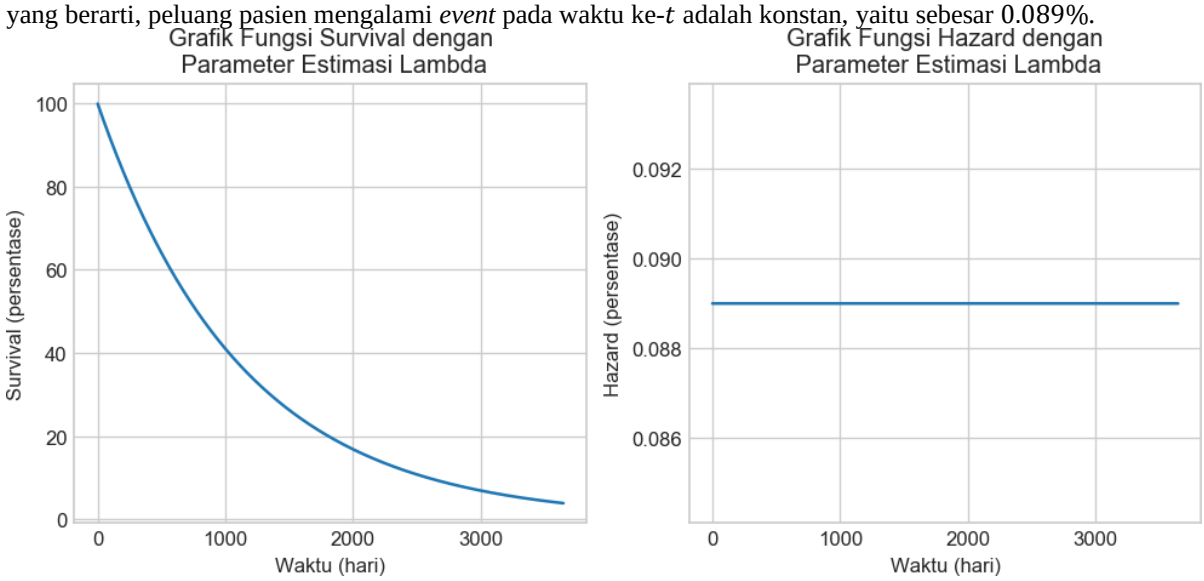
Sebagai contoh, misal ingin diketahui peluang pasien dapat bertahan hingga hari ke-1000 adalah:

$$\hat{S}(1000) = exp(-0.00089 \times 1000) = 0.41066$$

Artinya, pasien yang telah melakukan operasi transplantasi jantung, memiliki peluang untuk tetap bertahan hingga hari ke-1000 adalah 41.066%. Adapun peluang pasien mengalami kegagalan setelah bertahan selama t hari adalah:

$$\hat{h}(t) = 0.00089$$

yang berarti, peluang pasien mengalami *event* pada waktu ke-t adalah konstan, yaitu sebesar 0.089%.



GAMBAR 2. Grafik Fungsi Survival dan Fungi *Hazard*

Gambar 2 merupakan grafik dari fungsi survival dan fungsi *hazard* ketika diaplikasikan nilai t mulai dari 1 sampai dengan 3650 (mendekati 10 tahun). Berdasarkan kedua grafik tersebut, dapat dilihat bahwa peluang pasien dapat bertahan relatif tinggi, hal ini ditandakan dengan peluang pasien dapat bertahan hingga hari ke-2000, masih di atas 15%, dan peluang pasien dapat bertahan hingga hari ke-3000 setelah operasi, adalah sekitar 10%. Selain itu peluang pasien mengalami *event* di setiap harinya adalah sama, yaitu sebesar 0.089%.

PENUTUP

Kesimpulan

Berdasarkan hasil yang didapat, distribusi Chi-Square bukan merupakan distribusi konjugat dari distribusi eksponensial, meskipun keduanya merupakan bentuk spesial dari distribusi Gamma, dan distribusi Gamma merupakan distribusi konjugat dari distribusi eksponensial. Dengan menggunakan metode estimasi Bayesian didapatkan *posterior*, yaitu:

$$f(\lambda|t) = \frac{(\beta + \Sigma_{i=1}^{n} t_i)^{\alpha+\Sigma_{i=1}^{n}\delta_i}}{\Gamma(\alpha + \Sigma_{i=1}^{n}\delta_i)} \lambda^{(\alpha+\Sigma_{i=1}^{n}\delta_i)-1} e^{-\lambda(\beta+\Sigma_{i=1}^{n}t_i)}$$

Ketika digunakan SELF, didapatkan fungsi nilai parameter estimasi λ ($\hat{\lambda}$) yaitu $\frac{\alpha+\Sigma_{i=1}^{n}\delta_i}{\beta+\Sigma_{i=1}^{n}t_i}$ dan ketika diaplikasikan ke analisis survival pada data tersensor yang sebenarnya, didapatkan fungsi survival ($S(t)$) dan fungsi *hazard* ($h(t)$) sebagai berikut:

$$\hat{S}(t) = exp(-\hat{\lambda}) = exp\left(-\frac{\alpha + \Sigma_{i=1}^{n}\delta_i}{\beta + \Sigma_{i=1}^{n}t_i}t\right)$$

$$\hat{h}(t) = \hat{\lambda} = \frac{\alpha + \Sigma_{i=1}^{n}\delta}{\beta + \Sigma_{i=1}^{n}t_i}$$

serta, didapatkan nilai parameter estimasi, $\hat{\lambda} = 0.00089$, yang berarti peluang bertahan objek penelitian relatif tinggi dan peluang gagal yang rendah.

Saran

Penelitian ini masih dapat dikembangkan dengan mengaitkan variabel-variabel lain untuk menduga peluang bertahan suatu individu. Selain itu, penelitian selanjutnya dapat menggunakan *prior* yang lain, atau menggunakan data tersensor yang berdistribusi lain seperti distribusi Weibull, Gamma, log-normal, dll. Menggunakan fungsi kerugian yang berbeda untuk menduga parameter dengan *prior* yang sama juga dapat dijadikan pilihan untuk penelitian selanjutnya.

REFERENSI

Alzaatreh, A., Carl, L. E. E., Famoye, F., 2016, Family of generalized gamma distributions: Properties and applications, Hacettepe Journal of Mathematics and Statistics, 45(3): 869-886

Bishop, C. M., 2006, Pattern Recognition and Machine Learning, Edisi ke-1, Springer, New York

Carlin, B. P., Louis, T. A., 2008, Bayesian methods for data analysis, Edisi ke-3, CRC Press, New York

Dey, S., Dey, T., Kundu, D., 2014, Two-parameter Rayleigh distribution: different methods of estimation, American Journal of Mathematical and Management Sciences, 33(1): 55-74

Diaconis, P., Ylvisaker, D, 1979, Conjugate priors for exponential families, The Annals of statistics, 7(2): 269-281

Dobson, A. J., Barnett, A. G., 2018, An introduction to generalized linear models, Edisi ke-3, Chapman and Hall/CRC, New York

Escobar, L. A., Meeker Jr, W. Q., 1992, Assessing influence in regression analysis with censored data, Biometrics, 48(2): 507-528

Gompertz, B, 1825, XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies, In a letter to Francis Baily, Esq. FRS &c, Philosophical transactions of the Royal Society of London, 115: 513-583

Hald, A, 1998, A History of Mathematical Statistics from 1750 to 1930, Volume ke-2, Wiley, New York

Ieren, T. G., Oguntunde, P. E., 2018, A comparison between maximum likelihood and Bayesian estimation methods for a shape parameter of the weibull-exponential distribution, Asian Journal of Probability and Statistics, 1(1): 1-12

Islam, A. F. M. (2011), Loss functions, utility functions and Bayesian sample size determination, Doctoral dissertation di University of London, tidak diterbitkan

Kleinbaum, D. G., Klein, M., 2010, Survival analysis, Edisi ke-3, Springer, New York

Lynch, S. M., 2005, Encyclopedia of Social Measurement: Bayesian Statistics, by Kimberly Kempf-Leonard, Edisi ke-1, Elsevier, New York

Mohamed, S., Ghahramani, Z., Heller, K. A., 2008, Bayesian exponential family PCA, Advances in neural information processing systems, 21: 1089-1096

Myung, I. J., 2003, Tutorial on maximum likelihood estimation, Journal of mathematical Psychology, 47(1): 90-100

Ningrum, A. F., Satyahadewi, N., Rizki, S. W., Metode Bayesian SELF untuk Estimasi Parameter Model Survival Distribusi Rayleigh, BIMASTER, 9(1)

Ozguven, E. E., Ozbay, K., 2008, Nonparametric Bayesian estimation of freeway capacity distribution from censored observations, Transportation Research Record, 2061(1): 20-29

Patti, S., Biganzoli, E., Boracchi, P., 2007, Review of the Maximum Likelihood Functions for Right Censored Data. A New Elementary Derivation, COBRA Preprint Series, 21

Triana, Y., Purwadi, J., 2019, Exponential Distribution Parameter Estimation with Bayesian SELF Method in Survival Analysis, Journal of Physics: Conference Series, 1373(1): 012050

Walpole, R. E., 1993, Pengantar Statistika Edisi 3 (Introduction to Statistics 3rd Edition), Edisi ke-3, PT. Gramedia Pustaka Utama, Jakarta