BuletinIlmiah Mat. Stat. danTerapannya (Bimaster)
Volume 09, No. 2 (2020), hal 293-300.

ESTIMASI PARAMETER MODEL *SURVIVAL* DISTRIBUSI PARETO DENGAN METODE BAYESIAN GELF UNTUK DATA TERSENSOR

Nila Handayani, Setyo Wira Rizki

INTISARI

Data survival adalah data yang menunjukan waktu suatu individu atau objek dapat bertahan hidup hingga terjadinya suatu kegagalan atau kejadian tertentu. Data survival dikatakan tersensor apabila objek pada penelitian hilang atau sampai akhir penelitian objek tersebut belum mengalami kejadian tertentu. Tujuan dari penelitian ini adalah menentukan estimasi parameter model survival distribusi Pareto data tersensor dengan metode Bayesian GELF. Estimasi parameter fungsi survival dengan metode Bayesian GELF dari distribusi Pareto diperoleh $\hat{\theta}_{BG} = 1.224745$. *Nilai estimasi parameter digunakan untuk menghitung nilai estimasi peluang seorang individu untuk bertahan hidup* $\hat{s}(t)_{BG}$ *pada penderita kanker paru-paru. Nilai MAPE yang diperoleh sebesar 22.44%, hal ini menunjukkan bahwa metode Bayesian GELF memiliki kemampuan peramalan yang cukup baik sebagai suatu metode estimasi dalam kasus survival.*

Kata kunci**:** *Pareto,Prior Gamma, Likelihood*

PENDAHULUAN

Analisis *survival* adalah salah satu teknik statistika yang digunakan untuk menganalisis data yang bertujuan untuk mengetahui hasil dari variabel yang mempengaruhi suatu awal kejadian sampai akhir kejadian[1]. Pada analisis *survival* terdapat konsep penyensoran yaitu pengamatan tersensor dan pengamatan tidak tersensor. Pengamatan dikatakan tersensor apabila data tidak dapat diamati secara lengkap karena subjek penelitian hilang atau mengundurkan diri atau sampai akhir penelitian subjek tersebut belum mengalami kejadian tertentu. Sedangkan pengamatan dikatakan tidak tersensor apabila semua subjek penelitian atau unit data yang diteliti mati atau gagal atau mengalami suatu kejadian.

Terdapat dua model yang digunakan untuk menganalisis data survival yaitu model parametrik dan model non parametrik. Jika distribusi yang mendasari data *survival* tidak diketahui, artinya data tidak mengikuti suatu distribusi tertentu yang sudah ada maka digunakan model non parametrik. Pada model non parametrik terdapat dua metode yang terkenal yaitu metode Kaplan-Meier dan metode Nelson-Aalen. Model parametrik adalah suatu model *survival* dengan data *survival* yang mengikuti asumsi distribusi tertentu. Beberapa model parametrik terdiri dari distribusi eksponensial, distribusi Pareto, distribusi Weibull, distribusi log-normal, distribusi log-logistik, dan distribusi Gamma[2].

Pada analisis *survival*, terdapat salah satu distribusi yang cukup sering digunakan dalam analisis data uji hidup adalah distribusi Pareto. Keunikan distribusi Pareto adalah memiliki karakteristik ekor tebal (*heavy-tailed*). Terdapat beberapa metode estimasi parameter pada model parametrik, salah satunya adalah metode Bayesian.Sebelum menarik sampel dari suatu populasi biasanya diperoleh informasi mengenai parameter yang mengikuti suatu distribusi tertentu yang disebut dengan informasi prior. Distribusi prior adalah distribusi awal yang memberi informasi tentang suatu parameter. Adapun informasi prior yang digunakan dalam pengamatan ini berdistribusi Gamma. Metode Bayesian merupakan metode estimasi yang mengabungkan distribusi prior dan fungsi *likelihood*[3]. Fungsi *likelihood* yang digabung dengan distribusi prior akan menghasilkan suatu distribusi baru yaitu distribusi posterior yang menjadi dasar untuk inferensi didalam metode Bayesian. Ada beberapa

pendekatan dalam metode Bayesian yang dapat digunakan untuk mengestimasi parameter antara lain *general non-informatif prior, Lindley aproximation, general entroppy loss function (GELF), dan squared error loss function (SELF).*

Penelitian ini bertujuan untuk menentukan estimasi parameter model *survival* berdistribusi Pareto dengan prior gamma menggunakan metode Bayesian GELF pada kasus penderita kanker paru-paru. Data yang digunakan dalam penelitian ini adalah data kanker paru-paru yang diambil dari program R versi 3.3.0 dengan jumlah data sebanyak 71 orang. Distribusi yang digunakan untuk model *survival* data tersensor adalah distribusi Pareto. Langkah pertama yang dilakukan pada data kanker paru-paru dimulai dengan menguji kecocokan model untuk mengetahui apakah data yang digunakan berdistribusi Pareto atau tidak dengan uji *Kolmogorov-Smirnov*, dari distribusi Pareto ditentukan fungsi kepadatan peluang, fungsi distribusi kumulatif, fungsi *survival* dan fungsi *hazard*. Selanjutnya menentukan fungsi *likelihood* dari fungsi kepadatan peluang dan fungsi *survival.* Fungsi *likelihood* dan prior Gamma digunakan untuk membentuk distribusi posterior. Fungsi kepadatan peluang dari distribusi posterior digunakan untuk mengestimasi parameter dengan pendekatan Bayesian GELF. Langkah terakhir dalam penelitian ini adalah menginterpretasikan data pasien kanker paru-paru dari nilai yang diperoleh.

DISTRIBUSI SURVIVAL

Distribusi survival digunakan untuk mengestimasi distribusi waktu survival. Distribusi survival biasanya ditandai dengan tiga fungsi yaitu: fungsi kepadatan peluang, fungsi survival, dan fungsi hazard. Hal ini bearti jika salah satu dari persamaan fungsi diketahui maka kedua fungsi lainnya dapat ditentukan[4].

1. Fungsi kepadatan peluang (*probability density function)*

Fungsi kepadatan peluang adalah peluang suatu individu mati atau mengalami kejadiaan sesaat dalam interval waktu t sampai $(t + \Delta t)$. Fungsi kepadatan peluang $f(t)$ dirumuskan sebagai berikut:

$$f(t) = \lim_{n\to\infty}\left[\frac{P(t < T < (t + \Delta t))}{\Delta t}\right] = \lim_{n\to\infty}\left[\frac{F(t + \Delta t) - F(t)}{\Delta t}\right]$$

Jika T merupakan variabel random non negatif pada interval$[0, \infty]$, maka $F(t)$ merupakan fungsi distribusi kumulatif kontinu dari T. Didefinisikan sebagai peluang suatu individu mengalami kejadian kurang dari sama dengan waktu t, yaitu :

$$F(t) = P(T \le t) = \int_0^t f(t)dt \qquad (1)$$

2. Fungsi Survival

Fungsi *Survival* $S(t)$ didefinisikan sebagai peluang suatu individu dapat bertahan hidup waktu survival sampai dengan waktu t dengan $(t > 0)$ yaitu sebagi berikut:

$$S(t) = 1 - P(T \le t) = 1 - F(t), t \ge 0 \qquad (2)$$

3. Fungsi *Hazard*

Fungsi *hazard* (*hazard function*) adalah probabilitas kematian selama interval waktu $(t, t + \Delta t)$ dengan asumsi individu tetap hidup pada interval waktu tersebut dan biasanya dinotasikan dengan $\mathrm{h}(t)$. Fungsi *hazard* dinyatakan sebagai berikut :

$$h(t) = \lim_{n\to\infty}\left[\frac{P(t \le T < (t + \Delta t)|T \ge t}{\Delta t}\right]$$

Fungsi *hazard* juga dapat dinyatakan dalam distribusi fungsi kumulatif F(t) dan fungsi kepadatan peluang f(t) sebagai berikut :

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)} \quad (3)$$

DISTRIBUSI PARETO

Distribusi pareto berasal dari nama seorang ekonom yaitu Vilfredo Pareto (1848-1923) yang mengamati bahwa 80% kekayaan di Milan dimiliki oleh hanya 20% dari penduduknya. Distribusi Pareto disebut juga dengan distribusi *power-law*. Jika sebuah kumpulan data memiliki distribusi *power-law*, maka dikatakan bahwa data-data tersebut tidak sensitive terhadap rata-rata atau standar deviasi dari data tersebut atau dengan kata lain, data itu tidak bersifat acak. Fungsi kepadatan peluang Distribusi Pareto dapat dinyatakan sebagai berikut:

$$f(t) = \begin{cases} \frac{\alpha\theta^{\alpha}}{t^{\alpha+1}} & ; t \geq \theta \\ 0 & ; t < \theta \end{cases} \quad (4)$$

Fungsi distribusi Kumulatif untuk distribusi Pareto ialah:

$$F(t) = \begin{cases} 1 - \left(\frac{\theta}{t}\right)^{\alpha} & ; t \geq \theta \\ 0 & ; t < \theta \end{cases} \quad (5)$$

Fungsi survival dari distribusi Pareto sebagai berikut :

$$S(t) = \left(\frac{\theta}{t}\right)^{\alpha} \quad (6)$$

Sehingga fungsi *hazard* ialah:

$$h(t) = \frac{\alpha}{t} \quad (7)$$

ANALISIS METODE BAYESIAN

Pada Metode Bayesian, parameter dipandang sebagai variabel random dengan mengikuti suatu distribusi tertentu yang disebut distribusi prior. Penggabungan distribusi prior dan fungsi *likelihood* akan membentuk suatu distribusi posterior. Fungsi kepadatan peluang dari distribusi posterior menjadi dasar dalam proses estimasi parameter metode Bayesian. Berikut langkah-langkah yang dilakukan dalam analisis metode Bayesian, yaitu :

1. Menentukan Fungsi L*ikelihood* untuk Data Tersensor

Fungsi *Likelihood* dari distribusi Pareto untuk data tersensor adalah sebagai berikut:

$$L = (t_i; \theta, \delta_i) = \prod_{i=1}^{n} [f(t_i; \theta)]^{\delta i} [s(t_i; \theta)]^{1-\delta i} \quad (8)$$

Dengan δ_i adalah indikator penyensoran, bernilai 1 jika data tersensor dan bernilai 0 jika data tidak tersensor. Nilai t_i diperoleh dari min $(T_i, C_i), i = 1,2, \ldots, n$ dengan T_i adalah waktu hidup individu ke i dengan $i = 1,2, \ldots, n,$ dan C_i adalah waktu penyensoran individu ke $i$dengan$= 1,2, \ldots, n$. Sehingga fungsi *likelihood* dari distribusi Pareto untuk data tersensor rmemiliki bentuk sebagai berikut:

$$L(t_i;\theta,\delta_i) = \prod_{i=1}^{n}\left[\frac{\alpha\theta^{\alpha}}{t^{\alpha+1}}\right]^{\delta_i}\left[\left(\frac{\theta}{t}\right)^{\alpha}\right]^{1-\delta_i}$$

$$= \prod_{i=1}^{n}\left[\frac{\alpha\theta^{\alpha}}{t^{\alpha+1}}\right]^{\delta i}\left[\left(\frac{\theta}{t}\right)^{\alpha}\right]^{1-\delta i}$$

$$= \left[\left(\frac{\alpha\theta^{\alpha}}{{t_1}^{\alpha+1}}\right)^{\delta_1}\left(\frac{\alpha\theta^{\alpha}}{{t_2}^{\alpha+1}}\right)^{\delta_2}\cdots\left(\frac{\alpha\theta^{\alpha}}{{t_n}^{\alpha+1}}\right)^{\delta_n}\right]\left[\left(\left(\frac{\theta}{t_1}\right)^{\alpha}\right)^{1-\delta_1}\left(\left(\frac{\theta}{t_2}\right)^{\alpha}\right)^{1-\delta_2}\cdots\left(\left(\frac{\theta}{t_n}\right)^{\alpha}\right)^{1-\delta_n}\right]$$

$$= \left[\left(\frac{\alpha}{t_1}\right)^{\delta_1}\left(\frac{\alpha}{t_2}\right)^{\delta_2}\cdots\left(\frac{\alpha}{t_n}\right)^{\delta_n}\right]\left[\left(\frac{\theta}{t_1}\right)^{\alpha}\left(\frac{\theta}{t_2}\right)^{\alpha}\cdots\left(\frac{\theta}{t_n}\right)^{\alpha}\right]$$

$$= \frac{\alpha^{\sum_{i=1}^{n}\delta_i}}{\prod_{i=1}^{\mathrm{n}}(t_i)^{\delta_i}}\frac{\theta^{\mathrm{n}\alpha}}{\prod_{i=1}^{\mathrm{n}}(t_i)^{\alpha}}$$

$$= \frac{\theta^{n\alpha}\alpha^{\sum_{i=1}^{n}\delta_i}}{\prod_{i}^{n}(t_i)^{\delta_i+\alpha}} \qquad (9)$$

2. Menentukan Distribusi Prior

Metode Bayesian memberikan pilihan keyakinan terhadap suatu parameter dari sebuah distribusi. Ketika populasi mengikuti distribusi tertentu dengan suatu parameter didalamnya (dalam kasus ini θ), maka parameter θ mengikuti suatu distribusi peluang yang disebut distribusi prior. Salah satu distribusi prior yang dapat digunakan adalah prior sekawan, dalam penelitian ini prior sekawan dari distribusi Pareto adalah distribusi Gamma:

$$f(\theta) = \frac{B^A}{\Gamma(A)}\theta^{A-1}e^{-B\theta} \ , \theta \geq 0 \qquad (10)$$

3. Membentuk Distribusi Posterior

Dalam estimasi Bayesian, setelah informasi sampel diambil dan prior telah detentukan maka distribusi posteriornya dicari dengan menggabungkan priornya dengan informasi sampel yang diperoleh dalam bentuk fungsi *likelihood*, dimana prior ini indenpenden terhadap fungsi *likelihood*. Formula posterior $f(\theta|t_i)$ merupakan penggabungan darifungsi kepadatan peluang distribusi prior Gamma $f(\theta)$ dengan fungsi *likelihood* $L(t_i;\theta,\delta_i)$ dari distribusi Pareto. Berdasarkan persamaan (9) dan (10) maka fungsi kepadatan peluang distribusi posterior sebagai berikut :

$$f(\theta|t_i) = \frac{f(\theta).L(t_i;\theta,\delta_i)}{\int_0^{\infty} f(\theta)\,L(t_i;\theta,\delta_i)\,d\theta}$$

$$= \frac{\frac{B^A}{\Gamma(A)}\theta^{A-1}\,e^{-B\theta}\,\frac{\theta^{n\alpha}\alpha^{\sum_{i=1}^{n}\delta i}}{\prod_{i=1}^{n}(t_i)^{\delta i+\alpha}}}{\int_0^{\infty}\frac{B^A}{\Gamma(A)}\theta^{A-1}e^{-B\theta}.\frac{\theta^{n\alpha}\alpha^{\sum_{i=1}^{n}\delta i}}{\prod_{i=1}^{n}(t_i)^{\delta i+\alpha}}\,d\theta}$$

$$= \frac{\theta^{A-1+n\alpha}e^{-B\theta}}{\left(\frac{1}{B}\right)^{A-1+1-n\alpha}\int e^{-u}\,U^{A-1+n\alpha}\,du}$$

$$= \frac{\theta^{A+n\alpha} e^{-B\theta} B^{A+n\alpha}}{\Gamma(A+n\alpha)} \quad (11)$$

ESTIMASI PARAMETER BAYESIAN GELF

Estimasi parameter dengan pendekatan *general entropy loss function* (GELF) dapat didefinisikan sebagai berikut :

$$\mathcal{L}(\theta, \hat{\theta}_{BG}) = \left(\frac{\hat{\theta}_{BG}}{\theta}\right)^k - k\, ln\left(\frac{\hat{\theta}_{BG}}{\theta}\right) - 1\,, \qquad k \neq 0, \qquad 0 < \theta < \infty$$

Dengan $\hat{\theta}_{BG}$ merupakan estimator Bayesian GELF untuk parameter θ. Estimasi Bayesian GELF dari θ pada distribusi Pareto untuk data tersensor diperoleh dengan meminimumkan ekspetasi *loss function* yang diformulasikan sebagai berikut:

$$\hat{\theta}_{BG} = \mathrm{E}[\theta^k]^{\frac{-1}{k}} = \left(\frac{\Gamma(A+n\alpha)}{\Gamma(A+n\alpha-k)}\right)^{\frac{1}{k}} . \left(\frac{1}{B}\right) \quad (12)$$

Sehingga diperoleh estimasi fungsi survival $\left(\hat{S}_{BG}\right)$ dan fungsi hazard $\left(\hat{h}_{BG}\right)$dari distribusi Pareto pada data tersensor adalah:

$$\hat{S}_{BG} = \left(t_i; \hat{\theta}_{BG}\right) = \left(\frac{\hat{\theta}_{BG}}{t}\right)^{\alpha} \quad (13)$$

$$\hat{h}_{BG} = \left(t_i; \hat{\theta}_{BG}\right) = \frac{\alpha}{t} \quad (14)$$

STUDI KASUS

Data yang digunakan adalah data waktu *survival* pasien penderita kanker paru-paru sebanyak 71. Data tersebut merupakan data sekunder dari program R. Dari data 71 pasien penderita kanker paru-paru, selanjutnya dilakukan uji distribusi data untuk mengetahui data berdistribusi Pareto atau tidak. Selanjutnya estimasi parameter metode Bayesian GELF untuk penderita kanker paru-paru. Penjelasan lebih lengakap adalah sebagai berikut :

1. **Uji Distribusi Data**

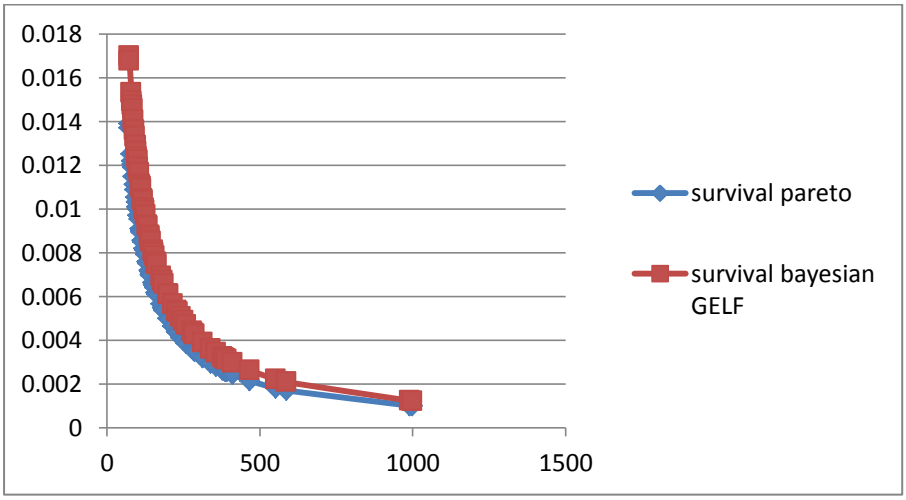
   Uji kecocokan (*goodness of fit*) digunakan untuk mengetahui ada atau tidaknya kesesuaian (kecocokan) model sebaran yang diasumsikan atau apakah satu variabel bias didekati dengan menggunkan distribusi atau tidak.Taraf nyata (*sig*) yang digunakan adalah $\alpha = 5\% = 0{,}05$. Dalam menentukan keputusan akhir untuk menolak atau menerima H_0 didasarkan pada wilayah kritis α dengan nilai *p-value* yang mendukung suatu uji dalam bentuk peluang, jika nilai *p-value* $\leq \alpha$ maka H_0 ditolak H_1 diterima. Dengan menggunakan software Easy-fit dapat diketahui bahwa nilai *P-value* untuk distribusi Pareto Bayesian GELF adalah 0,19547 sehingga dapat disimpulkan bahwa data tersebut berdistribusi Pareto.

   H_0 = data berdistribusi Pareto

   H_1 = data tidak berdistribusi Pareto

2. **Estimasi Parameter Metode Bayesian GELF**

   Dari data kasus penderita kanker paru-paru, dengan menggunakan program R diambil nilai $\alpha = 0{,}05$, $n = 71$, $\theta = 1{,}000239$, $A = 0{,}45$, $B = 2$, $k = 2$, $\hat{\theta}_{BG} = 1{,}224745$ sehingga diperoleh grafik fungsi *survival* dengan metode Bayesian GELF sebagai berikut :



Gambar 1. Grafik fungsi perbandingan nilai *survival* Pareto dan nilai *survival* Bayesian GELF

Dari gambar 1 terlihat bahwa grafik nilai *survival* pasien kanker paru-paru yang diestimasi dengan metode Bayesian GELF lebih tinggi daripada nilai *survival* sebenarnya. Hal ini disebabkan pengaruh estimasi parameter dari fungsi *survival* distribusi Pareto yang mempunyai nilai lebih besar dari nilai parameter sebenarnya:

Tabel 1 Hasil perhitungan Nilai *Survival* dan Nilai *Survival* Bayesian GELF.

t	Nilai *Survival* $S(t_i)$	Nilai Survival Bayesian GELF $\widehat{S}_{BG}(t_i)$
72	0,013892206	0,017010345
73	0,013701902	0,016777327
80	0,012502986	0,015309311
.	.	.
.	.	.
.	.	.
999	0,00100124	0,001225971

Dari hasil perhitungan nilai *Survival* dan nilai *Survival* dengan menggunakan metode Bayesian GELF perhitungan peluang hidup pada kasus penderita kanker paru-paru menjadi lebih tinggi.

3. Perhitungan MAPE

Mean absolute percentage error (MAPE) dihitung menggunakan kesalahan absolut pada tiap periode dibagi dengan nilai observasi yang nyata pada periode itu[7]. Nilai MAPE dapat dihitung dengan persamaan berikut :

$$MAPE = \frac{\sum|(S(t_i) - \widehat{S}_{BG}(t_i))/S(t_i)|}{n} x\ 100\% \quad (15)$$

Dimana $S(t_i)$ merupakan nilai *survival* awal, $\widehat{S}_{BG}(t_i)$ merupakan nilai *survival* yang telah diestimasi dan n adalah jumlah data . Kriteria nilai MAPE jika dilihat dengan tabel adalah sebagai berikut:

Tabel 2 Kriteria MAPE

MAPE	Status
$< 10\%$	Sangat Baik
$10\% - 20\%$	Baik
$20\% - 50\%$	Cukup
$> 50\%$	Buruk

Berdasarkan persamaan (15) diperoleh nilai MAPE sebagai berikut :

$$MAPE = \frac{|14{,}365|}{64} x\ 100 = 22.44\%$$

KESIMPULAN

1. Formula estimasi parameter dan nilai *survival* berdistribusi Pareto pada data tersensor menggunakan metode Bayesian GELF adalah sebagai berikut :

$$\hat{\theta}_{BG} = E[\theta^k]^{\frac{-1}{k}} = \left(\frac{\Gamma(\mathrm{A} + n\alpha)}{\Gamma(\mathrm{A} + n\alpha - k)}\right)^{\frac{1}{k}} . \left(\frac{1}{B}\right)$$

$$\hat{S}_{BG} = \left(t_i, \hat{\theta}_{BG}\right) = \left(\frac{\left(\frac{\Gamma(A+n\alpha)}{\Gamma(A+n\alpha-k)}\right)^{\frac{1}{k}}}{t}\right)^{\alpha}$$

2. Berdasarkan hasil estimasi metode Bayesian GELF diperoleh nilai survival yang lebih tinggi daripada nilai aktual survival.
3. Berdasarkan nilai MAPE yang diperoleh dari nilai *survival* distribusi Pareto dengan menggunakan pendekatan Bayesian GELF adalah sebesar 22,44%. Hal ini berarti bahwa metode Bayesian GELF memiliki kemampuan yang cukup dalam mengestimasi peluang hidup pasien penderita kanker paru-paru.

DAFTAR PUSTAKA

[1]. Kleinbaum, D.G., and Klein, M. *Survival Analysis* : *A Self-Learning Text Second Edition,* Springer Science Business Media, Inc, New York, 2005.

[2]. Fitria, S., Helmi, &Rizki, S.W. Estimasi Parameter Model *Survival* distribusi Eksponensial Data Tersensor Dengan Metode Maksimum Likelihood dan Bayesian SELF.*Buletin Ilmiah Math. Stat.* Dan Terapannya (*Bimaster),* 3: 213-220, 2016.

[3]. Box, G.E.P and Tiao, G.C. *Bayesian Inference in Statistical Analysis,* Addison Wesley Publishing Company, London, 1973

[4]. Lee. E.T. and Wang, J.W. *Statistical Methods for Survival Data Analysis Third Edition,* Canada : A Jhon Wiley & Sons, Inc, 2003.

[5]. Bolstad, W.M. *Introduction to Bayesian Statistical Second Edition,* A.John Wiley & Sons, Inc, Amerika, 2007.

NILA HANDAYANI :Jurusan Matematika FMIPA UNTAN,Pontianak
Nilayani161096@gmail.com
SETYO WIRA RIZKI :Jurusan Matematika FMIPA UNTAN,Pontianak
setyo.wirarizki@math.untan.ac.id