PENAKSIRAN PARAMETER DISTRIBUSI WEIBULL DENGAN METODE BAYESIAN SURVIVAL DAN MAKSIMUM LIKELIHOOD

(Assessment of The Weibull Distribution Parameter with Bayesian Survival and Maximum Likelihood Methods)

Sri Astuti Thamrin[1*], Azhar[2], dan Andi Kresna Jaya[3]

[1*] Sri Astuti Thamrin, Program Studi Statistika, Universitas Hasanuddin, Makassar
[2] Azhar, Program Studi Statistika, Universitas Hasanuddin, Makassar
[3] Andi Kresna Jaya, Program Studi Statistika, Universitas Hasanuddin, Makassar
[*] email Penulis Korespondensi: tuti@unhas.ac.id

ABSTRAK

Analisis kelangsungan hidup *(survival)* merupakan suatu prosedur statistik yang digunakan untuk menganalisis data dimana variabel yang diteliti adalah waktu sampai terjadinya suatu kejadian. Salah satu distribusi yang banyak digunakan dalam menangani masalah *survival* adalah distribusi Weibull. Distribusi ini memiliki parameter skala *(scale)* dan bentuk *(shape)* yang mampu menyajikan keakuratan kegagalan dengan sampel yang kecil. Untuk mendapatkan model Weibull *survival*, kedua parameter tersebut perlu ditaksir. Menaksir parameter dapat dilakukan dengan dua pendekatan yaitu Bayesian dan klasik (*frequentist*). Tujuan dari artikel ini adalah untuk mengestimasi parameter distribusi Weibull dengan metode Bayesian dan maksimum *likelihood.* Kedua metode tersebut dibandingkan dan diaplikasikan pada data simulasi *survival* yang berdistribusi Weibull. Hasil yang diperoleh menunjukkan bahwa estimasi parameter dengan menggunakan metode *Bayesian* memberikan nilai taksiran yang lebih baik dibandingkan dengan metode maksimum likelihood berdasarkan nilai *mean squared error* (*MSE*).

Kata Kunci : Analisis *Survival*, *Bayesian*, Distribusi Weibull, Maksimum *Likelihood, Mean Squared Error*.

ABSTRACT

Analysis of survival is a statistical procedure used to analyze data where the variables studied are the time until the occurrence of an event. One distribution that is widely used in dealing with survival problems is Weibull distribution. This distribution has scale and shape parameters that are able to present the accuracy of failures with small samples. To get a Weibull survival model, these two parameters need to be estimated. Estimating parameters can be done with two approaches namely Bayesian and classic (frequentist). The purpose of this article is to estimate the parameters of Weibull distribution with the Bayesian method and maximum likelihood. Both methods are compared and applied to survival simulation data with Weibull distribution. The result obtained indicates that parameter estimation using the Bayesian method provides a better estimated value than the maximum likelihood method based on the mean squared error (MSE) value.

Keywords: *Bayesian, maximum likelihood, mean squared error, survival analysis, Weibull distribution.*

I. PENDAHULUAN

Distribusi Weibull banyak dipakai untuk memodelkan dan menganalisa data waktu kegagalan. Data seperti ini banyak ditemui pada bidang kesehatan, biologi, teknik, dan ekonomi. Distribusi Weibull adalah distribusi probabilitas yang kontinu. Kelebihan dari distribusi Weibull adalah bentuk fungsionalnya yang mudah sehingga mudah diaplikasikan di beberapa kejadian. Distribusi Weibull memiliki parameter skala *(scale)* dan bentuk *(shape)*. Parameter-parameter tersebut harus diketahui agar dapat dicari lebih lanjut sifat dan karakteristik data yang berdistribusi Weibull.

Dalam statistika, cabang utama statistika inferensi adalah estimasi. Ada beberapa prosedur dalam estimasi. Prosedur yang tergantung pada sejumlah sampel disebut metode klasik. Metode ini mengasumsikan distribusi populasi diketahui. Metode yang digunakan untuk menaksir nilai parameter bila distribusi populasi diketahui adalah maksimum *likelihood*. Metode ini hanya mendasarkan inferensinya pada sampel. Prosedur lainnya yang tergantung pada informasi sebelumnya disebut metode *Bayesian*. Pada metode ini perlu diketahui bentuk distribusi awal (*prior*) dari populasi. Informasi ini kemudian digabungkan dengan informasi dari sampel untuk digunakan dalam mengestimasi parameter populasi, yang disebut *posterior*.

Omurlu dkk (2009) juga telah membandingkan kinerja analisis *survival Bayesian* dan analisis *Cox* regresi pada data simulasi dan kanker. Studi data simulasi dengan dua algoritma yang berbeda menggunakan prior informatif dan non-informatif. Selanjutnya, Ahmed dkk (2010) telah menggunakan informasi Jeffery prior pada estimasi parameter distribusi Weibull. Mereka menggunakan data tersensor dengan pendekatan *Bayesian*. Selain itu, Abdulabaas dkk (2013) telah membahas bagaimana estimasi dua parameter distribusi Weibull dengan menggunakan metode maksimum *likelihood* (MLE) dan *Bayesian* serta membandingkan hasil numerik dengan simulasi dalam program MATLAB. Thamrin dkk (2013) menaksir parameter distribusi Weibull pada data ekspresi gen menggunakan metode *Bayesian*. Kedua metode Bayesian dan klasik memegang peranan yang sangat penting dalam menaksir parameter. Oleh karena itu dalam artikel ini akan di bandingkan kedua metode tersebut dalam menaksir paremeter pada data simulasi *survival* yang berdistribusi Weibull. Dalam melakukan simulasi data dan menganalisis data simulasi *survival* ini digunakan perangkat lunak R versi 3.3.3.

II. METODOLOGI PENELITIAN

2.1 Simulasi data survival

Data yang digunakan dalam artikel ini adalah data simulasi. Adapun prosedur dalam melakukan simulasi data diberikan berikut ini.

a. Menetapkan jumlah data, $n = 30, 60, 100, 200$.
b. Menetapkan nilai parameter α dan γ dari distribusi Weibull, yaitu α = 2 dan γ = 1.
c. Membangkitkan suatu nilai dari variabel acak $u_i \sim U(0,1)$.
d. Membangkitkan waktu *survival* t_i, untuk $i = 1,2, \dots, n$.
   Nilai acak dari $u_i \sim U(0,1)$ dapat digunakan untuk membangkitkan waktu *survival* t_i:
$$t_i = \left(\frac{-\ln(1-u_i)}{\gamma}\right)^{1/\alpha}$$
e. Membangkitkan waktu sensor T_i dari suatu nilai variabel acak $T_i \sim U(0,1)$ dengan asumsi bahwa data tersensor kanan dan δ_i adalah indikator sensor.
$$\delta_i = \begin{cases} 0, \text{sensor jika } t_i > T_i \\ 1, \text{tidak sensor jika } t_i \le T_i \end{cases}$$
f. Menggabungkan data waktu *survival* t_i dan indikator sensor δ_i.

2.1.1 Konstruksi Fungsi likelihood

Fungsi kepadatan peluang dari distribusi Weibull (Collet, 2015) dituliskan sebagai berikut:

$$f(t;\alpha,\gamma)=\alpha\gamma t^{\alpha-1}\exp(-\gamma t^{\alpha}) \ , \quad t\geq 0$$

dimana $\alpha > 0$ adalah parameter bentuk dan $\gamma > 0$ adalah parameter skala. Jika $F(t) = P(T \leq t)$ untuk setiap bilangan t, maka $F(t)$ disebut fungsi distribusi kumulatif peubah acak T. Fungsi distribusi kumulatif $F(t)$ dari suatu peubah acak kontinu dengan distribusi peluang $P(T \leq t)$ dinyatakan dengan:

$$F(t)=\int_0^t f(x)dx \qquad (1)$$

Fungsi distribusi kumulatif untuk distribusi Weibull (α,β) pada data kelangsungan hidup dituliskan sebagai berikut:

$$F(t)=\int_0^t f(x)dx$$

$$=\int_0^t \alpha\gamma x^{\alpha-1}\exp(-\gamma x^{\alpha})\,dx \qquad (2)$$

Fungsi *survival* dapat dinotasikan dengan $S(t)$, yaitu peluang suatu individu bertahan hidup lebih dari waktu t, berdasarkan persamaan (2), maka diperoleh fungsi *survival* untuk distribusi Weibull sebagai berikut:

$$S(t)=P(T>t)$$

$$=\int_t^{\infty} f(x)dx \ . \qquad (3)$$

Dengan menggunakan definisi fungsi distribusi kumulatif pada persamaan (1), $F(t) = P(T \leq t)$, fungsi *survival* dapat dituliskan sebagai:

$$S(t)=\exp\ (-\gamma t^{\alpha}\ ) \qquad (4)$$

Dengan menggunakan teorema peluang bersyarat maka diperoleh fungsi hazard dari distribusi Weibull berdasarkan persamaan (3) dan (4) sebagai berikut:

$$h(t)=\frac{f(t)}{S(t)}= \ \alpha\gamma t^{\alpha-1} \qquad (5)$$

2.1.2 Spesifikasi Fungsi Likelihood

Dalam tulisan ini tipe sensor yang digunakan adalah sensor kanan. Misalkan δ_i dinotasikan sebagai variabel indikator dimana waktu kelangsungan hidup dikategorikan sebagai tidak tersensor δ_i = 1 jika individu ke-i (i= 1, 2, ..., n) mengalami suatu kejadian atau gagal dalam batas periode dan δ_i = 0 jika tidak gagal.

Jika diketahui data waktu hidup suatu objek pengamatan yang mengalami sensor kanan untuk n pasien adalah T = (s_1, s_2, ... , s_n) dimana s_i=(t_i,δ_i) adalah waktu pengamatan dan δ_i adalah fungsi indikator T_i, maka fungsi *likelihood* dari data tersebut adalah:

$$L(\ t_i;\ \alpha,\gamma\ )=\prod_{i=1}^{n}[f\ \ [s(t_i)]^{1-\delta_i}(t_i;\alpha,\gamma)]^{\delta_i}$$

dimana k adalah jumlah data yang tidak tersensor.

2.1.3 Penaksir maksimum likelihood

Misalkan $T_1, T_2, \ldots, T_n$ adalah peubah acak yang saling bebas dari populasi dengan fungsi kepadatan peluangnya dinyatakan oleh $f(t_i,\theta)$, dengan $\theta=(\alpha,\gamma)$ adalah parameter yang akan ditaksir dengan metode maksimum *likelihood*. Turunan-turunan parsial dari fungsi *log-likelihood* tersebut jika dinyatakan dalam bentuk matriks menjadi:

$$\boldsymbol{A}=\begin{bmatrix}\frac{\partial \ln L(\ t_i;\ \alpha,\gamma)}{\partial\alpha}\\ \frac{\partial \ln L(\ t_i;\ \alpha,\gamma)}{\partial\gamma}\end{bmatrix}=$$

$$\begin{bmatrix}\frac{k}{\alpha}+\sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} ln\ t_i-\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\gamma t_i{}^{\alpha}\ln t_i\\ \frac{k}{\gamma}-\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_i^{\alpha}\end{bmatrix} \qquad (6)$$

Terlihat bahwa estimasi parameter α dan γ tidak dapat diselesaikan secara analitik atau dengan kata lain menghasilkan estimator yang berbentuk implisit seperti pada persamaan (6). Oleh karena itu dalam pengaplikasian pada data diperlukan penyelesaian secara numerik. Metode numerik yang dapat digunakan adalah metode *Newton-Raphson* untuk menaksir parameter dari distribusi Weibull. Langkah-langkah dalam metode Newton Raphson, sebagai berikut:

1. Tentukan nilai awal dari α dan γ yaitu $\begin{bmatrix}\alpha\\ \gamma\end{bmatrix}_n$

2. Tentukan taksiran α dan γ yaitu $\begin{bmatrix}\alpha\\ \gamma\end{bmatrix}_{n+1}$ dengan rumus;
$$\begin{bmatrix}\alpha\\ \gamma\end{bmatrix}_{n+1} = \begin{bmatrix}\alpha\\ \gamma\end{bmatrix}_{n} - \boldsymbol{H^{-1}A}$$

2.1.4 Pendekatan Bayesian

Dalam estimasi parameter dengan menggunakan metode Bayes diperlukan distribusi prior, fungsi *likelihood* dan distribusi posterior. Distribusi prior dan fungsi *likelihood* digunakan untuk membentuk distribusi posterior dan distribusi posterior diperlukan untuk mengestimasi nilai parameter.

Dalam menentukan distribusi prior yang berkaitan dengan parameter pada pola distribusi, terdapat dua cara yaitu prior informatif dan prior non-informatif. Ketika pemilihan prior menjadi sulit dikarenakan tidak adanya informasi dari data sebelumnya atau distribusi priornya yang tidak mengandung informasi tentang parameter maka banyak peneliti yang memilih untuk menggunakan prior non-infomatif. Prior yang digunakan dalam penelitian ini adalah prior non-informatif. Dipilih parameter α berdistribusi Gamma, $\alpha \sim Gamma(10, 0.0001)$ dan parameter γ berdistribusi Gamma, $\gamma \sim Gamma(1,1)$.

Dalam estimasi *Bayesian* dikenal distribusi posterior. Distribusi posterior dapat dinyatakan dengan perbandingan antara fungsi kepadatan bersama dan fungsi marginal. Fungsi kepadatan bersama dapat ditulis dalam bentuk distribusi prior dan fungsi *likelihood.* Fungsi marginal ditulis dengan mengintegralkan bentuk distribusi prior dan fungsi *likelihood.* Oleh karena itu distribusi posterior dapat dituliskan sebagai:
$$f(\alpha,\gamma|t_i) = \frac{L(t_i|\alpha,\gamma)f_1(\alpha)f_2(\gamma)}{\int_0^\infty\int_0^\infty L(t_i|\alpha,\gamma)f_1(\alpha)f_2(\gamma)\,d\gamma\,d\alpha} \quad (7)$$

Bentuk akhir dari distribusi posterior pada persamaan (7) cukup rumit untuk diselesaikan secara analitik. Untuk mengatasi masalah ini akan digunakan algoritma *Gibbs Sampling* dengan membangkitkan sampel dari distribusi posterior bersama dengan terlebih dahulu menghitung prior dan *likelihood* (Gilks dkk 1996).

III. HASIL DAN PEMBAHASAN

3.1 Estimasi Parameter dengan metode MLE

Berdasarkan hasil perhitungan numerik dengan menggunakan metode MLE dan Newton Raphson, maka diperoleh hasil taksiran parameter bentuk α dan skala γ dengan 100 kali perulangan untuk data hasil simulasi *survival* berdistribusi Weibull.

Tabel 1. Nilai estimasi parameter α dan γ dengan metode MLE dan Newton-Raphson untuk data simulasi *survival.*

Para meter	Ukuran sampel	Nilai sebe narnya	Rata-rata nilai estimasi	Rata-rata nilai MSE
Bentuk α	30	2	1,829	0,127
	60	2	1,832	0,076
	100	2	1,831	0,052
	200	2	1,832	0,036
Skala γ	30	1	0,974	0,037
	60	1	0,922	0.018
	100	1	0,941	0.012
	200	1	0,938	0.007

Tabel 1 memperlihatkan hasil rata-rata nilai taksiran parameter bentuk α untuk beberapa ukuran sampel yaitu 30, 60, 100 dan 200 berturut-turut 1,829, 1,832, 1,831 dan 1,832. Selanjutnya hasil rata-rata nilai taksiran parameter skala γ dengan berbagai ukuran sampel yaitu 30, 60, 100 dan 200 berturut-turut 0,974, 0,922, 0,941 dan 0,938. Nilai tersebut telah konvergen pada iterasi ke-6 dengan perolehan nilai error < 0.001, dimana nilai 0,001 ini merupakan toleransi galat yang telah ditentukan sebelumnya sebagai kriteria penetapan konvergensi nilai yang diperoleh dalam tiap iterasi pada metode Newton Raphson.

Selanjutnya rata-rata nilai MSE dari masing-masing hasil taksiran parameter parameter bentuk α berturut-turut 0,127, 0,076, 0,052 dan 0,036 serta untuk rata-rata nilai MSE parameter skala γ berturut-turut 0,037, 0,018, 0,012 dan 0,007.

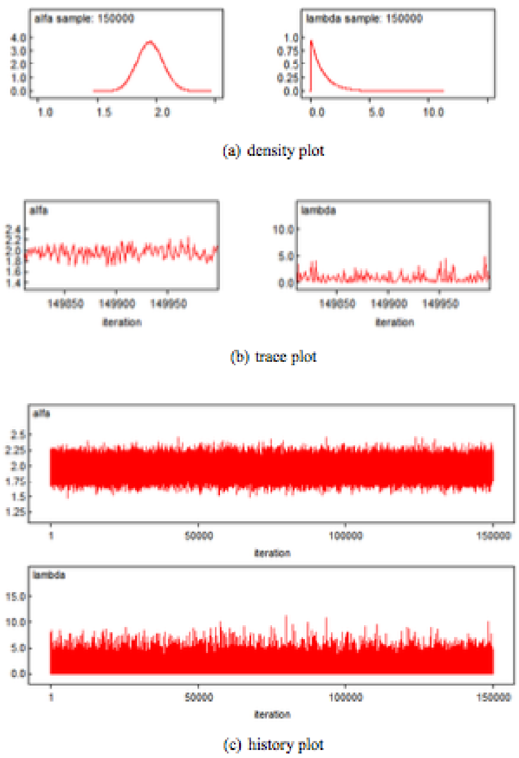
3.2 Estimasi Parameter dengan metode Bayesian

Berdasarkan hasil perhitungan dengan menggunakan algoritma *Gibbs Sampling* dengan pendekatan *Bayesian*, maka diperoleh rata-rata nilai parameter bentuk α dan parameter skala γ dengan 100 kali perulangan untuk data hasil simulasi *survival* berdistribusi Weibull pada Tabel 2.

Berdasarkan Tabel 2, diperoleh hasil rata-rata nilai taksiran parameter bentuk α untuk beberapa ukuran sampel yaitu 30, 60, 100 dan 200 berturut-turut 2,086, 1,960, 1,906 dan 1,845. Kemudian hasil taksiran parameter skala γ dengan ukuran sampel 30, 60, 100 dan 200 berturut-turut 0,999, 0,999, 0,999 dan 1,001. Nilai tersebut telah konvergen dengan iterasi sebanyak 150.000 yang ditunjukkan dengan plot kepadatan (*density plot*), *trace plot* dan *history plot* pada Gambar 1.

Tabel 2. Nilai estimasi posterior parameter α dan γ dengan metode *Bayesian* untuk data simulasi *survival*.

Parameter	Ukuran sampel	Nilai sebenarnya	Rata-rata nilai estimasi	Rata-rata nilai MSE
Bentuk α	30	2	1,829	0,127
	60	2	1,832	0,076
	100	2	1,831	0,052
	200	2	1,832	0,036
Skala γ	30	1	0,974	0,037
	60	1	0,922	0.018
	100	1	0,941	0.012
	200	1	0,938	0.007



Gambar 1. Plot Diagnostik parameter bentuk α dan skala γ untuk data simulasi *survival*

Setelah kondisi konvergen telah terpenuhi, maka langkah selanjutnya yaitu mencari nilai estimasi tiap parameter. Selanjutnya rata-rata nilai MSE yang diperoleh dari masing-masing hasil taksiran parameter bentuk α berturut-turut 0,075, 0,049, 0,030 dan 0,032 serta untuk nilai MSE parameter skala γ berturut-turut 8,7E-6, 5,1E-6, 7,5E-5 dan 7,5E-6.

IV. KESIMPULAN

Telah dilakukan perbandingan nilai estimasi menggunakan metode MLE dan metode Bayesian pada data simulasi *survival* berdistribusi Weibull. Berdasarkan nilai MSE yang diperoleh menunjukkan bahwa metode Bayesian merupakan penaksir yang lebih baik. Hal ini karena metode Bayesian memiliki nilai MSE yang lebih kecil dibandingkan dengan metode MLE dengan menggunakan beberapa ukuran sampel yang berbeda-beda.

DAFTAR PUSTAKA

Abdulabaas F., Yahya M., Neama I., (2013), A Comparison Between the Bayesian and the Classical estimator of Weibull Distribution. *Journal of Kufa Mathematics and Computer*, 1(8):21-28.

Ahmed, A., Hadel S, Noor A. I., (2010), Comparison of the Bayesian and Maximum Likelihood Estimation for Weibull Distribution. *Journal of Mathematics and Statistic*, 6(2):100-102.

Collet, D. (2015). *Modelling Survival Data in Medical Reseach*. 3rd Ed. London: Chapman and Hall.

Gilks, W,R., Richardson, S dan Spiegelhalter, D.G. (1996) *Markov chain monte carlo in practice*. Chapman and Hall, 1996.

Omurlu I., Kazim O., Mevlut T., (2009), Comparison of Bayesian survival analysis and Cox regression analysis in simulated and breast cancer data sets. *Expert System with Applications,* 36: 11341-11346.

Thamrin, S.A., Mengersen, K., dan McGree, J. (2013). *Bayesian Weibull Survival for Gene Expression Data*. In: Alston CL, Mengersen KL, Pettitt AN (eds) Case studies in Bayesian Statistical Modelling and Analysis, 1st edn.