Analisis Survival Parametrik Pada Data *Tracer Study* Universitas Sriwijaya

Alfensi Faruk
Jurusan Matematika, Fakultas MIPA, Universitas Sriwijaya
e-mail: alfensifaruk@unsri.ac.id

Abstract: In this study, we aimed to (1) show whether the Sriwijaya University tracer study data follow some survival distributions, (2) find the best survival distribution to represent the data, and (3) estimate the survival probability and hazard rate of the data. The tracer study was conducted from January 1, 2012 to December 31, 2012. There were 637 alumni who participated in the study. The result showed that the data follow the normal distribution, logistic distribution, and SEV distribution, in which the normal distribution was the best in representing the data. Based on the estimation procedure, the lowest probability of finding the first job was before graduation and the highest probability was about two years after graduation.

Keywords: *Parametric Survival Model, Tracer Study, Survival Distribution*

1. Pendahuluan

Masalah waktu tunggu kerja pertama alumni merupakan salah satu contoh penerapan dari analisis *survival*. Hal ini dikarenakan syarat-syarat agar suatu fenomena dikatakan sebagai waktu *survival* telah terpenuhi, yaitu (1) adanya suatu peristiwa yang diperhatikan (mendapatkan pekerjaan pertama), (2) adanya waktu awal pengamatan (hari kelulusan), dan (3) adanya satuan waktu pengamatan (biasanya dalam bulan atau tahun). Data mengenai waktu tunggu kerja pertama alumni tersebut biasanya diperoleh melalui *tracer study*, yaitu suatu studi penjajakan mengenai situasi terkini dari pekerjaan para alumni. Informasi dari *tracer study* tersebut sangat penting bagi pihak universitas sebagai salah satu landasan dalam membuat kebijakan akademik di masa yang akan datang.

Menggunakan model-model yang ada dalam analisis *survival*, peluang seorang alumni mendapatkan pekerjaan pertama dapat diketahui. Salah satu metode standar yang biasanya digunakan dalam mengestimasi peluang tersebut adalah metode Kaplan-Meier (Kaplan *et al*., [5]). Akan tetapi, menurut Sun [8] metode-metode standar dalam analisis *survival* tidak dapat digunakan lagi apabila data yang diperoleh adalah data yang tersensor interval. Hal ini juga berlaku pada sebagian besar data yang diperoleh

dalam *tracer study* di berbagai Universitas di Indonesia, karena data waktu tunggu kerja pertama para alumni tersebut biasanya berupa interval waktu (dalam bulan atau tahun).

Faruk [3] telah mengestimasi peluang-peluang *survival* berdasarkan data *tracer study* Universitas Sriwijaya (Unsri). Pendekatan yang digunakan dalam studi tersebut adalah pendekatan nonparametrik, yaitu menggunakan metode *nonparametric maximum likelihood estimate* untuk data tersensor interval. Walaupun pendekatan nonparametrik cukup populer, namun apabila ternyata menggunakan metode statistik tertentu (misalnya, metode grafik) dapat ditunjukkan bahwa data *survival* mengikuti suatu distribusi tertentu, maka pendekatan parametrik lebih tepat untuk digunakan terhadap data tersebut (Lee *et al*. [7]). Beberapa contoh penelitian yang membahas mengenai pendekatan parametrik dalam analisis *survival* antara lain adalah Akram *et al*. [1], Hayat *et al*. [4], dan Faruk [2].

Apabila dalam data tersensor interval juga memuat data tersensor kiri, maka tidak semua distribusi *survival* dapat digunakan pada data tersebut. Karena hanya distribusi *survival* yang memuat waktu negatif yang dapat digunakan. Distribusi tersebut antara lain adalah distribusi normal, distribusi logistik, dan distribusi *Smallest Extreme Values* (SEV). Data waktu tunggu kerja pertama alumni Unsri, juga memuat peristiwa dimana alumni telah mendapatkan pekerjaan pertama sebelum hari kelulusan. Hal ini berarti, waktu tunggu kerja pertama alumni Unsri juga memuat data *survival* yang tersensor kiri. Adapun, tujuan dari penelitian ini adalah (1) mengkaji bagaimana kesesuaian data *tracer study* Unsri dengan beberapa distribusi *survival*, (2) membandingkan distribusi mana yang paling sesuai dengan data, dan (3) mengestimasi fungsi *survival* dan fungsi *hazard* dari data berdasarkan distribusi *survival* tersebut.

2. Tinjauan Pustaka

Analisis Survival

Analisis *survival* adalah suatu cabang dalam statistika yang mempelajari tentang waktu *survival*, yaitu waktu hingga terjadinya suatu peristiwa tertentu (Kleinbaum *et al*. [6]). Apabila T adalah variabel acak kontinu yang melambangkan waktu *survival* dan waktu amatan (*observed time*) $t \geq 0$ adalah realisasi dari T, maka fungsi kepadatan peluang (fkp) dari T didefinisikan sebagai

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t<T<t+\Delta t)}{\Delta t}. \quad (1)$$

Fungsi *survival*, $S(t)$, dapat didefinisikan sebagai peluang suatu individu mengalami suatu peristiwa pada waktu lebih dari t yang dituliskan sebagai

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx\ . \quad (2)$$

Selanjutnya, fungsi penting lainnya dalam analisis *survival* adalah fungsi *hazard* atau *hazard rate* yang berbentuk

$$h(t) = \frac{\lim_{\Delta t \to 0} P\left[\frac{\Pr(t \le T < t+\Delta t | T \ge t)}{\Delta t}\right]}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t(1-F(t))} = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)}, \quad (3)$$

fungsi *hazard* dapat didefinisikan sebagai peluang terjadinya peristiwa dalam selang waktu yang sangat kecil, yaitu pada saat $\Delta t \to 0$.

Data Tersensor Interval

Data *survival* dikatakan sebagai data tersensor interval apabila waktu amatannya tidak diketahui secara eksak namun interval waktu yang memuat waktu amatan tersebut masih dapat diketahui. Misalkan dalam suatu populasi terdapat n buah subjek yang saling bebas dan jika T_i melambangkan waktu *survival* atau waktu amatan dari subjek ke-i, dengan $i = 1,2, \dots, n$, maka data tersensor interval dari waktu T_i adalah

$$O = \{(L_i, R_i]; i = 1, \dots, n\}, \quad (4)$$

dimana L_i = batas kiri interval, R_i = batas kanan interval, dan $(L_i, R_i]$ = data tersensor yang memuat waktu amatan T_i (Sun, [8]).

Fungsi Survival Berdistribusi Normal

Sebagian peneliti berpendapat bahwa distribusi normal kurang cocok digunakan dalam pemodelan data *survival* karena limit kiri dari distribusi ini menuju ke negatif tak hingga. Akan tetapi, karena distribusi normal memiliki nilai rata-rata yang relatif tinggi serta nilai simpangan baku yang relatif kecil, maka persoalan waktu *survival* negatif tersebut seharusnya tidak menjadi masalah. Oleh karena itulah distribusi normal dapat digunakan pada data *survival* tersensor interval yang memuat data tersensor kiri.

Fungsi kepadatan peluang (fkp) dari distribusi normal dengan parameter μ dan σ diberikan oleh

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{2}\right)^2}, \quad (5)$$

dimana μ dan σ berturut-turut adalah rata-rata dan simpangan baku dari waktu *survival*. Menggunakan persamaan (1), maka dapat diperoleh fungsi *survival* berdistribusi normal yang berbentuk

$$S(t) = \int_t^\infty f(x)dx = \int_t^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{2}\right)^2} dx, \quad (6)$$

sehingga fungsi *hazard* berdistribusi normal dapat dituliskan sebagai

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{2}\right)^2}}{\int_t^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{2}\right)^2} dx} . \quad (7)$$

Fungsi Survival Berdistribusi Logistik

Fungsi kepadatan peluang dari waktu *survival* T yang berdistribusi logistik adalah

$$f(t) = \frac{e^z}{\sigma(1+e^z)^2} , \quad (8)$$

dengan $z = \frac{t-\mu}{\sigma}$, $-\infty < t < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, μ adalah rata-rata (sebagai parameter lokasi), dan σ adalah simpangan baku (sebagai parameter skala), sedangkan fungsi *survival* berdistribusi logistik diberikan oleh

$$S(t) = \int_t^{\infty} \frac{e^{\frac{x-\mu}{\sigma}}}{\sigma(1+e^{\frac{x-\mu}{\sigma}})^2} dx , \quad (9)$$

dan fungsi *hazard*nya berbentuk

$$h(t) = \frac{\frac{e^{z\frac{t-\mu}{\sigma}}}{\sigma(1+e^{\frac{t-\mu}{\sigma}})^2}}{\int_t^{\infty} \frac{e^{\frac{x-\mu}{\sigma}}}{\sigma(1+e^{\frac{x-\mu}{\sigma}})^2} dx} . \quad (10)$$

Fungsi Survival Berdistribusi Smallest Extreme Values (SEV)

Bentuk umum fungsi kepadatan peluang dari waktu *survival* T yang berdistribusi SEV adalah

$$f(t) = \frac{1}{\sigma} e^{\frac{x-\mu}{\sigma}} e^{-e^{\frac{x-\mu}{\sigma}}} , \quad (11)$$

dimana μ dan σ berturut-turut adalah parameter lokasi dan parameter skala dari waktu *survival*. Selanjutnya, fungsi *survival* berdistribusi SEV diberikan oleh

$$S(t) = \int_t^{\infty} \frac{1}{\sigma} e^{\frac{x-\mu}{\sigma}} e^{-e^{\frac{x-\mu}{\sigma}}} dx , \quad (12)$$

sedangkan fungsi *hazard*nya dapat dituliskan sebagai berikut

$$h(t) = \frac{\frac{1}{\sigma} e^{\frac{x-\mu}{\sigma}} e^{-e^{\frac{x-\mu}{\sigma}}}}{\int_t^{\infty} \frac{1}{\sigma} e^{\frac{x-\mu}{\sigma}} e^{-e^{\frac{x-\mu}{\sigma}}} dx} . \quad (13)$$

3. Metode Penelitian

Data yang digunakan dalam penelitian ini diperoleh dari program *tracer study* yang dilaksanakan oleh Unsri pada periode 1 Januari Tahun 2012 hingga 31 Desember Tahun 2012. Instrumen dalam pengumpulan data berupa kuesioner yang memuat berbagai pertanyaan seputar pekerjaan terkini dari para alumni sebagai responden.

Responden yang menjadi subjek dalam penelitian ini sebanyak 637 orang. Karena di dalam data tersebut terdapat data tersensor kiri, maka tidak semua distribusi *survival* dapat digunakan. Oleh karena itu, hanya tiga distribusi saja yang digunakan dalam penelitian ini, yaitu distribusi normal, distribusi logistik, dan distribusi SEV. Untuk menguji kesesuaian data dengan ketiga distribusi tersebut, digunakan metode *probability plotting*, sedangkan uji Anderson Darling (AD) digunakan untuk mendapatkan distribusi *survival* yang terbaik dalam merepresentasikan data.

4. Hasil dan Pembahasan

Deskripsi Data

Dalam penelitian ini, data waktu tunggu kerja pertama alumni Unsri berupa interval waktu yang satuannya dalam bulan, dimana waktu *survival* tersebut dihitung sejak alumni tersebut diwisuda hingga mendapatkan pekerjaan pertama. Terdapat 8 interval waktu, termasuk satu interval untuk alumni yang telah mendapatkan pekerjaan pertama sebelum diwisuda (interval ke-1), yang dalam hal ini data tersebut dikategorikan sebagai data tersensor kiri (tabel 1).

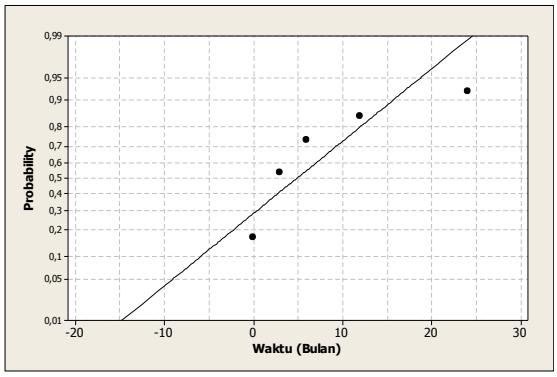
Tabel 1. Waktu Tunggu Mendapatkan Pekerjaan Pertama Alumni Unsri

Interval ke-	Interval Waktu (Dalam Bulan)	Jumlah Responden
1	$(-\infty, 0]$	106
2	$(0,3]$	222
3	$(0, \infty)$	31
4	$(3,6]$	117
5	$(6,12]$	69
6	$(12, \infty)$	13
7	$(12,24]$	39
8	$(24, \infty)$	40
	Total	637

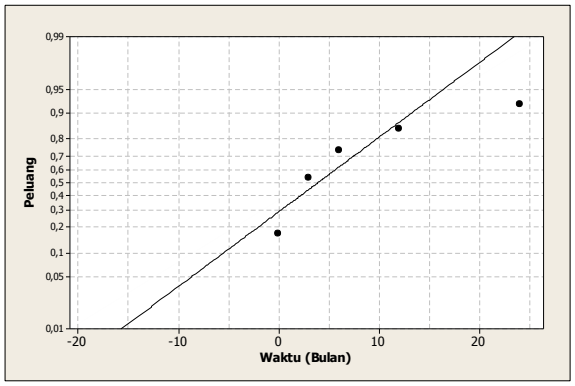
Pencocokan Distribusi Menggunakan Metode Grafik

Metode grafik dapat digunakan untuk mengetahui bagaimana kesesuaian data dengan suatu distribusi *survival*. Terdapat dua jenis metode grafik, yaitu *probability plotting* dan *hazard plotting*. *Probability plotting* dilakukan dengan membuat plot

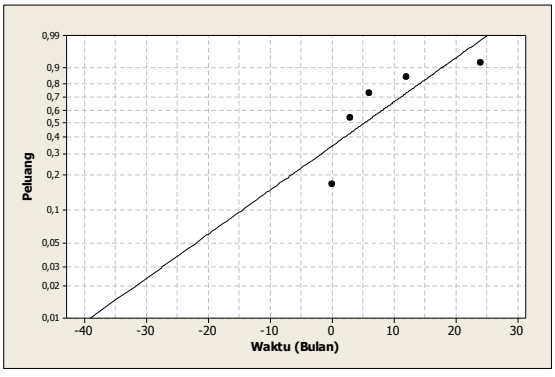
antara waktu (sebagai sumbu-x) terhadap nilai estimasi fungsi distribusi kumulatif (sebagai sumbu-y), sedangkan *hazard plotting* dilakukan dengan membuat plot antara waktu sebagai sumbu-x dengan fungsi *hazard* sebagai sumbu-y. Cara pengambilan kesimpulan kedua metode tersebut adalah sama, yaitu jika plot yang terbentuk berada disekitar suatu garis lurus maka dapat disimpulkan bahwa data mengikuti distribusi *survival* tersebut. Dalam penelitian ini, metode grafik yang digunakan hanya salah satu dari kedua metode tersebut, yaitu metode *probability plotting*.



Gambar 1. Plot Probabilitas Distribusi Normal



Gambar 2. Plot Probabilitas Distribusi Logistik



Gambar 3. Plot Probabilitas Distribusi SEV

Gambar 1, gambar 2, dan gambar 3 berturut-turut adalah hasil plot dari distribusi normal, distribusi logistik, dan distribusi SEV. Berdasarkan plot yang ditampilkan oleh ketiga gambar tersebut, terlihat bahwa plot dari estimasi fungsi distribusi kumulatif terhadap waktu (dalam bulan) berada di sekitar garis lurus, sehingga dapat disimpulkan bahwa data waktu tunggu kerja pertama alumni Unsri yang diperoleh dari hasil *tracer study* tahun 2012 mengikuti distribusi normal, distribusi logistik, dan distribusi SEV.

Uji Anderson Darling (AD)

Salah satu uji *goodness of fit* yang biasa digunakan adalah uji Anderson Darling (AD). Uji AD digunakan untuk menguji apakah data mengikuti suatu distribusi tertentu dengan hipotesis awal dan hipotesis alternatif yang berbentuk

H_0: Data mengikuti suatu distribusi tertentu
H_a: Data tidak mengikuti suatu distribusi tertentu,

jika nilai *p-value* untuk uji AD lebih kecil dari pada taraf signifikansi (biasanya 0,05 atau 0,1) maka hipotesis awal (H_0) ditolak, dengan kata lain data tersebut tidak mengikuti suatu distribusi tertentu. Dalam prakteknya, nilai *p-value* dari uji AD tidak selalu dapat dihitung karena untuk beberapa kasus secara matematis nilainya tidak ada.

Skor AD juga dapat digunakan untuk menentukan distribusi *survival* mana yang paling baik dalam merepresentasikan data. Distribusi *survival* yang terbaik adalah distribusi dengan nilai skor AD yang paling kecil. Adapun, bentuk umum dari statistik uji AD adalah

$$A^2 = -n - S, \tag{14}$$

dengan

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} [\ln F(Y_i) + \ln(1 - F(Y_{n+1-i}))], \tag{15}$$

dalam hal ini n adalah banyaknya data, $i = 1,2, \dots, n$, dan $F(Y_i)$ adalah fungsi distribusi kumulatif untuk data Y_i, dengan $Y_i \in \{Y_1, Y_2, \dots, Y_n\}$ dan $Y_1 \leq Y_2 \leq \cdots \leq Y_n$.

Hasil uji AD terhadap data pada tabel 1 menghasilkan skor AD dari setiap distribusi *survival* yang diperiksa. Perhitungan skor-skor AD tersebut dibantu oleh *software* Minitab 16, yang dalam kasus ini tidak diperoleh nilai *p-value*. Dalam tabel 2, terlihat bahwa skor terkecil adalah skor dari distribusi normal (sebesar 1,237), kemudian berturut-turut disusul oleh distribusi logistik (sebesar 1,312) dan distribusi SEV (sebesar (1,37). Hal ini berarti bahwa distribusi yang paling sesuai dengan data adalah distribusi normal, yang berturut-turut diikuti oleh distribusi logistik dan distribusi SEV.

Tabel 2. Hasil Uji Anderson Darling

Distribusi	Skor Anderson Darling
Normal	1,237
Logistik	1,312
Smallest Extreme Values (SEV)	1,37

Estimasi Model Survival

Langkah pertama yang dilakukan dalam mengestimasi model *survival* parametrik adalah mengestimasi nilai-nilai parameter dari setiap distribusi. Metode estimasi yang digunakan dalam penelitian ini adalah metode *Maximum Likelihood Estimation*. Hasil dari estimasi parameter dari distribusi normal, logistik, dan SEV diberikan dalam tabel 3. Selanjutnya, nilai-nilai estimasi parameter tersebut digunakan untuk mengestimasi fungsi *survival* dan fungsi *hazard*, yang hasilnya ditampilkan dalam tabel 4. Langkah terakhir adalah mengestimasi kurva *hazard* dan kurva *survival* dari ketiga distribusi tersebut yang hasilnya ditampilkan oleh gambar 1 dan gambar 2. Dalam penelitian ini, perhitungan dan penggambaran grafik dari semua estimasi tersebut dibantu oleh *software* Minitab 16.

Tabel 3. Hasil Estimasi Parameter

Distribusi	**Nilai Estimasi Parameter**	
	$\widehat{\boldsymbol{\mu}}$	$\widehat{\boldsymbol{\sigma}}$
Normal	4,89	8,44
Logistik	3,91	4,26
Smallest Extreme Value	9,11	10,47

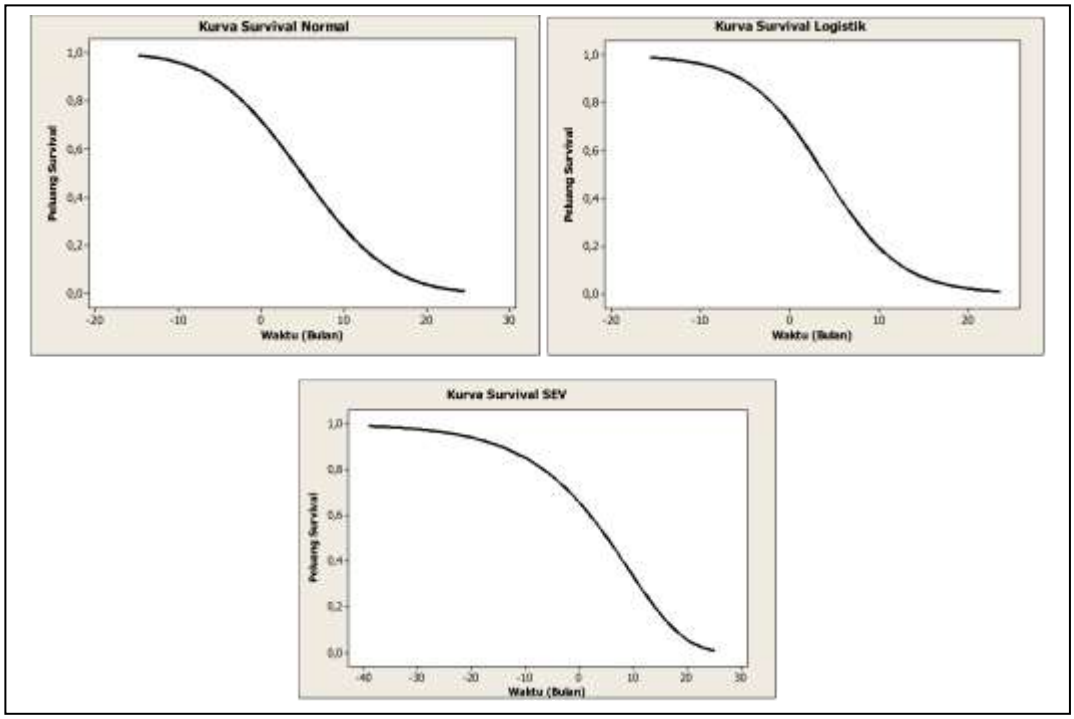
Tabel 4. Hasil Estimasi Fungsi Survival dan Fungsi Hazard

Bulan (t)	**Estimasi Fungsi Survival** $\left(\widehat{\boldsymbol{S}}(\boldsymbol{t})\right)$			**Estimasi Fungsi Hazard** $\left(\widehat{\boldsymbol{h}}(\boldsymbol{t})\right)$		
	Normal	**Logistik**	**SEV**	**Normal**	**Logistik**	**SEV**
-12	0,977	0,977	0,875	0,007	0,005	0,013
-6	0,902	0,911	0,790	0,023	0,021	0,023
-3	0,825	0,835	0,730	0,037	0,039	0,030
0	0,719	0,715	0,658	0,056	0,067	0,040
3	0,589	0,553	0,573	0,078	0,105	0,054
6	0,448	0,379	0,476	0,105	0,146	0,071
12	0,200	0,130	0,268	0,166	0,204	0,126
24	0,012	0,009	0,016	0,309	0,233	0,396

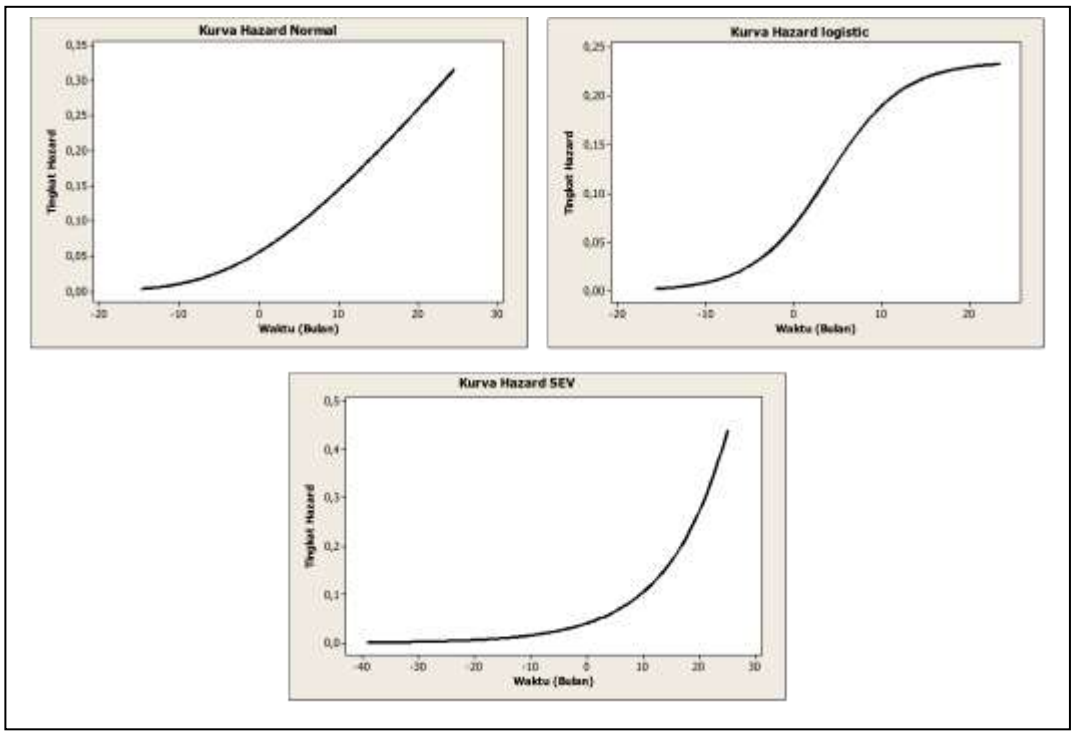
Dalam tabel 4, diperlihatkan hasil estimasi dari fungsi *survival* $\hat{S}(t)$ dan fungsi *hazard* $\hat{h}(t)$ untuk beberapa nilai t. Fungsi *survival* diartikan sebagai besarnya peluang seorang alumni Unsri mendapatkan pekerjaan pertamanya lebih dari waktu t. Sebagai contoh, nilai $\hat{S}(3)$ untuk distribusi normal adalah 0,589 yang artinya peluang seorang alumni Unsri mendapatkan pekerjaan pertamanya lebih dari bulan ke-3 adalah sebesar 0,589.

Sementara itu, fungsi *hazard* memiliki makna sebagai besarnya peluang seorang alumni Unsri mendapatkan pekerjaan pertamanya pada waktu t. Contohnya, nilai $\hat{h}(t)$ untuk distribusi SEV adalah 0,126 yang artinya peluang seorang alumni mendapatkan pekerjaan pertama pada bulan ke-12 setelah diwisuda adalah sebesar 0,126.

Terlihat juga bahwa peluang terendah dan peluang tertinggi seorang alumni mendapatkan pekerjaan pertama berdasarkan ketiga distribusi tersebut terletak pada waktu yang sama, dimana peluang terendah terjadi ketika alumni masih kuliah (sebelum lulus), sedangkan peluang tertinggi terjadi pada saat bulan ke-24 (dua tahun) setelah lulus.



Gambar 4. Estimasi Kurva Survival



Gambar 5. Estimasi Kurva Hazard

Dalam gambar 4, diperlihatkan estimasi kurva *survival* dari data waktu mendapatkan pekerjaan pertama alumni Unsri untuk setiap distribusi. Kurva *survival* dapat merepresentasikan tren dari nilai-nilai fungsi *survival* sepanjang waktu. Terlihat bahwa estimasi kurva *survival* dari ketiga distribusi *survival* (normal, logistik, dan SEV) memiliki tren yang hampir sama. Sedangkan, tren dari nilai-nilai estimasi fungsi *hazard* sepanjang waktu direpresentasikan oleh estimasi kurva *hazard* (gambar 5). Walaupun semua estimasi kurva *hazard* dari setiap distribusi dalam gambar 5 merupakan fungsi naik, akan tetapi tren kenaikan dari setiap distribusi berbeda-beda.

5. Kesimpulan

Menggunakan metode *probability plotting*, dapat ditunjukkan bahwa ketiga distribusi yang diperiksa (yaitu distribusi normal, distribusi logistik, dan distribusi SEV) sesuai dengan data waktu mendapatkan pekerjaan pertama alumni Unsri. Selanjutnya, berdasarkan uji Anderson Darling dapat disimpulkan juga bahwa distribusi yang terbaik dalam merepresentasikan data tersebut adalah distribusi normal, yang berturut-turut diikuti oleh distribusi logistik dan distribusi SEV. Berdasarkan estimasi yang telah dilakukan, peluang terendah seorang alumni Unsri mendapatkan pekerjaan

pertama adalah ketika masih kuliah, sedangkan peluang tertinggi terjadi pada saat dua tahun setelah kelulusan.

Ucapan Terimakasih

Terimakasih sebesar-besarnya diucapkan kepada Lembaga Penelitian (Lemlit) Universitas Sriwijaya, yang telah membiayai penelitian ini melalui skim penelitian Sains dan Teknologi untuk tahun anggaran 2015, sehingga penelitian ini dapat berjalan dengan baik dan diselesaikan tepat pada waktunya.

Daftar Pustaka

[1] M Akram, A. M Ullah, & R Taj. 2007. *Survival* Analysis of Cancer Patients Using Parametric and Non-Parametric Approaches. *Pakistan Vet.J*., 27(4), 194-198.

[2] Faruk, Alfensi. 2014. Estimasi Parameter Data Tersensor Tipe I Berdistribusi Log-Logistik Menggunakan Maximum Likelihood Estimate dan Iterasi Newton-Rhapson. *Prosiding Seminar Nasional MIPA*, 21-25.

[3] Faruk, Alfensi. 2015. Analisis Data Tersensor Interval Dalam Pemodelan Waktu Mendapatkan Pekerjaan Pertama Alumni Universitas Sriwijaya. *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UNY 2015*, 123-130.

[4] A. E Hayat, A Suner, B Uyar., O Dursun, N. M Orman, & G Kitapcioglu. 2010. Comparison of Five Survival Models: Breast Cancer Registry Data from Ege University Cancer Research Center. *Turkiye Klinikleri J Med Sci*, 30(5), 1665-1674.

[5] E. L Kaplan & P Meier. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53, 457-481.

[6] D. G Kleinbaum & M Klein. 2005. *Survival Analysis A Self-Learning Text Second Edition*. New York: Springer

[7] E. T Lee & J. W Wang. 2003. *Statistical Methods for Survival Data Analysis Third Edition*. New Jersey: John Wiley & Sons.

[8] Sun, Jianguo. 2006. *The Statistical Analysis of Interval-censored Failure Time Data*. New York: Springer.