

BAB II

TINJAUAN PUSTAKA

2.1 Variabel Acak Diskrit

Variabel acak diskrit adalah variabel acak yang tidak mengambil seluruh nilai yang ada dalam sebuah interval atau variabel hanya memiliki nilai tertentu. Variabel acak diskrit merupakan variabel acak yang nilainya dihasilkan dari sebuah perhitungan (pencacahan). Nilainya berupa bilangan bulat dan asli, tidak berbentuk pecahan. Distribusi Bernoulli, distribusi binomial, dan distribusi multinomial termasuk kedalam distribusi peluang diskrit.

2.1.1 Distribusi Bernoulli

Menurut Hajarisman (2009), distribusi Bernoulli adalah distribusi yang bersumber dari percobaan Bernoulli. Percobaan Bernoulli adalah percobaan yang menghasilkan dua kemungkinan hasil, yaitu “sukses” dan “gagal”. Contohnya adalah pelemparan satu buah mata uang logam, dimana terdapat 2 kemungkinan hasil yang bisa diperoleh dari satu kali pelemparan, yaitu “angka” dan “gambar”. Misalkan munculnya “angka” dianggap kejadian yang “sukses” dimana peluang munculnya adalah π dan munculnya “gambar” dianggap kejadian yang “gagal” di mana peluang munculnya adalah $1 - \pi$. Selanjutnya, variabel acak “T” terkait percobaan tersebut diberi nilai 1 dengan peluang π jika “sukses” terjadi dan diberi nilai 0 jika “gagal” terjadi dengan peluang $1 - \pi$. Dengan demikian, variabel acak “T” dikatakan berdistribusi Bernoulli.

Berikut ini adalah fungsi peluang dari distribusi Bernoulli:

$$p(t) = \begin{cases} \pi^t(1 - \pi)^{t-1}; & t = 0,1 \\ 0 & ; t \text{ yang lainnya} \end{cases} \quad \dots(2.1)$$

Rata-rata dan varians dari suatu variabel acak yang mengikuti distribusi Bernoulli adalah $E(T) = \pi$ dan $V(T) = \pi(1 - \pi)$.

2.1.2 Distribusi Binomial

Menurut Hajarisman (2009), banyak aplikasi statistika dilakukan pada sejumlah m buah observasi yang tetap (*fixed*). Misalkan t_1, t_2, \dots, t_m menyatakan respon untuk buah percobaan yang identik dan saling bebas sedemikian rupa sehingga: $P(T_i = 1) = \pi$ dan $P(T_i = 0) = 1 - \pi$

Di sini digunakan label yang sangat umum untuk menyatakan angka “1” sebagai peristiwa “sukses” dan “0” untuk menyatakan peristiwa “gagal”. Percobaan yang identik mempunyai makna bahwa peluang sukses π adalah sama untuk setiap percobaan. Sedangkan percobaan yang saling bebas mempunyai makna bahwa $\{T_i\}$ merupakan variabel acak yang saling bebas. Hal ini seringkali disebut sebagai percobaan Bernoulli. Total banyaknya peristiwa sukses, $T = \sum_{i=1}^m T_i$, mempunyai distribusi binomial dengan indeks m dan parameter π , yang dapat disingkat sebagai $T \sim \text{binomial}(m, \pi)$.

Fungsi peluang untuk variabel acak t_1, t_2, \dots, t_m yang berdistribusi binomial sebagai berikut:

$$p(t) = \binom{m}{t} \pi^t (1 - \pi)^{m-t}, \text{ untuk } t = 0, 1, 2, \dots, m \quad \dots(2.2)$$

Dimana koefisien binomial $\binom{m}{t} = \frac{m!}{t!(m-t)!}$

Rata-rata dan varians dari suatu variabel acak yang mengikuti distribusi binomial adalah $E(T) = m\pi$ dan $\text{var}(T) = m\pi(1 - \pi)$.

2.2 Distribusi Multinomial

Distribusi multinomial adalah perluasan dari distribusi binomial. Distribusi binomial muncul sebagai distribusi banyaknya kejadian "sukses" dalam uji coba saling bebas Bernoulli, dengan peluang sukses (π) yang sama dalam setiap percobaan. $T \sim \text{Binomial}(m, \pi)$ menyatakan jumlah T didistribusikan sebagai distribusi binomial

dengan banyaknya percobaan (m) dan peluang sukses (π). Distribusi multinomial muncul sebagai lanjutan dasar dari distribusi binomial ketika setiap percobaan saling bebas dan memiliki lebih dari dua kemungkinan hasil yang bersifat mutual eksklusif. Memperhatikan keseluruhan pengulangan dari banyaknya percobaan (m) saling bebas dari sebuah eksperimen, maka masing-masing akan menghasilkan suatu kejadian hingga $(d + 1)$ kejadian yang bersifat mutual eksklusif. Dalam setiap pengulangan percobaan, peluang kejadian E_i sama dengan banyaknya peluang sukses $\sum_{i=1}^{d+1} \pi_i$. Misalkan $\mathbf{T} = (T_1, T_2, \dots, T_{d+1})^t$ menunjukkan vektor acak dari banyaknya jumlah kejadian yang muncul E_1, E_2, \dots, E_{d+1} hingga banyaknya percobaan (m), maka $\sum_{i=1}^{d+1} T_i = m$. Jika $\mathbf{t} = (t_1, t_2, \dots, t_{d+1})^t$ menyatakan \mathbf{T} , $\sum_{i=1}^{d+1} t_i = m$. Sehingga, vektor acak \mathbf{T} memiliki distribusi multinomial dengan parameter $(m, \boldsymbol{\pi})$, dimana $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_{d+1})^t$. Gabungan fungsi peluang dari \mathbf{T} adalah:

$$P(\mathbf{T} = \mathbf{t}) = \frac{m!}{t_1! t_2! \dots t_{d+1}!} \prod_{i=1}^{d+1} \pi_i^{t_i} \quad \dots(2.3)$$

Kejadian $d = 1$, memiliki dua hasil kejadian yang bersifat mutual eksklusif untuk setiap percobaan, sesuai dengan distribusi binomial. Sehingga memiliki vektor acak dua dimensi $\mathbf{T} = (T, m - T)^t$, dimana $m - T$ sebagai distribusi binomial $(1 - \pi, \pi)$. Karena m diketahui, variabel bebas $m - T$ tidak memberikan informasi tambahan, maka dari itu angka satu (dalam distribusi binomial) dapat mengurangi dimensi \mathbf{T} . Sedemikian rupa sehingga $\sum_{i=1}^{d+1} T_i = m$ dan $\sum_{i=1}^{d+1} \pi_i = 1$, maka dapat mengurangi dimensi \mathbf{T} dan $\boldsymbol{\pi}$ dengan menghilangkan kategori terakhir. Kemudian tentukan $\mathbf{T} = (T_1, T_2, \dots, T_{d+1})^t$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_d)^t$, dan dinyatakan sebagai \mathbf{T} dengan $\mathbf{t} = (t_1, t_2, \dots, t_d)^t$. Dengan demikian, tanpa memperhatikan rumus umum, dapat dinyatakan bahwa \mathbf{T} memiliki distribusi multinomial dengan parameter $(m, \boldsymbol{\pi})$, dengan gabungan fungsi peluang pada persamaan (2.3) dimana $t_{d+1} = m - \sum_{i=1}^d t_i$

dan $\pi_{d+1} = 1 - \sum_{i=1}^d \pi_i$. Kemudian diterapkan pada distribusi multinomial, menggunakan \mathbf{T} , \mathbf{t} , dan $\boldsymbol{\pi}$ tanpa melibatkan kategori terakhir. $\mathbf{T} \sim \text{multinomial}(m, \boldsymbol{\pi})$ menunjukkan bahwa jumlah vektor \mathbf{T} berdistribusi multinomial dengan banyaknya percobaan (m) dan peluang sukses ($\boldsymbol{\pi}$). Rata-rata dan varians dari \mathbf{T} , sebagai berikut:

$$E(\mathbf{T}) = m \boldsymbol{\pi} \quad \dots(2.4)$$

$$\text{Var}(\mathbf{T}) = m \Delta(\boldsymbol{\pi}) \quad \dots(2.5)$$

dimana $\Delta(\boldsymbol{\pi}) = \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}^t$. Perlu diingat bahwa $\Delta(\boldsymbol{\pi})$ adalah sebuah $d \times d$ matriks varians kovarians dimana diagonal elemen i nya adalah $\pi_i(1 - \pi_i)$ dan elemen *off*-diagonalnya adalah $-\pi_i \pi_j$ dimana $i \neq j$.

2.3 Penaksiran Parameter Distribusi Multinomial

Fungsi peluang untuk variabel acak t_1, t_2, \dots, t_{d+1} yang mengikuti distribusi multinomial adalah sebagai berikut:

$$f(t_1, t_2, \dots, t_{d+1} | \pi_1, \pi_2, \dots, \pi_{d+1}) = \frac{n!}{\prod_{i=1}^{d+1} t_i!} \prod_{i=1}^{d+1} \pi_i^{t_i} \quad \dots(2.6)$$

Fungsi *log-likelihood* untuk fungsi peluang dari variabel acak yang berdistribusi multinomial adalah

$$l(\pi_1, \pi_2, \dots, \pi_{d+1}) = \log n! - \sum_{i=1}^{d+1} \log t_i! + \sum_{i=1}^{d+1} t_i \log \pi_i \quad \dots(2.7)$$

Berdasarkan fungsi *log-likelihood* diatas, maka penaksir kemungkinan maksimum untuk parameter $\pi_1, \pi_2, \dots, \pi_{d+1}$ adalah sebagai berikut:

$$\pi_i = \frac{t_i}{n} = \frac{t_1}{n}, \frac{t_2}{n}, \dots, \frac{t_{d+1}}{n} \quad \dots(2.8)$$

Hasil turunan penaksir kemungkinan maksimum selengkapnya dilampirkan pada lampiran 1.

2.4 Uji Kecocokan Distribusi

Salah satu cara alternatif untuk mengukur kecocokan model chi-kuadrat Pearson dengan hipotesis pengujian adalah

H_0 = Data cocok terhadap distribusi multinomial.

H_1 = Data tidak cocok terhadap distribusi multinomial

dan statistik uji yang didefinisikan sebagai berikut:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}, i = 1, 2, \dots, n \quad \dots(2.9)$$

dimana O_i adalah nilai observasi ke i dan E_i adalah frekuensi harapan dari observasi ke i . Adapun derajat bebas untuk statistik chi-kuadrat Pearson adalah $n - p - 1$, dimana p adalah banyaknya parameter yang ditaksir dari distribusi. dengan kriteria uji Tolak H_0 jika $\chi_{hitung}^2 > \chi_{tabel}^2$.

2.5 Overdispersi

Mempertimbangkan suatu studi di mana unit eksperimennya adalah kluster dan setiap unit elemen dalam kluster dikelompokkan menjadi lebih dari dua kategori yang saling bebas, model yang mengikuti distribusi multinomial diasumsikan bahwa satuan unsurnya saling bebas dengan peluang yang sama. Pengklasteran menyebabkan adanya korelasi antar elemen. Peluang sukses dapat bervariasi dari satu unit elemen ke elemen lainnya dalam kluster. Akibatnya, data biasanya menunjukkan varians yang lebih besar daripada varians yang seharusnya dari model multinomial. Fenomena seperti ini dikenal sebagai overdispersi atau variasi ekstra. Sehingga istilah overdispersi ini dapat diartikan bahwa variasi dari respon T melebihi varians multinomial, $m \Delta(\boldsymbol{\pi})$.

Menurut McCullagh & Nelder, 1989 cara sederhana untuk mendeteksi adanya overdispersi dapat dilihat melalui rasio antara nilai devians atau nilai chi-kuadrat

Pearson terhadap derajat bebasnya, apabila rasio antara nilai devians atau chi-kuadrat Pearson dengan derajat bebas lebih dari 1 (satu) maka dalam kasus tersebut terindikasi adanya overdispersi, mungkin dalam mekanisme paling umum adalah adanya pengelompokan dalam populasi.

2.6 Distribusi *Random-Clumped* Multinomial

Model multinomial overdispersi kedua yang akan dibahas diusulkan oleh Morel & Nagaraj (1993) dan Neerchal & Morel (1998). Ini adalah perluasan multivariat dari distribusi *random-clumped* Binomial. Model ini dapat diturunkan sebagai berikut:

Misalkan Y, Y_1^0, \dots, Y_m^0 saling bebas dan secara identik menyebarkan variabel acak multinomial dengan parameter $(\boldsymbol{\pi}; 1)$. Dan jika U_1, U_2, \dots, U_m variable acak Uniform (0,1) dan $I(\cdot)$ menunjukkan fungsi indikator. Untuk setiap $i, i = 1, 2, \dots, m$, tentukan nilai Y_i sebagai Y dengan peluang (ρ) , Y_i^0 dengan peluang $(1 - \rho)$. Untuk $p, 0 < \rho < 1$, dapat dijelaskan sebagai berikut:

$$Y_i = Y \text{ jika } U_i \leq \rho, i = 1, 2, \dots, m$$

$$Y_i = Y_i^0 \text{ jika } U_i > \rho, i = 1, 2, \dots, m$$

Dengan kata lain, setiap Y_i dapat direpresentasikan sebagai:

$$Y_i = Y I(U_i \leq \rho) + Y_i^0 I(U_i > \rho), i = 1, 2, \dots, m \quad \dots(2.10)$$

Dalam model untuk data berkelompok ini, respon dari setiap anggota klaster sama seperti Y , dengan peluang ρ . Juga, untuk $i \neq j$ dapat ditunjukkan bahwa $P_r(Y_i = 1 | Y_j = 1) = \rho^2$. Artinya peluang Y_i sukses dengan syarat Y_j juga sukses adalah ρ^2 . Demikian pula, ternyata bahwa korelasi $(Y_i, Y_j) = \rho^2$.

Jika $T = \sum_{i=1}^m Y_i$, dan Y_i didefinisikan sebagai persamaan (2.14), maka rata-rata dan matriks kovarians dari T berikut:

$$E(T) = m\boldsymbol{\pi} \quad \dots(2.11)$$

$$V(\mathbf{T}) = m \{ \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^t \} \{ 1 + \rho^2(m-1) \}. \quad \dots(2.12)$$

Oleh karena itu, variabel T yang berasal dari model campuran pada persamaan (2.14) menunjukkan adanya overdispersi terhadap distribusi multinomial. Kemudian, dengan $\mathbf{T} = \sum_{i=1}^m \mathbf{Y}_i$ maka:

$$\mathbf{Y}_i = \mathbf{Y} \sum_{i=1}^m I(U_i \leq \rho) + \mathbf{Y}_i^0 \sum_{i=1}^m I(U_i > \rho), i = 1, 2, \dots, m \quad \dots(2.13)$$

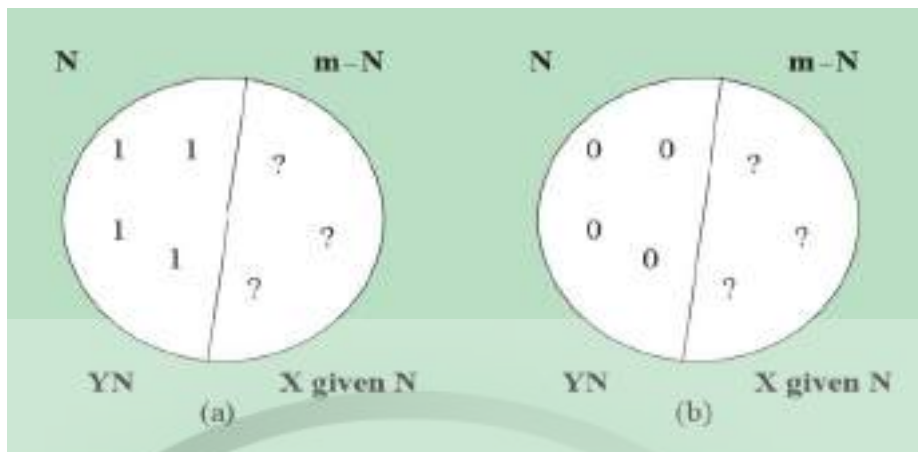
Hal ini menyebabkan representasi dari \mathbf{T} sebagai berikut:

$$\mathbf{T} = (\mathbf{Y}N) + (\mathbf{X}|N) \quad \dots(2.14)$$

dimana $N \sim \text{binomial}(\rho, m)$, $\mathbf{Y} \sim \text{multinomial}(\boldsymbol{\pi}; 1)$, N dan \mathbf{Y} saling bebas, dan $(\mathbf{X}|N) \sim \text{multinomial}(\boldsymbol{\pi}; m - N)$, jika $N < m$.

Persamaan (2.14) mengilustrasikan bagaimana terjadinya keragaman ekstra yang disebabkan oleh pengelompokan dalam sampling. Variabel acak cacahan yang dinyatakan oleh N ditambahkan kedalam kategori \mathbf{Y} vektor cacahan dalam persamaan (2.14) mempunyai dua bagian. Pada bagian pertama yang dinotasikan oleh $\mathbf{Y}N$, menghasilkan respon \mathbf{Y} secara berulang sebanyak N kali. Hal ini mencerminkan bahwa sampling kluster dimana beberapa respon didalam kluster adalah mirip. Sedangkan pada bagian kedua yang dinyatakan dalam bentuk $\mathbf{X}|N$ terdiri dari $m - N$ respon yang saling bebas.

Gambar 2.1 mengilustrasikan distribusi *random-clumped* multinomial akan dengan sebuah kluster $m = 7$ dengan respon biner. Misalkan Kluster mempunyai dua buah sub kelompok. Sub kelompok pertama berukuran 4 ($N; 0 \leq N \leq 7$), berisi individu-individu yang sangat dipengaruhi satu sama lainnya dan menghasilkan respon indentik. Kemudian tiga individu sisanya pada sub kelompok kedua memberikan respon yang saling bebas dan sub kelompok kedua saling bebas dengan sub kelompok pertama. Respon yang dinotasikan dengan “?” ini mengindikasikan bahwa responnya bisa jadi “0” atau “1”.



Gambar 2.1 Dua kemungkinan hasil untuk sekelompok ukuran $m = 7$ di bawah distribusi *random-clumped* multinomial

Pada Gambar 2.1 terlihat bahwa beberapa unit elemen memberikan tanggapan yang sama, baik di “1” pada (a) atau “0” pada (b). Simbol (?) menunjukkan unit unsur/elemen merespon secara independen.

Dapat ditunjukkan bahwa vektor acak \mathbf{T} menunjukkan pada persamaan (2.14) memiliki rata-rata dan varians sama seperti persamaan (2.11) dan (2.12). Akibatnya persamaan (2.14) itu memberikan distribusi alternatif untuk memodelkan keragaman ekstra atau overdispersi. Hal ini menunjukkan bahwa fungsi peluang dari vektor acak (2.14) dapat dituliskan sebagai berikut:

$$P_{RCM}(\mathbf{t}; \boldsymbol{\pi}, \rho) = \sum_{i=1}^k \pi_i P(\mathbf{t}; \mathbf{x}_i, m) \quad \dots(2.15)$$

dimana $P(\mathbf{t}; \mathbf{x}_i, m)$ menunjukkan fungsi peluang berdistribusi multinomial $M_k(\mathbf{x}_i; m)$; $\mathbf{x}_i = (1 - \rho)\boldsymbol{\pi} + \rho \mathbf{e}_i$ untuk $i = 1, 2, \dots, d + 1$, $\mathbf{x}_k = (1 - \rho)\boldsymbol{\pi}$; dan \mathbf{e}_i adalah kolom ke i dari matriks identitas dimensi $(d + 1)$.

Pada persamaan (2.15) merupakan fungsi peluang dari campuran dari distribusi multinomial. Kemudian, distribusi campuran merupakan distribusi marginal \mathbf{T} , dimana $\mathbf{T}|\mathbf{X} \sim \text{multinomial}(\mathbf{X}; m)$ dan \mathbf{X} merupakan distribusi diskrit dengan sampel acak $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d$ yang dihubungkan oleh peluang kejadian π_i . Karena distribusi Dirichlet multinomial merupakan distribusi campuran dari variabel acak multinomial,

dapat dilihat bahwa dua distribusi tersebut menggambarkan hasil overdispersi dari peubah acak X .

2.7 Penaksir Parameter Distribusi *Random-Clumped* Multinomial

Parameter dari distribusi *random-clumped* multinomial ditaksir dengan menggunakan metode *Fisher scoring*. Misal $l(\boldsymbol{\theta}) = \sum_{j=1}^n \ln \{P(\mathbf{t}_j; \boldsymbol{\theta})\}$ merupakan fungsi kemungkinan, dimana $\boldsymbol{\theta} = (\boldsymbol{\pi}^t, \rho)^t$. Berikut merupakan algoritma *Fisher scoring*:

$$\hat{\boldsymbol{\theta}}_{(l+1)} = \hat{\boldsymbol{\theta}}_{(l)} + \{n I(\hat{\boldsymbol{\theta}}_{(l)})\}^{-1} \left\{ \frac{\partial L(\hat{\boldsymbol{\theta}}_{(l)})}{\partial \hat{\boldsymbol{\theta}}_{(l)}} \right\}, l = 0, 1, 2 \dots \infty \quad \dots(2.13)$$

dimana $\hat{\boldsymbol{\theta}}_{(l)}$ dan $\hat{\boldsymbol{\theta}}_{(l+1)}$ masing-masing menunjukkan nilai taksiran parameter yang dihasilkan pada iterasi ke- l dan $(l + 1)$, dan $I(\hat{\boldsymbol{\theta}}_{(l)})$ adalah matriks informasi *Fisher* untuk pengamatan tunggal yang dievaluasi pada iterasi ke- l .

Besaran $I(\hat{\boldsymbol{\theta}}_{(l)})$ merupakan matriks informasi *Fisher* berukuran $p \times p$ yang dievaluasi pada iterasi ke- l . Misal $\boldsymbol{\theta} = (\pi_1, \pi_2, \dots, \pi_{k-1}, \rho)^t$, sehingga diperoleh matriks informasi sebagai berikut:

$$I_{RDM}(\hat{\boldsymbol{\theta}}_{(l)}) = \begin{pmatrix} I_{RDM}(\pi_1, \pi_1) & I_{RDM}(\pi_1, \pi_2) & \dots & I_{RDM}(\pi_1, \pi_{k-1}) & I_{RDM}(\pi_1, \rho) \\ I_{RDM}(\pi_2, \pi_1) & I_{RDM}(\pi_2, \pi_2) & \dots & I_{RDM}(\pi_2, \pi_{k-1}) & I_{RDM}(\pi_2, \rho) \\ \vdots & \vdots & \dots & \vdots & \vdots \\ I_{RDM}(\pi_{k-1}, \pi_1) & I_{RDM}(\pi_{k-1}, \pi_2) & \dots & I_{RDM}(\pi_{k-1}, \pi_{k-1}) & I_{RDM}(\pi_{k-1}, \rho) \\ I_{RDM}(\rho, \pi_1) & I_{RDM}(\rho, \pi_2) & \dots & I_{RDM}(\rho, \pi_{k-1}) & I_{RDM}(\rho, \rho) \end{pmatrix} \quad \dots(2.14)$$

Nilai awal untuk memulai iterasi pada parameter π diperoleh dengan metode momen pada persamaan sebagai berikut:

$$\hat{\boldsymbol{\pi}}_{(0)} = \frac{1}{nm} \sum_{j=1}^n \mathbf{t}_j \quad \dots(2.15)$$

dimana \mathbf{t}_j merupakan vektor yang terdiri dari komponen pertama $(k - 1)$ dari vektor \mathbf{t}_j dan nilai awal untuk memulai iterasi pada parameter ρ diperoleh persamaan sebagai berikut:

$$\hat{\rho}_{(0)} = \frac{\left(\text{Diag}\{\sum_{j=1}^n \mathbf{t}_j - m \hat{\pi}_{(0)}\} (\mathbf{t}_j - m \hat{\pi}_{(0)}) \times [\text{Diag}\{\text{Diag}(\hat{\pi}_{(0)}) - \hat{\pi}_{(0)}(\hat{\pi}_{(0)})^t\}]^{-1} \right)}{\{m(n-1)(k-1) = \{1 + \rho^2(m-1)\}} \quad \dots(2.16)$$

Menurut Nagaraj dkk (2005) proses iterasi akan konvergen ketika nilai taksiran yang diperoleh sudah memenuhi kriteria sebagai berikut:

$$\frac{L(\hat{\theta}_{[l,s(l)]}) - L(\hat{\theta}_{[l-1,s(l-1)]})}{|L(\hat{\theta}_{[l,s(l)]})| + 10^{-6}} < \varepsilon, \text{ dimana } \varepsilon = 10^{-8} \quad \dots(2.17)$$