

Łukasz Kuźmiński, Anna Nikodem

ESTYMACJA PARAMETRÓW ROZKŁADÓW W MIESZANCE DWÓCH ROZKŁADÓW POISSONA

1. Wstęp

Rozkład skończonej mieszanki jednowymiarowych rozkładów jest określony za pomocą funkcji gęstości wyrażającej się następującym wzorem:

$$f(x) = \sum_{j=1}^r \alpha_j f_j(x), \quad (1)$$

przy czym

$$\sum_{j=1}^r \alpha_j = 1, \quad (2)$$

gdzie: r – liczba rozkładów w mieszance,

α_j – parametr udziału j -tego rozkładu w mieszance zwany parametrem mieszającym,

$f_j(x)$ – gęstość j -tego rozkładu.

W przypadku mieszanki dwóch rozkładów Piossona funkcja rozkładu (1) jest liniową kombinacją funkcji rozkładu Poissona z parametrem λ_1 i λ_2 , tj.

$$f(x|\lambda) = \sum_{j=1}^2 \alpha_j f_j(x|\lambda_j) = \alpha_1 \frac{\lambda_1^x}{x!} e^{-\lambda_1} + \alpha_2 \frac{\lambda_2^x}{x!} e^{-\lambda_2}, \quad (3)$$

gdzie λ jest wektorem $[\lambda_1 \ \lambda_2]$, a wagi spełniają warunek $\alpha_1 + \alpha_2 = 1$.

W artykule przedstawione zostaną estymatory parametrów rozkładów Poissona oraz estymatory parametrów mieszających przy użyciu metody największej wiarygodności i metody momentów.

2. Metoda największej wiarygodności

Estymując wektor nieznanych parametrów λ metodą największej wiarygodności, należy znaleźć takie wartości oceny szacowanych parametrów, aby wiarygodność próby była największa. Funkcja wiarygodności dla n -elementowej próby $(x_1, x_2, \dots, x_n)$ jest równa

$$L(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta), \quad (4)$$

przy czym $f(x_i | \theta)$ jest wartością funkcji gęstości dla obserwacji x_i .

Rozważając mieszkankę składającą się z dwóch rozkładów Poissona, otrzymuje się funkcję wiarygodności postaci

$$L(x_1, x_2, \dots, x_n | \lambda) = \prod_{i=1}^n f(x_i | \lambda) = \prod_{i=1}^n \left(\alpha_1 \frac{\lambda_1^{x_i}}{x_i!} e^{-\lambda_1} + \alpha_2 \frac{\lambda_2^{x_i}}{x_i!} e^{-\lambda_2} \right), \quad (5)$$

gdzie logarytm naturalny funkcji (5) jest równy

$$\ln L(x_1, x_2, \dots, x_n | \lambda) = \sum_{i=1}^n \ln f(x_i | \lambda) = \sum_{i=1}^n \ln \left(\alpha_1 \frac{\lambda_1^{x_i}}{x_i!} e^{-\lambda_1} + \alpha_2 \frac{\lambda_2^{x_i}}{x_i!} e^{-\lambda_2} \right). \quad (6)$$

Warunki konieczne istnienia maksimum logarytmu funkcji wiarygodności są następujące:

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^n \hat{p}(j | x_i), \quad (7)$$

$$\sum_{i=1}^n \hat{p}(j | x_i) \frac{\partial \ln f_j(x_i | \hat{\lambda}_j)}{\partial \hat{\lambda}_j} = 0, \quad (8)$$

gdzie $\hat{\alpha}_j$, $\hat{\lambda}_j$ jest estymatorem α_j , λ_j , a prawdopodobieństwo $\hat{p}(j | x_i)$, interpretowane jako prawdopodobieństwo *a posteriori*, że obserwacja x_i pochodzi z j -tego rozkładu mieszanki, wyraża się za pomocą wzoru Bayesa

$$\hat{p}(j | x_i) = \frac{\hat{\alpha}_j f_j(x_i | \hat{\lambda}_j)}{\sum_{l=1}^2 \hat{\alpha}_l f_l(x_i | \hat{\lambda}_l)}. \quad (9)$$

Z warunku (8) wynika, że ocena parametru mieszającego $\hat{\alpha}_j$ jest średnią arytmetyczną prawdopodobieństw *a posteriori* dla j -tego rozkładu. We wzorze (9) zaś do określenia prawdopodobieństw *a posteriori* potrzebna jest znajomość ocen parametrów rozkładów mieszanki. Oceny tych parametrów zależą z kolei od prawdopodobieństw *a posteriori*. Ten problem rozwiązywany jest za pomocą algorytmu iteracyjnego wyznaczania prawdopodobieństw *a posteriori*, który zostanie przedstawiony w dalszej części artykułu.

2.1. Algorytm iteracyjny wyznaczania prawdopodobieństw *a posteriori*

Algorytm rozpoczyna się od wyznaczenia wartości początkowych prawdopodobieństw *a posteriori* $\hat{p}^0(j|x_i)$. Każda r -ta iteracja algorytmu polega na wyznaczeniu kolejno:

1. Ocen parametrów rozkładu przy wykorzystaniu prawdopodobieństwa *a posteriori* otrzymanego w poprzedniej iteracji.

2. Nowych wartości prawdopodobieństw *a posteriori*, liczonych według wzoru (9) oraz oceny parametrów rozkładów mieszanki obliczone w tej iteracji. Prawdopodobieństwa te oznaczają się $\hat{p}^r(j|x_i)$, gdzie r oznacza numer iteracji, a $\hat{p}^r(j|x_i)$ to prawdopodobieństwo *a posteriori*, że obserwacja x_i pochodzi z j -tego rozkładu mieszanki, wyznaczone w r -tej iteracji.

Procedurę iteracyjną zatrzymuje się, gdy prawdopodobieństwa *a posteriori* otrzymane w dwóch kolejnych iteracjach różnią się nieznacznie. Można przyjąć następujący warunek

$$\max_{i,j} \left| \hat{p}^r(j|x_i) - \hat{p}^{r-1}(j|x_i) \right| < \varepsilon, \quad (10)$$

po spełnieniu którego procedurę zatrzymuje się. W warunku (10) ε jest niewielką liczbą dodatnią, równą np. 0,005.

W algorytmie problem wyznaczania wartości początkowych prawdopodobieństw *a posteriori* $\hat{p}^0(j|x_i)$ można rozwiązać, stosując m.in. następujące sposoby:

1) ustalenie prawdopodobieństwa *a posteriori* na podstawie wiedzy merytorycznej badacza,

2) losowe określenie prawdopodobieństw *a posteriori* na podstawie nK liczb losowych o rozkładzie jednostajnym na odcinku $[0; 1)$ według wzoru

$$\hat{p}^0(j|x_i) = \frac{los_{ij}}{\sum_{l=1}^K los_{il}}, \quad (11)$$

gdzie los_{ij} jest liczbą losową o numerze $(i-1)K + j$.

Opisany algorytm wyznaczania prawdopodobieństw *a posteriori* jest zbieżny w skończonej liczbie iteracji.

Przykład 1

Do badania pobrana została próba losowa o liczebności 10, pochodząca z dwóch rozkładów Poissona z parametrami λ_1 i λ_2 . Udział (parametr mieszający) obserwacji w próbie pochodzących z rozkładu pierwszego oznaczony zostanie przez α_1 i odpowiednio – pochodzący z rozkładu drugiego przez α_2 . Parametry te są jednocześnie parametrami udziału obu rozkładów w mieszance.

Obserwacje podane są w tab. 1.

Tabela 1. Wartości obserwacji w próbie

Numer obserwacji	1	2	3	4	5	6	7	8	9	10
x_i	4	3	3	1	2	2	2	16	14	13

Źródło: opracowanie własne.

W tabeli 2 przedstawione są wartości początkowe prawdopodobieństw *a posteriori*, wyznaczone zgodnie z opisanym wcześniej sposobem II.

Tabela 2. Wartości początkowe prawdopodobieństw *a posteriori*

<i>A posteriori</i> dla rozkładu 1	<i>A posteriori</i> dla rozkładu 2
0,459880577	0,540119423
0,568164066	0,431835934
0,346013488	0,653986512
0,401588297	0,598411703
0,759941576	0,240058424
0,238314711	0,761685289
0,370712149	0,629287851
0,412762457	0,587237543
0,587779509	0,412220491
0,108949226	0,891050774
0,425410605	0,574589395

Źródło: opracowanie własne.

Wartości początkowe estymatorów parametrów mieszających wynoszą $\hat{\alpha}_1^0 = 0,42$, $\hat{\alpha}_2^0 = 0,58$. Funkcje gęstości prawdopodobieństwa dla rozpatrywanych rozkładów Poissona przybierają postać:

$$f_1(x_i|\lambda_1) = \frac{\lambda_1^{x_i}}{x_i!} e^{-\lambda_1}, \quad f_2(x_i|\lambda_2) = \frac{\lambda_2^{x_i}}{x_i!} e^{-\lambda_2}.$$

Logarytmy naturalne z tych funkcji oraz pierwsze pochodne liczone po parametrach λ_j wynoszą:

$$\ln f_1(x_i|\lambda_1) = x_i \ln \lambda_1 + \ln x_i - \lambda_1, \ln f_2(x_i|\lambda_2) = x_i \ln \lambda_2 + \ln x_i - \lambda_2,$$

$$\frac{\partial \ln f_1(x_i|\lambda_1)}{\partial \lambda_1} = \frac{x_i}{\lambda_1} - 1, \frac{\partial \ln f_2(x_i|\lambda_2)}{\partial \lambda_2} = \frac{x_i}{\lambda_2} - 1.$$

Wykorzystując powyższe obliczenia oraz warunek (8), otrzymamy równania

$$\sum_{i=1}^n \hat{p}(1|x_i) \left(\frac{x_i}{\hat{\lambda}_1} - 1 \right) = 0, \sum_{i=1}^n \hat{p}(2|x_i) \left(\frac{x_i}{\hat{\lambda}_2} - 1 \right) = 0.$$

Przekształcając powyższe równania, uzyskuje się estymatory $\hat{\lambda}_1^r, \hat{\lambda}_2^r$ parametrów rozkładów Poissona z mieszanki obliczonej w r -tej iteracji na podstawie prawdopodobieństw *a posteriori* $\hat{p}^{r-1}(1|x_i), \hat{p}^{r-1}(2|x_i)$ z poprzedniej iteracji. Estymatory te wynoszą:

$$\hat{\lambda}_1^r = \frac{\sum_{i=1}^n \hat{p}^{r-1}(1|x_i) x_i}{\sum_{i=1}^n \hat{p}^{r-1}(1|x_i)}, \hat{\lambda}_2^r = \frac{\sum_{i=1}^n \hat{p}^{r-1}(2|x_i) x_i}{\sum_{i=1}^n \hat{p}^{r-1}(2|x_i)}. \quad (12)$$

Iteracja I

Wykorzystując wartości prawdopodobieństw *a posteriori* z tab. 2, wartości obserwacji z tab. 1 oraz wzory (12), oblicza się ocenę parametrów dla obu rozkładów mieszanki. Wartości ocen w pierwszej iteracji wynoszą $\hat{\lambda}_1^1 = 5,63, \hat{\lambda}_2^1 = 5,27$.

Uwzględniając powyższe wartości estymatorów parametrów, średnie arytmetyczne początkowych prawdopodobieństw *a posteriori* z tab. 1 $\hat{\alpha}_1^0 = 0,42, \hat{\alpha}_2^0 = 0,58$ oraz wzór (9), oblicza się wartości prawdopodobieństw *a posteriori* dla tej iteracji. Wyniki przedstawia tab. 3.

Przybliżone wartości estymatorów parametrów mieszających iteracji I wynoszą odpowiednio $\hat{\alpha}_1^1 = 0,43, \hat{\alpha}_2^1 = 0,57$. W kolejnych iteracjach schemat postępowania jest analogiczny. Poniżej prezentowane są wyniki dla kolejnych iteracji.

Iteracja II

Wartości estymatorów parametrów rozkładów mieszanki obliczone w iteracji II wynoszą odpowiednio $\hat{\lambda}_1^2 = 4,29$ i $\hat{\lambda}_2^2 = 7,3$. Wartości prawdopodobieństw *a posteriori* obliczone w tej iteracji z uwzględnieniem powyższych wartości estymatorów parametrów przedstawia tab. 4.

Estymatory parametrów mieszających w tej iteracji wynoszą $\hat{\alpha}_1^2 = 0,62, \hat{\alpha}_2^2 = 0,38$.

Tabela 3. Wartości prawdopodobieństw *a posteriori* uzyskane w iteracji I

<i>A posteriori</i> dla rozkładu 1	<i>A posteriori</i> dla rozkładu 2
0,476886497	0,523113503
0,503589022	0,496410978
0,503589022	0,496410978
0,556781230	0,443218770
0,530271090	0,469728910
0,530271090	0,469728910
0,530271090	0,469728910
0,201807517	0,798192483
0,238434264	0,761565736
0,258379820	0,741620180
0,433028064	0,566971936

Źródło: opracowanie własne.

Tabela 4. Wartości prawdopodobieństw *a posteriori* uzyskane w iteracji II

<i>A posteriori</i> dla rozkładu 1	<i>A posteriori</i> dla rozkładu 2
0,649146180	0,350853820
0,758855056	0,241144944
0,758855056	0,241144944
0,901024608	0,098975392
0,842578109	0,157421891
0,842578109	0,157421891
0,842578109	0,157421891
0,003146842	0,996853158
0,009049498	0,990950502
0,015294777	0,984705223
0,623090174	0,376909826

Źródło: opracowanie własne.

Iteracja III

Wartości estymatorów parametrów rozkładów mieszanki obliczone w iteracji III wynoszą odpowiednio $\hat{\lambda}_1^3 = 2,4$, $\hat{\lambda}_2^3 = 10,6$. Wartości prawdopodobieństw *a posteriori* uzyskane w iteracji III prezentuje tab. 5.

Estymatory parametrów mieszających w tej iteracji wynoszą $\hat{\alpha}_1^3 = 0,69$, $\hat{\alpha}_2^3 = 0,31$.

Iteracja IV

Wartości estymatorów parametrów rozkładów mieszanki obliczone w iteracji IV wynoszą odpowiednio $\hat{\lambda}_1^4 = 2,41$, $\hat{\lambda}_2^4 = 14$. Wartości prawdopodobieństw *a posteriori* uzyskane w iteracji IV prezentuje tab. 6.

Tabela 5. Wartości prawdopodobieństw *a posteriori* uzyskane w iteracji III

<i>A posteriori</i> dla rozkładu 1	<i>A posteriori</i> dla rozkładu 2
0,941422770	0,058577230
0,986157427	0,013842573
0,986157427	0,013842573
0,999286140	0,000713860
0,996843368	0,003156632
0,996843368	0,003156632
0,996843368	0,003156632
2,79237E-07	0,999999721
5,48678E-06	0,999994513
2,43211E-05	0,999975679
0,690358396	0,309641604

Źródło: opracowanie własne.

Tabela 6. Wartości prawdopodobieństw *a posteriori* uzyskane w iteracji IV

<i>A posteriori</i> dla rozkładu 1	<i>A posteriori</i> dla rozkładu 2
0,995290639	0,004709361
0,999184638	0,000815362
0,999184638	0,000815362
0,999975729	2,42706E-05
0,999859286	0,000140714
0,999859286	0,000140714
0,999859286	0,000140714
1,46319E-07	0,999999854
4,91941E-06	0,999995081
2,8524E-05	0,999971476
0,699324709	0,300675291

Źródło: opracowanie własne.

Estymatory parametrów mieszających w tej iteracji wynoszą $\hat{\alpha}_1^4 = 0,699$, $\hat{\alpha}_2^4 = 0,301$.

Iteracja V

Wartości estymatorów parametrów rozkładów mieszanki obliczone w iteracji V wynoszą odpowiednio $\hat{\lambda}_1^5 = 2,43$, $\hat{\lambda}_2^5 = 14,3$.

Estymatory parametrów mieszających w tej iteracji wynoszą $\hat{\alpha}_1^5 = 0,6995$, $\hat{\alpha}_2^5 = 0,3005$.

Wartości prawdopodobieństw *a posteriori* z VI i V iteracji spełniają warunek (10), dlatego procedurę iteracyjną zatrzymuje się, stąd wartości estymatorów parametrów mieszających wynoszą

Tabela 7. Wartości prawdopodobieństw *a posteriori* uzyskane w iteracji V

<i>A posteriori</i> dla rozkładu 1	<i>A posteriori</i> dla rozkładu 2
0,996422758	0,003577242
0,999391335	0,000608665
0,999391335	0,000608665
0,999982473	1,75271E-05
0,999896692	0,000103308
0,999896692	0,000103308
0,999896692	0,000103308
1,58252E-07	0,999999842
5,49885E-06	0,999994501
3,24132E-05	0,999967587
0,699491605	0,300508395

Źródło: opracowanie własne.

$$\hat{\alpha}_1 = 0,6995, \hat{\alpha}_2 = 0,3005,$$

natomiast wartości estymatorów parametrów rozkładów mieszanki równają się

$$\hat{\lambda}_1 = 2,43, \hat{\lambda}_2 = 14,3.$$

3. Metoda momentów

W metodzie momentów za estymator nieznanego parametru będącego k -tym momentem rozkładu populacji przyjmuje się k -ty moment z próby. Dla skończonej mieszanki rozkładów dyskretnych o rozkładzie określonym w (1) i przy warunku (2), k -ty moment rozkładu mieszanki wyraża się jako liniowa kombinacja k -tych momentów j -tych rozkładów wchodzących w skład tej mieszanki, tzn.

$$E(X^k) = \sum_{x=0}^{\infty} x^k f(x) = \sum_{x=0}^{\infty} x^k \sum_{j=1}^r \alpha_j f_j(x) = \sum_{j=1}^r \alpha_j \sum_{x=0}^{\infty} x^k f_j(x) = \sum_{j=1}^r \alpha_j E(X_j^k), \quad (13)$$

gdzie $\sum_{j=1}^r \alpha_j = 1$. W przypadku mieszanki dwóch rozkładów Poissona z parametrami

mi λ_1 i λ_2 , pierwszy, drugi oraz trzeci moment wynoszą odpowiednio:

$$E(X_j) = \lambda_j, E(X_j^2) = \lambda_j^2 + \lambda_j, E(X_j^3) = \lambda_j^3 + 3\lambda_j^2 + \lambda_j. \quad (14)$$

Zakładając, że $\alpha_1 = \alpha$ i $\alpha_2 = 1 - \alpha$, estymatory $\hat{\alpha}$, $\hat{\lambda}_1$, $\hat{\lambda}_2$ parametrów α , λ_1 , λ_2 metodą momentów uzyskuje się, rozwiązując układ równań

$$\begin{cases} \bar{x} = \alpha\lambda_1 + (1-\alpha)\lambda_2, \\ \overline{x^2} = \alpha(\lambda_1^2 + \lambda_1) + (1-\alpha)(\lambda_2^2 + \lambda_2), \\ \overline{x^3} = \alpha(\lambda_1^3 + 3\lambda_1^2\lambda_1) + (1-\alpha)(\lambda_2^3 + 3\lambda_2^2\lambda_2) \end{cases} \quad (15)$$

lub równoznacznie

$$\begin{cases} \alpha\lambda_1 + (1-\alpha)\lambda_2 = \bar{x}, \\ \alpha\lambda_1^2 + (1-\alpha)\lambda_2^2 = \overline{x^2} - \bar{x}, \\ \alpha\lambda_1^3 + (1-\alpha)\lambda_2^3 = \overline{x^3} - 3\bar{x}^2 + 2\bar{x}. \end{cases} \quad (16)$$

W celu porównania metody momentów z metodą największej wiarygodności w przykładzie 2 zostaną znalezione estymatory parametru mieszającego oraz parametrów rozkładów Poissona pierwszą metodą do danych z przykładu 1.

Przykład 2

Na podstawie wylosowanej próby w przykładzie 1 średnia, drugi oraz trzeci moment z próby wynoszą odpowiednio $\bar{x} = 6$, $\overline{x^2} = 66,8$, $\overline{x^3} = 918$. Rozwiązując układ równań z trzema niewiadomymi

$$\begin{cases} \alpha\lambda_1 + (1-\alpha)\lambda_2 = 6, \\ \alpha(\lambda_1^2 + \lambda_1) + (1-\alpha)(\lambda_2^2 + \lambda_2) = 66,8, \\ \alpha(\lambda_1^3 + 3\lambda_1^2\lambda_1) + (1-\alpha)(\lambda_2^3 + 3\lambda_2^2\lambda_2) = 918, \end{cases}$$

otrzymuje się wartości oceny szacowanych parametrów, tj. $\hat{\alpha} = 0,6313$, $\hat{\lambda}_1 = 2,1939$, $\hat{\lambda}_2 = 12,5158$.

4. Podsumowanie

Do podsumowania wyników uzyskanych dwiema opisywanymi w tej pracy metodami potrzebna będzie informacja dotycząca pochodzenia danych zawartych w tab. 1. Obserwacje od 1-7 pochodzą z populacji o rozkładzie Poissona z $\lambda_1 = 2$, natomiast obserwacje od 8-10 z populacji o rozkładzie Poissona z $\lambda_2 = 15$. Z tej informacji wywnioskować można również, że parametry udziału (parametry mieszające) dla danych w badanej próbie pochodzących z pierwszego rozkładu i dla danych z drugiego rozkładu wynoszą odpowiednio $\alpha_1 = 0,7$ i $\alpha_2 = 0,3$.

Porównując wyniki uzyskane metodą największej wiarygodności oraz metodą momentów, można stwierdzić, że oszacowanie wartości parametrów mieszających

uzyskane przy użyciu metody największej wiarygodności jest bardziej dokładne aniżeli w metodzie momentów. Wartości estymatorów parametrów mieszających uzyskane metodą największej wiarygodności są bliższe rzeczywistym wartościom parametrów udziału dla badanej próby.

Jeżeli chodzi o oszacowanie wartości parametrów λ dla obu rozkładów Poissona w mieszance, to wyniki uzyskane obiema metodami wyraźnie różnią się od rzeczywistych wartości parametrów. Na uwagę zasługuje jednak to, iż wartości estymatorów $\hat{\lambda}_1$ i $\hat{\lambda}_2$ uzyskane za pomocą metody największej wiarygodności są bliższe rzeczywistym wartościom parametrów λ_1 i λ_2 , aniżeli te uzyskane za pomocą metody momentów, która daje mało efektywne estymatory.

Ostatecznie można powiedzieć, że metoda największej wiarygodności z algorytmem iteracyjnym jest dobrym narzędziem do estymacji parametrów udziału rozkładów w mieszance rozkładów.

Literatura

- [1] Bartos J., Dyczka W., Królikowska K., Kryszewski W., Wasilewska M., *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach – część II. Statystyka matematyczna*, Wydawnictwo Naukowe PWN, Warszawa 2002.
- [2] Everitt B.S., Hand D.J., *Finite Mixture Distributions*, Chapman and Hall, London 1981.
- [3] Feller W., *Wstęp do rachunku prawdopodobieństwa*, PWN, Warszawa 1966.
- [4] Fisz M., *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN, Warszawa 1967.
- [5] Jajuga K., *Statystyczna analiza wielowymiarowa*, PWN, Warszawa 1993.

PARAMETER ESTIMATION FOR A MIXTURE OF TWO POISSON DISTRIBUTION

Summary

For a heterogeneous portfolio we assume that the number of claims has the Poisson distribution with a variable parameter λ . The claim intensity λ is the outcome of a mixing variable Λ . Then the claim number distribution is the mixed Poisson distribution with a mixing parameter Λ .

Two estimation methods for the case of the mixture of two Poisson distributions will be compared in this paper: the maximum likelihood and the method of moments.

Łukasz Kuźmiński – mgr inż., doktorant w Katedrze Statystyki Akademii Ekonomicznej we Wrocławiu.

Anna Nikodem – mgr, asystentka w Katedrze Statystyki Akademii Ekonomicznej we Wrocławiu.