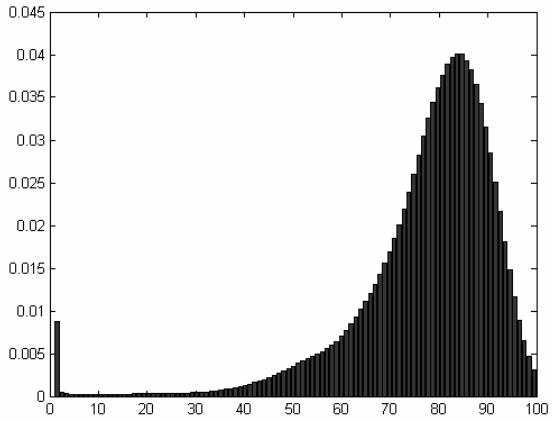
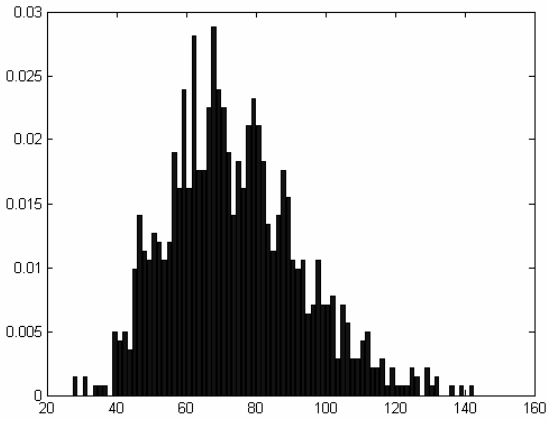
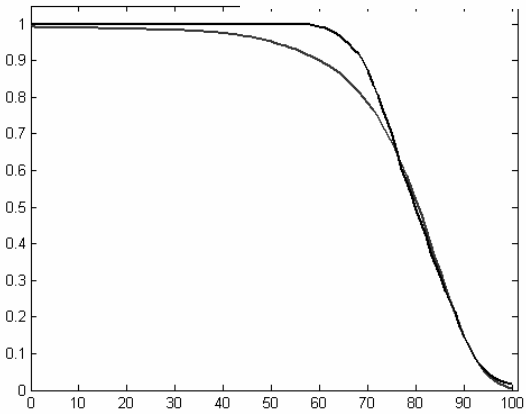
Modelowanie stochastyczne I

Procesy Poissona a podział komórek







Paweł Cibis
#136008
11 stycznia 2006

Zadanie 3. – Podział komórek (lista dodatkowa)

1. Zarys zagadnienia

Zadanie polega na obliczeniu statystyk długości życia przy założeniu prawdziwości pewnej teorii. Głosi ona, iż w podziale ludzkich komórek następują pomyłki, których pojawianie się można opisać procesem Poissona o intensywności $\lambda = 2{,}5$ w ciągu roku. Gdy liczba pomyłek osiągnie wartość $C = 196$ następuje zgon.

W teorii tej jest jeden wspólny model dla kobiet i mężczyzn. Analiza tablic umieralności (z roku 1998) i wykreślonych na ich podstawie krzywych przeżycia sugeruje, iż takie podejście jest błędne, gdyż przeciętny czas trwania życia jest dla mężczyzn znacznie krótszy niż dla kobiet. Dlatego należy spróbować dopasować model osobno dla kobiet i mężczyzn tak, aby rzeczywiste i teoretyczne kształty krzywych były do siebie jak najbardziej zbliżone. W tym celu, w dalszej części zostaną wyestymowane nowe parametry modelu (intensywność i krytyczna ilość błędnych komórek).

2. Obliczenia dla $\lambda = 2{,}5$ i $C = 196$

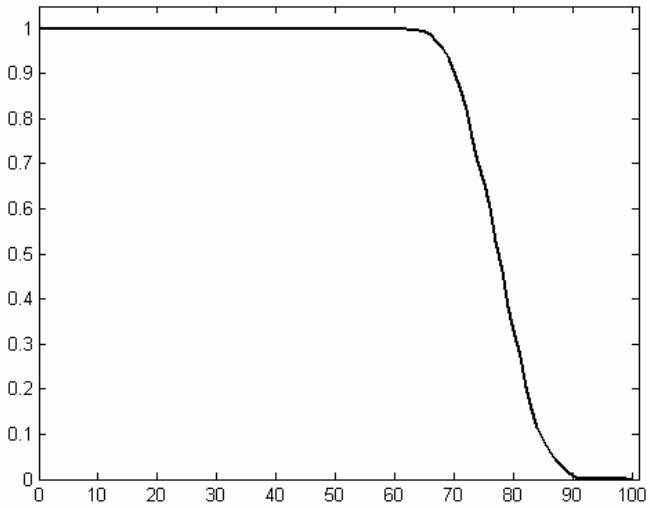
Oznaczenie: X – długość życia (w ukończonych latach)

Tabela 1. Parametry statystyczne modelu.

$\overline{X}$	S^2	S	$P(X < 67{,}2)$
78,045	29,02	5,39	0,02

Tabela 2. Prawdopodobieństwo dożycia danego wieku.

wiek	$P(X >$ wiek$)$	wiek	$P(X >$ wiek$)$	wiek	$P(X >$ wiek$)$
0	1	34	1	68	0,95
1	1	35	1	69	0,95
2	1	36	1	70	0,91
3	1	37	1	71	0,89
4	1	38	1	72	0,84
5	1	39	1	73	0,79
6	1	40	1	74	0,74
7	1	41	1	75	0,66
8	1	42	1	76	0,58
9	1	43	1	77	0,50
10	1	44	1	78	0,39
11	1	45	1	79	0,30
12	1	46	1	80	0,22
13	1	47	1	81	0,19
14	1	48	1	82	0,16
15	1	49	1	83	0,15
16	1	50	1	84	0,11
17	1	51	1	85	0,07
18	1	52	1	86	0,05
19	1	53	1	87	0,03
20	1	54	1	88	0,03
21	1	55	1	89	0,01
22	1	56	1	90	0,01
23	1	57	1	91	0,01
24	1	58	1	92	0,01
25	1	59	1	93	0,01
26	1	60	1	94	0,01
27	1	61	1	95	0,01
28	1	62	1	96	0,01
29	1	63	1	97	0
30	1	64	1	98	0
31	1	65	0,99	99	0
32	1	66	0,98	100	0
33	1	67	0,96		



Rysunek 1. Krzywa przeżycia dla modelu z $\lambda = 2{,}5$ i C = 196.

3. Metoda estymacji

Dane w tablicach długości życia są wyznaczane dla całkowitych wartości wieku człowieka. Krzywa w modelu została obliczona dla tych samych wartości, gdyż wyznaczanie punktów dla współrzędnych niecałkowitych w takiej sytuacji nie miało większego sensu (nie można by było ich porównać z rzeczywistymi wynikami). W większości przypadków rozbieżność pomiędzy krzywymi na odcinku równym jeden można policzyć dokładnie ze wzoru na pole trapezu o wysokości jeden (odległość pomiędzy kolejnymi latami), ponieważ wzór ten jest prawidłowy także dla równoległoboków (więc tym bardziej prostokątów i kwadratów). Niedokładność pojawia się jedynie wtedy, gdy krzywe przecinają się pomiędzy przyjętymi węzłami – w punkcie odpowiadającym niecałkowitej wartości wieku. Obliczony błąd jest wtedy nieco większy niż w rzeczywistości. Nie są to jednak sytuacje na tyle częste, a wartości na tyle duże, by zaburzyło istotnie wyniki.

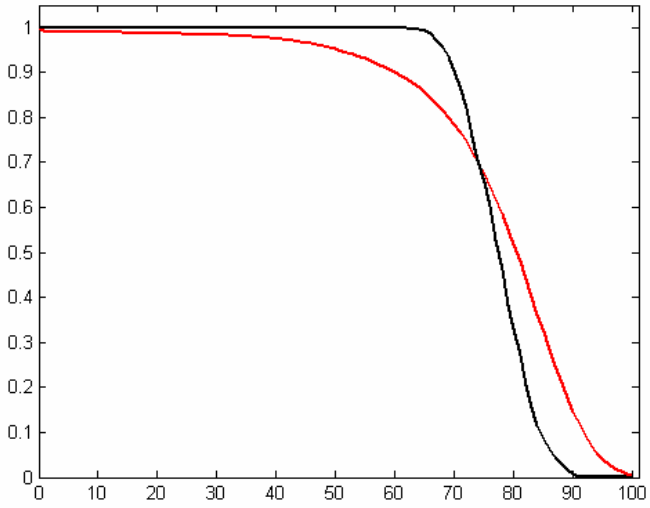
Estymacja optymalnych parametrów procesu Poissona polegała na wybraniu do analizy przedziałów parametrów, zdyskretyzowaniu a następnie obliczeniu błędów dopasowania dla każdego powstałego w ten sposób węzła. Okazało się, iż dla każdej wartości intensywności procesu istnieje pewna wartość krytycznej liczby pomyłek (oczywiście całkowita), dla której błąd jest minimalny. Jednakże wraz ze wzrostem intensywności minima te są coraz większe. Dlatego optymalne wartości parametrów były wyszukiwane dla niskich intensywności poprzez zmniejszanie i równoczesne zagęszczanie badanych przedziałów.

Dla każdej pary parametrów błąd dopasowania był liczony dla krzywej utworzonej na podstawie 100 przebiegów procesu. Dodatkowo dla każdego punktu, w którym pojawiło się minimum dostatecznie niskie, by zostało zaakceptowane (jako kryterium przyjęte zostało niedopasowanie niższe od 4%), w ramach weryfikacji błąd był wyliczany ponownie, kilka razy – dla krzywej utworzonej na podstawie 1000 przebiegów procesu.

Można w tym momencie postawić pytanie o to, jak dalece dokładnie można dopasowywać wykres. Skoro minima rosną wraz ze wzrostem intensywności, to może maleją do zera wraz z jej zmniejszaniem się? Niestety, idealne dopasowanie nie jest możliwe – dla bardzo małych intensywności (mniejszych od 0,1), przy bardzo gęsto rozmieszczonych węzłach błąd ponownie wzrastał – zwiększała się dokładność przybliżenia początkowego odcinku krzywej, ale znacznie bardziej zwiększał się błąd na końcowym fragmencie (co jest widoczne na rys. 7).

4. Dopasowanie krzywej dla kobiet

Dla „oryginalnych" parametrów modelu błąd dopasowania wynosi 6,291%. Wyraźne niedopasowanie jest widoczne na środkowym i końcowym odcinku krzywych (rys. 2).

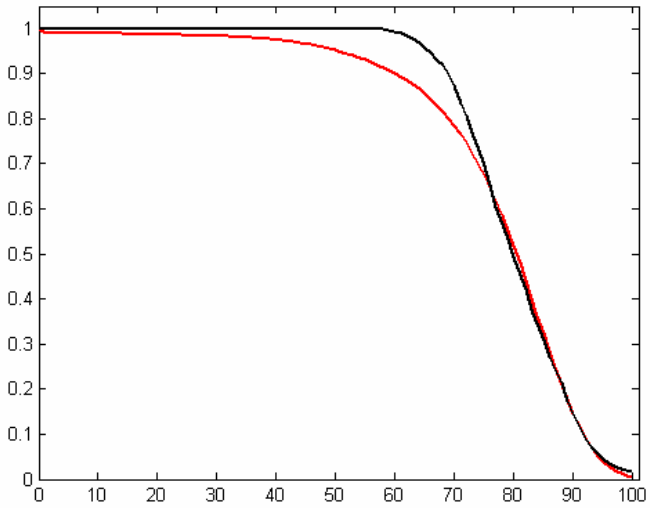


Rysunek 2. Teoretyczna krzywa przeżycia dla modelu z $\lambda = 2{,}5$ i C = 196 (kolor czarny) oraz rzeczywista krzywa przeżycia dla kobiet (kolor czerwony).

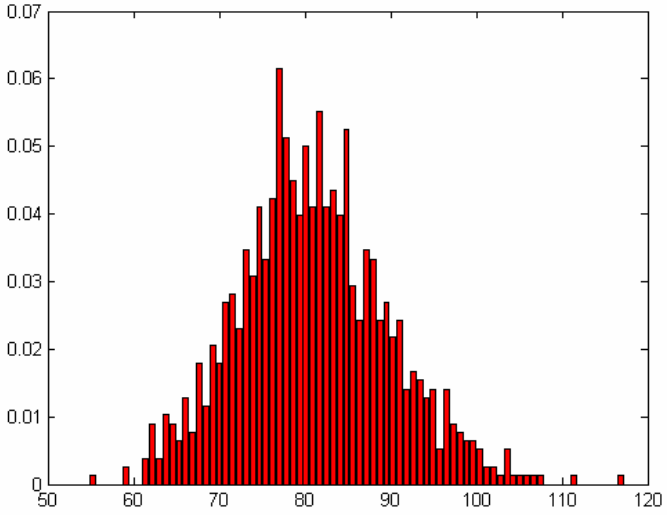
Wyestymowane zostały następujące parametry: $\lambda = 1$ i $C = 81$. Otrzymany błąd wyniósł 3,1388% (rys. 3).

Dzięki estymacji parametrów udało się zmniejszyć błąd dopasowania krzywych do ok. 3%. Na sporym odcinku różnice wciąż są jednak dość znaczne (różnica prawdopodobieństw dożycia wynosi czasem prawie 0.1). W modelu widoczny jest gwałtowny spadek prawdopodobieństwa dożycia po przekroczeniu wieku 65 lat. W rzeczywistości następuje on ok. 10 lat później – wcześniej prawdopodobieństwo to maleje wyraźnie już od wieku 30 lat, ale nie aż tak gwałtownie. Natomiast w modelu do wieku 60 lat nie następuje praktycznie żaden spadek. Wynika to z postaci przyjętego modelu – przy zadanych parametrach zgon powinien nastąpić przeciętnie koło 81. roku życia (wynik rzeczywisty: 77,7), ale odchylenia od średniej są dużo mniejsze niż w rzeczywistości, ponieważ trudno by, przy intensywności pojawiania się pomyłek podczas podziału komórek równej 1 na rok, osiągnięte zostało w ciągu 30-40 lat 81 błędów w na tyle licznym odsetku przebiegów procesu, by wyraźnie wpływał on na oszacowane prawdopodobieństwo.

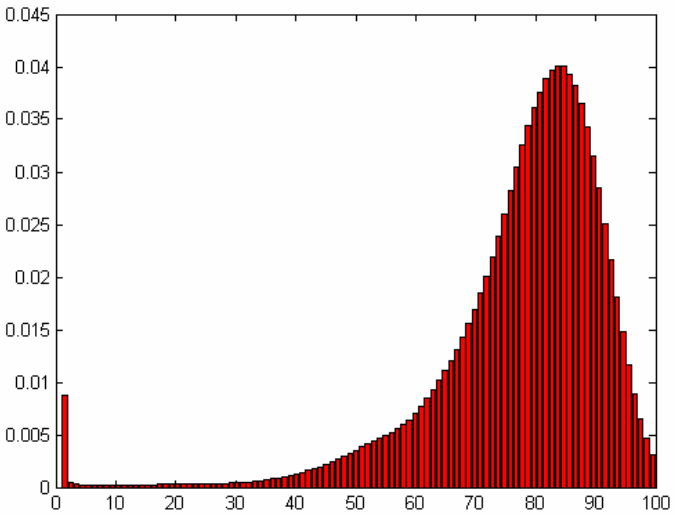
Unormowany histogram przedstawiający rozkład długości życia w analizowanym modelu wskazuje na rozkład normalny (rys. 4). Odchylenie standardowe wyniosło 8,7, a średni wiek – 81. Wobec tego zgodnie z regułą trzech sigm ponad 95% obserwacji (czyli momentów zgonu) mieści się w przedziale [54,9; 107,1]. Poza tym przedziałem nie następuje już na tyle duża ilość zgonów, by miało to istotny wpływ na kształt krzywej przeżycia. W rzeczywistości odchylenie standardowe jest znacznie większe (ok. 15), a rozkład jest wyraźnie lewostronnie asymetryczny, gdyż model przewiduje jako jedyną przyczynę śmierci przekroczenie krytycznej liczby błędnych komórek (co następuje dość późno i gwałtownie zarazem), nie biorąc pod uwagę różnego rodzaju „wypadków losowych", które mogą się przytrafić w życiu dużo wcześniej, w tym podwyższonego ryzyka śmierci noworodków (obrazuje to histogram rozkładu rzeczywistego na rys. 5).



Rysunek 3. Teoretyczna krzywa przeżycia dla modelu z $\lambda = 1$ i $C = 81$ (kolor czarny) oraz rzeczywista krzywa przeżycia dla kobiet (kolor czerwony).



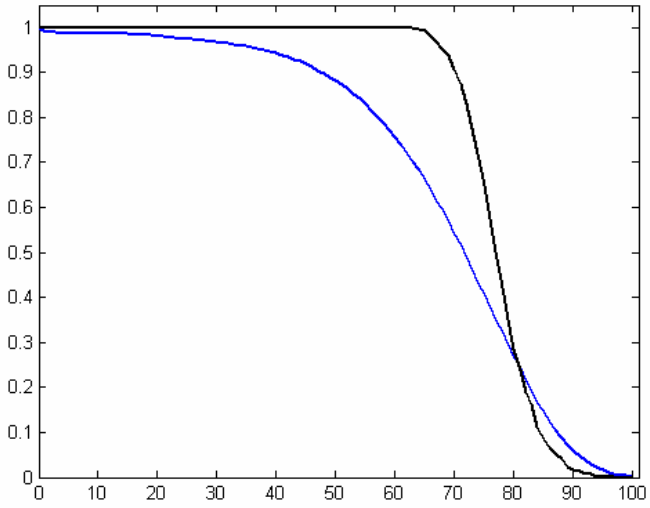
Rysunek 4. Unormowany histogram długości życia dla modelu z $\lambda = 1$ i $C = 81$.



Rysunek 5. Unormowany histogram rzeczywistej długości życia dla kobiet.

5. Dopasowanie krzywej dla mężczyzn

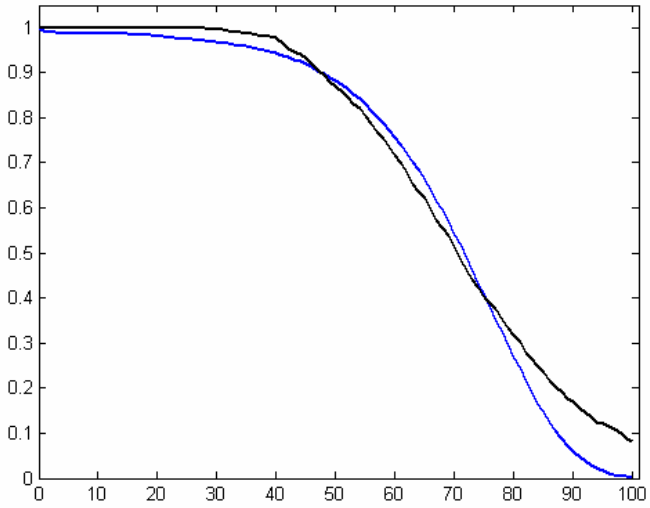
Dla „oryginalnych” parametrów modelu błąd dopasowania wynosi 9,5358%. Bardzo duże niedopasowanie jest widoczne na środkowym odcinku krzywych (rys. 6).



Rysunek 6. Teoretyczna krzywa przeżycia dla modelu z $\lambda = 2{,}5$ i C = 196 (kolor czarny) oraz rzeczywista krzywa przeżycia dla mężczyzn (kolor niebieski).

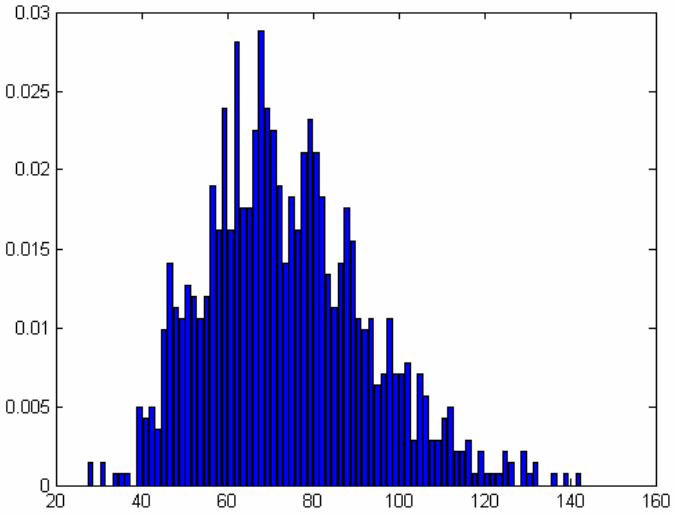
Rzeczywista krzywa przeżycia dla mężczyzn zaczyna maleć szybciej i bardziej gwałtowniej niż kobieca. Wiąże się to m.in. z wykonywaniem przez mężczyzn pewnych zarezerwowanych dla nich (lub przez nich zdominowanych) zawodów o wyższym stopniu ryzyka oraz większej eksploatacji organizmu – takich jak np. zawód górnika czy żołnierza. Dlatego dopasowanie do znacznie później malejącej krzywej z „oryginalnego” modelu jest tu dużo gorsze – błąd sięga prawie 10%. Aczkolwiek zauważalne jest lepsze dopasowanie dla wyższego wieku.

Wyestymowane zostały następujące parametry: $\lambda = 0{,}19$ i $C = 14$. Otrzymany błąd wyniósł 3,6898% (rys. 7).

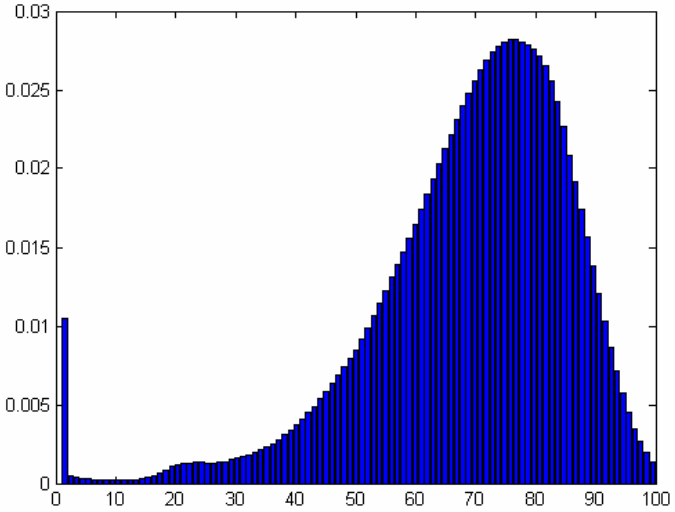


Rysunek 7. Teoretyczna krzywa przeżycia dla modelu z $\lambda = 0{,}19$ i C = 14 (kolor czarny) oraz rzeczywista krzywa przeżycia dla mężczyzn (kolor niebieski).

W tym przypadku również po estymacji udało się zmniejszyć błąd dopasowania. Dla obu krzywych wartość przeciętna jest zbliżona. Wyraźnie widać jednak różnice w kształcie krzywych w porównaniu z przypadkiem szacowanym dla kobiet. Zmniejszenie błędu poniżej 4% wymagało poprawienia dokładności na początkowym i środkowym odcinku krzywej. Spowodowało to jednak zwiększenie się błędów w jej końcowym fragmencie. Generalnie jednak niedopasowanie okazało się mniejsze.



Rysunek 8. Unormowany histogram długości życia dla modelu z $\lambda = 0{,}19$ i C = 14.



Rysunek 9. Unormowany histogram rzeczywistej długości życia dla kobiet.

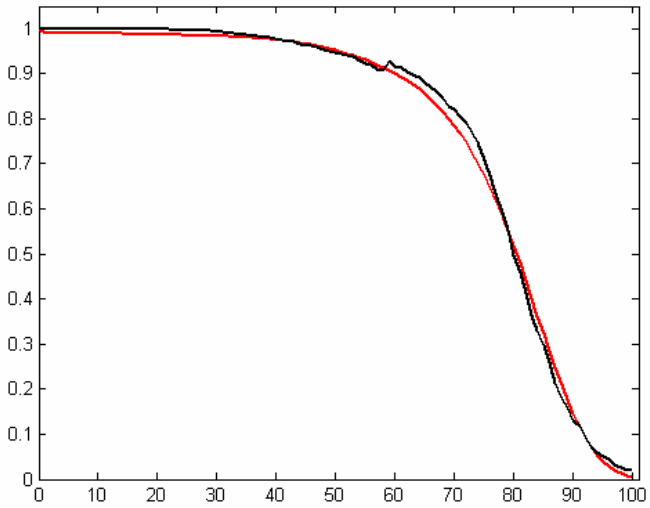
Przeciętna długość życia wg modelu wyniosła 73,8 (wartość rzeczywista: 69,3), a odchylenie standardowe 19,2 (wartość rzeczywista: 17). Zaobserwować można, że rozproszenie wartości wieku zgonu jest bardziej zbliżone niż w modelu dla kobiet. Równocześnie jednak zwiększa się nieco różnica w średnich długościach życia, a kierunek asymetrii rozkładów jest przeciwny (rys. 8 i 9).

6. Wnioski końcowe

Mimo iż udaje się otrzymać dość niskie błędy dopasowań podczas estymacji parametrów modelu, nie przybliża on zbyt dobrze rzeczywistych krzywych przeżycia, ponieważ poprawienie dokładności na jednym odcinku skutkuje utratą precyzji na innym. Rozwiązaniem tego problemu może być zbudowanie modelu złożonego z kilku modeli cząstkowych. Na przykład – jeden przybliżałby początkowy fragment krzywej, inny środkowy, a jeszcze inny końcowy. W odniesieniu do teorii oznaczałoby to, iż intensywność pomyłek przy podziałach komórek nie jest stała i/lub w różnych okresach życia inna ilość takich błędów może spowodować śmierć.

W przykładowym modelu złożonym dla kobiet (rys. 10) otrzymany błąd dopasowania wyniósł 1,4%, zaś intensywność i krytyczna ilość pomyłek dana była odpowiednio funkcjami $\Lambda(t)$ i $C(t)$ postaci:

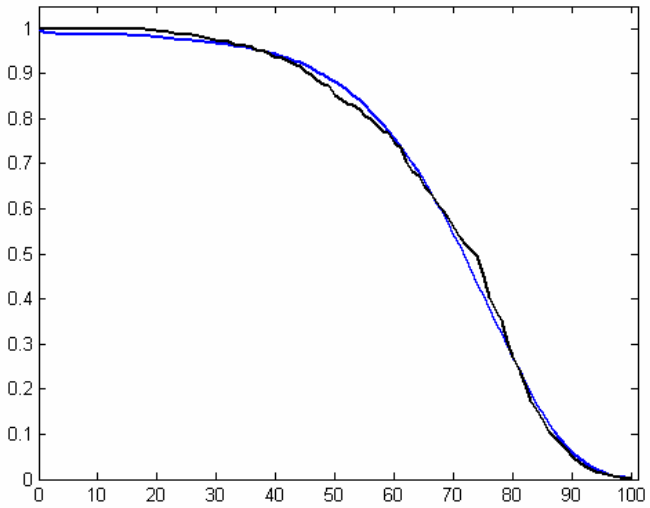
$$\Lambda(t) = \begin{cases} 0,065 & \text{dla } 0 \le t \le 59 \\ 0,165 & \text{dla } 59 \le t \le 74 \\ 1 & \text{dla } 75 \le t \le 100 \end{cases} \qquad C(t) = \begin{cases} 7 & \text{dla } 0 \le t \le 59 \\ 15 & \text{dla } 59 \le t \le 74 \\ 81 & \text{dla } 75 \le t \le 100 \end{cases}$$



Rysunek 10. Teoretyczna krzywa przeżycia dla modelu z parametrami $\Lambda(t)$ i $C(t)$ (kolor czarny) oraz rzeczywista krzywa przeżycia dla kobiet (kolor czerony).

Podobnie, w przykładowym modelu złożonym dla mężczyzn (rys. 11) błąd dopasowania wyniósł 1,3%, zaś intensywność i krytyczna ilość pomyłek dana była odpowiednio funkcjami $\Lambda(t)$ i $C(t)$ postaci:

$$\Lambda(t) = \begin{cases} 0,0675 & \text{dla } 0 \le t \le 59 \\ 0,1825 & \text{dla } 59 \le t \le 74 \\ 0,9075 & \text{dla } 75 \le t \le 100 \end{cases} \qquad C(t) = \begin{cases} 6 & \text{dla } 0 \le t \le 59 \\ 14 & \text{dla } 59 \le t \le 74 \\ 68 & \text{dla } 75 \le t \le 100 \end{cases}$$



Rysunek 11. Teoretyczna krzywa przeżycia dla modelu z parametrami $A(t)$ i $C(t)$ (kolor czarny) oraz rzeczywista krzywa przeżycia dla kobiet (kolor niebieski).

W modelu złożonym widoczne jest o wiele lepsze dopasowanie krzywych niż w modelach prostych – sumaryczny błąd jest mniejszy i składa się z drobnych niedopasowań na poszczególnych odcinkach. Uzyskane w ten sposób krzywe są bardziej wiarygodne i dzięki temu mogą mieć szersze zastosowanie.