Stanisława OSTASIEWICZ
AE Wrocław
Paulina UCIEKLAK-JEŻ
Akademia im. Jana Długosza w Częstochowie

Statystyczne podstawy metody Sullivana

1. Cel pracy

W pracy zaprezentowano oraz przeprowadzono analizę statystyczną statystyki $\hat{e}_x^{DF}$ zaproponowanej w r. 1971 przez D.F. Sullivana jako estymatora przeciętnego dalszego trwania życia w pełno sprawności. Estymator ten określony został na połączonym zbiorze danych dotyczących przeciętnego trwania życia w populacji oraz danych dotyczących niepełnosprawności w populacji. Jeżeli obie populacje są stacjonarne to okazało się, że statystyka $\hat{e}_x^{DF}$ jest estymatorem zgodnym i nieobciążonym przeciętnego dalszego trwania życia w populacji. Znana jest również wariancja tak skonstruowanego estymatora.

2. Elementy analizy przeżycia

Średnia długość trwania życia jest jedną z miar charakteryzujących jakość życia.

Na bazie tego miernika skonstruowany został przez D.F. Sullivana w roku 1971 miernik charakteryzujący przeciętne dalsze trwanie życia bez niepełnosprawności czyli w dobrej kondycji fizycznej i psychicznej. Miernik ten wykorzystywany był do badania wielu populacji.

Podstawowym źródłem danych wykorzystywanych przy jego konstrukcji są demograficzne tablice trwania życia i dane dotyczące niepełnosprawności, które mogą pochodzić ze statystyki publicznej dotyczącej tego zjawiska lub też z ba-

dań ankietowych. Zaletą tej metody jest to, że dane dotyczące umieralności i niepełnosprawności mogą pochodzić z różnych źródeł.

Podstawowym parametrem charakteryzującym umieralność jest funkcja intensywności zgonów, którą zwykle oznacza się $\mu(x,y)$ (por.[1,2]). Funkcja ta zależy od wieku x oraz od momentu urodzenia y. Innym parametrem charakteryzującym trwanie życia jest funkcja przeżycia $s(x,y)$. Między funkcją przeżycia i funkcją intensywności zgonów zachodzi następujący związek:

$$s(x,y) = e^{-\mu(x,y)} \tag{1}$$

Jeżeli przez $l(x,y)$ oznaczymy liczebność grupy wiekowej x w kohorcie urodzonej w momencie y (czyli liczbę dożywających wieku x) to liczba ta może być wyrażona następująco:

$$l(x,y) = l(0,y)\ \exp(-\int_0^x \mu(t,y)dt) \tag{2}$$

gdzie $l(0,y)$ jest to początkowa liczebność kohorty urodzonej w momencie y.

Przeciętne dalsze trwanie życia osoby w wieku x, która urodziła się w momencie y określone jest następująco:

$$e(x,y) = \frac{1}{l(x,y)} \int_x^\infty s(t,y)dt \tag{3}$$

Oznacza to, że czas życia osoby x letniej urodzonej w momencie y będzie jeszcze wynosił średnio $e(x,y)$ lat.

Wśród osób dożywających wieku x lat są osoby cieszące się dobrym zdrowiem jak również osoby niepełnosprawne. Przyjmijmy, że odsetek takich osób wśród osób dożywających wieku x wynosi $\pi(x,y)$. Przeciętne dalsze trwanie życia w zdrowiu, które oznaczamy $e^{DF}(x,y)$ obliczamy według wzoru (por.[3,4]):

$$e^{DF}(x,y) = \frac{1}{l(x,y)} \int_x^\infty (1-\pi(t,y))s(t,y)dt \tag{4}$$

Aby określić funkcję przeżycia $s(t,y)$ potrzebna jest znajomość funkcji intensywności zgonów $\mu(x,y)$, której na ogół nie znamy.

W praktyce posługujemy się empirycznym modelem trwania życia, który nazywany jest tablicami trwania życia.

Tablice trwania życia mogą być dwojakiego rodzaju:

— Tablice kohortowe,

— Tablice przekrojowe.

Tablice kohortowe opisują trwanie życia kohorty czyli grupy ludzi urodzonych w tym samym momencie czasowym. Grupa ta obserwowana jest od momentu urodzenia do momentu śmierci ostatniej osoby z kohorty.

Wiek dzielony jest na przedziały a wartości funkcji biometrycznych podawane są w punktach początkowych tych przedziałów.

Jeżeli wiek podzielony jest na przedziały o długości rocznej, wówczas tablice nazywane są pełnymi, wartości funkcji biometrycznych podawane są w punktach x = 0, 1, 2, ω. W przypadku tablic skróconych wiek podzielony jest na przedziały o równych lub różnych długościach a wartości funkcji biometrycznych podobnie jak w przypadku tablic pełnych podawane są na początku przedziałów. Bardzo często konstruowane są tablice skrócone o 5 letnich przedziałach wieku. Wyjątek stanowi najmłodsza grupa wieku (0 lat), której rozpiętość jest równa 1 rok. Wartość funkcji $l(x,y)$ w punkcie x nazywamy liczbą dożywających wieku x spośród osób urodzonych w momencie y. Wartość funkcji intensywności zgonów $\mu(x,y)$ w punkcie początkowym przedziału nazywamy prawdopodobieństwem zgonu w wieku x i oznaczamy q_x, natomiast wartość funkcji przeżycia w punkcie początkowym przedziału o początku x nazywamy prawdopodobieństwem dożycia do wieku x i oznaczamy p_x.

Dla pełnych tablic trwania życia prawdopodobieństwo zgonu a ciągu roku osoby w wieku x oblicza się następująco (por.[6]):

$$q_x = \frac{d_x}{l_x} \tag{5}$$

gdzie l_x oznacza liczbę osób z kohorty, które dożyły wieku x lat i d_x liczbę osób z kohorty które zmarły w wieku x lat.

Prawdopodobieństwo zgonu w przedziale wieku o długości n lat określone jest następująco:

$${}_nq_x = \frac{d_{x+n}}{l_x} \tag{6}$$

W praktyce częściej posługujemy się tablicami przekrojowymi. Opisują one trwanie życia kohorty utworzonej w danym roku z ludności urodzonej w różnych momentach czasowych.

Tablice przekrojowe mogą być budowane tylko wtedy gdy badana ludność jest ludnością zastojową (por[1,6]).

Populacja zastojowa charakteryzuje się tym, że łączna liczba ludności jak też jej rozkład według wieku nie zmieniają się w czasie. Taką hipotetyczną kohortę można by otrzymać gdyby przez długi okres czasu liczba urodzeń w ciągu roku była stała i w każdej rocznej kohorcie nowonarodzonych umieralność byłaby taka sama w ciągu całego życia i równa umieralności aktualnie obserwowanej. W takiej populacji intensywność umieralności zależy tylko od wieku i nie zależy od momentu urodzenia to znaczy, że funkcja umieralności i funkcja przeżycia spełniają warunki:

$\mu(x,y) = \mu(x)$ dla każdego y

$s(x,y) = s(x)$ dla każdego y.

Z warunku stacjonarności wynika, że rozkład wieku w każdym przedziale wieku hipotetycznej kohorty jest stały w czasie i proporcjonalny do wartości funkcji przeżycia. Oznacza to, że dla każdego $s \in [x, x+n_x)$ rozkład wieku jest określony za pomocą następującej funkcji gęstości:

$$\frac{S(s)}{\int_x^{x+n_x} S(t)dt} \tag{7}$$

gdzie n_x oznacza rozpiętość przedziału wiekowego o początku x. W przypadku tablic pełnych $n_x = 1$.

Prawdopodobieństwo zgonu ${}_{n_x}q_x$ w przedziale wiekowym $[x, x+n_x)$ wyznacza się na podstawie współczynnika zgonów ${}_{n_x}M_x$ w tym przedziale wieku. Współczynnik ten dla danego przedziału wieku w badanym roku obliczany jest na podstawie danych gromadzonych przez USC w następujący sposób:

$${}_{n_x}M_x = \frac{{}_{n_x}D_x}{{}_{n_x}P_x} = \frac{{}_{n_x}D_x}{{}_{n_x}\bar{L}_x} \tag{8}$$

gdzie ${}_{n_x}D_x$ oznacza liczbę zgonów w przedziale wieku $[x, x+n_x)$ i ${}_{n_x}P_x$ oznacza średnią liczbę ludności w przedziale wieku $[x, x+n_x)$.

Jeżeli liczebność próby, na podstawie której wyznaczono współczynnik zgonów ${}_{n_x}M_x$ jest duża to można przyjąć że:

$$ {}_{n_x}M_x = {}_{n_x}m_x $$

gdzie ${}_{n_x}m_x$ oznacza współczynnik zgonów w populacji i określony jest wzorem:

$$ {}_{n_x}m_x = \frac{\int_x^{x+n_x} s(t)\mu(t)dt}{\int_x^{x+n_x} s(t)dt} = \frac{{}_{n_x}d_x}{{}_{n_x}\overline{L}_x} \tag{9} $$

dla każdego x.

Mianownik wzoru (9) określa średnią liczbę lat przeżytych przez ludność w wieku $[x, x+n_x)$.W praktyce jako ${}_{n_x}\overline{L}_x$ przyjmuje się wielkość $n_x \cdot {}_{n_x}P_x$ gdzie ${}_{n_x}P_x$ oznacza liczbę ludności w punkcie środkowym przedziału $[x, x+n_x)$ a n_x rozpiętość tego przedziału.

Liczba lat przeżytych w przedziale $[x,x + n_x)$ przez osoby które dożyły do momentu x czyli osiągnęły wiek x jest sumą dwóch wielkości:

— liczby lat przeżytych przez osoby które dożyły do końca przedziału $[x,x + n_x)$,

— liczby lat przeżytych przez osoby które przeżyły tylko część przedziału $[x,x + n_x)$.

Część przedziału która przeżywana jest przez osobę zmarłą w tym przedziale jest oznaczona ${}_{n_x}a_x$ a jej wielkość średnia określona jest wzorem (por. [1,6])

$$ {}_{n_x}a_x = \frac{{}_{n_x}L_x - n_x \cdot l_x + n_x}{n_x \cdot {}_{n_x}d_x} \tag{10} $$

gdzie ${}_{n_x}L_x$ oznacza łączną liczbę lat przeżytych w przedziale $[x, x+n_x)$ przez osoby dożywające wieku x.

Przekształcając to równanie otrzymuje się następujący wzór na średnią liczbę osobo lat przeżytych w przedziale $[x, x+n_x)$.

$$ {}_{n_x}L_x = n_x \cdot l_{x+n_x} + n_x \cdot {}_{n_x}a_x \cdot {}_{n_x}d_x = n_x \cdot l_{x+n_x} + {}_{n_x}q_x \cdot l_x \cdot {}_{n_x}a_x \tag{11} $$

Korzystając ze wzorów (9) i (11) można napisać:

$$ {}_{n_x}m_x = \frac{{}_{n_x}d_x}{{}_{n_x}L_x} = \frac{{}_{n_x}d_x}{{}_{n_x}l_x - n_x(1-{}_{n_x}a_x)\cdot{}_{n_x}d_x} = \frac{\frac{{}_{n_x}d_x}{l_x}}{\frac{n_x[1-(1-{}_{n_x}a_x)\cdot{}_{n_x}d_x}{l_x}} = \frac{{}_{n_x}q_x}{n_x[1-(1-{}_{n_x}a_x)q_x]} $$

Stąd można wyznaczyć wzór na prawdopodobieństwo zgonu ${}_{n_x}q_x$, który będzie miał następującą postać:

$$ {}_{n_x}q_x = \frac{{}_{n_x}m_x}{[1+n_x(1-{}_{n_x}a_x)_{n_x}m_x]} \tag{12} $$

Znacznie prostszym wzorem na obliczenie wielkości ${}_{n_x}q_x$ jest wzór (6). Jednakże występuje w nim wielkość l_x czyli liczba dożywających wieku x, a takiej wielkości w statystykach USC nie ma.

Całkowita liczba lat przeżytych przez osoby w wieku $[x,x + n_x)$ jest dana wzorem (por. [1,3]):

$$ {}_{n_x}L_x = n_x \cdot l_{x+n_x} + l_{x\,n_x}q_x \cdot {}_{n_x}a_x \tag{13} $$

Jest to łączna liczba lat przeżytych przez osoby dożywające końca przedziału $[x, x+n_x)$ powiększona o liczbę lat przeżytych łącznie przez osoby, które zmarły w przedziale $[x, x+n_x)$ przeżywszy tylko część tego przedziału. Liczba osób, które zmarły w tym przedziale wynosi $l_x \cdot {}_{n_x}q_x$ i każda z nich przeżyła w tym przedziale średnio ${}_{n_x}a_x$ lat.

Korzystając z definicji przeciętnego dalszego trwania życia mamy:

$$ e_x = \frac{1}{l_x}\sum_{i\in A_x} {}_{n_i}L_i \tag{15} $$

gdzie: $A_x = \{i \in A : x \le i\}$.

Suma $\sum_{i\in A_x} {}_{n_i}L_i$ oznacza łączną liczbę lat którą przeżyją jeszcze osoby w wieku x, czyli jest to suma lat do przeżycia przez osoby w wieku x i starsze. Na każdą osobę przypada więc e_x lat.

Można pokazać, że jeśli spełnione są warunki stacjonarności to e_x policzone według wzoru (15) jest równe teoretycznej wartości dalszego przeciętnego trwania życia $e(x)$ określonego wzorem (3).

Zauważmy, że w notacji demograficznej ciągłej l_x oznacza liczbę ludności w wieku x, natomiast w notacji dyskretnej liczba ludności w wieku x oznaczona jest l_x. Oba symbole mają jednak dokładnie taką samą interpretację i oznaczają liczbę ludności dożywającej wieku x. Warunkowe prawdopodobieństwo zgonu w przedziale $[x,x + n_x)$ równe jest liczbie zgonów w przedziale $[x,x + n_x)$ przypadającej na jedną osobę w wieku x. W notacji ciągłej może być ono wyrażone następująco:

$$ {}_{n_x}q_x = \frac{\int_x^{n_x} l(t)\mu(t)dt}{l(x)} \tag{16} $$

Podobnie ${}_{n_x}a_x$ może być zapisane następująco (por.[3]):

$$ {}_{n_x}a_x = \frac{\int_x^{x+n_x} l(t)\mu(t)dt}{l(x)} \tag{17} $$

Podstawiając do wzoru (12) wyrażenie na ${}_{n_x}q_x$ i ${}_{n_x}a_x$ otrzymujemy:

$$ {}_{n_x}L_x = n_x l_{x+n_x} + l_x \cdot \frac{\int_x^{x+n_x} l(t)\mu(t)dt}{l_x} \cdot \frac{\int_x^{x+n_x} l(t)\mu(t)(t-x)dt}{\int_x^{x+n_x} l(t)\mu(t)dt} $$

Po przeprowadzeniu obliczeń po prawej stronie powyższego wzoru otrzymujemy:

$$ {}_{n_x}L_x = \int_x^{x+n_x} l(t)dt $$

Stąd wynika, że:

$$ e_x = \frac{1}{l_x}\int_x^{x+n_x} l(t)dt . $$

Wzór ten jest identyczny ze wzorem (3) i określa przeciętne dalsze trwanie życia dla kohorty.

Tak więc e_x policzone dla tablic przekrojowych jest równe wielkości przeciętnego dalszego trwania życia dla kohorty.

3. Metoda Sullivana

Omówione przeciętne dalsze trwanie życia wykorzystane zostanie do oceny przeciętnego dalszego trwania życia w zdrowiu e_x^{DF} określonego wzorem (4). Estymator tego parametru oznaczony będzie $\hat{e}_x^{DF}$ (por. [7]). Dane potrzebne do oszacowania interesującej nas wielkości pochodzić będą z przekrojowych tablic trwania życia i statystyki publicznej dotyczącej niepełnosprawności. Dane dotyczące niepełnosprawności mogą pochodzić również ze specjalnie przeprowadzonej ankiety. W każdym przypadku współczynnik niepełnosprawności obliczany jest w grupach wiekowych jako udział niepełnosprawnych danej grupy wiekowej w ogólnej liczbie ludności danej grupy. Oznacza to, że przedziały wiekowe występujące w tablicy trwania życia i przedziały wiekowe, w których badana jest niepełnosprawność muszą być takie same. Estymator parametru e_x^{DF} określonego wzorem (4) jest następującej postaci:

$$\hat{e}_x^{DF} = \frac{1}{l_x} \sum_{i \in A_x} (1 - {}_{n_i}\hat{\pi}_i)\, {}_{n_i} L_i$$

gdzie ${}_{n_i}\hat{\pi}_i$ jest udziałem ludności niepełnosprawnej w ludności grupy wiekowej $[i, i+n_i)$ policzonym na podstawie pobranej próby statystycznej. Współczynnik niepełnosprawności obliczamy według wzoru:

$${}_{n_i}\hat{\pi}_i = \frac{1}{{}_{n_i}N_i} \sum_{j=1}^{{}_{n_i}N_i} Y_{ij}(t_{ij}) \tag{18}$$

gdzie ${}_{n_i}N_i$ oznacza liczbę jednostek statystycznych w wieku $[i, i+n_i)$ w pobranej próbie statystycznej i $Y_{ij}(t_{ij})$ jest zmienną indykatorem. Przyjmuje wartość 1 jeśli j-ta jednostka statystyczna w pobranej próbie jest w wieku $[i, i+n_i)$ i jest niepełnosprawna i wartość 0 w przypadku przeciwnym.

Zakłada się, że współczynnik niepełnosprawności w populacji $\pi(x, y)$ zależy tylko od wieku x i nie zależy od momentu urodzenia y. Oznacza to, że spełniony jest warunek:

$\pi(x, y) = \pi(x)$ dla każdego y.

Przeciętne dalsze trwanie życia w populacji (por. wzór (4)) może być określone następująco:

$$e^{DF}(x)=\frac{1}{l_x}\left[\sum_{i\in A_x}[_{n_i}L_i-\int_i^{i+n_i}\pi(t)l(t)dt\right] \tag{19}$$

gdzie $\int_i^{i+n_i}\pi(t)l(t)dt$ oznacza liczbę osób niepełnosprawnych w wieku $[i,i+n_i)$ w populacji.

Zmienna losowa $Y(t)$ przyjmuje dwie wartości 1 i 0 odpowiednio z prawdopodobieństwami $\pi(t)$ i $1-\pi(t)$. Rozkład zmiennej losowej Y w przedziale $[x,x+n_x)$ jest określony wzorem:

$$P(Y(t)=1)=_{n_x}\pi_x$$

$P(Y(t)=0)=1-_{n_x}\pi_x$ Wariancja zmiennej losowej Y(t) w przedziale $[x,x+n_x)$ jest równa $_{n_x}\pi_x\cdot(1-\pi_x)$ co oznacza, że w każdym przedziale jest mniejsza od 1. Ponadto wariancja $\sigma^2(_{n_x}\hat{\pi}_x)$ statystyki $_{n_x}\hat{\pi}_x$ w każdym przedziale $[x,x+n_x)$ jest określona wzorem:

$$\sigma^2(_{n_x}\hat{\pi}_x)=\frac{_{n_x}\pi_x(1-_{n_x}\pi_x)}{_{n_x}N_x}.$$

Jeżeli $_{n_x}N_x$ dąży do nieskończoności to wariancja statystyki $_{n_x}\hat{\pi}_x$ dąży do zera.

Ponadto $_{n_x}\hat{\pi}_x$ jest estymatorem nieobciążonym parametru $_{n_x}\pi_x$. Stąd wynika, że $_{n_x}\hat{\pi}_x$ jest nieobciążonym i zgodnym estymatorem parametru $_{n_x}\pi_x$. Z tego wynika, że estymator $\hat{e}_x^{DF}$ zaproponowany przez Sullivana jest estymatorem zgodnym i nieobciążonym parametru e_x^{DF}.

Zmienna losowa występująca we wzorze (18) ma rozkład dwumianowy ponieważ jest sumą niezależnych zmiennych losowych o rozkładzie zero-jedynkowym. Stąd dla każdego przedziału $[x,x+n_x)$ zmienna losowa $\sum_{j=1}^{n_iN_i}Y_{ij}(t_{ij})$ ma rozkład $B(_{n_i}N_i,_{n_x}p_x)$. Stąd $V(_{n_i}\hat{\pi}_i)=\frac{_{n_x}p_x(1-_{n_x}p_x)}{_{n_i}N_i}$ w każdym przedziale $[x,x+n_x)$.

Wariancja estymatora $\hat{e}_x^{DF}$ wyraża się następująco:

$$V(\hat{e}_x^{DF}) = \frac{1}{l_x^2} \sum_{i \in A_x} \frac{{}_{n_x}p_x(1-{}_{n_x}p_x)}{{}_{n_i}N_i} {}_{n_i}L_i^2 \tag{20}$$

Z przeprowadzonego badania wynika, że e_x jest zgodnym i nieobciążonym estymatorem parametru e_x^{DF}. Efektywności estymatora nie badano.

Literatura

[1] Balicki A., *Analiza przeżycia i tablice wymieralności*, PWE,Warszawa 2006.
[2] Błaszczyszyn B., Rolski T., *Podstawy matematyki ubezpieczeń na życie*, WNT, Warszawa 2004.
[3] Imai K., Soneji S., *On the estimation of disability-free life expectancy: Sullivan method and its extension*, "Journal of the American Statistical Association" 2007/01/31.
[4] Jagger C., Cox B.,Le Roy S., and the EHEMU team, *Health expectancy calculation by the Sullivan Method: A practical guide*, EHEMU Technical Report September 2006.
[5] Kędelski M., Paradysz J., *Demografia*, Wydawnictwo AE w Poznaniu, Poznań 2006.
[6] Newton I., Bowers, Jr. Hans, Gerber U., James C. Hickman, Donald A., Hickman, Donald A. Jones, Cecil J. Nesbit, *Actuarial mathematics, the Society of Actuaries*, 1986.
[7] Ostasiewicz W., *Propedeutyka probabilistyki*, Wydawnictwo AE we Wrocławiu. Wrocław 2000.

Individual and cohort life tables for heterogeneous populations

Summary

The novelty of this paper is the method for construction individual life tables accounting for individual characteristics which determines the vitality. The presented method is based on the frailty models which were developed at the end of 1970s. Illustrative life tables are constructed with the use of Gompertz's function for modeling intensity of morality and gamma distribution for modeling individual propensity for mortality.