ZESZYTY NAUKOWE Instytutu Zarządzania i Marketingu Akademii im. Jana Długosza w Częstochowie
Seria: Pragmata tes Oikonomias 2009, z. III

Stanisława Ostasiewicz

Ocena prawdopodobieństwa przeżycia dla danych niepełnych

1. Wprowadzenie

W przypadku szacowania parametrów Tablic Trwania Życia (TTŻ) próba losowa ustalana jest i pobierana w tym samym momencie czasu (por. [3]). Próbę stanowi grupa noworodków (generacja), która obserwowana jest od momentu urodzenia aż do momentu zgonu ostatniej jednostki populacji, czyli około 100 lat. Liczebność populacji sukcesywnie zmniejsza się, co jest powodowane wymieraniem populacji. W tym badaniu śmierć jest jedynym powodem ubytku jednostki z populacji. Parametry tablic obliczane są na podstawie liczby zgonów i liczby dożywających określonego wieku.

W pracy niniejszej rozpatrywane jest badanie (por. [2, 5]), w którym ustalony jest okres jego trwania, natomiast badane jednostki mogą się zgłaszać na badanie w dowolnym momencie jego trwania i obserwowane są do czasu śmierci lub zakończenia badania, w zależności – co zdarzy się wcześniej. Zdarzeniem obserwowanym jest czas przeżycia. Powodem zmniejszania się populacji w tym przypadku jest śmierć, ale również może to być wycofanie się jednostki z badania z innych niż śmierć powodów. Taką jednostkę tracimy z pola widzenia, a informację o jej przeżyciu nazywa się informacją utraconą lub cenzurowaną. Można by oczywiście z takich obserwacji zrezygnować, ale byłaby to strata. Traktuje się je więc jako niepełne i wykorzystuje, ponieważ niosą z sobą pewne informacje. W badaniach medycznych taka sytuacja zdarza się bardzo często, na przykład jeśli badany jest czas przeżycia populacji, która poddana była skomplikowanej operacji.

W tego rodzaju badaniach oceniane jest prawdopodobieństwo przeżycia rocznych przedziałów lub funkcja przeżycia.

Niech N_x oznacza liczbę jednostek, które dożyły do x rocznicy udziału w badaniu, czyli do początku przedziału $(x, x+1)$, i niech p_x oznacza prawdo-

podobieństwo, że jednostka, która dożyła do momentu x, przeżyje 1 rok, to znaczy do końca przedziału $(x, x+1)$. Grupę osób dożywających momentu x można podzielić na dwie rozłączne podgrupy o liczebnościach odpowiednio równych m_x i n_x. Pierwszą grupę tworzą osoby, które przystąpiły do badania wcześniej niż $x+1$ lat przed jego zakończeniem, czyli są obserwowane przez cały okres czasu $(x, x+1)$. Do drugiej grupy należą badani, którzy przystąpili do badania mniej niż $x+1$ lat przed zakończeniem badania, ale dłużej niż x lat. Dla tych osób data zakończenia obserwacji poprzedza ich $x+1$ rocznicę przystąpienia do badania. Spośród m_x osób grupy pierwszej, s_x przeżyje do końca przedziału i d_x umrze, nie dożywając $x+1$ rocznicy. W grupie drugiej początkowa liczba badanych wynosiła n_x. Spośród tych badanych, $d_x^{'}$ umrze przed końcem przedziału i w_x przeżyje do końca przedziału. Ogólna liczba zgonów D_x w przedziale $(x, x+1)$ jest równa $d_x^{'} + d_x$. W grupie, która obserwowana jest przez cały przedział, liczba zgonów wynosi d_x i w grupie, która wcześniej zrezygnowała z badania, wynosi $d_x^{'}$. Zarówno liczba zgonów, jak i liczba dożywających są to zmienne losowe, które będą wykorzystane do estymacji prawdopodobieństwa przeżycia p_x i prawdopodobieństwa zgonu q_x.

Liczba osób dożywających do końca przedziału, spośród osób obserwowanych przez cały okres $(x, x+1)$, oznaczona jest jako zmienna losowa S_x. Zmienna ta ma rozkład dwumianowy $B(m_x, p_x)$. Prawdopodobieństwo, że s_x osób z tej grupy dożyje do końca przedziału, jest równe:

$$P(S_x = s_x) = \binom{m_x}{s_x} p_x^{s_x} (1-p_x)^{m_x - s_x} = \binom{m_x}{s_x} p_x^{s_x} (1-p_x)^{d_x}.$$

Oznaczmy przez $p_x(\frac{1}{2})$ prawdopodobieństwo dożycia do momentu wycofania się osoby z badania, który jest momentem losowym należącym do przedziału $(x, x+1)$. Grupa osób, które zrezygnowały z badania, liczy n_x osób. Zmienną losową, oznaczającą liczbę osób dożywających do momentu wycofania się z badania, oznaczymy W_x. Zmienna ta ma rozkład $B(n_x, p_x(\frac{1}{2}))$. Prawdopodobieństwo, że w_x osób z tej grupy dożyje do momentu wycofania się z badania, jest równe:

$$P(W_x = w_x) = \binom{n_x}{w_x} p_x(\tfrac{1}{2})^{w_x} (1 - p_x(\tfrac{1}{2}))^{d_x^{'}};$$

gdzie $d_x^{'} = n_x - w_x$.

Grupy jednostek wycofujących się z badania i pozostających pod obserwacją przez cały okres badania są niezależne, tak więc rozkład prawdopodobieństwa

liczby dożywających jest dwuwymiarowym rozkładem dwumianowym o niezależnych składowych. Funkcja wiarogodności dla estymacji parametru p_x ma postać:

$$L(x, p_x) = \binom{m_x}{s_x} p_x^{s_x} (1 - p_x)^{d_x} \binom{n_x}{w_x} p_x(\tfrac{1}{2})^{w_x} (1 - p_x(\tfrac{1}{2}))^{d_x'} \qquad (1)$$

Jeżeli do obliczenia $p_x(\frac{1}{2})$ przyjmiemy konkretną procedurę obliczeniową, to otrzymamy funkcję wiarogodności dla estymacji parametru p_x.

Głównym problemem jest wyrażenie prawdopodobieństwa dożycia do momentu wycofania się z badania $p_x(\frac{1}{2})$ przez prawdopodobieństwo p_x przeżycia roku.

2. Metoda aktuarialna

W metodzie tej (por. [1, 2, 5]) nie wyróżnia się dwóch grup badanych: tych obserwowanych przez cały okres badania i tych którzy są obserwowani przez część okresu. Estymator prawdopodobieństwa przeżycia okresu $(x, x+1)$ ma następującą postać:

$$\hat{p}_x = 1 - \frac{D_x}{N_x - \frac{1}{2} W_x}$$

gdzie D_x oznacza ogólną liczbę zgonów w przedziale $(x, x+1)$, N_x liczbę dożywających początku przedziału i W_x liczbę dożywających do końca przedziału jednostek, które wycofały się z badania.

Z postaci estymatora wynika, że jednostki, które wycofują się z badania, obserwuje się tylko przez połowę okresu. Jeżeli $\hat{p}_x$ potraktujemy jako prawdopodobieństwo sukcesu w $N_x - \frac{1}{2} W_x$ próbach, to wariancja estymatora $\hat{p}_x$ określona jest wzorem:

$$V(\hat{p}_x) = \frac{1}{N_x - \frac{1}{2} w_x} \hat{p}_x \cdot \hat{q}_x$$

3. Estymator A

Przyjmuje się założenie (por. [1, 2]), że moment śmierci jednostek biorących udział w badaniu jest losowy i ponadto moment wycofania się z badania też jest losowy z przedziału $(x, x+1)$.

Jeżeli przyjmiemy, że intensywność zgonów wśród osób wycofujących się z badania jest taka sama jak osób biorących udział w badaniu cały czas i ponadto, jest ona stała w całym przedziale, to prawdopodobieństwo $p_x(\frac{1}{2})$ można wyrazić następująco:

$$p_x(\tfrac{1}{2}) = \int_x^{x+1} \exp(-\int_x^t \mu(\tau)d\tau)dt$$

gdzie $\mu(\tau)$ jest to intensywność zgonów osób wycofujących się z badania. Jeżeli

$$\mu(\tau) = \mu_x$$

gdzie μ_x jest intensywnością zgonów osób biorących udział w badaniu cały czas i ponadto intensywność zgonów jest stała w całym przedziale, to

$$\int_x^t \mu(\tau)d\tau = \int_x^t \mu_x d\tau = -\mu_x(t-x)$$

stąd

$$p_x(\tfrac{1}{2}) = \int_x^{x+1} \exp(-(t-x)\mu_x dt = \frac{1}{\mu_x}(1-e^{-\mu_x}) = -\frac{1}{\log p_x}(1-p_x)$$

Korzystając z powyższej zależności między $p_x(\frac{1}{2})$ i p_x oraz z postaci funkcji wiarogodności (por. wzór 1), otrzymujemy:

$$L_A(x, p_x) = p_x^{s_x}(1-p_x)^{d_x+w_x}\ (\log p_x^{-n_x}[(1-p_x)+\log p_x]^{d_x'}$$

Logarytmując powyższą funkcję, a następnie różniczkując ją względem p_x i przyrównując pochodną do zera, otrzymujemy następujące równanie:

$$\frac{s_x}{p_x} - \frac{d_x+w_x}{1-p_x} - \frac{n_x}{p_x \log p_x} + \frac{d_x'(1-p_x)}{[(1-p_x)+\log p_x]p_x} = 0$$

Rozwiązanie tego równania jest estymatorem największej wiarogodności parametru p_x.

Analityczne rozwiązanie tego równania jest bardzo trudne, ale można w sposób prosty znaleźć rozwiązanie numeryczne.

4. Estymator B

Kiedy osoba wycofuje się z badania w sposób losowy, wówczas średni czas jej przebywania w przedziale $(x, x+1)$ wynosi połowę długości tego przedziału (por. [2]), czyli przedział $(x, x+\frac{1}{2})$. Prawdopodobieństwo dożycia do momentu wycofania, które oznacza się $p_x(\frac{1}{2})$, jest następujące:

$$p_x(\tfrac{1}{2}) = \exp(-\tfrac{1}{2}\mu_x) = \exp(-\mu_x)^{\frac{1}{2}} = p_x^{\frac{1}{2}}$$

Przyjmując $p_x^{\frac{1}{2}}$ jako prawdopodobieństwo dożycia do wycofania się z badania i $1-p_x^{\frac{1}{2}}$ jako prawdopodobieństwo zgonu przed wycofaniem się z badania, rozkład zmiennej losowej W_x jest następujący:

$$P(W_x = w_x) = \binom{n_x}{w_x} p_x^{\frac{w_x}{2}} (1-p_x^{\frac{1}{2}})^{d_x'}.$$

Jak widać, liczba dożywających do połowy przedziału ma rozkład dwumianowy $B(p_x^{\frac{1}{2}}, n_x)$.

Oczekiwana liczba dożywających i oczekiwana liczba zgonów są odpowiednio równe:

$$E(W_x) = n_x p_x^{1/2} \text{ i } E(D_x') = n_x(1-p_x^{1/2})$$

Funkcja wiarogodności ma postać:

$$L_B(x, p_x) = p_x^{s_x+(1/2)w_x}(1-p_x)^{d_x}(1-p_x^{1/2})^{d_x'}$$

Estymatorem $\hat{p}_x$ prawdopodobieństwa przeżycia przedziału $(x, x+1)$ jest rozwiązanie następującego równania kwadratowego:

$$((N_x - \tfrac{1}{2}n_x)\hat{p}_x + \tfrac{1}{2}d_x'\hat{p}_x^{1/2} - (s_x + \tfrac{1}{2}w_x) = 0$$

Rozwiązanie to jest następujące (por. [2]):

$$\hat{p}_x = \left[\frac{-\frac{1}{2}d_x' + \sqrt{\frac{1}{4}d_x'^2 + 4(N_x - \frac{1}{2}n_x)(s_x + \frac{1}{2}w_x)}}{2(N_x - \frac{1}{2}n_x)}\right]^2.$$

Jest to ocena prawdopodobieństwo zgonu w przedziale $(x, x+1)$.

Asymptotyczna wariancja estymatora $\hat{p}_x$ dana jest w postaci wzoru:

$$V(\hat{p}_x) = \frac{p_x q_x}{M_x} \quad \text{gdzie} \quad M_x = m_x + n_x(1+\hat{p}_x^{1/2})^{-1}$$

5. Estymator Elvebecka

Jeżeli rozkład prawdopodobieństwa zgonów jest jednostajny w rozpatrywanym przedziale, oznacza to, że liczba dożywających w każdym punkcie przedziału jest funkcją liniową pomiędzy punktami x i $x+1$. W przypadku krótkich przedziałów interpolacja liniowa jest dość dokładna (por. [5, s. 97]).
Prawdopodobieństwo przeżycia do środka przedziału jest następujące:

$$p_x(\tfrac{1}{2}) = 1 - \tfrac{1}{2}q_x = \tfrac{1}{2}(1+p_x)$$

Podstawiając do funkcji wiarogodności określonej wzorem (1) powyższe wyrażenie, otrzymuje się funkcję wiarogodności:

$$L_E(x, p_x) = p_x^{s_x}(1-p_x)^{D_x}(1+p_x)^{w_x}.$$

Na podstawie funkcji tej wyznaczany jest estymator $\hat{p}_x^E$ prawdopodobieństwa przeżycia w przedziale $(x, x+1)$:

$$\hat{p}_x^E = \frac{w_x - D_x + \sqrt{(w_x - D_x)^2 + 4N_x s_x}}{2N_x}.$$

Asymptotyczna wariancja tego estymatora jest równa:

$$V(\hat{p}_x^E) = \frac{\hat{p}_x^E(1-(\hat{p}_x^E)^2)}{(N_x+n_x)[1+\hat{p}_x^E - n_x/(N_x+n_x)]}.$$

6. Estymator Droletta

W metodzie tej (por. [1]) nie rozpatruje się obserwacji cenzurowanych. Obserwacje dotyczące przeżycia i śmierci rozpatruje się w momencie końcowym przedziału $(x, x+1)$. Stąd też $n_x = 0$ i $m_x = N_x$. Funkcja wiarogodności w tym przypadku ma postać:

$$L_D(x, p_x) = p_x^{s_x}(1-p_x)^{d_x}$$

Estymator prawdopodobieństwa przeżycia $\hat{p}_x^D$ i jego wariancja wyrażają się następująco:

$$\hat{p}_x^D = \frac{s_x}{m_x},$$

$$V(\hat{p}_x^D) = \frac{\hat{p}_x^D \hat{q}_x^D}{m_x}.$$

7. Estymator Kaplana–Meiera

Estymator ten (por. [1, 4, 5]) umożliwia dokładniejsze oszacowanie funkcji przeżycia niż estymatory poprzednio omówione. Przy jego konstrukcji wykorzystuje się uporządkowane czasy trwania poszczególnych jednostek biorących udział w badaniu. Przez czas trwania jednostki rozumie się długość okresu, który minął od momentu przystąpienia do badania do momentu śmierci lub do momentu wycofania się z badania. W tym drugim przypadku czas przeżycia nazywa się czasem cenzurowanym. Załóżmy, że w badaniu wzięło udział l_0 osób i że ich obserwowane czasy trwania równe są $t_1, t_2, t_3, ..t_{l_0}$ (niektóre z tych czasów są cenzurowane). Czasy trwania mogą się powtarzać, wówczas uporządkowany ciąg czasów przeżycia jest krótszy i przyjmijmy, że liczy on w elementów. Uporządkowany ciąg czasów przeżycia jest wówczas następujący: $t_0 < t_{(1)} < t_{(2)} < ... < t_{(w)}$ gdzie $t_{(k)}$ oznacza k-ty co do wielkości ukończony czas przeżycia. Uporządkowane czasy przeżycia wyznaczają granice przedziałów, przy czym zakłada się, że zgon następuje na początku przedziału czasu trwania (oznacza to przedziały są lewostronnie otwarte). Oznacza to, że w pierwszym przedziale $< t_0, t_{(1)})$ nie ma zgonów. Pierwszy zgon ma miejsce w momencie $t_{(1)}$. Niech l_i oznacza liczbę jednostek, dla których czas przeżycia jest większy niż $t_{(i)}$ lub równy $t_{(i)}$. Liczbę tych jednostek, dla których zgon nastąpił w momencie $t_{(i)}$ oznacza się $d_{(i)}$. Tak więc estymatorem zgonu w przedziale czasu $(t_i - \varepsilon, t_i)$ może być statystyka $\frac{d_i}{l_i}$, a estymatorem, że osoba nie umrze w tym przedziale, statystyka $\frac{l_i - d_i}{l_i}$. W przedziale $(t_i\ ,\ t_{i+1} - \varepsilon)$ nie ma zgonów, a więc prawdopodobieństwo przeżycia jest równe 1.

Prawdopodobieństwo przeżycia obu przedziałów łącznie, czyli $(t_{(i)} - \varepsilon, t_{(i)} + \varepsilon)$, jest to iloczyn prawdopodobieństw przeżycia przedziałów $(t_i - \varepsilon, t_i >$ i $(t_i\ ,\ t_{i+1} - \varepsilon)$, czyli wynosi $\left(\frac{l_i - d_i}{l_i}\right)$.

Jeśli $\varepsilon \to 0$, to

$$\left(\frac{l_i - d_i}{l_i}\right)$$

jest estymatorem prawdopodobieństwa przeżycia w przedziale $(t_{(i)}, t_{(i+1)} >$.

Estymatorem funkcji przeżycia w całym badanym okresie jest iloczynem wartości funkcji przeżycia w poszczególnych przedziałach $(t_{(i)}, t_{(i+1)} >$ i określony jest następująco:

$$\hat{S}_i^{KM} = \prod_{k=1}^{i}\left(\frac{l_k - d_k}{l_k}\right).$$

Gdzie l_k oznacza łączną liczbę osób dla których czas przeżycia przekracza $t_{(k)}$ lub jest równy $t_{(k)}$, i d_k jest liczbą zmarłych w momencie $t_{(k)}$. Wariancja estymatora Kaplana–Meiera określona jest wzorem:

$$V(\hat{S}_i^{KM}) = \left(\hat{S}_i^{KM}\right)^2 \left\{\sum_{k=1}^{i} \frac{d_k}{l_k(l_k - d_k)}\right\}$$

Sposób konstrukcji estymatora Kaplana–Meiera przedstawiony zostanie na przykładzie czasów zaniku choroby (wyrażonych w tygodniach) u 18 jednostek chorych na chorobę nowotworową (por. [1]). Uporządkowane czasy trwania są następujące:

$$10, 13^*, 18^*, 19, 23^*, 30, 36, 38^*, 54^*, 56^*, 59, 75, 93, 97, 104^*, 107, 107^*, 107^*.$$

Liczby z gwiazdkami oznaczają obserwacje cenzurowane.

Zaobserwowane czasy trwania podzielimy na przedziały tak, że zaobserwowane wartości ukończonego czasu trwania wyznaczają granice tych przedziałów i zgony występują na początku przedziału. Przedziały te są następujące:

$$< 0,10), < 10,19), < 19,30), < 30,36), < 36,59), < 59,75), < 75,93), < 93,97), < 97,107).$$

Wartość estymatora funkcji przeżycia w pierwszym przedziale $<t_0,t_{(1)})$ jest następująca:

$$\hat{S}_1^{KM} = \frac{l_1 - d_1}{l_1} = \frac{18-0}{18} = 1.$$

Oznacza to, że prawdopodobieństwo przeżycia w tym przedziale jest równe 1. Wariancja estymatora jest natomiast równa:

$$V(\hat{S}_1^{KM}) = 1 \cdot \frac{d_1}{l_1(l_1 - 1_1)} = 1\frac{0}{18(18-0)} = 0$$

Prawdopodobieństwo przeżycia przedziału $<10,19)$ oblicza się podobnie.
Jest to liczba obliczona następująco:

$$\frac{l_2 - d_2}{l_2} = \frac{18-1}{18} = \frac{17}{18}.$$

Prawdopodobieństwo przeżycia obu przedziałów (jeżeli jednostki przeżywają poszczególne przedziały niezależnie), czyli wartość funkcji przeżycia poza moment czasu 10, jest iloczynem prawdopodobieństw przeżycia obu tych przedziałów. Czyli:

$$\hat{S}_2^{KM} = \frac{l_1 - d_1}{l_1} \cdot \frac{l_2 - d_2}{l_2} = 1 \cdot \frac{18-1}{18} = \frac{17}{18}$$

Wariancja estymatora $\hat{S}_2^{KM}$ jest następująca:

$$V(\hat{S}_2^{KM}) = \left(\frac{17}{18}\right)^2 \cdot \left\{\frac{d_1}{l_1(l_1 - d_1)} + \frac{d_2}{l_2(l_2 - d_2)}\right\} = \left(\frac{17}{18}\right)^2 (0 + \frac{1}{18(18-1)} = 0{,}892 \cdot 0{,}00327 = 0{,}00291$$

Funkcja przeżycia jest funkcją schodkową przyjmującą takie same wartości w całym przedziale. Wartość ta jest równa wartości funkcji w momencie w którym nastąpił zgon.
Obecnie policzone zostaną wartości estymatorów w punktach 19, 30, 36, 59, 75, 93, 97 i 107. Wartości estymatorów oznaczone będą $\hat{S}_3, \hat{S}_4, \hat{S}_5, \hat{S}_6, \hat{S}_7, \hat{S}_8$

$$\hat{S}_3 = \frac{l_1 - d_1}{l_1} \cdot \frac{l_2 - d_2}{l_2} \cdot \frac{l_3 - d_3}{l_3} = 1 \cdot \frac{18-1}{18} \cdot \frac{15-1}{15} = 1 \cdot \frac{17}{18} \cdot \frac{14}{15} = 0{,}8815$$

$$\hat{S}_4 = \frac{l_1 - d_1}{l_1} \cdot \frac{l_2 - d_2}{l_2} \cdot \frac{l_3 - d_3}{l_3} \cdot \frac{l_4 - d_4}{l_4} = 1 \cdot \frac{18-1}{18} \cdot \frac{15-1}{15} \cdot \frac{13-1}{13} = 1 \cdot \frac{17}{18} \cdot \frac{14}{15} \cdot \frac{12}{13} = 0{,}8137$$

$$\hat{S}_6 = \frac{l_1 - d_1}{l_1} \cdot \frac{l_2 - d_2}{l_2} \cdot \frac{l_3 - d_3}{l_3} \cdot \frac{l_4 - d_4}{l_4} \cdot \frac{l_5 - d_5}{l_5} \cdot \frac{l_6 - d_6}{l_6} = \hat{S}_5 \cdot \frac{8-1}{8} = 0{,}6525$$

$$\hat{S}_7 = \hat{S}_6 \cdot \frac{7-1}{7} = 0{,}5594$$

$$\hat{S}_8 = \hat{S}_7 \cdot \frac{6-1}{6} = 0{,}4662$$

$$\hat{S}_9 = \hat{S}_8 \cdot \frac{5-1}{5} = 0{,}3729$$

$$\hat{S}_{10} = \hat{S}_9 \cdot \frac{3-1}{3} = 0{,}2486$$

Teraz policzone zostaną wariancje estymatorów $\hat{S}_3$ i $\hat{S}_4$.

$$V(\hat{S}_3) = 0{,}8815^2 \left\{ \frac{d_1}{l_1(l_1 - d_1)} + \frac{d_2}{l_2(l_2 - d_2)} + \frac{d_3}{l_3(l_3 - d_3)} \right\} =$$

$$= 0{,}8815^2 \left(0 + \frac{1}{18(18-1)} + \frac{1}{15(15-1)} \right) = 0{,}00624$$

$$V(\hat{S}_4) = 0{,}8137^2 \left\{ \frac{d_1}{l_1(l_1 - d_1)} + \frac{d_2}{l_2(l_2 - d_2)} + \frac{d_3}{l_3(l_3 - d_3)} + \frac{d_4}{l_4(l_4 - d_4)} \right\} =$$

$$= 0{,}8137^2 \left(0 + \frac{1}{18(18-1)} + \frac{1}{15(15-1)} + \frac{1}{13(13-1)} \right) = 0{,}0000901$$

Wariancje pozostałych estymatorów podane zostaną bez obliczeń.

$$V(\hat{S}_5) = 0{,}012$$
$$V(\hat{S}_6) = 0{,}017$$
$$V(\hat{S}_7) = 0{,}0199$$
$$V(\hat{S}_8) = 0{,}02108$$
$$V(\hat{S}_9) = 0{,}0205$$
$$V(\hat{S}_{10}) = 0{,}01943$$

8. Estymator C

Niech t_j oznacza moment zgonu (por. [2]) w przedziale $(x, x+1)$ j-tej osoby $j = 1, 2, 3, ..., D_x$. Ponieważ zgon może mieć miejsce w dowolnym punkcie przedziału, moment zgonu jest zmienną losową ciągłą o funkcji gęstości określonej wzorem:

$$f(t_j, x) = e^{-t_j \mu_x} \cdot \mu_x = e^{-\mu_x \cdot t_j} \cdot \mu_x = -p_x^{t_j} \cdot \ln p_x$$

gdyż

$$e^{-\mu_x} = -p_x \text{ i } \mu_x = \ln p_x,\ 0 \le t_j \le 1.$$

Moment wycofania się z badania z przedziału $(x, x+1)$ i-tej jednostki (który oznaczamy τ_i) spośród w_x jednostek jest również zmienną losową ciągłą przyjmującą wartości z przedziału między zero i jeden. Funkcja gęstości tej zmiennej losowej jest następująca:

$$g(\tau_i, x) = e^{-\tau_i \mu_x} = e^{-\mu_x \cdot \tau_i} = p_x^{\tau_i} \qquad 0 \le \tau_i \le 1 \qquad i = 1, 2, ..., w_x$$

Zakłada się, że intensywność zgonów jest taka sama jak intensywność wycofywania się z badania obserwowanych jednostek. Prawdopodobieństwo tego, że każda jednostka z grupy s_x przeżyje przedział $(x, x+1)$, jest równe p_x. Funkcja wiarogodności dla wejściowej grupy N_x jednostek jest następująca:

$$L_C(x, p_x) = p_x^{s_x} \left(\prod_{i=1}^{w_x} p_x^{\tau_i} \right) \prod_{j=1}^{D_x} (p_x^{t_j} \log p_x) = p_x^{T_x} (\log p_x)^{D_x}$$

gdzie

$$T_x = s_x + \sum_{i=1}^{w_x} \tau_i + \sum_{j=1}^{D_x} t_j \ .$$

Estymator prawdopodobieństwa przeżycia p_x otrzymany na podstawie powyższej funkcji wiarogodności jest następujący:

$$\hat{p}_x = e^{-\frac{D_x}{T_x}} = e^{-\hat{\mu}_x}$$

gdzie

$$\hat{\mu}_x = -\frac{D_x}{T_x}.$$

$\hat{\mu}_x$ jest to estymator intensywności zgonów.
Wariancja estymatora prawdopodobieństwa przeżycia w przedziale $(x, x+1)$ jest równa:

$$Var(\hat{p}_x) = \hat{p}_x^2 (\frac{D_x}{T_x^2}).$$

Inne własności podane są w publikacji [2].

Literatura

[1] Balicki A., *Analiza przeżycia i tablice wymieralności*, Polskie Wydawnictwo Ekonomiczne, Warszawa 2006.
[2] Chiang C.L., *The life table and its aplications*, Robert E. Krieger Publishing Company Malabar, Florida 1984.
[3] Holzer J., *Demografia*, PWE, Warszawa 1999.
[4] Kleinbaum D.G., *Survival analysis*, Springer-Verlag, New York 1996.
[5] *Metody oceny i porządkowania ryzyka w ubezpieczeniach życiowych*, red. S. Ostasiewicz, Wydawnictwo Akademii Ekonomicznek, Wrocław 2000.

Summary

Estimation of the Probability of Survival for Incomplete Data

The aim of this paper is to analyse the basic methods of estimation of probability of survival in the case of follow-up study, when some individuals left the cohort before the end of study.