DZIAŁ NAUKOWY

MATEMATYKA STOSOWANA 8, 2007

Ilona Kopocińska (Wrocław)
Bolesław Kopociński (Wrocław)

Zagadnienie Steinhausa o szacowaniu strat wojennych na podstawie analizy nekrologów prasowych

Józefowi Łukaszewiczowi
w osiemdziesiątą rocznicę Urodzin

Streszczenie. Przedmiotem noty jest próba rekonstrukcji, wzmiankowanej we wspomnieniach Hugona Steinhausa, metody oszacowania strat wojennych w armii Niemieckiej Rzeszy podczas drugiej wojny światowej przy wykorzystaniu ówczesnych nekrologów prasowych. Zakładając dla wielkości rodziny niemieckiej rozdęty rozkład geometryczny i wprowadzając pewne przekształcenia tej zmiennej losowej oraz biorąc pod uwagę rozkład treści nekrologów, wskazujemy na możliwość oszacowania frakcji poległych w armii.
Słowa kluczowe: metoda klepsydr Steinhausa, rozdęty rozkład geometryczny, rozrzedzenie dwumianowe, estymacja parametrów.

1. Wprowadzenie. W przekazach o dziełach Profesora Hugona Steinhausa znajdujemy wzmiankę o sugerowanej przez Niego możliwości oszacowania strat wojennych Niemieckiej Rzeszy poniesionych podczas drugiej wojny światowej, na podstawie analizy treści nekrologów prasowych. Problem był wielokrotnie przywoływany przez Profesora na seminarium zastosowań matematyki we Wrocławiu na przełomie lat pięćdziesiątych i sześćdziesiątych ubiegłego stulecia. Jest o nim wzmianka we *Wspomnieniach i zapiskach* (patrz [4], s. 520): *23 IX 1963. Parę dni temu zjawił się u mnie pan Gdula, któremu dałem jako temat pracy magisterskiej zbadanie strat niemieckich w zabitych, chciałem przekonać się, jak moja metoda klepsydr (Osiczyna, styczeń 1942) wygląda, gdy ma się tych klepsydr więcej. Pan Gdula znalazł we Wrocławiu komplety „Völkischer Beobachter” obejmujące kilkaset klepsydr. W dziele Schirera (The Rise and Fall of the Third Reich) autor podaje liczby z diariusza feldmarszałka Holdera, który objął po Brauchitschu komendę nad całym rosyjskim frontem; liczby te obejmują czas od*

22 VI 1941 do 28 II 1942 roku i tylko front rosyjski (nie licząc Włochów i Rumunów czy też Węgrów). Okazuje się niekonsekwencja tych dat, tak że można raczej kontrolować je moim sposobem, niż przeciwnie.

Nie są nam znane informacje o rozwiązaniu wspomnianego zagadnienia przez Hugona Steinhausa, Edmunda Gdulę czy też kogoś innego. Niniejsza nota jest więc próbą rekonstrukcji matematycznego modelu prowadzącego do pewnego pozytywnego rezultatu, jednakże bez weryfikacji wyników na podstawie danych liczbowych.

Idąc tropem Steinhausa, przyjmując założenia modelu, zobowiązani jesteśmy uwzględnić ograniczone możliwości dostępu do danych demograficznych Niemiec w warunkach okupowanej Polski, co zmusza do daleko idących uproszczeń. Przyjmiemy zatem, naszym zdaniem, rozsądne założenia dotyczące rodziny i reprezentatywność próby nekrologów. Dalej przy opisie probabilistycznym komentujemy zakładaną równość prawdopodobieństw i niezależność rozważanych zdarzeń losowych.

Podstawowym obiektem naszych rozważań jest rodzina niemiecka. Przyjmujemy, że ma ona D dzieci, w tym S synów, spośród których W wcielono do wojska, a z nich Z poległo. Ogólna liczba żołnierzy N może być nieznana – przedmiotem estymacji jest liczba M żołnierzy poległych do ustalonej daty albo lepiej frakcja $z = M/N$ poległych. W tym celu, zgodnie z ideą Steinhausa, mamy wykorzystać nekrologi prasowe. Nekrologi zawierają następujące treści: A_0 – nasz jedyny syn poległ, A_j – nasz j-ty ($j \geq 1$) syn poległ. Odnotujmy, że zdarzenie A_1, czyli „nasz pierwszy syn poległ” zawiera A_0, czyli „nasz jedyny syn poległ”. Wyróżnienie A_0 i wprowadzenie $A_1^* = A_1 \setminus A_0$, co pozostało w naszej pamięci, pochodzi od Steinhausa. Zwiększa ono liczbę zdarzeń, które w praktyce mogą być wzięte pod uwagę.

2. Model. Główną rolę w naszych rozważaniach gra zmienna losowa $G = G_{\alpha,p}$ mająca rozkład rozdęty geometryczny:

$$P(G_{\alpha,p} = 0) = 1 - \frac{\alpha p}{q},$$
$$P(G_{\alpha,p} = n) = \alpha p^n, \;\; 0 < p < 1, \;\; q = 1 - p, \;\; 0 < \alpha < q/p, \;\; n \geq 1.$$

Niech X będzie zmienną losową o rozkładzie $P(X = n) = p_n$, $n \geq 0$. Weźmy pod uwagę operator B z parametrem b, który działa na X dając zmienną B_bX o rozkładzie $P(B_bX = k) = \sum_{n=k}^{\infty} p_n b(k; n, b)$, gdzie $b(k; n, b) = \binom{n}{k} b^k (1-b)^{n-k}$. Łatwo sprawdzić, że $B_aB_bX = B_{ab}X$, $B_bG_{\alpha,p} =^d G_{\alpha',p'}$, gdzie $\alpha' = \frac{\alpha}{q+bp}$, $p' = \frac{bp}{q+bp}$. Odnotujmy, że

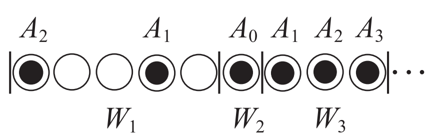
$$P(B_bG_{\alpha,p} = 0) = 1 - \frac{\alpha bp}{q(q + bp)},$$

$$P(B_b G_{\alpha,p} = n) = \frac{\alpha}{q+bp}\Big(\frac{bp}{q+bp}\Big)^n, \quad n \geq 1.$$

Przechodzimy do analizy zagadnienia Steinhausa. Założenia modelu są następujące: za Lotką [1] przyjmujemy, że $D = G_{\alpha,p}$, $S = B_s D$, $W = B_w S =^d B_{sw} D$, $Z = B_z W =^d B_{swz} D$, gdzie $0 < s < 1$, $0 < w < 1$, $0 < z < 1$. Operator B (patrz [3]) zwany rozrzedzeniem dwumianowym, zawiera *implicite* założenie prób Bernoulliego. Zakładamy więc w szczególności, że śmierć spotyka żołnierza niezależnie od liczebności rodziny i liczby żołnierzy służących w armii, należących do owej rodziny. Są to być może daleko idące uproszczenia, ale ich zmiana wymagałaby znajomości nowych, trudnych do wyestymowania parametrów. W szczególności przyjęcie jednakowego w dla braci wynika z braku wiedzy o strukturze wieku w rodzinach.

Ogół żołnierzy numerujemy liczbami $1, 2, \ldots, N$; braciom w wojsku przypisujemy numery kolejne. Zatem żołnierze będący w grupach rodzinnych utworzą dyskretny proces odnowy $1, 2, \ldots, W_1, W_1+1, \ldots, W_1+W_2, W_1+W_2+1, \ldots$, gdzie cykle procesu odnowy $W_1, W_2, \ldots$ są niezależnymi zmiennymi losowymi o rozkładzie G z odpowiednimi parametrami. Każdemu żołnierzowi zabitemu przypisujemy jako markę nekrolog jemu poświęcony. Cykl charakteryzujemy jego długością W oraz markami: liczbą synów S, liczbą zabitych Z i treścią ewentualnego nekrologu A od tej rodziny, losowanego ze wszystkich nekrologów możliwych w tym cyklu. Mamy więc do czynienia z markowanym procesem odnowy (patrz [2]). Marka S jest użyteczna przy analizowaniu nekrologów treści A_0. Zauważmy, że odczytany nekrolog umieszczamy w kontekście stanu jego grupy rodzinnej.

Rys. 1 przedstawia fragment omawianego procesu odnowy. Kółko oznacza żołnierza – poległych zaczerniono. Są tam trzy cykle $W_1 = 5$, $W_2 = 1$, $W_3 = 3$, dla nich $Z_1 = 2$, $Z_2 = 1$, $Z_3 = 3$. Na rysunku umieszczono ewentualne nekrologi. Nekrolog, który do nas dociera, jest losowany spośród tych, które są w cyklu.



Rys. 1. Fragment sekwencji żołnierzy.

Niech $H_j(N)$ oznacza oczekiwaną liczbę nekrologów o treści A_j w zbiorze żołnierzy $\{1, 2, \ldots, N\}$. W [2] znajdujemy asymptotykę tych wielkości. Nam wystarczy pierwszy wyraz podanego tam rozwinięcia asymptotycznego:

$$H_j(N) = \frac{N}{\mathrm{E}(W)} P(A = A_j) + O(1), \quad N \to \infty, \quad j \geq 0.$$

Pozostają więc do obliczenia prawdopodobieństwa $P(A = A_j) = a_j$, $j \geq 0$. Niech $a_1^* = P(A_1^*) = a_1 - a_0$. Oznaczmy przez π prawdopodobieństwo zamieszczenia w prasie nekrologu przez rodziców i odnalezienie jego przez nas. Nie będziemy identyfikowali rodzin, gdyby powtarzały się one w odczytywanych nekrologach. Można sprawdzić (patrz Dodatek 1), że

$$a_0 = \frac{\pi\alpha swzp}{(q+sp)^2}, \quad a_j = \pi\alpha_z \sum_{k=j}^{\infty} \frac{1}{k} p_z^k, \quad j \geq 1, \tag{1}$$

gdzie $\alpha_z = \dfrac{\alpha}{q + swzp}$, $p_z = \dfrac{swzp}{q + swzp}$.

3. Estymacja parametrów modelu. Zbieranie nekrologów daje możliwość oszacowania frakcji zdań określonej treści. Empirycznie możemy bowiem oszacować frakcje $\hat{a}_j = H_j(N)/\sum_{j=1}^{\infty} H_j(N) = a_j/a$, $j \geq 0$, gdzie $a = \sum_{j=1}^{\infty} a_j$. Można sprawdzić (patrz Dodatek 2), że

$$a = \frac{\pi\alpha_z p_z}{q_z}, \text{ gdzie } q_z = 1 - p_z. \tag{2}$$

Odnotujmy, że wyrażenia $\hat{a}_j$, $j \geq 0$, nie zależą od α, π; parametry $s, p/q$ oraz w, z zawsze pozostają w iloczynie, zatem nie da się ich estymować oddzielnie. W analizie naszego problemu, zakładając niewielką dokładność obliczeń, można wziąć $s = 0.5$. Należy przypuścić, że Steinhaus poznał w w jakiś inny sposób. My tutaj przyjęliśmy $w = 0.5$ całkowicie arbitralnie, a następnie w swoich doświadczeniach numerycznych estymowaliśmy parametry p, z przyjmując jako kryterium minimum kwadratów odchyleń prawdopodobieństw $\hat{a}_j$ od ich wartości obserwowanych.

Wiadomo, że Hugon Steinhaus w realnych zastosowaniach matematyki cenił prostotę formuł, nawet kosztem ich optymalności. Przy znajomości $s = 0.5$, p, w, Steinhausowskiego typu estymatorem z może być

$$\tilde{z} = \frac{\tilde{p}_z}{w\tilde{r}(1-\tilde{p}_z)}, \text{ gdzie } \tilde{p}_z = 2\frac{\hat{a}_2 - \hat{a}_3}{\hat{a}_1 - \hat{a}_2} \text{ jest estymatorem } p_z,$$

natomiast estymator $\tilde{r}$ parametru $p/2q$ spełnia równanie

$$\hat{a}_0 = \frac{1 + w\tilde{z}\tilde{r}}{(1+\tilde{r})^2}.$$

Uogólnienie. Dotychczas zakładaliśmy, że nekrologi obserwujemy od początku wojny do określonego dnia. Można rozważać również badania okresowe, dotyczące zbierania nekrologów ogłoszonych w pewnym przedziale czasu. Problem badań okresowych poruszamy w Dodatku 2.

4. Próby numeryczne. Aby zbadać praktyczną użyteczność obliczeniową przytoczonych wzorów, przyjęliśmy pewien (nieuzasadniony jednak realnymi przesłankami) zestaw wartości parametrów: $p = 0.75$, $\alpha = 0.25$, $s = w = 0.5$, $z = 0.3$, $\pi = 1$. Obliczyliśmy: $a_0 = 0.0360$, $a_1^* = 0.1297$,

$a_2 = 0.0157$, $a_3 = 0.0020$, $a_4 = 0.0003$, przy czym $a = 0.1837$. Widzimy, że a_4 jest empirycznie raczej niedostępne. Namiastkę danych empirycznych utworzyliśmy modyfikując nieco te liczby. Sugerując się możliwością emocjonalnej reakcji rodziców w ekstremalnych dla nich przypadkach, zwiększyliśmy liczby a_0 i a_3 biorąc po 5% od sąsiednich prawdopodobieństw, a_4 zaniedbaliśmy. Przyjęliśmy więc dalej jako empiryczne prawdopodobieństwa $a\hat{a}_0 = 0.0378$, $a\hat{a}_1 = 0.1279$, $a\hat{a}_2 = 0.0156$, $a\hat{a}_3 = 0.0021$. Ustalmy $s = w = 0.5$. Estymując p, z otrzymaliśmy: $p = 0.744$, $z = 0.320$, przy czym teraz jest $a_0 = 0.0377$, $a_1^* = 0.1279$, $a_2 = 0.0162$, $a_3 = 0.0021$, a więc metoda najmniejszych kwadratów estymacji działa sprawnie. Trzeba dodać, że zadanie numeryczne nie jest trudne, albowiem w procedurze obliczeniowej mamy dobre przybliżenie wstępne. Zniekształcenie parametrów po zniekształceniu danych okazało się raczej duże, co daje pewną wskazówkę o skuteczności metody nekrologów w praktyce. Wykorzystaliśmy cztery dane liczby (sumujące się do jedności) do estymowania dwóch parametrów. Ponieważ nie ma praktycznie szans na sensowne oszacowanie a_4, więc możliwości uogólnienia modelu i szacowania większej liczby parametrów są nikłe.

Dodatek 1. *Dowód wzoru* (1). Weźmy pod uwagę zmienną losową S oraz W i Z odpowiednio warunkowane. Prawdopodobieństwo a_j znajdujemy biorąc pod uwagę zdarzenia $\{S = m\}$, $\{W = n|S = m\}$, $\{Z = k|W = n\}$, faktu napisania nekrologu i jego odnalezieniu przez nas. Obliczamy:

$$a_0 = P(S=1)P(W=1|S=1)P(Z=1|W=1)\pi = \frac{\alpha sp}{(q+sp)^2} wz\pi,$$

$$a_j = \sum_{m=j}^{\infty} P(S=m) \sum_{n=j}^{m} P(W=n|S=m) \sum_{k=j}^{n} P(Z=k|W=n) \frac{1}{k}\pi,$$

gdzie $P(W = n|S = m) = b(n; m, w)$, $P(Z = k|W = n) = b(k; n, z)$.

Dalej, wykorzystując operator B, obliczamy

$$\begin{aligned} a_j &= \pi \sum_{n=j}^{\infty} \sum_{m=n}^{\infty} P(S=m) b(n; m, w) \sum_{k=j}^{n} \frac{1}{k} b(k; n, z) \\ &= \pi \sum_{n=j}^{\infty} P(W=n) \sum_{k=j}^{n} \frac{1}{k} b(k; n, z) = \pi \sum_{k=j}^{\infty} \frac{1}{k} \sum_{n=k}^{\infty} P(W=n) b(k; n, z) \\ &= \pi \sum_{k=j}^{\infty} \frac{1}{k} P(Z=k) = \pi \alpha_z \sum_{k=j}^{\infty} \frac{1}{k} p_z^k. \end{aligned}$$

Dodatek 2. *Dowód wzoru* (2). Obliczamy

$$a = a_0 + a_1^* + \sum_{j=2}^{\infty} a_j = \sum_{j=1}^{\infty} a_j = \pi\alpha_z \sum_{j=1}^{\infty}\sum_{k=j}^{\infty} \frac{1}{k} p_z^k$$

$$= \pi\alpha_z \sum_{k=1}^{\infty} \frac{1}{k} \sum_{j=1}^{k} p_z^k = \pi\alpha_z \sum_{k=1}^{\infty} p_z^k = \frac{\pi\alpha_z p_z}{1-p_z}.$$

Dodatek 3. *Badanie okresowe.* Przypuśćmy, że zbieramy nekrologi w pewnym przedziale czasu i niech M_0 będzie liczbą poległych do początku i M_1 będzie liczbą poległych potem do końca tego przedziału. Teraz mamy do wyestymowania dwa interesujące parametry: $z_0 = M_0/N$ i $z_1 = M_1/N$. Nietrudno zauważyć, że wzór (1) przyjmie teraz postaci

$$\begin{aligned} a_0 &= \pi P(S=1)P(W=1|S=1)P(Z_0=0|W=1)\times \\ &\times P(Z_1=1|W=1,Z_0=0) = \frac{\alpha sp}{(q+sp)^2} w(1-z_0)z_1, \end{aligned} \tag{3}$$

$$\begin{aligned} a_j &= \pi \sum_{m=j}^{\infty} P(S=m) \sum_{n=j}^{m} P(W=n|S=m) \sum_{i=0}^{n-j} P(Z_0=i|W=n)\times \\ &\times \sum_{k=j}^{n-i} \frac{1}{k} P(Z_1=k|W=n,Z_0=i),\ j \geq 1, \end{aligned}$$

gdzie

$$P(Z_0=i|W=n) = b(i;n,z_0),\ P(Z_1=k|W=n,Z_0=i) = b(k;n-i,z_1).$$

Stąd, po przekształceniu otrzymujemy

$$a_j = \pi\frac{\alpha}{q+zp} \sum_{k=j}^{\infty} \frac{1}{k}\left(\frac{zp}{q+zp}\right)^k,\ j \geq 1,\ \text{gdzie } z = sw(1-z_0)z_1. \tag{4}$$

W badaniach okresowych mamy więc do czynienia z sytuacją podobną do badań ciągłych, ale będących tu parametrów w iloczynie $w(1-z_0)z_1$ nie da się rozdzielić.

Prace cytowane

[1] A. J. Lotka, *Théorie analytique des associations biologiques II*, Actualités scientifique et industrielles, **780**, Paris 1939 [cytowane za W. Fellerem, *Wstęp do rachunku prawdopodobieństwa*, wyd. II, PWN, Warszawa 1966, s. 127].

[2] I. Kopocińska, *Estimation of multivariate cumulative processes*, Complex Systems, **13**, 2 (2001), 177–183.

[3] I. Kopocińska, B. Kopociński and A. Okulewicz, A class of distributions applicable to the description of the number of nematodes parasitizing birds, *Mathematical Medicine and Biology*, **21** (2004), 35–48.
[4] H. Steinhaus, *Wspomnienia i zapiski*, Aneks, Londyn 1990.

Instytut Matematyczny
Uniwersytet Wrocławski
Plac Grunwaldzki 2/4, 50-384 Wrocław
E-mail: ibk@math.uni.wroc.pl

Hugo Steinhaus problem of estimation of the war casualities on the base of contemporary press obituaries

Abstract. In the memoirs of Professor Hugo Steinhaus the problem of estimation of the war casualties of German Army during the Second World War on the base of contemporary press obituaries is mentioned. Assuming the inflated geometric probability distribution function for the size of German family, introducing some transformations of this random variable and taking into account the probability distribution function of the contens of obituaries we propose in the note a solution of the problem.
Key words: Steinhaus of obituaries method, inflated geometric probability distribution, binomial dilution, parameter estimation.

(*wpłynęło 31 maja 2007 r.*)