DZIAŁ NAUKOWY

MATEMATYKA STOSOWANA 9, 2008

Ryszard Zieliński (Warszawa)

Estymacja frakcji

Streszczenie. W populacji składającej się z N elementów jest nieznana liczba M elementów wyróżnionych. W artykule w przystępny sposób prezentuję różne problemy związane z estymacją frakcji $\theta = M/N$.
Słowa kluczowe: Frakcja, prawdopodobieństwo sukcesu w doświadczeniu Bernoulliego, estymator nieobciążony, estymator o jednostajnie minimalnym błędzie średniokwadratowym, estymator Bayesowski, losowanie warstwowe, randomizowane odpowiedzi, przedział ufności.

1. Wstęp. Ten artykuł ma być popularno-naukową prezentacją tytułowego zagadnienia estymacji frakcji. Obiektem naszego zainteresowania jest ustalony zbiór (statystycy lubią termin populacja, więc i ja będę używał tego terminu), w którym niektóre elementy są jakoś wyróżnione. Sam zbiór (populację) będę oznaczał przez Ω, a zbiór elementów wyróżnionych przez W. Liczbę elementów w populacji Ω będę oznaczał przez N, a liczbę elementów wyróżnionych (liczbę elementów w zbiorze W) przez M. Frakcją nazywam ułamek M/N. Będę również używał terminologii związanej z następującą interpretacją: jeżeli z populacji wylosuję pewien element X, to prawdopodobieństwo zdarzenia polegającego na tym, że jest to jeden z elementów wyróżnionych jest równe $P(W) = M/N$. Dla zwięzłości to prawdopodobieństwo, czyli frakcję, będę oznaczał literą θ. Rozważam sytuację, gdy θ nie jest znane i potrzebujemy je oszacować na podstawie badania reprezentacyjnego: losujemy pewną liczbę n elementów z populacji Ω; w tej próbie losowej zliczamy liczbę elementów wyróżnionych, będę ją oznaczał przez K, i na podstawie tej obserwacji chcemy możliwie dokładnie oszacować (wyestymować) θ.

Postawione zadanie nie jest „wydumane w ciszy gabinetu statystyka-matematyka". Prostym przykładem „z życia" jest zadanie oszacowania frekwencji wyborczej na podstawie badania reprezentacyjnej próby z populacji potencjalnych elektorów albo zadanie oszacowania frakcji zwolenników danej opcji politycznej. W statystycznej kontroli jakości takim zadaniem

jest oszacowanie wadliwości (frakcji sztuk wadliwych) w partii produktów lub w procesie produkcyjnym. Medycyna interesuje się szacowaniem frakcji tych pacjentów z udarem mózgu, u których wcześniej wystąpił określony zespół symptomów. Przypuszczam, że każdy może podać kilka tego rodzaju przykładów.

2. Rozwiązanie podstawowe. Niech K będzie liczbą elementów wyróżnionych w n-elementowej próbie, tzn. liczbą tych elementów X_i w próbie $X_1, X_2, \ldots, X_n$, które spełniają warunek $X_i \in W$. Można to zapisać jeszcze inaczej. Niech ξ będzie zmienną losową, określoną wzorem

$$\xi = \begin{cases} 1, & \text{jeżeli } X \in W, \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Wtedy $K = \sum_{i=1}^{n} \xi_i$ (czasami używa się terminologii: K jest liczbą sukcesów w ciągu $\xi_1, \xi_2, \ldots, \xi_n$ prób Bernoulliego). Wielkość K jest oczywiście zmienną losową o rozkładzie dwumianowym

$$P_\theta\{K = k\} = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad k = 0, 1, \ldots, m. \tag{1}$$

Naturalnym estymatorem nieznanej frakcji θ wyróżnionych elementów w populacji jest K/n, czyli frakcja wyróżnionych elementów w próbie losowej $X_1, X_2, \ldots, X_n$, czyli średnia arytmetyczna liczb $\xi_1, \xi_2, \ldots, \xi_n$. Zanim przystąpimy do urealniania problemu przez wprowadzanie do zadania różnych dodatkowych elementów związanych z różnymi zastosowaniami odnotujmy, że przy tak ogólnym sformułowaniu zadania, jak to, które wyżej przedstawiliśmy, estymator K/n jest optymalny ze względu na cały szereg różnych kryteriów. Oto jego podstawowe własności.

Estymator K/n jest estymatorem nieobciążonym. Oznacza to, że wartość oczekiwana zmiennej losowej K/n jest równa temu, co ten estymator ma szacować:

$$\Big(\forall\theta \in (0,1)\Big)\ E_\theta \frac{K}{n} = \frac{1}{n}\sum_{k=0}^{n} k\binom{n}{k}\theta^k(1-\theta)^{n-k} = \theta.$$

Estymator K/n jest estymatorem największej wiarogodności. Przypomnijmy to pojęcie na konkretnym przykładzie. Przypuśćmy, że w próbie o liczności $n = 10$ zaobserwowaliśmy $K = 2$. Rozumujemy: gdyby $\theta = 0.1$, to, zgodnie z wzorem (1), prawdopodobieństwo zdarzenia losowego $\{K = 2\}$ byłoby równe $P_\theta\{K = 2\} = 0.19371$; gdyby $\theta = 0.5$, to mielibyśmy $P_\theta\{K = 2\} = 0.04395$. W tym sensie wartość 0.1 parametru θ jest bardziej wiarogodna niż wartość 0.5 tego parametru. I konsekwentnie w tym sensie, najbardziej wiarogodna w rozważanym przykładzie jest ta wartość parametru θ, która maksymalizuje $P_\theta\{K = 2\} = \binom{10}{2}\theta^2(1-\theta)^8$, czyli wartość 2/10.

Estymator K/n jest estymatorem uzyskanym metodą momentów: frakcja θ jest wartością średnią zmiennej losowej ξ, a estymator K/n jest wartością średnią tej zmiennej losowej w losowej próbie $\xi_1, \xi_2, \ldots, \xi_n$. Ogólnie: jeżeli interesujący nas parametr jest pewnym funkcjonałem na przestrzeni rozkładów prawdopodobieństwa opisujących badaną populację, to estymatorem uzyskanym metodą momentów jest wartość tego funkcjonału na rozkładzie empirycznym z próby.

Estymator K/n jest estymatorem nieobciążonym o jednostajnie minimalnej wariancji. Wariancja estymatora nieobciążonego jest miarą jego dokładności; opisuje ona to, jak bardzo losowe wartości estymatora koncentrują się wokół estymowanej wartości badanego parametru. Jeżeli nieobciążony estymator parametru θ oznaczymy ogólnie przez $\hat{\theta}$, a jego wariancję przez $Var_\theta(\hat{\theta})$:

$$Var_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2,$$

to zalety estymatora o małej wariancji wynikają natychmiast chociażby z najprostszej nierówności Czebyszewa (Jakubowski i in. 2001)

$$P_\theta\{|\hat{\theta} - \theta| \geq \varepsilon\} \leq \frac{Var_\theta(\hat{\theta})}{\varepsilon^2}. \tag{2}$$

Z definicji, nieobciążony estymator $\tilde{\theta}$ jest estymatorem nieobciążonym o jednostajnie minimalnej wariancji, jeżeli

$(\forall\theta) \quad Var_\theta(\tilde{\theta}) \leq Var_\theta(\hat{\theta}),$ dla wszystkich estymatorów nieobciążonych $\hat{\theta}$.

Estymator K/n jest właśnie takim estymatorem frakcji; wariancja tego estymatora jest równa

$$Var_\theta(\frac{K}{n}) = E_\theta(\frac{K}{n} - \theta)^2 = \frac{1}{n}\sum_{k=0}^{n}\left(\frac{k}{n} - \theta\right)^2 \binom{n}{k}\theta^k(1-\theta)^{n-k} = \frac{\theta(1-\theta)}{n}.$$

Jeżeli w nierówności (2) położymy $\varepsilon = t\sqrt{Var_\theta(\hat{\theta})}$, otrzymamy

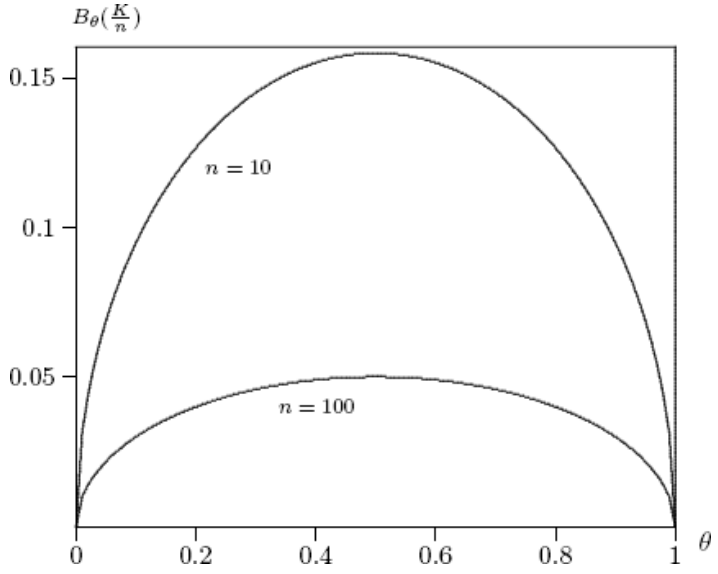
$$P_\theta\left\{|\hat{\theta} - \theta| < t\sqrt{Var_\theta(\hat{\theta})}\right\} \geq 1 - \frac{1}{t^2}$$

i wtedy $\left(\hat{\theta} - t\sqrt{Var_\theta(\hat{\theta})}, \hat{\theta} + t\sqrt{Var_\theta(\hat{\theta})}\right)$ traktuje się jako coś w rodzaju przedziału ufności dla nieznanej frakcji θ, na poziomie ufności $1 - 1/t^2$. Powiedziałem „coś w rodzaju przedziału ufności", bo skoro nie znamy frakcji θ, to tego przedziału nie możemy obliczyć. Niektórzy statystycy upierają się przy tej konstrukcji podstawiając do wzoru na wariancję $\hat{\theta}$ zamiast θ (taki przedział można oczywiście łatwo obliczyć), ale wtedy poziom ufności nie jest już $1 - 1/t^2$. Niektórzy szacują wtedy poziom ufności na podstawie Centralnego Twierdzenia Granicznego, ale nie polecam takiego postępowania (szczególnie przy niezbyt dużych licznościach próby n) bo Centralne

Twierdzenie Granicznego dla schematu Bernoulliego nie zachodzi jednostajnie względem $\theta \in (0,1)$ i taka procedura przy wartościach θ bliskich jednemu z końców tego przedziału prowadzi do bezsensownych wyników. Istnieją łatwe konstrukcje dokładnych przedziałów ufności dające się łatwo realizować za pomocą programów komputerowych łatwo dostępnych w pakietach statystycznych, a nawet w niektórych kalkulatorach kieszonkowych. Opowiem o tym później a teraz wrócę do naszego głównego wątku: estymacji (punktowej) frakcji θ. Naszym podstawowym estymatorem jest zatem $\hat{\theta} = K/n$, a dokładność estymatora będę opisywał za pomocą jego błędu średniokwadratowego $B_\theta(\hat{\theta})$, zdefiniowanego jako pierwiastek z jego wariancji

$$B_\theta(\hat{\theta}) = \sqrt{Var_\theta(\hat{\theta})}.$$

Błąd średniokwadratowy estymatora zależy od (nieznanej!) frakcji θ oraz od liczności próby, co dla $n = 10$ oraz $n = 100$ jest przedstawione na Rys. 1.



Rys. 1

Wydaje się, że postawione na wstępie zadanie estymacji frakcji zostało w pełni rozwiązane: mamy estymator K/n, który jest nieobciążony, wśród wszystkich estymatorów nieobciążonych ma jednostajnie minimalny błąd (umówiliśmy się, że chodzi o błąd średniokwadratowy), a ponadto przemawiają za nim wszystkie te argumenty, które przemawiają na korzyść estymatorów największej wiarogodności oraz estymatorów konstruowanych metodą momentów. Zastanówmy się jednak przez chwilę nad tym „jednostajnie" minimalnym błędem.

3. Jednostajnie minimalna wariancja – czy na pewno o to chodzi? Wiadomo, że frakcja może być jedną z liczb z przedziału $(0,1)$. „Jednostajnie" minimalny błąd estymatora oznacza, że jest on minimalny przy każdej wartości $\theta \in (0,1)$. Ale jeżeli z góry wiemy, że estymowana frakcja mieści się w pewnym przedziale (t_1, t_2), $0 < t_1 < t_2 < 1$, to może nam wcale nie zależeć na małym błędzie estymatora dla frakcji o wartościach poza tym przedziałem. Czy zyskujemy coś na minimalizowaniu błędu estymatora tylko na tym wyróżnionym przedziale?

Powiemy, że estymator $\hat{\theta}_1$ jest lepszy od estymatora $\hat{\theta}_2$ na przedziale (t_1, t_2), jeżeli jego średnia wariancja (a zatem i średni błąd) na tym przedziale jest mniejsza, tzn. jeżeli

$$\int_{t_1}^{t_2} Var_\theta(\theta_1)d\theta < \int_{t_1}^{t_2} Var_\theta(\theta_2)d\theta.$$

Rozważamy estymatory $\hat{\theta} = \hat{\theta}(K)$, które są funkcją liczby K obserwacji wyróżnionych w próbie. Dla takich estymatorów mamy

$$Var_\theta\left(\hat{\theta}(K)\right) = \sum_{k=0}^{n} \left[\hat{\theta}(k) - \theta\right]^2 \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

zatem

$$\int_{t_1}^{t_2} Var_\theta\left(\hat{\theta}(K)\right) d\theta = \sum_{k=0}^{n} \binom{n}{k} \int_{t_1}^{t_2} \left[\hat{\theta}(k) - \theta\right]^2 \theta^k (1-\theta)^{n-k} d\theta$$

$$= \sum_{k=0}^{n} \binom{n}{k} \Big[\hat{\theta}(k)^2\, c(k,n;t_1,t_2) - 2\hat{\theta}(k)\, c(k+1,n;t_1,t_2) + c(k+2,n;t_1,t_2)\Big],$$

gdzie

$$c(k,n;t_1,t_2) = \int_{t_1}^{t_2} \theta^k (1-\theta)^{n-k} d\theta.$$

Minimalizując, dla każdego k oddzielnie, wyrażenia w nawiasach kwadratowych otrzymujemy optymalny estymator w postaci

$$\hat{\theta}(K) = \frac{c(K+1,n;t_1,t_2)}{c(K,n;t_1,t_2)},$$

co łatwo można zapisać za pomocą standardowej, łatwo dostępnej w rożnych numerycznych pakietach komputerowych, niekompletnej funkcji beta

$$\hat{\theta}(K) = \frac{K+1}{n+2} \cdot \frac{I_{t_2}(K+2, n-K+1) - I_{t_1}(K+2, n-K+1)}{I_{t_2}(K+1, n-K+1) - I_{t_1}(K+1, n-K+1)}, \tag{3}$$

gdzie

$$I_x(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt$$

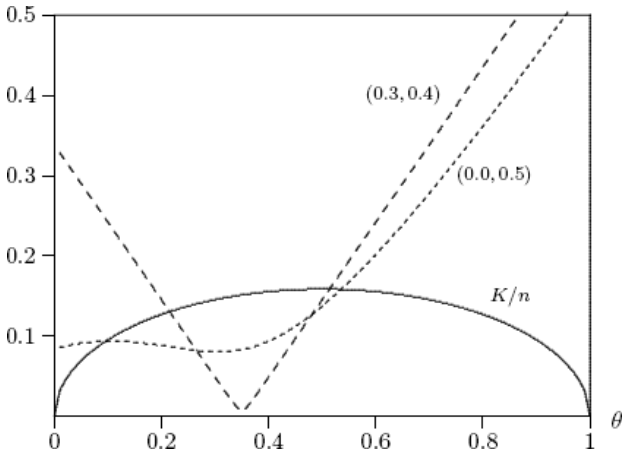
jest zwykłą niekompletną funkcją beta oraz $\Gamma(\alpha)$ jest funkcją gamma: $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$.

Dla ilustracji numerycznej, w pierwszej kolumnie TABELKI podano wszystkie możliwe wartości statystyki K w próbie o liczności $n = 10$, w drugiej kolumnie wartości standardowego estymatora K/n, a w trzeciej i czwartej kolumnie wartości estymatora (3), gdy z góry wiadomo, że estymowana frakcja mieści się w przedziale, odpowiednio, $(0, 0.5)$ lub $(0.3, 0.4)$. Zwróćmy uwagę na to, że zmodyfikowany estymator nigdy nie przyjmuje wartości poza przedziałem (t_1, t_2), dla którego został zaprojektowany.

TABELKA

K	Przedział (t_1, t_2)		
	(0, 1)	(0, 0.5)	(0.3, 0.4)
0	0.0	0.0837	0.3377
1	0.1	0.1644	0.3411
2	0.2	0.2396	0.3466
3	0.3	0.3030	0.3482
4	0.4	0.3519	0.3518
5	0.5	0.3872	0.3554
6	0.6	0.4121	0.3589
7	0.7	0.4296	0.3622
8	0.8	0.4422	0.3652
9	0.9	0.4514	0.3681
10	1.0	0.4583	0.3707

Błąd tych estymatorów kształtuje się tak, jak to przedstawiono na Rys. 2. Zależy on istotnie od tego, jak wybraliśmy przedział (t_1, t_2): im przedział jest krótszy, tym błąd wewnątrz tego przedziału jest mniejszy, ale jeżeli wybrany przez nas przedział nie pokrywa nieznanej, szacowanej wartości frakcji θ, to błąd może być bardzo duży. Dla porównania na tym samym rysunku narysowano także błąd standardowego estymatora K/n.



Rys. 2

Poszukując optymalnego estymatora frakcji w sytuacji, gdy nasza wiedza *a priori* o tej frakcji lokuje ją „gdzieś w przedziale (t_1, t_2)", minimalizowaliśmy

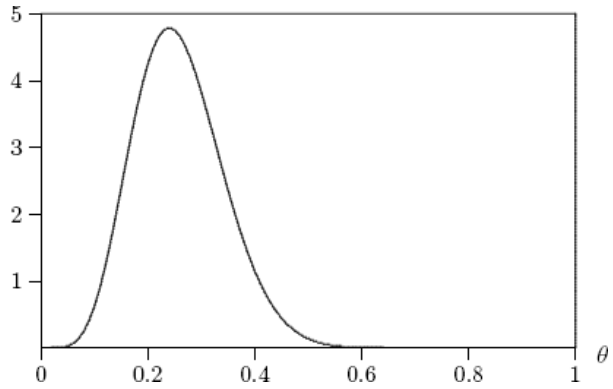
$$\int_{t_1}^{t_2} Var_\theta\left(\hat{\theta}(K)\right) d\theta = \int_0^1 1_{(t_1,t_2)}(\theta) Var_\theta\left(\hat{\theta}(K)\right) d\theta,$$

czyli wariancję uśrednioną wagą $1_{(t_1,t_2)}(\theta)$. Użyłem tutaj oznaczenia:

$$1_{(t_1,t_2)}(\theta) = \begin{cases} 1, & \text{gdy } t_1 \le \theta \le t_2, \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

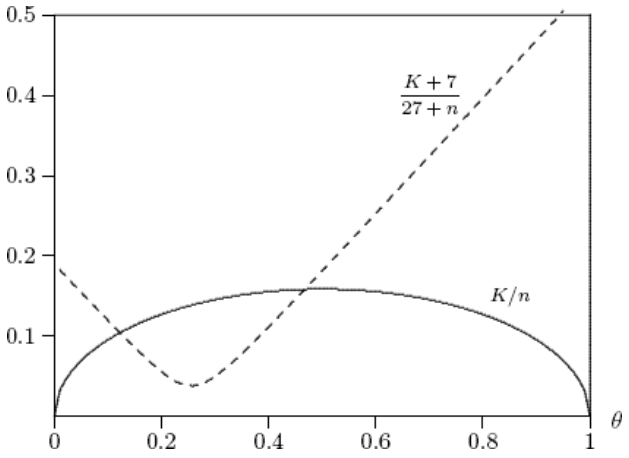
Łatwo można sobie wyobrazić, że moglibyśmy to uśrednienie dokonać dla innej niż $1_{(t_1,t_2)}(\theta)$ wagi, powiedzmy wagi $\pi(\theta), \theta \in (0,1)$, na przykład takiej, jaką przedstawia Rys. 3. Wygodnie jest wybierać wagę spośród gęstości rozkładów prawdopodobieństwa, a w naszym przypadku estymacji frakcji spośród gęstości rozkładu beta

$$\frac{\Gamma(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}. \tag{4}$$



Rys. 3

Rys. 3 przedstawia gęstość (4) dla $\alpha = 7$ i $\beta = 20$). Wybór wagi (4) jest wygodny z tego powodu, że możemy wtedy korzystać z rozbudowanego aparatu *statystyki Bayesowskiej* (Bartoszewicz 1996, DeGroot 1981). W statystyce Bayesowskiej wagę $\pi(\theta)$ interpretujemy jako *rozkład a priori*, a rozwiązaniem naszego zadania, tzn. optymalnym estymatorem frakcji θ, jest wtedy $(K + \alpha)/(\alpha + \beta + n)$ – jest to średnia w *rozkładzie a posteriori*. Błąd średniokwadratowy estymatora Bayesowskiego dla rozkładu *a priori* z Rys. 3 i dla liczności próby $n = 10$ przedstawiono na Rys. 4; dla porównania przedstawiono tam również błąd estymatora standardowego K/n.



Rys. 4

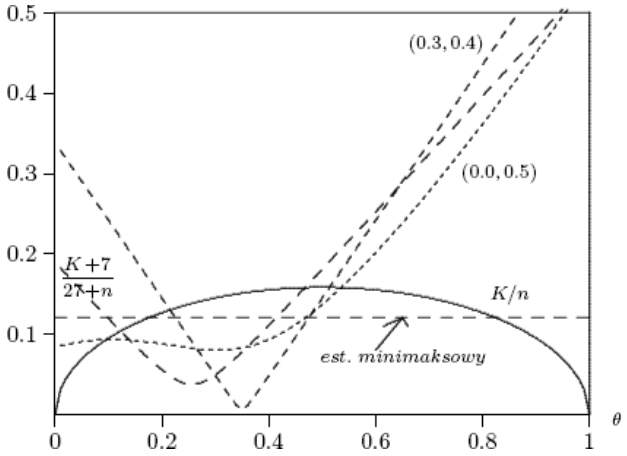
Odnotujmy jeszcze jedno podejście do modelowania naszej wiedzy *a priori* o estymowanym parametrze: w teorii zbiorów rozmytych krzywą z Rys. 3, po przeskalowaniu w taki sposób, żeby przyjmowała wartości w przedziale

$[0,1]$, nazywa się krzywą przynależności θ do przedziału $(0,1)$ – fuzzysetowcy nie lubią jednak odwoływania się do interpretacji probabilistycznych, więc i ja nie będę tutaj wnikał w ich interpretacje.

4. Estymator minimaksowy. Przyjrzyjmy się jeszcze raz błędom estymacji jako funkcji frakcji θ (Rys.2 i Rys.4). Wiemy już, że ten błąd zależy od nieznanej wartości frakcji i że możemy tak manipulować, żeby był on możliwie mały w obszarze o którym wiemy, że zawiera to nieznane θ. Ale jeżeli mamy pecha i prawdziwa, nieznana wartość tego parametru leży daleko poza wybranym przez nas obszarem, błąd może okazać się katastrofalnie duży. Można się przeciwko temu zaasekurować konstruując estymator, którego maksymalny błąd będzie możliwie mały. Takie estymatory nazywają się estymatorami minimaksowymi (Bartoszewicz 1996). W naszym przypadku takim estymatorem jest

$$\frac{K+\frac{1}{2}\sqrt{n}}{n+\sqrt{n}}.$$

Okazuje się, że estymatory minimaksowe mają stały błąd, zależny tylko od n, i że ten błąd jest równy $1/\left(2(1+\sqrt{n})\right)$. Na Rys. 5 pokazujemy wykresy błędów wszystkich rozważanych do tej pory estymatorów oraz estymatora minimaksowego, dla $n = 10$.



Rys. 5

5. Warstwy. Przypomnijmy sformułowanie zadania: ustalony jest pewien skończony zbiór Ω zawierający N elementów, a w nim jest pewna, nieznana liczba M elementów wyróżnionych. Zadanie polega na oszacowaniu frakcji $\theta = M/N$.

Można sobie wyobrazić, że w niektórych badaniach potrafimy rozbić

zbiór Ω na dwa (lub więcej, ale to „więcej" pozostawiam Czytelnikowi) podzbiory ("warstwy") bardziej jednorodne w tym sensie, że w każdym z nich „prawie wszystkie" elementy (a w każdym bądź razie znakomita większość elementów) są wyróżnione albo elementów wyróżnionych jest bardzo mało. Na przykład chcemy w danym społeczeństwie ocenić frakcję osób mających jedno z dwóch możliwych zdań na interesujący socjologa temat i z góry wiemy, że panie mają na ten temat przeważnie inne zdanie niż panowie. Inny przykład: w różnych badaniach sondażowych mieszkańcy małych wsi i małych miast mogą w większości mieć inne zdanie niż mieszkańcy wielkich metropolii. Sformalizujmy to w następujący sposób. Cały badany zbiór Ω zostaje rozbity na dwa rozłączne podzbiory A i B, o licznościach N_A i N_B ($N_A + N_B = N$), z liczbami M_A oraz M_B ($M_A + M_B = M$) elementów wyróżnionych w tych podzbiorach. Oznaczmy przez θ_A oraz θ_B frakcje elementów wyróżnionych w tych podzbiorach Zadanie, jak powiedzieliśmy, polega na oszacowaniu frakcji

$$\theta = \frac{M_A + M_B}{N_A + N_B} = \frac{N_A}{N}\theta_A + \frac{N_B}{N}\theta_B.$$

Z łatwością zauważamy, że naturalnym estymatorem frakcji θ mógłby być estymator

$$\hat{\theta} = \frac{N_A}{N}\hat{\theta}_A + \frac{N_B}{N}\hat{\theta}_B,$$

gdzie

$$\hat{\theta}_A = \frac{K_A}{n_A}, \quad \hat{\theta}_B = \frac{K_B}{n_B}, \quad n_A + n_B = n$$

są znanymi nam już, niezależnymi estymatorami frakcji θ_A i θ_B w warstwach na podstawie prób o licznościach n_A i n_B, w których zaobserwowano, odpowiednio, K_A i K_B elementów wyróżnionych. Dla wariancji estymatora $\hat{\theta}$ otrzymujemy wtedy

$$\begin{aligned} Var_\theta(\hat{\theta}) &= E_\theta\left(\frac{N_A}{N}\hat{\theta}_A + \frac{N_B}{N}\hat{\theta}_B - \theta\right)^2 \\ &= E_\theta\left(\frac{N_A}{N}(\hat{\theta}_A - \theta_A) + \frac{N_B}{N}(\hat{\theta}_B - \theta_B)\right)^2 \\ &= \left(\frac{N_A}{N}\right)^2 \frac{\theta_A(1-\theta_A)}{n_A} + \left(\frac{N_B}{N}\right)^2 \frac{\theta_B(1-\theta_B)}{n_B}. \end{aligned}$$

Przez odpowiednie rozbicie całej populacji Ω na rozłączne zbiory A i B oraz przez odpowiedni wybór wielkości prób z każdego z tych podzbiorów możemy istotnie zmniejszyć tę wariancję, czyli błąd estymacji. Nie będę tutaj rozwijał tego wątku: obszerne informacje na temat optymalnego *losowania warstwowego* można znaleźć w licznych podręcznikach metod reprezentacyjnych, np. Zasępa (1972) lub Bracha (1996). Idealne rozbicie polega na

tym, żeby w jednym z tych zbiorów, powiedzmy w zbiorze A, znalazły się wszystkie elementy wyróżnione i żadne inne: wtedy frakcja $\theta_A = 1$ i wariacja estymatora jest równa zeru. W praktyce jest to raczej niemożliwe, ale rozbicie całej populacji na możliwie jednorodne podzbiory, takie np. jak wyżej kobiety-mężczyźni lub wieś-miasto, w konkretnych przypadkach może doprowadzić do znacznej redukcji błędu. Oszacowanie tego błędu nie musi być jednak bardzo łatwe, jak możemy się o tym przekonać na podstawie przyglądania się post factum różnym wynikom badań sondażowych.

6. Randomizowane odpowiedzi. Wyobraźmy sobie, że celem badania jest oszacowanie w pewnym społeczeństwie frakcji osób, które mają pewną cechę lub popełniły pewien czyn, do których nie mają ochoty przyznać się, a jedyny sposób badania polega na bezpośrednim zapytaniu o to każdej wylosowanej do próby osoby. Trudno oczywiście w takiej sytuacji liczyć na prawdomówność respondenta. Te kłopotliwe pytania mogą dotyczyć np. nadużywania narkotyków, zwyczajów seksualnych, oszustw podatkowych (np. pytanie może brzmieć: czy złożyłeś kiedyś świadomie fałszywe oświadczenie podatkowe). W zadaniu estymacji, które rozważaliśmy do tej pory, pojawia się trudność w obliczeniu liczby K jednostek wyróżnionych w próbie, więc nie możemy zastosować żadnego z omawianych wyżej estymatorów.

Pewien sposób wybrnięcia z pojawiających się tutaj kłopotów zaproponował Warner (1965). W terminach naszego artykułu, pytanie zadane respondentowi brzmiałoby: *czy należysz do grupy wyróżnionej*; dla zwięzłości umówmy się, że brzmi ono *czy jesteś W?* Propozycja Warnera polegała na tym, żeby zadać respondentowi dwa pytania: P1) *czy jesteś W?* oraz P2) *czy nie jesteś W?* Respondent ma wylosować jedno z tych pytać i uczciwie na nie odpowiedzieć, nie informując jednak ankietera, na które pytanie odpowiada. Może on np. rzucić kostką do gry i odpowiedzieć na P1, gdy wyrzucił 1,2,3 lub 4 oczka, lub na P2, gdy wyrzucił 5 lub 6 oczek, przy czym tylko on zna wynik tego rzutu, a więc tylko on wie, na które pytanie odpowiada. Badanie organizuje się w taki sposób, że prawdopodobieństwo wylosowania pytania P1 jest nam znane; oznaczymy je przez p. W tej sytuacji respondent może udzielić uczciwej odpowiedzi, bo z tej odpowiedzi nikt nie będzie mógł niczego wywnioskować o przynależności respondenta do wyróżnionej grupy. Jeżeli θ jest interesującą nas frakcją w populacji, to prawdopodobieństwo usłyszenia odpowiedzi TAK wyraża się oczywistym wzorem

$$P\{TAK\} = p\theta + (1-p)(1-\theta).$$

Niech T oznacza liczbę odpowiedzi TAK w próbie n-elementowej. Wtedy estymatorem prawdopodobieństwa $P\{TAK\}$ jest T/n. Wstawiając ten estymator w miejsce $P\{TAK\}$ w powyższym wzorze i rozwiązując otrzymane

równanie względem θ, otrzymamy estymator – oznaczymy go przez $\hat{\theta}_W$:

$$\hat{\theta}_W = \frac{T/n - (1-p)}{2p-1}.$$

Wariancja tego estymatora wyraża się wzorem

$$Var_\theta(\hat{\theta}_W) = \frac{\theta(1-\theta)}{n} + \frac{p(1-p)}{n(2p-1)^2}.$$

Łatwo jest zauważyć, że wariancja estymatora $\hat{\theta}_W$ jest sumą wariancji w badaniu bezpośrednim bez dodatkowej randomizacji oraz składnika powiększającego tę wariancję o pewną wielkość związaną z randomizacją. Ten drugi składnik możemy, przy ustalonej liczności próby n, minimalizować wybierając p możliwie blisko 0 lub 1, ale taki wybór zbliża badanie do badania bez randomizacji pytań, przez co wprowadza pewien szkodliwy czynnik psychologiczny: respondent mógłby podejrzewać, że ankieter z dużą pewnością orientuje się, na jakie pytanie otrzymuje odpowiedź.

Pewien sposób udoskonalenia estymacji polega na tym, żeby pytanie P2 zastąpić jakimś innym, „neutralnym" pytaniem, takim jednak, dla którego znamy prawdopodobieństwo odpowiedzi TAK; oznaczymy to prawdopodobieństwo przez q. Może to zapytanie brzmieć np. „*Rzuć monetą. Czy otrzymałeś orła?*" (wtedy $q = 1/2$), albo np. „*Czy urodziłeś się w poniedziałek?*" (możemy przypuszczać, że wtedy $q = 1/7$). Teraz prawdopodobieństwo usłyszenia odpowiedzi TAK wyraża się wzorem

$$P\{TAK\} = p\theta + (1-p)q.$$

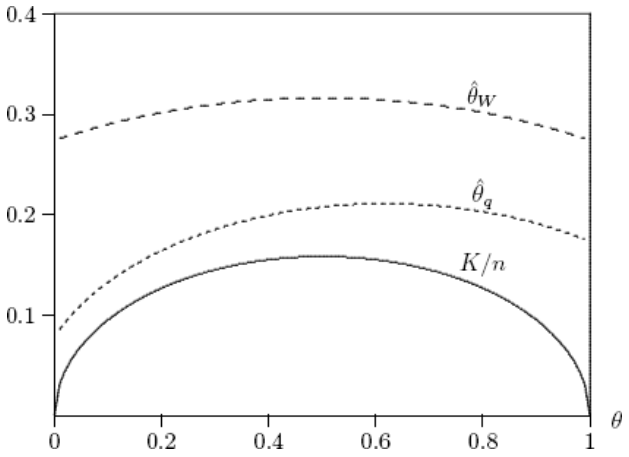
Postępując jak poprzednio otrzymamy estymator – oznaczymy go przez $\hat{\theta}_q$:

$$\hat{\theta}_q = \frac{1}{p}\left(\frac{T}{n} - (1-p)q\right).$$

Wariancja tego estymatora wyraża się wzorem

$$Var_\theta(\hat{\theta}_q) = \frac{\lambda(1-\lambda)}{np^2}, \qquad \lambda = p\theta + (1-p)q.$$

Wariancje estymatorów $\hat{\theta}_W$ i $\hat{\theta}_q$ są oczywiście większe od wariancji estymatora podstawowego K/n. Wielkością błędu tych estymatorów można manipulować przez odpowiedni wybór parametrów p oraz q (jak również, oczywiście, n). Wykresy błędów tych estymatorów dla $p = 0;75$, $q = 1/7$ oraz $n = 10$ przedstawiono na Rys 6. Można starać się tak dobrać te parametry, żeby, jak to już wcześniej robiliśmy, błąd był możliwie mały dla tych wartości frakcji θ, które *a priori* wydają się najbardziej oczekiwane. Jednak na naszym szczeblu ogólności wykładu nic bardziej rozsądnego na ten temat nie umiem powiedzieć.



Rys. 6

7. Przedział ufności. Wróćmy do sprawy przedziału ufności. Chociaż dokładna konstrukcja przedziału ufności dla frakcji jest od dawna znana, to była ona trudna do realizacji przez statystyka-praktyka. Miał on do dyspozycji albo obszerne tablice statystyczne z trudną interpolacją, albo wzory przybliżone, najczęściej oparte na przybliżeniu rozkładu dwumianowego rozkładem normalny. Teraz, gdy każdy praktyk ma na swoim stole komputer, możemy wrócić do rozwiązania dokładnego. Oto to rozwiązanie (Lehmann 1968). Jednostajnie najdokładniejszymi jednostronnymi przedziałami ufności na poziomie ufności 1–α są $\Big(0, b_{K+1,n-K}(1-\alpha)\Big)$ oraz $\Big(b_{K,n-K+1}(\alpha), 1\Big)$, gdzie $b_{p,q}(\gamma)$ jest kwantylem rzędu γ rozkładu beta $B(p,q)$ o gęstości proporcjonalnej do $x^{p-1}(1-x)^{q-1}$. Odpowiednio do tego $\Big(b_{K,n-K+1}(\alpha/2)$, $b_{K+1,n-K}(1-\alpha/2)\Big)$ jest dwustronnym przedziałem ufności na poziomie ufności $1-\alpha$. Kwantyle $b_{p,q}(\gamma)$ są łatwo dostępne w różnych numerycznych pakietach komputerowych, a nawet w bardziej zaawansowanych kalkulatorach kieszonkowych.

8. Wnioski. Wniosek z tego, co do tej pory powiedziałem, jest prosty: możemy w znacznym stopniu panować nad błędem estymacji frakcji i wcale nie musimy ograniczać się do uzyskanego metodą największej wiarogodności lub metodą momentów estymatora nieobciążonego o jednostajnie minimalnej wariancji, tzn. do klasycznego, i często traktowanego jako „jedynie słusznego", estymatora K/n. Dokładne przedziały ufności możemy łatwo obliczać bez uciekania się do ciągle i ciągle sugerowanych w różnych podręcznikach mało dokładnych, a czasami bezsensownych przybliżeń przez rozkład normalny.

Literatura

Podaję tylko te prace, które w artykule bezpośrednio cytowałem. Liczne szczegóły na temat zagadnień wyżej prezentowanych łatwo jest znaleźć w literaturze, w tym szczególnie polecam google. Może w tym być przydatna wskazówka, że używanemu przeze mnie terminowi „błąd" odpowiada tam „mean square error", terminowi „losowanie warstwowe" – „stratified sampling", a „randomizowanym odpowiedziom" – „randomized response".

[1] Bartoszewicz, J. (1996): Wykłady ze statystyki matematycznej. Warszawa, PWN
[2] Bracha, Cz. (1996): Teoretyczne podstawy metody reprezentacyjnej. Warszawa, WNT
[3] DeGroot, M.H. ((1981): Optymalne decyzje statystyczne. Warszawa, PWN
[4] Jakubowski J., Sztencel, R. (2001): Wstęp do teorii prawdopodobieństwa. Wyd. SCRIPT, Warszawa
[5] Lehmann, E.L. (1968): Testowanie hipotez statystycznych, Warszawa, PW
[6] Warner, S. (1965): Randomized response: a survey technique for eliminating evasive answer bias. JASA, March 1965, 63–69
[7] Zasępa, R. (1972): Metoda reprezentacyjna. Warszawa, PWE

Ryszard Zieliński
Instytut Matematyczny PAN
ul. Śniadeckich 8
00-956 Warszawa 1, Poland
E-mail: R.Zielinski@impan.gov.pl

Estimating proportion

Abstract. A population of N elements contains an unknown number M of marked units. Problems of estimating the fraction $\theta = M/N$ are discussed. The well known standard solution is $\hat{\theta} = K/n$ which is the uniformly minimum variance unbiased estimator, maximum likelihood estimator, estimator obtained by the method of moments, and in consequence it shares all advantages of such estimators. In the paper some versions of the estimator are considered which are more adequate in real situations. If we know in advance that the unknown fraction lies in a given interval (t_1, t_2) and we consider an estimator $\hat{\theta}_1$ as better than the estimator $\hat{\theta}_2$ if the average of its mean square error is smaller on that interval, then the optimal estimator is given by (3). The values of the estimator for $(t_1, t_2) = (0, 0.5)$ and for $(t_1, t_2) = (0.3, 0.4)$ in a sample of size $n = 10$ if the number of marked units in the sample equals K, are given in the table TABELKA and the mean square errors of these estimator, versus the error of the standard estimator $\hat{\theta} = K/n$ are presented in Rys. 2. Averaging the mean square error with a weight function, for example such as in Rys.3, gives us the Bayesian estimator with the mean square error like in Rys. 4 (for $n = 10$). If in some real situations we are interested in minimizing the mean square error "in the worst possible case", the adequate is the minimax estimator. Another situation appears if the population can be divided in some more homogenous subpopulations, for example in two subpopulations with fractions of marked units close to zero or close to one in each of them. Then stratified sampling is more effective; then the mean square error of estimation may be significantly reduced. In the paper the problem of randomized

responses is also presented, very shortly and elementarily. The problem arises if a unit in the sample can not be for sure recognized as “marked” or “not marked” and that can be done with some probability only. The situation is typical for survey interview: it allows respondents to respond to sensitive issues (such as criminal behavior or sexuality) while remaining confidential. The final section of the paper is devoted to some remarks concerning the confidence intervals for the fraction. The exact optimal solution is well known for mathematicians but it is probably not very easy for statistical practitioners to follow all theoretical details, and typically confidence interval based on asymptotic approximation of the binomial distribution by a normal distribution are used. That is neither sufficiently exact nor correct. The proper and exact solution is given by quantiles of a suitable Beta distribution which are easily computable in typical statistical and mathematical computer packages.

Key words: Fraction, probability of success in Bernoulli scheme, unbiased estimator, uniformly minimum variance estimator, Bayesian estimator, stratified sampling, randomized response, confidence interval.

(*wpłynęło 12 lutego 2007 r.*)