DZIAŁ NAUKOWY

MATEMATYKA STOSOWANA 10, 2009

Ryszard Zieliński (Warszawa)

Przedział ufności dla frakcji

Streszczenie. Przedziały ufności zostały wymyślone przez Jerzego Spławę–Neymana w 1934 [15]. Praktyczne zastosowanie teorii Neymana do przedziałowej estymacji prawdopodobieństwa sukcesu w schemacie Bernoulliego (parametru rozkładu dwumianowego) stwarzało jednak pewne trudności zarówno jeśli chodzi o ich konstrukcję (rozkład dyskretny!), jak i o ich numeryczne obliczanie. Jako panaceum wymyślono *asymptotyczne przedziały ufności* oparte na przybliżaniu rozkładu dwumianowego rozkładem normalnym: konstrukcja i rachunki stają się bardzo proste. Kłopot polega na tym, że w przypadku skończonej próby pojawiają się wtedy trudności z wyznaczeniem przedziału ufności na postulowanym poziomie ufności. Obecnie powszechny dostęp do komputerów i licznych prostych kalkulatorów „kieszonkowych” z funkcjami statystycznymi umożliwia łatwą realizację dokładnych konstrukcji Neymana.

Słowa kluczowe: frakcja, prawdopodobieństwo sukcesu w schemacie Bernoulliego, przedział ufności, przedział Walda, przedział asymptotyczny, najdokładniejsze przedziały ufności, przedziały jednostronne, przedziały dwustronne.

1. Wstęp. W poprzednim artykule *Estymacja frakcji* [21] przedstawiłem problem estymacji punktowej parametru θ w rozkładzie dwumianowym

$$P_\theta\{S_n = k\} = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad k = 0, 1, \ldots, n; \quad 0 < \theta < 1.$$

Niniejszy artykuł ma być popularno-naukową prezentacją tytułowego zagadnienia estymacji przedziałowej parametru θ. Artykuł jest w zasadzie adresowany do polskiego Czytelnika, więc większość cytowanej literatury jest polska, łatwo dostępna nie tylko w specjalistycznych bibliotekach matematycznych.

Przedziały ufności zostały wymyślone przez Jerzego Spławę-Neymana w 1934 [15]. Precyzyjny opis idei i konstrukcji można znaleźć np. u Bartoszewicza [2], Craméra [5], Zubrzyckiego [22] lub Lehmanna [13]. Obliczanie przedziału ufności dla prawdopodobieństwa sukcesu w schemacie Bernoulliego początkowo stwarzało pewne trudności numeryczne. Dla praktycznych aplikacji potrzebne były tablice lub nomogramy, które pojawiły się prawie natychmiast po wynalezieniu przedziałów ufności. Nie każdy miał takie tablice pod ręką, a gdy miał – nie zawsze potrafił się nimi posługiwać (niełatwa interpolacja!). Jako łatwy środek zastępczy wymyślono *asymptotyczne przedziały ufności* (patrz niżej), w swojej podstawowej wersji znane także jako przedziały Walda (patrz np. Magiera [14], s. 171). Rewolucja komputerowa z końca ubiegłego tysiąclecia pozwala dzisiaj odrzucić wszystkie te nieudane uproszczenia i również w praktycznych aplikacjach w ekonomii, medycynie, technice itd, obliczać dokładnie to, czego od przedziałów ufności oczekujemy.

2. Problem. Zmienna losowa X ma rozkład Bernoulliego z prawdopodobieństwem sukcesu θ, jeżeli

$$P_\theta\{X=1\}=\theta=1-P_\theta\{X=0\},\quad 0<\theta<1. \tag{1}$$

Niech $X_1, X_2, \ldots, X_n$ będzie ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie Bernoulliego z prawdopodobieństwem sukcesu θ, tzn. n-elementową próbą losową z rozkładu (1). Wtedy $S_n=\sum_{j=1}^n X_j$ jest minimalną i zupełną statystyką dostateczną, co oznacza, że wszystkie informacje o parametrze θ są zawarte w tej statystyce. W konsekwencji wszystkie funkcje obserwacji $X_1, X_2, \ldots, X_n$ będziemy traktowali jak funkcje statystyki S_n. Interesuje nas przedziałowa estymacja parametru θ, o którym wiemy tylko to, że „leży gdzieś w przedziale $(0,1)$".

Definicja. Losowy przedział $\left(\underline{\theta}(S_n), \overline{\theta}(S_n)\right)$ nazywamy *przedziałem ufności dla parametru θ na poziomie ufności γ*, jeżeli

$$P_\theta\{\underline{\theta}(S_n)\le\theta\le\overline{\theta}(S_n)\}\ge\gamma\quad \text{dla każdego } \theta\in(0,1). \tag{2}$$

Ta definicja wymaga komentarza.

Taką definicję jak wyżej podana można znaleźć u Lehmanna [13] w rozdz. III.5 (s. 104, definicja (17) rodziny zbiorów ufności), u Bartoszewicza [2] w rozdz. V.9 (s. 296, Def. 9.1), u Niemiry [16] w rozdz. 6 (s. 151, Def. 6.1), u Trybuły [19] w rozdz. III.13 (s. 179), u Magiery [14] w rozdz. *Teoria estymacji* (s. 107, 190).

W niektórych podręcznikach w formalnej definicji przedziału ufności zamiast (2) pojawia się warunek

$$P_\theta\{\underline{\theta}(S_n)\le\theta\le\overline{\theta}(S_n)\}=\gamma\quad \text{dla każdego } \theta\in(0,1). \tag{2'}$$

[Gajek i in. [7] rozdz. 4.5 (s. 82), Krzyśko [12] rozdz. 2.6 (s. 131, Def. 2.11), Plucińska i in. [17] rozdz. 5.9 (s. 268, Def. 5.62)] Pojawiają się jednak wtedy kłopoty. U Gajka i in. [7] kilka wierszy dalej na tej samej stronie mamy „uniwersalny przedział ufności" oparty na nierówności Czebyszewa, który w sensie podanej tam definicji nie jest przedziałem ufności. Krzyśko [12] swoje konstrukcje przedziałów ufności opiera na funkcjach centralnych i może konstruować zdefiniowane przez siebie przedziały ufności dopóty, dopóki te funkcje mają rozkłady ciągłe; wpada w konflikt ze swoją definicją, gdy konstruuje przedział ufności dla frakcji. Plucińska i in. [17] modyfikuje swoją definicję, gdy natrafia na zmienne losowe typu skokowego (podaje wtedy nową definicję 5.65; wprowadza w ten sposób dwie definicje tego samego obiektu). Fisz [6] jest bardziej pedantyczny: podaje definicję z nierównością „$\geq \gamma$", a zaraz w następnym wierszu komentuje, że w przypadku zmiennych losowych ciągłych można pisać „$=\gamma$" zamiast „$\geq \gamma$" (s. 509). Silvey [18] używa „$=\gamma$", gdy mówi o zbiorach ufności na poziomie ufności γ, ale także „$\geq \gamma$", mówiąc wtedy „na poziomie ufności co najmniej γ".

Czasami przedziały ufności są wprowadzane w sposób opisowy, bez jawnego formułowania definicji, ale za to z obszerniejszą interpretacją i przykładowymi konstrukcjami (Cramér [5] w rozdz. XI.34, Zubrzycki [22] w rozdz. VIII.50, Klonecki [10] w rozdz. 10, Koronacki i in. [11] w rozdz. 3.3). Można jednak znaleźć i taki oto tekst na temat przedziału ufności dla średniej rozkładu normalnego $N(\mu, \sigma^2)$ o znanej wariancji:

Zaobserwowawszy próbę losową $X_1, X_2, \ldots, X_n$, czyli mając realizację tej próby $x_1, x_2, \ldots, x_n$, możemy obliczyć realizację średniej w próbie, $\bar{x}$ i podać przedział ufności dla μ na poziomie ufności $1-\alpha$

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{3.22}$$

...Przedział (3.22) nie jest już przedziałem losowym, jest zaś zwykłym przedziałem na osi liczbowej i albo zawiera nieznaną liczbową wartość średnią μ, albo nie... Ścisłe znaczenie sformułowania „zadana doza przekonania", w statystyce zastępuje się pojęciem „zadanego poziomu ufności".

Chociaż ten cytat pochodzi z książki o statystyce *dla studentów kierunków technicznych i przyrodniczych*, potraktujmy to jako „tekst z innej bajki" i wróćmy do spraw poważnych.

Niektóre ze spotykanych w literaturze definicji są formalnie niepoprawne. W statystyce matematycznej mamy do czynienia z *przestrzeniami statystycznymi*, w których określona jest rodzina prawdopodobieństw, oznaczana zwykle przez $\mathcal{P}$ lub, gdy chcemy rozmawiać o parametrach, przez $\{P_\theta, \theta \in \Theta\}$, gdzie Θ jest zbiorem możliwych wartości parametru, a θ elementem tego zbioru. Nie jest więc formalnie poprawna definicja typu "*Przedział losowy $[L, U]$ o takiej własności, że*

$$P\big(L \leq \theta \leq G\big) \geq 1-\alpha,$$

nazywamy przedziałem ufności dla parametru θ." Dla formalnej poprawności wystarczyło by zamiast P napisać P_θ i dopisać kwantyfikator *dla każdego θ*. Posługiwanie się symbolem P w statystyce, bez wyraźnego wskazania o które $P \in \mathcal{P}$ chodzi, brak kwantyfikatora czy chodzi o jakieś jedno, szczególne P, czy o każde $P \in \mathcal{P}$, uważam za formalną, a nawet szkodliwą niepoprawność. Musimy ucząc innych statystyki stale pamiętać, że „prawdopodobieństwo" nie jest jakimś samodzielnym bytem, ale że w statystyce jest dużo różnych „prawdopodobieństw" i chodzi właśnie o to, żeby jakoś zidentyfikować to prawdopodobieństwo, które rządzi naszą próbą.

Reasumując ten fragment artykułu poświęcony definicji deklaruję, że wszędzie dalej będę rozważał przedziały ufności w sensie definicji (2). Za przyjęciem właśnie tej definicji przemawiają dwa następujące argumenty

Primo. Ponieważ zmienna losowa S_n jest dyskretna, w celu skonstruowania przedziału ufności $\left(\underline{\theta}(S_n), \overline{\theta}(S_n)\right)$ spełniającego postulat (2') potrzebna jest dodatkowa randomizacja [2]. Otrzymujemy wtedy elegancką konstrukcję teoretyczną z kłopotliwymi zastosowaniami w praktyce.

Secundo. Jeżeli $\left(\underline{\theta}(S_n), \overline{\theta}(S_n)\right)$ jest przedziałem ufności w sensie definicji (2) i jeżeli $d(n), g(n) \geq 0$, to $\left(\underline{\theta}(S_n)-d(n), \overline{\theta}(S_n)+g(n)\right)$ też jest przedziałem ufności. Mamy więc dużo przedziałów ufności i możemy wśród nich wybierać.

3. Przedział ufności. Jak już wspomniałem we Wstępie, przedziały ufności wymyślił Jerzy Spława-Neyman. Neyman [15] pisze, że rozwiązanie problemu estymacji, o którym mówił, „*consists in determining certain intervals, which I propose to call the confidence intervals*". W tej podstawowej pracy dokładnie opisał konstrukcję przedziału ufności, w szczególności dla frakcji, z czym możemy się teraz łatwo zapoznać dzięki, np., Cramérowi [5] lub Zubrzyckiemu [22]. Zacytuję ogólną, przejrzystą i świetnie nadającą się do dydaktyki nawet na elementarnym poziomie konstrukcję przedziału ufności podaną przez Zubrzyckiego [22]; w poniższym cytowaniu używam oryginalnych oznaczeń Zubrzyckiego, więc poszczególne symbole mogą oznaczać coś innego niż w podstawowym tekście tego artykułu:

Konstrukcja przedziałów ufności... jest bardzo ogólna i przy pewnych założeniach co do ciągłości rozkładów da się powtórzyć dla dowolnego parametru. Można ją też stosować w przypadku kilku parametrów jednocześnie i budować dla nich obszary ufności.

Niech bowiem X będzie wielowymiarową przestrzenią euklidesową punktów $x = (x_1, \ldots, x_n)$ reprezentujących wyniki obserwacji. Niech dalej Ω będzie przestrzenią wartości parametru θ (liczbowego lub wektorowego) wyznaczającego w X rozkład o gęstości $f_\theta(x)$. Ustalmy α z przedziału $0 < \alpha < 1$ i dla każdego $\theta \in \Omega$ wybierzmy zbiór $S_\theta \subset X$, taki że

$$\int\limits_{S_\theta} f_\theta(x)dx = \alpha.$$

Rozważmy teraz w przestrzeni $X \times \Omega$ zbiór D tych wszystkich punktów (x, θ), *dla których jednocześnie* $\theta \in \Omega$ *i* $x \in S_\theta$. *Wówczas (porównaj rys. 1) dla ustalonego* $\theta \in \Omega$ *zbiór*

$$\{(x, \theta) : x \in S_\theta\}$$

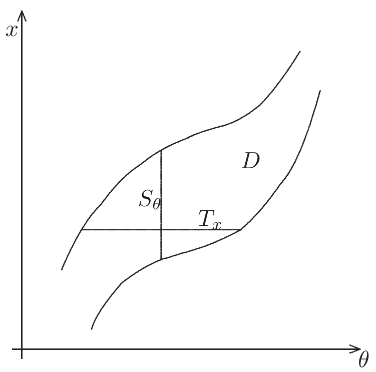
jest przekrojem zbioru D *równoległym do osi* x. *Zbior* D *ma tę własność, że niezależnie od tego, czy parametr* θ *ma ustaloną wartość, czy też uważamy go za zmienną losową o jakimś rozkładzie prawdopodobieństwa, losowy punkt* (x, θ) *będzie należał do* D *z prawdopodobieństwem* α. *A teraz zapiszmy przynależność punktu* (x, θ) *do* D *inaczej, biorąc pod uwagę przekroje zbioru* D *równolegle do osi* θ. *0znaczmy*

$$T_x = \{\theta : (x, \theta) \in D\}.$$

Wówczas trzy zapisy

$$\begin{gathered} \theta \in \Omega, \quad x \in S_\theta, \\ (x, \theta) \in D, \\ x \in X, \quad \theta \in T_x \end{gathered}$$

określają na trzy sposoby przynależność punktu (x, θ) *do zbioru* D. *Wobec tego* T_x *są poszukiwanymi przez nas przedziałami ufności o poziomie ufności* α, *mającymi tę własność, że niezależnie od tego, czy* θ *jest ustalone, czy losowe, z ustalonym prawdopodobieństwem* α *losowy przedział* T_x, *odpowiadający obserwacji* x, *zawiera wartość parametru* θ, *określającą rozkład, według którego losowano* x.



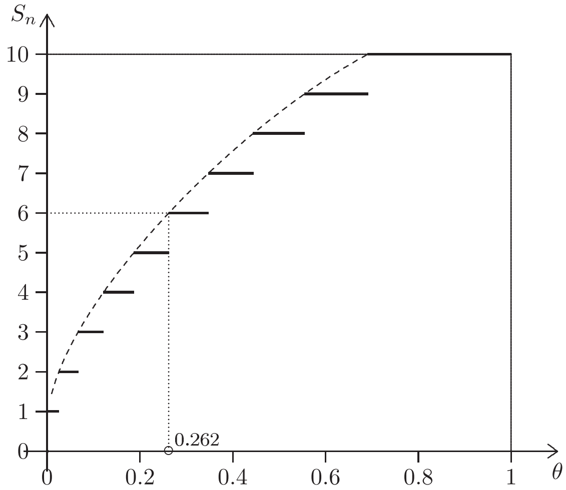
Rys. 1. Ogólna konstrukcja przedziału ufności.

Tu koniec cytatu. Pozostaje tylko wykonać odpowiednie rachunki. Pokażę dokładnie (w polskim piśmiennictwie nie udało mi się znaleźć opisu tego algorytmu), jak przedstawioną wyżej konstrukcję ogólną realizuje się w przypadku (dwustronnego) przedziału ufności na poziomie ufności γ dla parametru θ w rozkładzie dwumianowym – będę musiał uwzględnić fakt, że jest to rozkład dyskretny. Zacznę od granicy dolnej tego przedziału (tzn.

od górnej krzywej na Rys. 1). Wyznaczę ją jako najmniejszą (dlaczego?) funkcję $(0,1) \ni \theta \to d_\gamma(\theta) \in \{0,1,\ldots,n\}$ taką, że dla każdego ustalonego $\theta \in (0,1)$

$$P_\theta\{S_n \le d_\gamma(\theta)\} \ge \frac{1+\gamma}{2}.$$

Na przykład, dla $n = 10$ oraz $\gamma = 0.95$, funkcja $d_\gamma(\theta)$ wygląda jak krzywa schodkowa na Rys. 2



Rys. 2. Konstrukcja przedziału ufności dla frakcji ($n = 10, \gamma = 0.95$)

Znajdujemy, np. że jeżeli zaobserwowano $S_{10} = 6$, to wartość dolnej granicy dwustronnego przedziału ufności na poziomie ufności $\gamma = 0.95$ wynosi 0.262. Numerycznie dolną granicę tego przedziału wyznaczamy jako najmniejsze θ takie, że $P_\theta\{S_{10} \le 6\} \ge 0.975$. Rachunki bardzo się upraszczają, gdy skorzystamy z tożsamości

$$\sum_{j=0}^{k} \binom{n}{j} \theta^j (1-\theta)^{n-j} = B(n-k, k+1; 1-\theta), \quad k = 0,1,\ldots,n,$$

gdzie $B(\alpha,\beta;t)$ jest wartością dystrybuanty zmiennej losowej beta $B(\alpha,\beta)$ o parametrach (α,β), w punkcie $t \in [0,1]$:

$$B(\alpha,\beta;t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^t x^{\alpha-1}(1-x)^{\beta-1} dx.$$

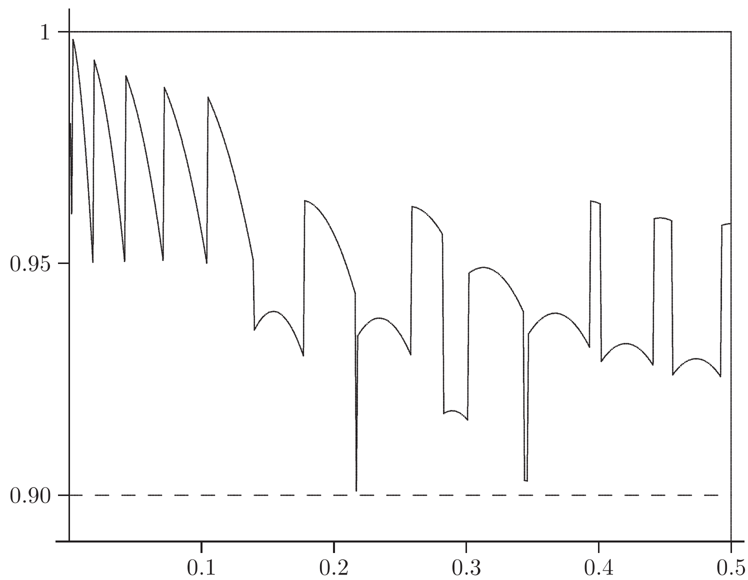
Wtedy $d_\gamma(\theta)$ jest rozwiązaniem, względem θ, równania

$$B(n - S_n + 1, S_n; 1-\theta) = \frac{1+\gamma}{2}.$$

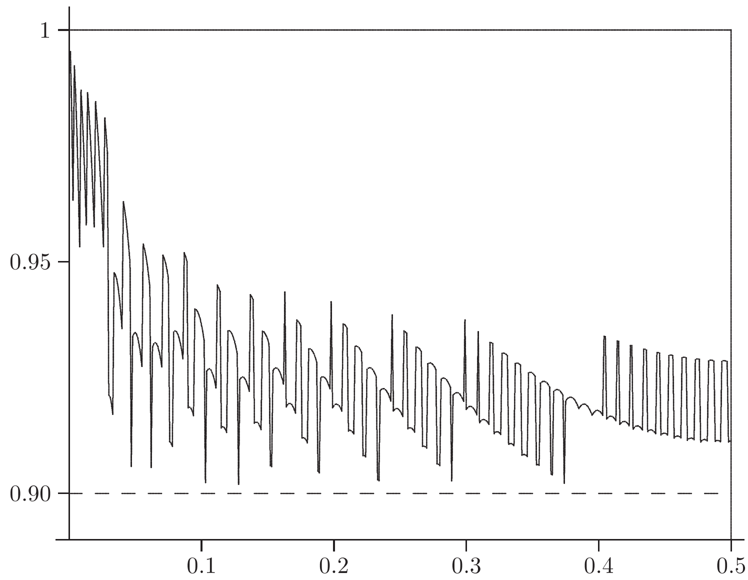
Inaczej mówiąc, $d_\gamma(\theta)$ jest kwantylem $B^{-1}(S_n, n-S_n+1; \frac{1-\gamma}{2})$ rzędu $(1-\gamma)/2$ rozkładu beta $B(S_n, n-S_n+1)$. Skorzystałem tu ze znanych związków $B(p,q;t) = 1-B(q,p;1-t)$. Krzywą $d_\gamma(\theta)$ pokazano na Rys. 2 linią przerywaną. Dolną granicą dwustronnego przedziału ufności na poziomie ufności γ jest więc $B^{-1}(S_n, n-S_n+1; \frac{1-\gamma}{2})$ (por. [2], s. 301). Górną granicę przedziału ufności wyznacza się analogicznie. Ostatecznie, dwustronny przedział ufności na poziomie ufności γ ma postać

$$\left(B^{-1}\Big(S_n, n-S_n+1; \frac{1-\gamma}{2}\Big),\ B^{-1}\Big(S_n+1, n-S_n; \frac{1+\gamma}{2}\Big)\right). \tag{3}$$

Ten przedział ufności, będę go nazywał przedziałem Neymana, spełnia warunek (2) Definicji: dla żadnej wartości θ prawdopodobieństwo jej pokrycia nie jest mniejsze od założonej wartości γ, ale na skutek dyskretności rozkładu dwumianowego, dla niektórych wartości parametru θ może być większe, nawet znacznie, od postulowanego poziomu ufności γ. Ilustrują to Rys. 3 i Rys. 4. Wszystkie przedziały, o których mówię w tym artykule, mają tę własność, że prawdopodobieństwa pokrycia θ i $1-\theta$ są sobie równe, więc wszystkie rysunki prezentują te prawdopodobieństwa tylko dla $\theta \in (0, 1/2)$.



Rys. 3. Prawdopodobieństwo pokrycia przedziałem Neymana ($n = 20,\ \gamma = 0.9$)



Rys. 4. Prawdopodobieństwo pokrycia przedziałem Neymana
($n = 100,\ \gamma = 0.9$)

Interesowałem się *dwustronnymi* przedziałami ufności postaci (2), takimi jakie konstruuje się na Rys. 1; przedziały jednostronne można otrzymać przyjmując $\underline{\theta}(S_n) \equiv 0$ lub $\overline{\theta}(S_n) \equiv 1$. Nie jest to jednak sprawa czysto techniczna: jednostronne przedziały ufności dla frakcji, konstruowane w wyżej opisany sposób, są *jednostajnie najdokładniejszymi przedziałami ufności* na danym poziomie ufności, czego nie da się ogólnie powiedzieć o przedziałach dwustronnych (Lehman [13], Bartoszewicz [2]); istnieje tu pełna analogia do istnienia lub nieistnienia testów jednostajnie najmocniejszych w problemie testowania jednostronnych lub dwustronnych hipotez o parametrze rozkładu dwumianowego.

Łatwo jest sobie wyobrazić, że w połowie ubiegłego wieku praktyczne obliczenia przedziału ufności (3) mogły sprawiać duże trudności. Pierwszymi, którzy wykonali rachunki umożliwiające praktyczne zastosowanie teorii Neymana byli Clopper i in. [4]. W cytowanym artykule podają nomogramy, pozwalające na obliczanie dwustronnych przedziałów ufności na poziomie ufności $\gamma = 0.95$ i $\gamma = 0.99$ (i niektórych przedziałów jednostronnych). Skonstruowano też tablice: najobszerniejsze, jakie udało nam się znaleźć, umieściliśmy w naszym zbiorze [20]. To jednak było w tych odległych czasach, gdy komputery jeszcze nie trafiły pod strzechy: jak dzisiaj oblicza się przedziały ufności (3) mówię w ostatniej części tego artykułu. W połowie XX wieku nie było jeszcze takich łatwych możliwości, w związku z czym wymyślono pewien sposób dla ominięcia trudności numerycznych: *asymptotyczne przedziały ufności.*

4. Asymptotyczne przedziały ufności. Jak wiadomo, odpowiednio unormowana statystyka S_n ma asymptotycznie rozkład normalny. Dokładniej, dla każdego $\theta \in (0,1)$ i dla każdego $x \in (-\infty, \infty)$

$$P_\theta\left\{\frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)/n}} \le x\right\} \to \Phi(x), \quad n \to \infty, \tag{4}$$

gdzie $\Phi(x)$ oznacza wartość dystrybuanty rozkładu normalnego $N(0,1)$ w punkcie x oraz, dla zwięzłości zapisu i w zgodzie z powszechnie przyjętymi oznaczeniami, $\hat{\theta}_n = S_n/n$. W kontekście naszego problemu, standardowo przyjęta interpretacja wzoru (4) brzmi: dla „dużych" n zmienna losowa $(\hat{\theta}_n - \theta)/\sqrt{\theta(1-\theta)/n}$ ma „w przybliżeniu" rozkład normalny $N(0,1)$. Kłopot polega na tym, że to „duże" i „przybliżenie" zależy od nieznanej wartości parametru θ, którą właśnie estymujemy.

Przyjrzyjmy się konstrukcji przedziałów ufności opartych na tym asymptotycznie normalnym przybliżeniu i popatrzmy, co z tego wychodzi. Są tu dwie szkoły.

Jedna szkoła uznaje, że $(\hat{\theta}_n - \theta)/\sqrt{\theta(1-\theta)/n}$ ma asymptotyczny rozkład normalny $N(0,1)$ i na tej podstawie uznaje, że dla „dużych n" i dla każdego $\theta \in (0,1)$, mamy „w przybliżeniu",

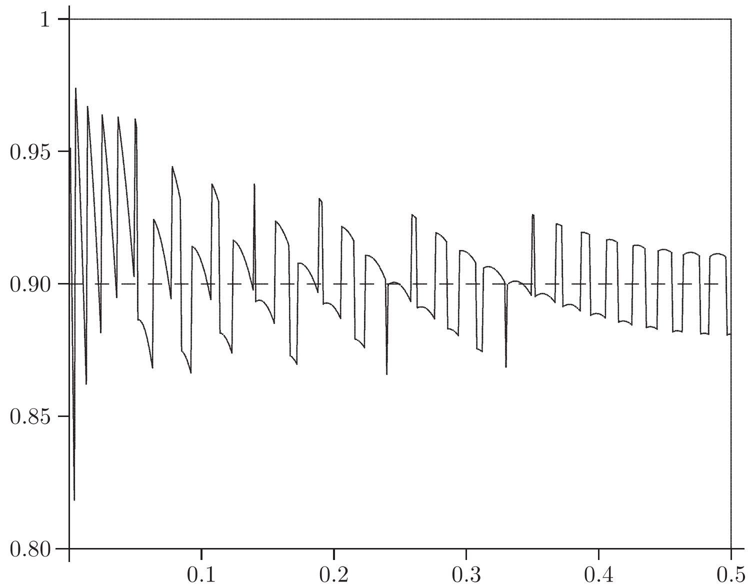
$$P_\theta\left\{\hat{\theta}_n - z_{(1+\gamma)/2}\sqrt{\frac{\theta(1-\theta)}{n}} \le \theta \le \hat{\theta}_n + z_{(1+\gamma)/2}\sqrt{\frac{\theta(1-\theta)}{n}}\right\} = \gamma, \tag{5}$$

gdzie z_β jest kwantylem rzędu β rozkładu normalnego $N(0,1)$.

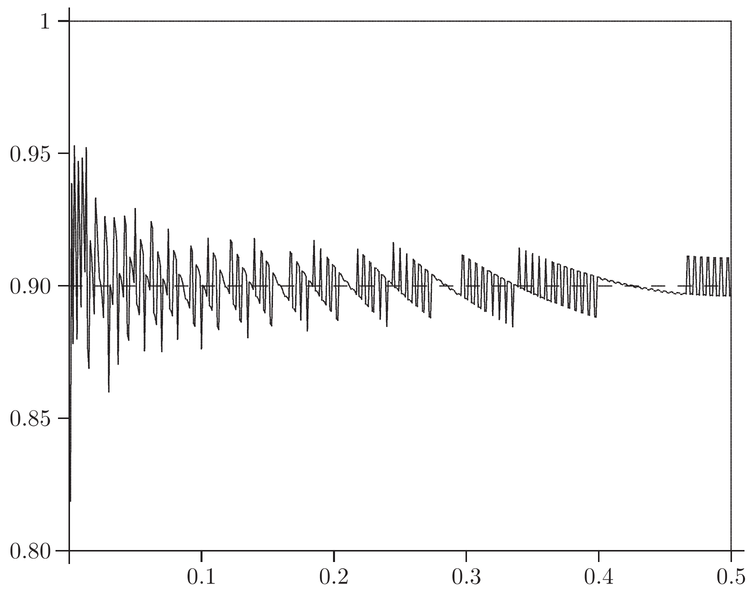
Rozwiązując nierówność w nawiasie klamrowym w (5) względem θ, otrzymujemy "asymptotyczny przedział ufności" (Cramér [5] s. 492, Fisz [6] s. 512, Niemiro [16] s. 155, Trybuła [19] s. 184, Krzyśko [12] s. 162):

$$\begin{aligned}\Bigg(&\frac{n}{n+z^2_{(1+\gamma)/2}}\left[\hat{\theta}_n + \frac{z^2_{(1+\gamma)/2}}{2n} - z_{(1+\gamma)/2}\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n} + \left(\frac{z_{(1+\gamma)/2}}{2n}\right)^2}\right],\\ &\frac{n}{n+z^2_{(1+\gamma)/2}}\left[\hat{\theta}_n + \frac{z^2_{(1+\gamma)/2}}{2n} + z_{(1+\gamma)/2}\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n} + \left(\frac{z_{(1+\gamma)/2}}{2n}\right)^2}\right]\Bigg).\end{aligned} \tag{6}$$

Inna szkoła uznaje, że $(\hat{\theta}_n - \theta)/\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)/n}$ ma asymptotyczny rozkład normalny $N(0,1)$ i wtedy, dla każdego $\theta \in (0,1)$, mamy „w przybliżeniu"



Rys. 5. Prawdopodobieństwo pokrycia przedziałem (6)
($n = 50,\ \gamma = 0.9$)



Rys. 6. Prawdopodobieństwo pokrycia przedziałem (6)
($n = 200,\ \gamma = 0.9$)

$$P_\theta\left\{\hat\theta_n - z_{(1+\gamma)/2}\sqrt{\frac{\hat\theta_n(1-\hat\theta_n)}{n}} \le \theta \le \hat\theta_n + z_{(1+\gamma)/2}\sqrt{\frac{\hat\theta_n(1-\hat\theta_n)}{n}}\right\} = \gamma,$$

co prowadzi do przedziału ufności

$$\left(\hat\theta_n - z_{(1+\gamma)/2}\sqrt{\frac{\hat\theta_n(1-\hat\theta_n)}{n}}, \hat\theta_n + z_{(1+\gamma)/2}\sqrt{\frac{\hat\theta_n(1-\hat\theta_n)}{n}}\right) \tag{7}$$

(Gajek i in. [7] s. 85, Grzegorzewski i in. [8] s. 113, Kala [9] s. 58, Koronacki i in. [11] s. 211, Krzyśko [12] s. 163).

Okazuje się, że można jeszcze inaczej. We wzorze (5) występuje pierwiastek, zależny od nieznanego parametru θ, co komplikuje konstrukcję przedziału ufności. Trudność tę pozwala ominąć spostrzeżenie, że

$$\sqrt{\frac{\theta(1-\theta)}{n}} \le \frac{1}{2\sqrt{n}}$$

i stąd przedział ufności

$$\left(\hat\theta_n - \frac{z_{(1+\gamma)/2}}{2\sqrt{n}}, \quad \hat\theta_n + \frac{z_{(1+\gamma)/2}}{2\sqrt{n}}\right) \tag{8}$$

(Klonecki [10] s. 189, Kala [9] s. 58).

Przyjrzyjmy się dokładniej dwu pierwszym propozycjom; ostatnią pozostawiam bez komentarza.

Przede wszystkim odnotujmy, że autorzy propagujący asymptotyczne przedziały ufności zwykle deklarują, że ich rozwiązania są tylko przybliżone, a czasami podają dodatkowe warunki ich stosowalności.

Podając przedział ufności (6), Fisz [6] (s. 512) i Niemiro [16] (s. 155) mówią, że próba ma być duża (Fisz [6]: *Ograniczymy się do dużych prób*, Niemiro [16]: *W praktyce, jeśli n jest „odpowiednio duże", oczekujemy, że nierówność (2) jest w przybliżeniu spełniona*). Trybuła [19] (s. 184) postuluje, że „n musi być dostatecznie wielkie ($n \ge 50$)". Krzyśko [12] (s. 163) mówi po prostu, że jest to przybliżony przedział ufności. Prawdopodobieństwo pokrycia tym przedziałem wartości parametru $\theta \in (0, 1)$ dla $n = 50$, a nawet dla $n = 200$, pokazują Rys. 5 i Rys. 6. W myśl definicji (2) to po prostu nie są przedziały ufności na założonym poziomie ufności!

Przedziały (7) są też przedziałami „przybliżonymi": Gajek i in. [7] (s. 85) oraz Grzegorzewski i in. [8] (s. 113) postulują, żeby $n \ge 100$. Kala [9] i Krzyśko [12] mówią tylko, że jest to przedział przybliżony. Koronacki i in.

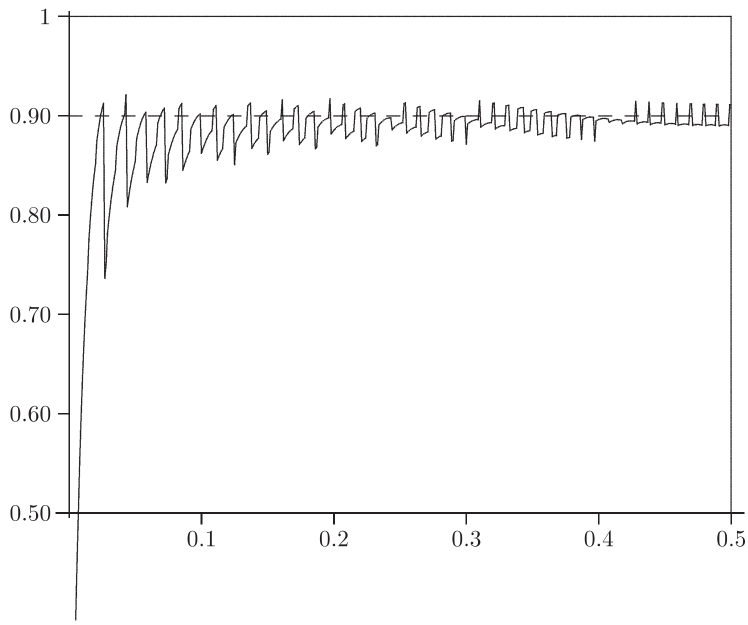
[11] (s. 211) pisze

$$P\left\{\hat{\theta}_n - z_{(1+\gamma)/2}\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} \leq \theta \leq \hat{\theta}_n + z_{(1+\gamma)/2}\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}\right\} \approx \gamma$$

i podobnie Niemiro [16] (s. 155) dla przedziału (6)

$$P_\theta\left\{\frac{n}{n+z^2_{(1+\gamma)/2}}\left[\hat{\theta}_n+\frac{z^2_{(1+\gamma)/2}}{2n}-z_{(1+\gamma)/2}\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}+\left(\frac{z_{(1+\gamma)/2}}{2n}\right)^2}\right]\leq\theta\right.$$
$$\left.\leq\frac{n}{n+z^2_{(1+\gamma)/2}}\left[\hat{\theta}_n+\frac{z^2_{(1+\gamma)/2}}{2n}+z_{(1+\gamma)/2}\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}+\left(\frac{z_{(1+\gamma)/2}}{2n}\right)^2}\right]\right\}\approx\gamma,$$

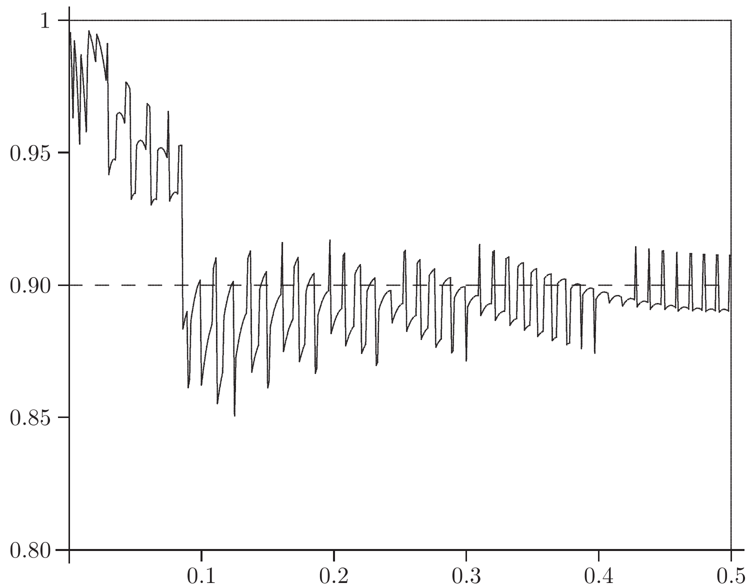
ale nie precyzują co oznacza ten podwójny wężyk „≈". Prawdopodobieństwo pokrycia przedziałem (7) wartości parametru $\theta \in (0,1)$ dla $\gamma = 0.9$ i $n = 100$ pokazuje Rys. 7.



Rys. 7. Prawdopodobieństwo pokrycia przedziałem (7)
($n = 100,\ \gamma = 0.9$)

Dla przedziałów (7) prawdopodobieństwo pokrycia maleje do zera, gdy $\theta \to 0$ oraz gdy $\theta \to 1$. Nie jest to nowym odkryciem: w wielu podręcznikach można znaleźć komentarz, że ten przedział można stosować, gdy $n\theta \geq 5$ oraz $n(1-\theta) \geq 5$. Takie zalecenie istotnie zmienia cały nasz problem estymacji: w naszym modelu statystycznym przestrzenią parametru θ jest przedział

$(0, 1)$. Przy przedstawionej sugestii przestrzenią parametrów (?) staje się $[5/n, 1 - 5/n]$. Oryginalny pomysł prezentuje Koronacki i in. [11] (s. 149): stosować ten przedział wtedy, gdy $n\hat{\theta}_n \geq 5$ oraz $n(1-\hat{\theta}_n) \geq 5$; w pozostałych przypadkach stosować przedział Neymana. Prawdopodobieństwo pokrycia dla tak zmodyfikowanego przedziału pokazuje Rys. 8. Jak widać, to nadal nie jest przedział ufności na postulowanym poziomie ufności.



Rys. 8. Prawdopodobieństwo pokrycia przedziałem Koronackiego-Mielniczuka ($n = 100,\ \gamma = 0.9$)

Jeszcze inny pomysł na „poprawienie" przedziału asymptotycznego zaproponował Baran [1]:

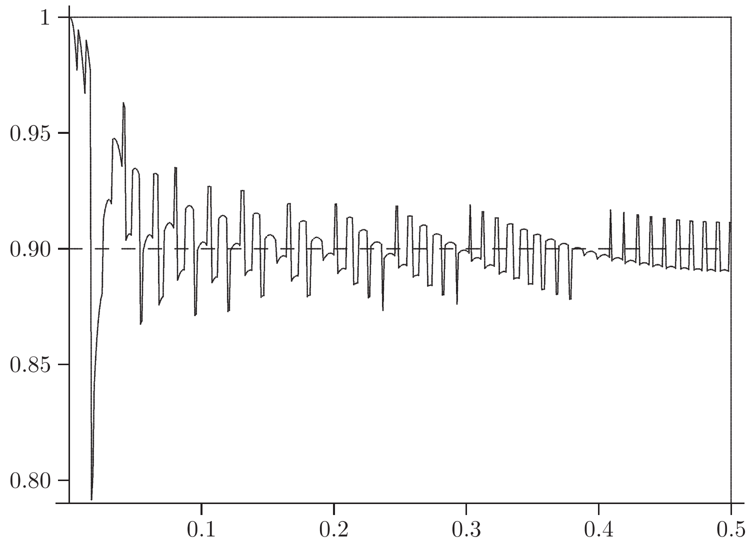
$$\left(\tilde{\theta} - z_{(1+\gamma)/2}\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n+b(S_n)}},\quad \tilde{\theta} + z_{(1+\gamma)/2}\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n+b(S_n)}}\right)$$

gdzie

$$\tilde{\theta} = \frac{S_n + a(S_n)}{n + b(S_n)}$$

oraz

$$(a,b)(S_n) = \begin{cases} (1/2, 5/4), & \text{gdy } S_n = 0, \\ (1, 7/4), & \text{gdy } S_n = 1, \\ (3/4, 7/4), & \text{gdy } S_n = n-1, \\ (3/4, 5/4), & \text{gdy } S_n = n, \\ (3/4, 3/2), & \text{poza tym.} \end{cases}$$



Rys. 9. Prawdopodobieństwo pokrycia przedziałem Barana ($n = 100,\ \gamma = 0.9$)

Prawdopodobieństwo pokrycia dla tego przedziału pokazuje Rys. 9. To znowu nie jest przedział ufności!

5. Podsumowanie. W całej dyskusji trzymałem się definicji, że $\left(\underline{\theta}(S_n), \overline{\theta}(S_n)\right)$ jest przedziałem ufności dla parametru $\theta \in (0,1)$ na poziomie ufności γ, jeżeli $P_\theta\{\underline{\theta}(S_n) \le \theta \le \overline{\theta}(S_n)\} \ge \gamma$ dla każdego $\theta \in (0,1)$. Uważam, że taki gwarantowany poziom ufności jest warunkiem *sine qua non* w teorii estymacji przedziałowej; ten warunek jest spełniony przez przedział ufności Neymana. Techniczne przeszkody numeryczne w stosowaniu tego przedziału, które jeszcze przed pół wiekiem stymulowały poszukiwanie prostszych rozwiązań, są już dawno pokonane i dzisiaj każdy inżynier, biolog, ekonomista i in., który potrafi stosować test Studenta i test chi-kwadrat, z pewnością sobie poradzi z poprawną estymacją przedziałową frakcji za pomocą przedziału Neymana.

Oto kilka prostych sposobów obliczania przedziałów ufności za pomocą powszechnie dostępnych narzędzi. Podane niżej wzory są aktualne dla $1 \le S_n \le n-1$. Jeżeli $S_n = 0$, to dolna granica przedziału ufności jest równa 0, a jeżeli $S_n = n$, to górna granica jest równa 1; pozostałą granicę obliczamy według podanych niżej reguł.

1. W Excelu (polska wersja Microsoft Excel 2002) robi się to na przykład tak: wpisuje się n do komórki A1, S_n do komórki A2, γ do komórki A3 i wtedy dolną granicę przedziału ufności otrzymuje się za pomocą for-

muły

$$\text{ROZKŁAD.BETA.ODW}((1 - A3)/2; A2; A1 - A2 + 1)$$

oraz górną za pomocą formuły

$$\text{ROZKŁAD.BETA.ODW}((1 + A3)/2; A2 + 1; A1 - A2)$$

2. W pakiecie Statistica obliczenia realizuje się za pomocą funkcji $VBeta((1-\gamma)/2, S_n, n - S_n + 1)$ oraz $VBeta((1+\gamma)/2, S_n + 1, n - S_n)$, odpowiednio dla dolnej i dla górnej granicy przedziału ufności.

3. W pakiecie Mathematica dla dolnej i górnej granicy mamy, odpowiednio, $Quantile[BetaDistribution[S_n, n - S_n + 1], (1-\gamma)/2]$ oraz $Quantile[BetaDistribution[S_n + 1, n - S_n], (1+\gamma)/2]$.

4. W środowisku R możemy to zrealizować za pomocą funkcji $qbeta((1-\gamma)/2, S_n, n - S_n + 1)$ oraz $qbeta((1+\gamma)/2, S_n + 1, n - S_n)$.

Wszędzie tam, gdzie nie mamy bezpośredniego dostępu do kwantyli rozkładu beta, możemy korzystać z kwantyli rozkładu F:

$$B^{-1}(\alpha, \beta; q) = \frac{\alpha}{\alpha + \beta F^{-1}(2\beta, 2\alpha; q)},$$

gdzie $F^{-1}(2\beta, 2\alpha; q)$ jest kwantylem rzędu q rozkładu F z $(2\beta, 2\alpha)$ stopniami swobody.

Jak widać, żadne kombinowanie z „przybliżonymi” i „asymtotycznymi” wzorami, które w dodatku nie dają poprawnych przedziałów ufności, nie jest potrzebne. Nie tylko ja jestem zdania, że z „asymptotycznymi” przedziałami ufności dla frakcji należy już raz na zawsze skończyć. *„Bye-bye, so long, farewell” to the Wald interval* (Casella [3]).

6. Komentarz techniczny. Prawdopodobieństwo $p(\theta)$ pokrycia przez przedział ufności (2) „prawdziwej” wartości θ było obliczane według wzoru

$$p(\theta) = \sum_{k=0}^{n} \mathbb{1}\{\underline{\theta}(k) \le \theta \le \overline{\theta}(k)\} P_\theta\{S_n = k\},$$

gdzie $\mathbb{1}\{A\} = 1$, gdy zdanie A jest prawdziwe oraz $\mathbb{1}\{A\} = 0$, gdy zdanie A nie jest prawdziwe. Tak samo obliczałem prawdopodobieństwa pokrycia dla wszystkich pozostałych przedziałów ufności, przy odpowiednich wzorach dla $\underline{\theta}(k)$ oraz $\overline{\theta}(k)$. Dla wykonania rysunków funkcja $p(\theta)$ była tablicowana dla wartości $\theta = 0.01(0.01)0.5$.

Podziękowanie. Dziękuję Wojciechowi Zielińskiemu za liczne dyskusje, a szczególnie za konsultacje w zakresie wykorzystania pakietów komputerowych.

Prace cytowane

[1] J. Baran, *Nowy przedział ufności dla prawdopodobieństwa sukcesu rozkładu dwumianowego*, XXXIII Konferencja „Statystyka Matematyczna Wisła 2007", 3–7 grudnia 2007.
[2] J. Bartoszewicz, *Wykłady ze statystyki matematycznej*, PWN (1996).
[3] G. Casella, *Statistical Science* 16 (2001), 2, p. 120
[4] C. J. Clopper i E. S. Pearson, *The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial*, Biometrika, Vol. 26 (1934), No. 4, pp. 404-413.
[5] H.Cramér, *Metody matematyczne w statystyce*, PWN (1958).
[6] M. Fisz, *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN (1967).
[7] L. Gajek i M. Kałuszka, *Wnioskowanie statystyczne*, Modele i metody, WNT (1996).
[8] P. Grzegorzewski, K. Bobecka, A. Dembińska, J. Pusz, *Rachunek prawdopodobieństwa i statystyka*, Wydanie czwarte, poprawione. Wyższa Szkoła Informatyki Stosowanej i Zarządzania, Warszawa (2003).
[9] R. Kala, *Statystyka dla przyrodników.* Wydawnictwo Akademii Rolniczej im. Augusta Cieszkowskiego w Poznaniu (2002).
[10] W. Klonecki, *Statystyka dla inżynierów*, PWN (1999).
[11] J. Koronacki i J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych.* Wydanie drugie. WNT (2004).
[12] M. Krzyśko, *Statystyka matematyczna*, Uniwersytet im. Adama Mickiewicza w Poznaniu (2004).
[13] E. L. Lehmann, *Testowanie hipotez statystycznych*, PWN (1968).
[14] R. Magiera, *Modele i metody statystyki matematycznej*, Wydanie drugie rozszerzone. Część II, Wnioskowanie statystyczne. Oficyna Wydawnicza GiS, s.c., Wrocław (2007).
[15] J. Neyman, *On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selecion*, Journal of the Royal Statistical Society, Vol. 97 (1934), No. 4, pp. 558-625.
[16] W. Niemiro, *Rachunek prawdopodobieństwa i statystyka matematyczna*, Szkoła Nauk Ścisłych (1999).
[17] A. Plucińska i E. Pluciński, *Rachunek prawdopodobieństwa. Statystyka matematyczna. Procesy stochastyczne*, WNT (2000).
[18] S. D. Silvey, *Wnioskowanie statystyczne*, PWN (1978).
[19] S. Trybuła, *Statystyka matematyczna z elementami teorii decyzji*, Oficyna Wydawnicza Politechniki Wrocławskiej (2001).
[20] R. Zieliński i W. Zieliński, *Tablice statystyczne*, PWN (1990).
[21] R. Zieliński, *Estymacja frakcji*, Matematyka Stosowana 9 (2008), s. 76-90.
[22] S. Zubrzycki, *Wykłady z rachunku prawdopodobieństwa i statystyki matematycznej*, PWN (1966).

Ryszard Zieliński
Instytut Matematyczny PAN
Warszawa, Poland
E-mail: R.Zielinski@impan.gov.pl

Confidence intervals for a binomial proportion

Abstract. Confidence intervals have been invented by Jerzy Spława-Neyman in 1934 (*J.Neyman, Journal of the Royal Statistical Society, Vol. 97, No.4. (1934), pp. 558-625*). To apply the theory to the binomial case some nomograms or vast tables were needed (*C.J.Clopper and E.S.Pearson: Biometrika, Vol. 26, No.4 (Dec.,1934), pp. 404-413*) which made the idea far from practical use. An easy to present and motivate and easy to compute remedy was to replace the binomial distribution by its normal approximation; the solution is currently in near universal use. Due to inadequate coverage probability many other alternative intervals have been recommended: an exhaustive and up to date review with a discussion one can find in *L.D.Brown, T.T.Cai and A.DasGupta: Statistical Science 2001, Vol. 16, No.2, 101-133*. We address the problem in the context of Polish textbooks.
Keywords: binomial proportion, confidence interval, Neyman interval, asymptotic approximation.

(*wpłynęło*)