PRZEDZIAŁ UFNOŚCI DLA FRAKCJI

Ryszard Zieliński

XXXVIII Ogólnopolska Konferencja Zastosowań Matematyki
Zakopane–Kościelisko 8-15 września 2009

ESTYMACJA FRAKCJI

W populacji składającej się z N elementów jest nieznana liczba M elementów wyróżnionych

W statystycznej kontroli jakości: oszacowanie wadliwości (frakcji sztuk wadliwych) w partii produktów lub w procesie produkcyjnym

Medycyna, np., szacowanie frakcji tych pacjentów z udarem mózgu, u których wcześniej wystąpił określony zespół symptomów

...

Problem. Zmienna losowa X ma rozkład Bernoulliego z nieznanym prawdopodobieństwem sukcesu θ :

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

$X_1, X_2, \ldots, X_n$ – próba losowa, $S_n=\sum_{j=1}^{n} X_j$ – minimalna zupełna statystyka dostateczna

Obserwacja: S_n

Interesuje nas **ESTYMACJA PRZEDZIAŁOWA** parametru θ

DEFINICJA. Losowy przedział

$$\left(\underline{\theta}(S_n), \overline{\theta}(S_n) \right)$$

nazywamy

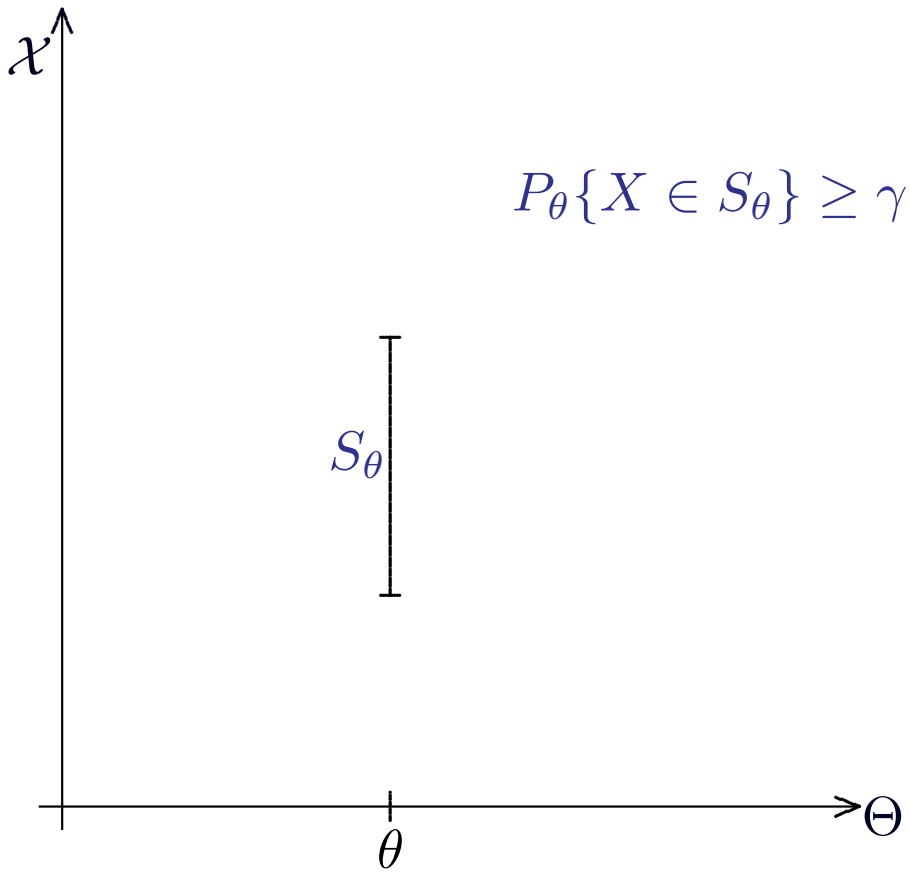
przedziałem ufności dla parametru θ na poziomie ufności γ

jeżeli

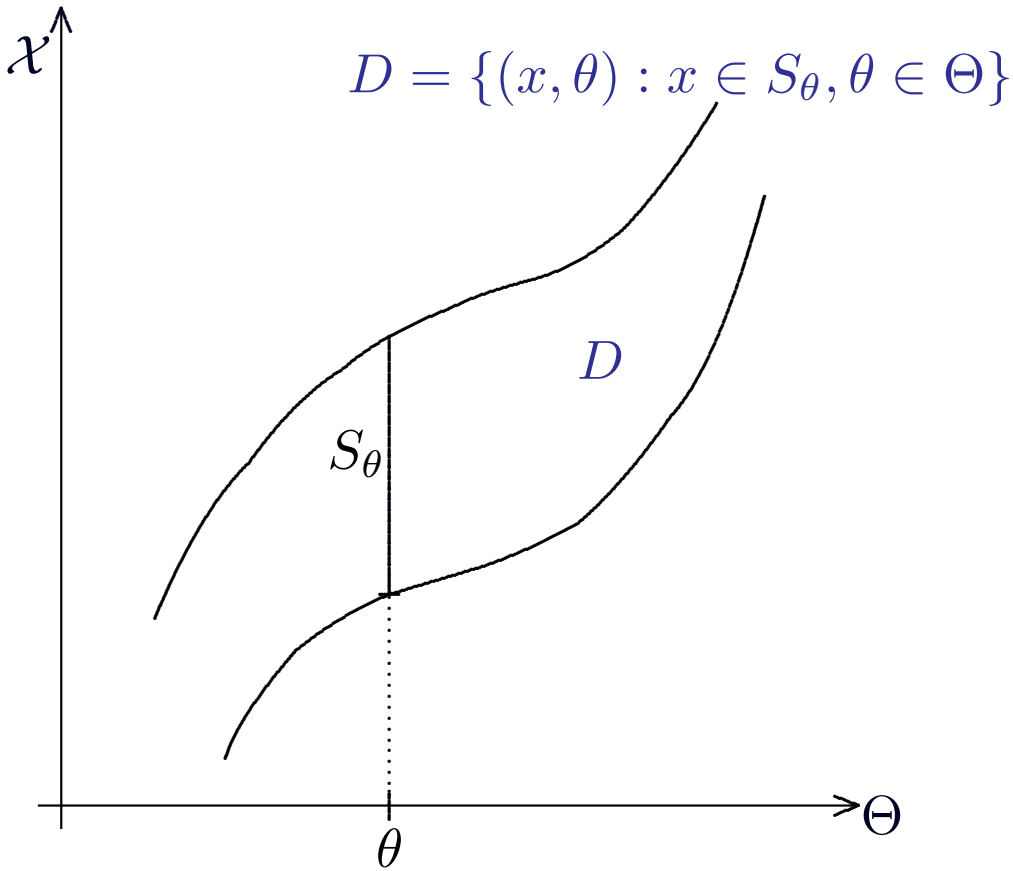
$$P_\theta\{\underline{\theta}(S_n) \leqslant \theta \leqslant \overline{\theta}(S_n)\} \geqslant \gamma \quad \text{dla każdego } \theta \in (0,1)$$

OGÓLNA KONSTRUKCJA PRZEDZIAŁU UFNOŚCI
w modelach parametrycznych

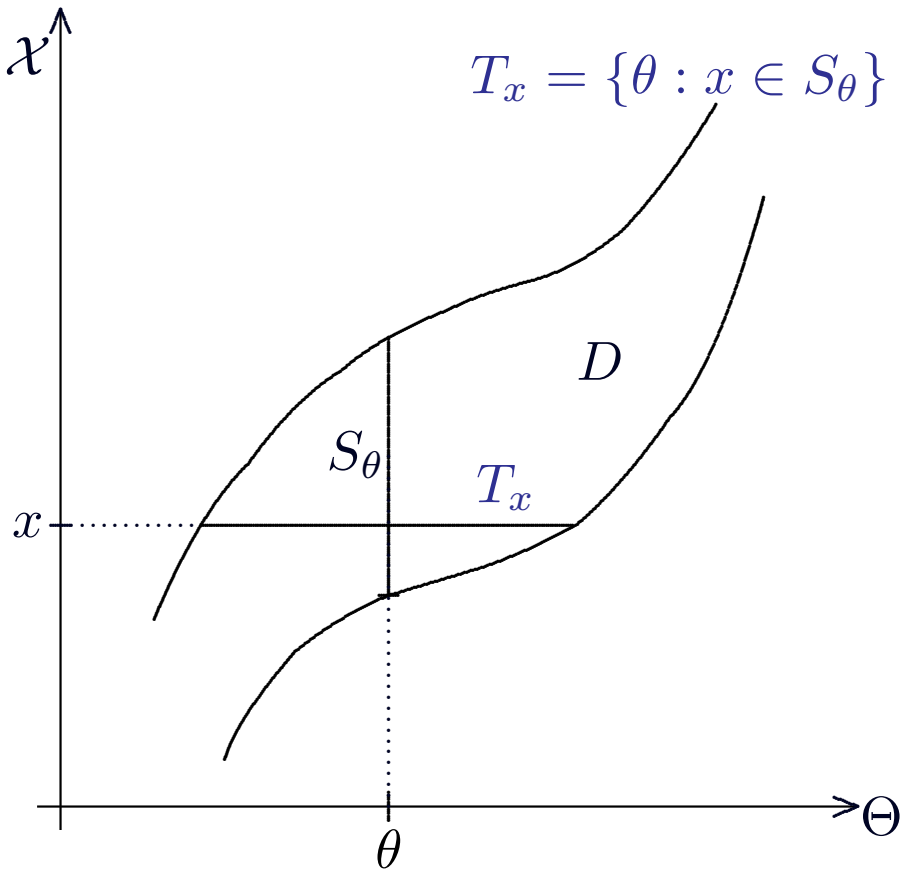
Obserwacja X ma rozkład z parametrem θ



Ogólna konstrukcja przedziału ufności



Ogólna konstrukcja przedziału ufności



Ogólna konstrukcja przedziału ufności

Mamy z tej konstrukcji

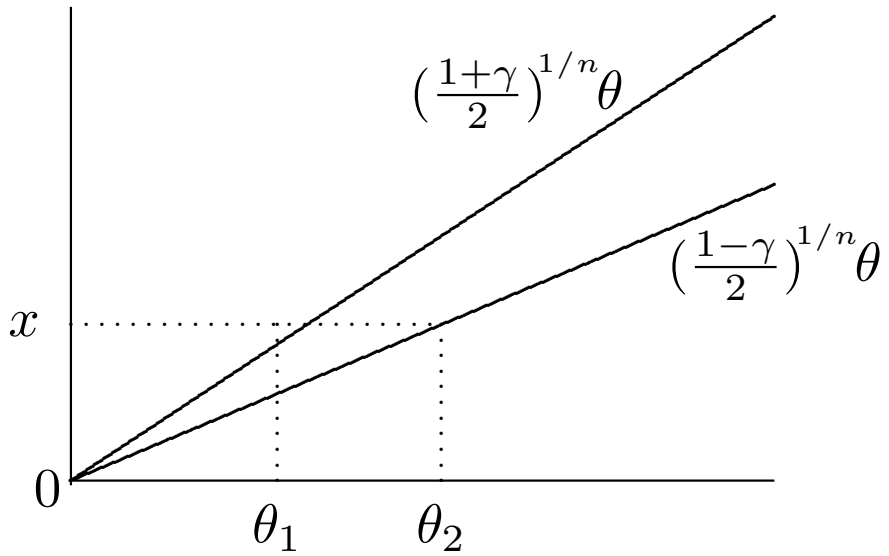
$$\theta \in T_x \quad \Longleftrightarrow \quad x \in S_\theta$$

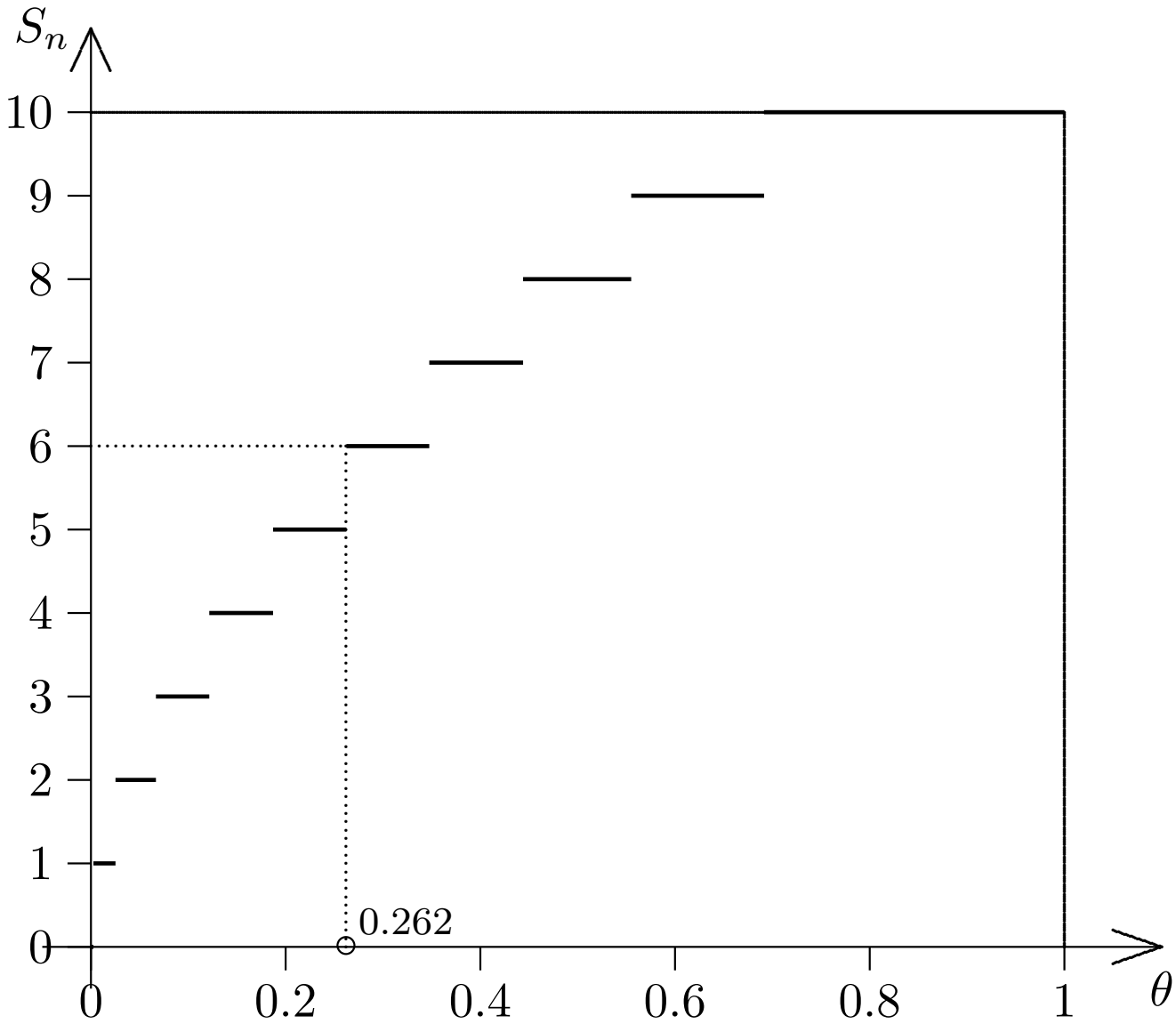
Zatem dla każdego $\theta \in \Theta$:

$$P_\theta\{\theta \in T_X\} = P_\theta\{X \in S_\theta\} \geqslant \gamma$$

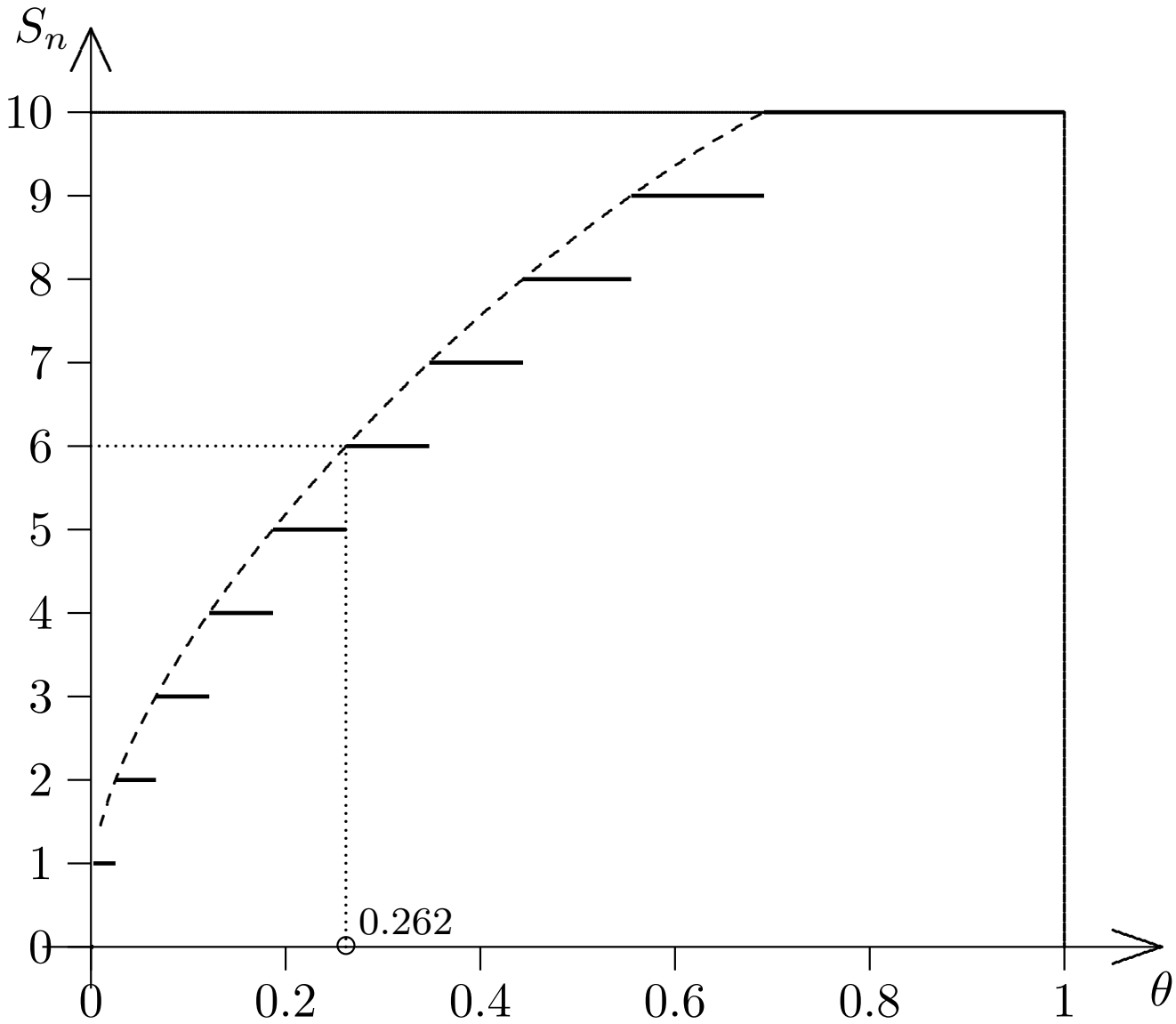
Przykład: rozkład jednostajny $U(0, \theta)$

$$\theta_1 = \left(\frac{1+\gamma}{2}\right)^{-1/n} x, \quad \theta_2 = \left(\frac{1-\gamma}{2}\right)^{-1/n} x$$





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



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

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

Rozkład beta

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

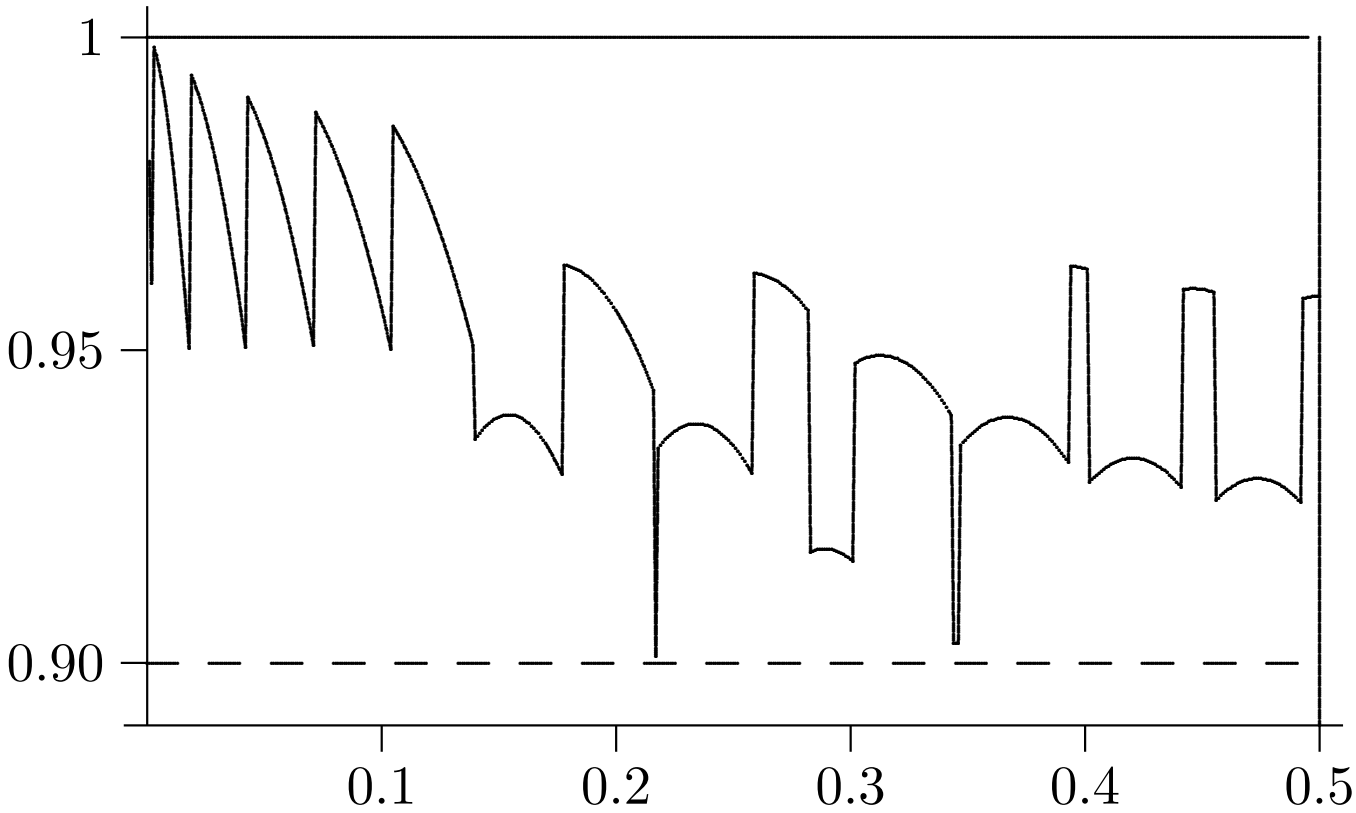
Rozkład beta

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

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^{B^{-1}(\alpha,\beta;\gamma)} u^{\alpha-1}(1-u)^{\beta-1}du=\gamma$$

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

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



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

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

- *Tablice rozkładu beta*

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

- *Tablice rozkładu dwumianowego*
- *Nomogramy przedziałów ufności*
- *Nomogramy rozkładu beta*

PRZEDZIAŁY ASYMPTOTYCZNE

$$(\forall x) \qquad P_\theta \left\{ \frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)/n}} \leqslant x \right\} \to \Phi(x), \quad n \to \infty \quad \left(\hat{\theta}_n = S_n/n \right)$$

PRZEDZIAŁY ASYMPTOTYCZNE

$$(\forall x) \qquad P_\theta\left\{\frac{\hat{\theta}_n-\theta}{\sqrt{\theta(1-\theta)/n}} \leqslant x\right\} \to \Phi(x), \quad n\to\infty \quad \left(\hat{\theta}_n = S_n/n\right)$$

Dla "dużych"n zmienna losowa $(\hat{\theta}_n-\theta)/\sqrt{\theta(1-\theta)/n}$ *ma*
W PRZYBLIŻENIU *rozkład normalny* $N(0,1)$

PRZEDZIAŁY ASYMPTOTYCZNE

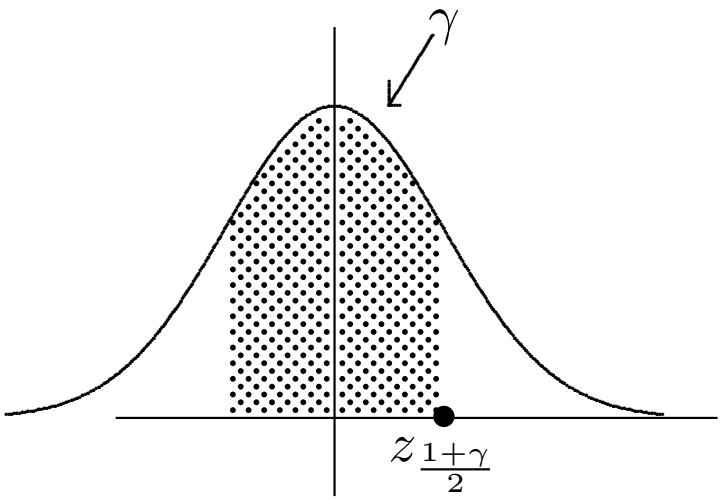
$$(\forall x) \qquad P_\theta \left\{ \frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)/n}} \leqslant x \right\} \to \Phi(x), \quad n \to \infty \quad \left(\hat{\theta}_n = S_n/n \right)$$

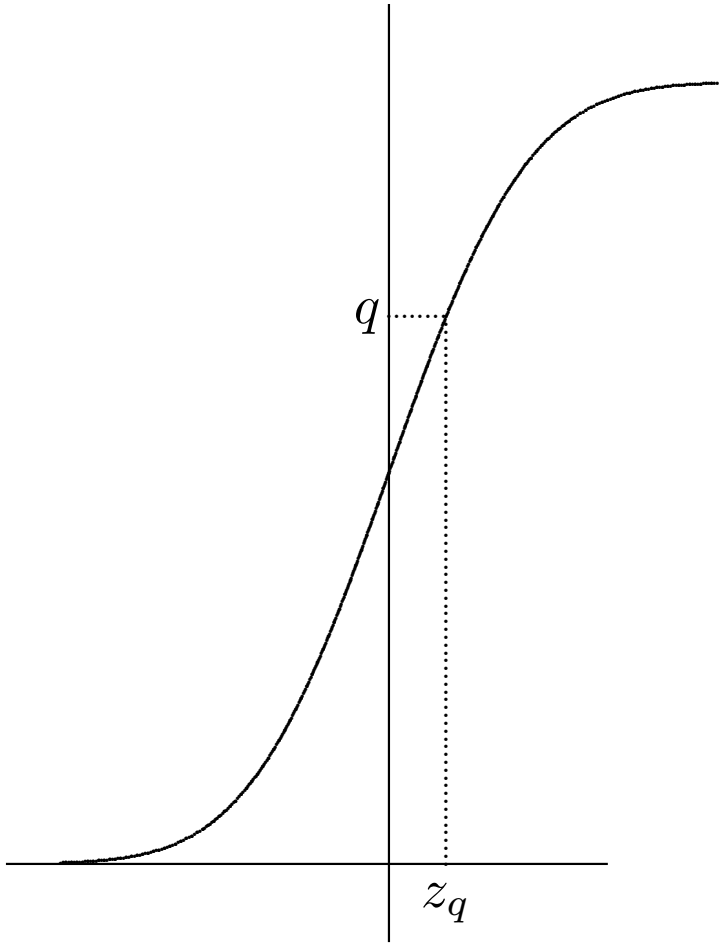
Dla "dużych"n zmienna losowa $(\hat{\theta}_n - \theta)/\sqrt{\theta(1-\theta)/n}$ *ma* W PRZYBLIŻENIU *rozkład normalny* $N(0,1)$

Inna szkoła uznaje, że $(\hat{\theta}_n - \theta)/\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)/n}$ ma asymptotyczny rozkład normalny $N(0,1)$

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

Oznaczenia





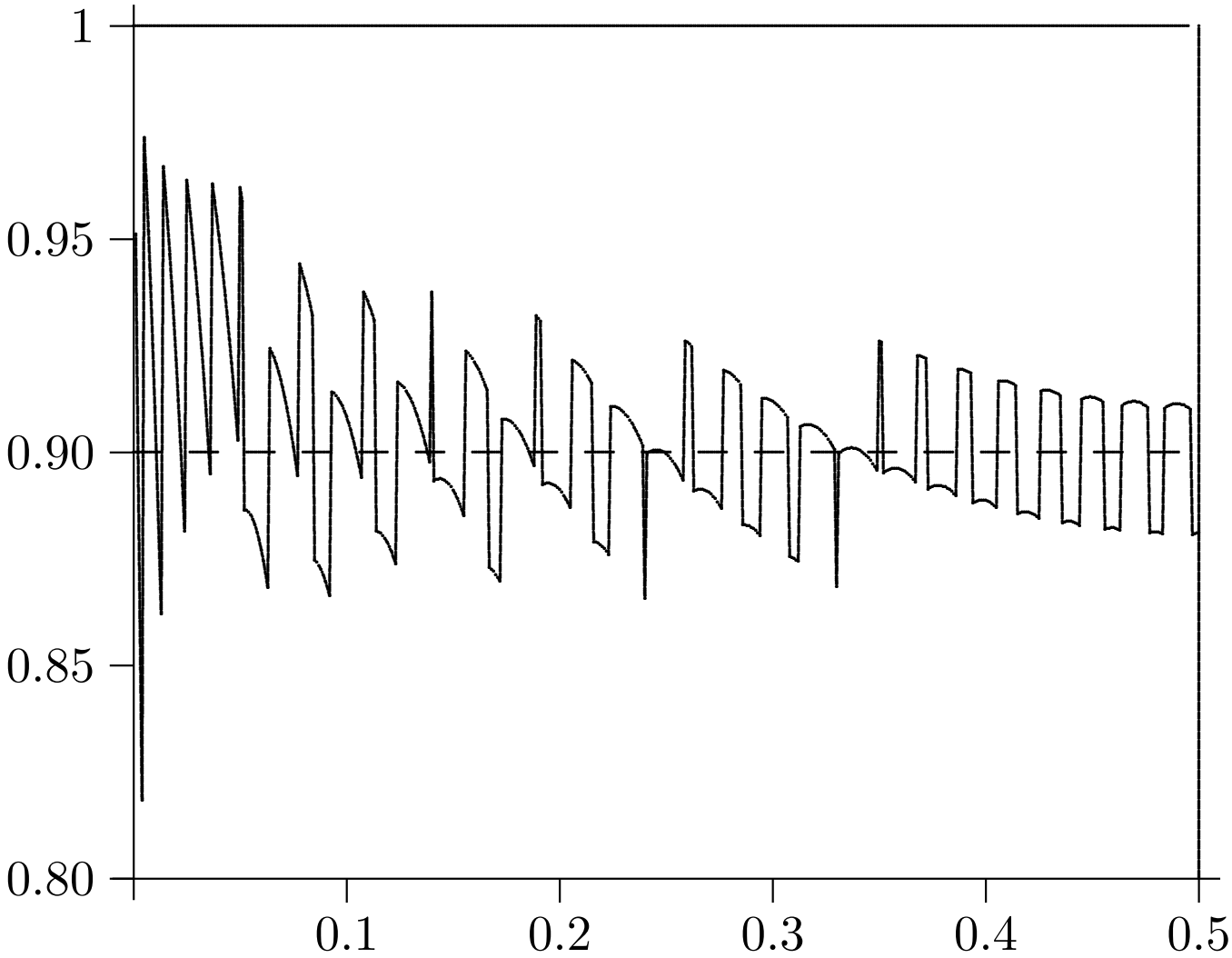
$$P_{N(0,1)}\left\{\left|\frac{\hat{\theta}_n-\theta}{\sqrt{\theta(1-\theta)/n}}\right|\leqslant z_{\frac{1+\gamma}{2}}\right\}=\gamma$$

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

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

$$P_{N(0,1)}\left\{\left|\frac{\hat{\theta}_n-\theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)/n}}\right|\leqslant z_{\frac{1+\gamma}{2}}\right\}=\gamma$$

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



Typowe prawdopodobieństwo pokrycia przedziałem asymptotycznym

"Udoskonalenia":

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

gdzie

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

oraz

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

"Udoskonalenia"

Stosować przedział asymptotyczny Walda wtedy, gdy $n\hat{\theta}_n \geqslant 5$ oraz $n(1-\hat{\theta}_n) \geqslant 5$

Kłopot z przybliżeniem asymptotycznym

Wprawdzie

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

ale

$$\forall n \; \exists \theta \qquad \left| P_\theta \left\{ \frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)/n}} \leqslant 0 \right\} - \Phi(0) \right| > \frac{1}{4}$$

DOKŁADNE PRZEDZIAŁY UFNOŚCI

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 formuły

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

oraz górną za pomocą formuły

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

DOKŁADNE PRZEDZIAŁY UFNOŚCI

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

DOKŁADNE PRZEDZIAŁY UFNOŚCI

W pakiecie Mathematica dla dolnej i górnej granicy mamy, odpowiednio,

$$Quantile[BetaDistribution[S_n, n - S_n + 1],\ (1-\gamma)/2]$$

oraz

$$Quantile[BetaDistribution[S_n + 1, n - S_n],\ (1+\gamma)/2]$$

DOKŁADNE PRZEDZIAŁY UFNOŚCI

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:

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

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

Jarosław Bartoszewicz (1996): *Wykłady ze statystyki matematycznej*. PWN

Dobiesław Bobrowski, Krystyna Maćkowiak-Łybacka (2006): *Wybrane metody wnioskowania statystycznego*. Wydawnictwo Politechniki Poznańskiej