Przedział ufności dla odsetka pytań drażliwych w modelu Item Count Technique

Stanislaw Jaworski
Wojciech Zieliński

Katedra Ekonometrii i Statystyki
SGGW

IV Kongres Statystyki Polskiej
04 lipca 2024

Problem

Zadanie

Oszacować odsetek odpowiedzi na pytanie drażliwe.

Problem

Zadanie

Oszacować odsetek odpowiedzi na pytanie drażliwe.
Problemy drażliwe

- skala łapówkarstwa
- oszustwa podatkowe
- szara strefa
- używanie narkotyków
- LGBT
- przemoc wobec dzieci

Problem

Zadanie

Niech Y będzie zmienną losową:

$$P\{Y = 1\} = \pi = 1 - P\{Y = 0\}.$$

Liczba $\pi \in (0, 1)$ jest prawdopodobieństwem odpowiedzi *YES* na pytanie drażliwe. Estymujemy prawdopodobieństwo π.

Model dla próby $Y_1, \ldots, Y_n$

$$(\{0, 1, \ldots, n\}, \{Bin(n, \pi), \pi \in (0, 1)\})$$

Problem

Kłopot

Zmienne losowe $Y_1, \ldots, Y_n$ nie są obserwowalne.

Problem

Kłopot

Zmienne losowe $Y_1, \ldots, Y_n$ nie są obserwowalne.

Antidotum

Odpowiedzi na pytanie drażliwe są „ukrywane” przez zadawanie pytania „neutralnego".

Odpowiedź na pytanie neutralne jest liczbą naturalną.

- Ile średnio śpisz kwadransów na dobę.
- Ile razy byłeś w kinie w ostatnim miesiącu.
- . . .

Antidotum

Antidotum

Badana grupa osób jest losowo dzielona na dwie części.

Antidotum

Antidotum

Badana grupa osób jest losowo dzielona na dwie części.

Każdej osobie zadawane są te same dwa pytania.

Antidotum

Antidotum

Badana grupa osób jest losowo dzielona na dwie części.

Każdej osobie zadawane są te same dwa pytania.

Wynikiem ankiety jest:
w pierwszej grupie: neutralne - drażliwe
w drugiej grupie: neutralne + drażliwe

Dane

Modele

- Pytanie drażliwe: $P\{Z = 1\} = \pi = 1 - P\{Z = 0\}$
- Pytanie neutralne: $X \sim Po(\lambda)$

Dane

Modele

- Pytanie drażliwe: $P\{Z = 1\} = \pi = 1 - P\{Z = 0\}$
- Pytanie neutralne: $X \sim Po(\lambda)$

Dane

- $Y_{1i} = X_{1i} - Z_{1i},\ i = 1, \cdots, n_1$
- $Y_{2i} = X_{2i} + Z_{2i},\ i = 1, \cdots, n_2$

Dane

Modele

- Pytanie drażliwe: $P\{Z=1\}=\pi=1-P\{Z=0\}$
- Pytanie neutralne: $X \sim Po(\lambda)$

Dane

- $Y_{1i}=X_{1i}-Z_{1i},\ i=1,\cdots,n_1$
- $Y_{2i}=X_{2i}+Z_{2i},\ i=1,\cdots,n_2$

Zmienna losowa

$$T=\sum_{i=1}^{n_2}Y_{2i}-\sum_{i=1}^{n_1}Y_{1i}=\left(\sum_{i=1}^{n_2}X_{2i}-\sum_{i=1}^{n_1}X_{1i}\right)+\left(\sum_{i=1}^{n_2}Z_{2i}+\sum_{i=1}^{n_1}Z_{1i}\right)$$

Zmienna losowa

$$T = \sum_{i=1}^{n_2} Y_{2i} - \sum_{i=1}^{n_1} Y_{1i} = \left(\sum_{i=1}^{n_2} X_{2i} - \sum_{i=1}^{n_1} X_{1i} \right) + \left(\sum_{i=1}^{n_2} Z_{2i} + \sum_{i=1}^{n_1} Z_{1i} \right)$$

$$\left(\frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i} - \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} \right) + \left(\frac{1}{n_2} \sum_{i=1}^{n_2} Z_{2i} + \frac{1}{n_1} \sum_{i=1}^{n_1} Z_{1i} \right)$$

Wartość oczekiwana

$$(\lambda - \lambda) + (\pi + \pi) = 2\pi$$

Rozkład prawdopodobieństwa

Zmienna losowa T

$$F_{\pi;\lambda;n_1,n_2}(t) = \sum_{k=0}^{n_1+n_2} \left[1 - Q_{-\lfloor t-k \rfloor}\left(\sqrt{2}\sqrt{n_1\lambda}, \sqrt{2}\sqrt{n_2\lambda}\right)\right] \times \binom{n_1+n_2}{k} \pi^k (1-\pi)^{n_1+n_2-k}$$

$Q_x(\cdot,\cdot)$ jest funkcją Q-Marcuma

Przedział ufności dla π

Przedzial ufności

Przedział ufności na poziomie 2γ jest rozwiązaniem dla danego t dwóch równań względem π

$$F_{\pi;\lambda;n_1,n_2}(t) = 1 - \gamma$$
$$F_{\pi;\lambda;n_1,n_2}(t) = \gamma$$

Przedział ufności dla π

Przedzial ufności

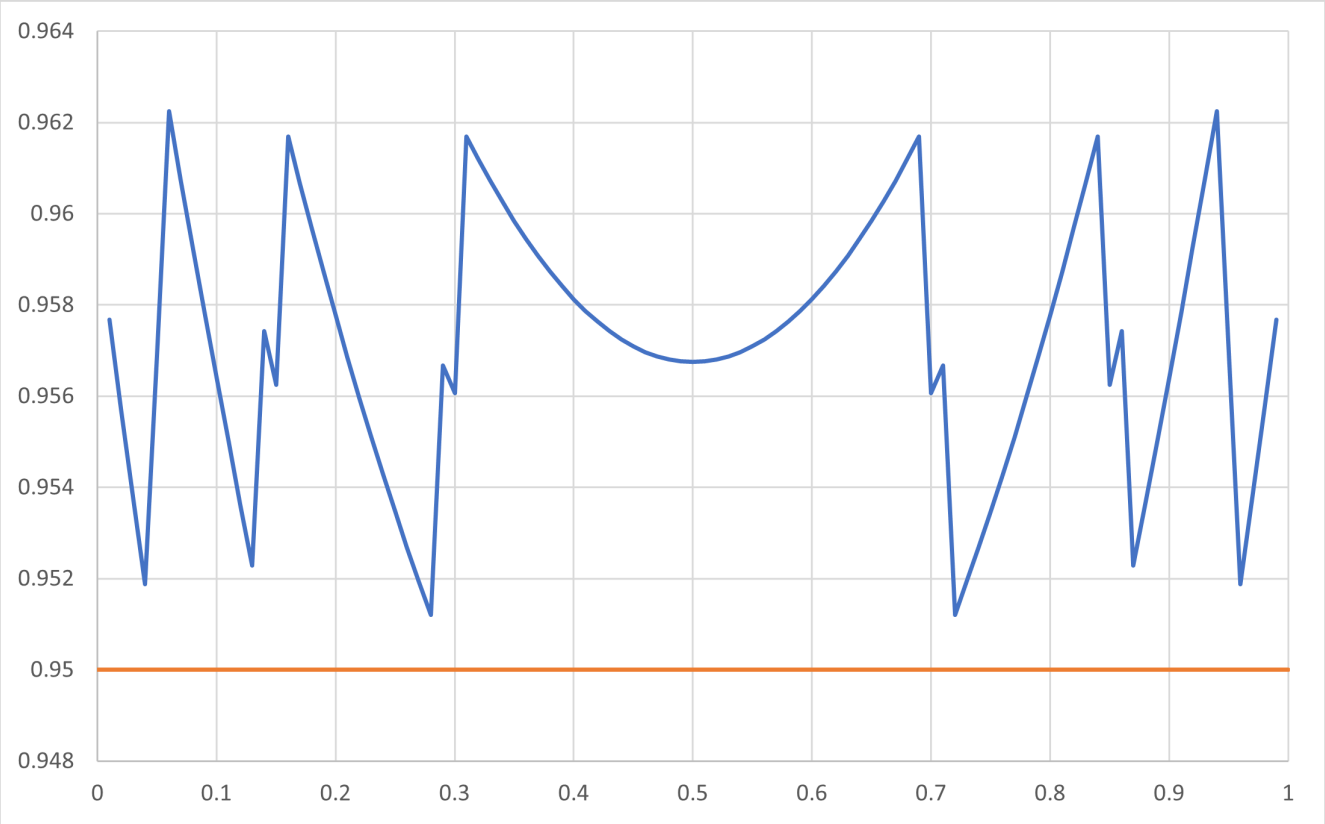
W zależności od zaobserwowanego t przedział może być jedno- lub dwustronny:

$$F_{0;\lambda;n_1,n_2}(t) = \begin{cases} < 1-\gamma, & \text{lewy koniec} = 0 \\ \geq 1-\gamma, & \text{lewy koniec} > 0 \end{cases}$$

$$F_{1;\lambda;n_1,n_2}(t) = \begin{cases} > \gamma, & \text{prawy koniec} = 1 \\ \leq \gamma, & \text{prawy koniec} < 1 \end{cases}$$

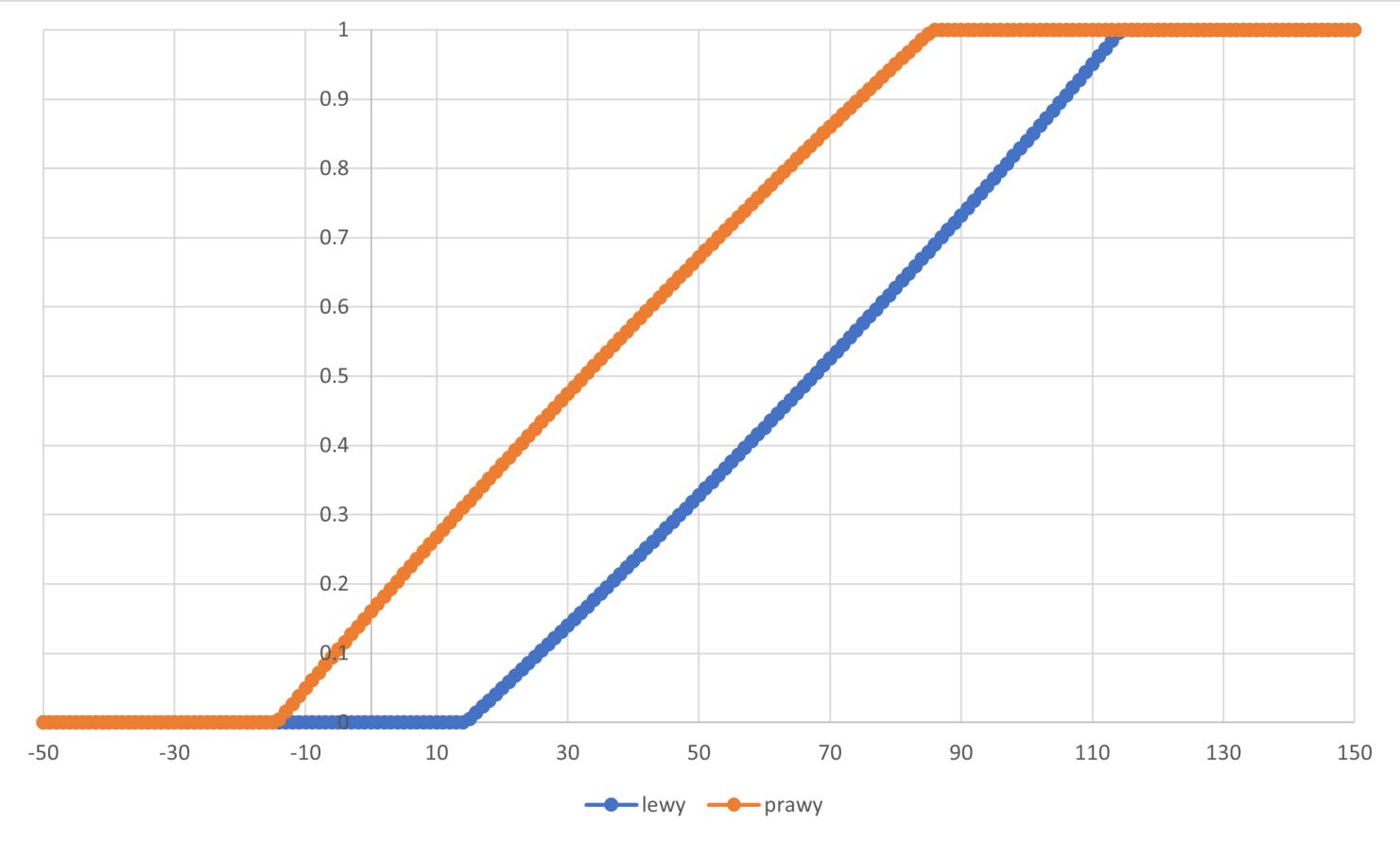
Przedział ufności dla π; $\lambda = 2$; $n_1 = n_2 = 100$

Pokrycie vs π



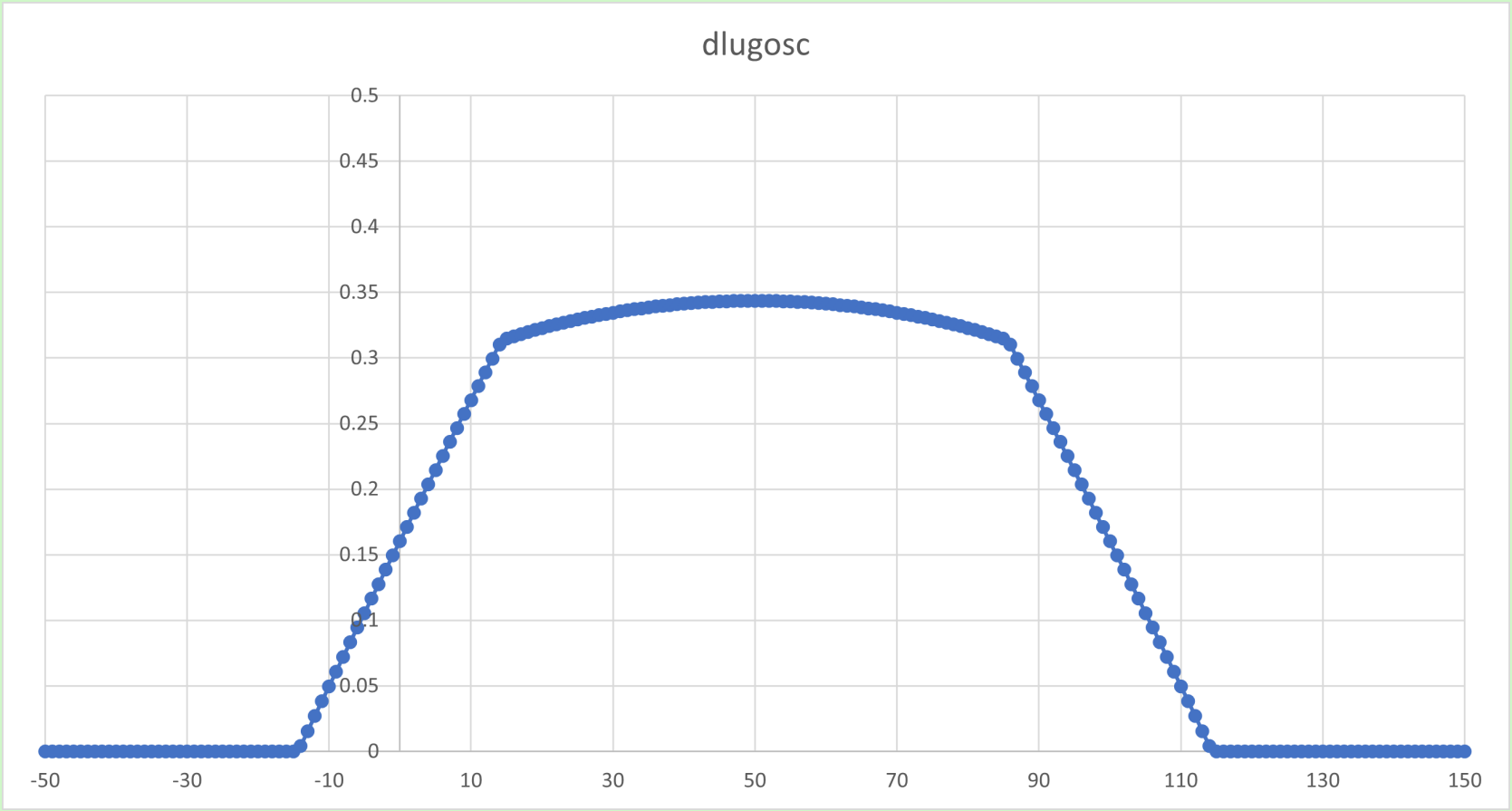
Przedział ufności dla π; $\lambda = 2$; $n_1 = n_2 = 100$

Przedziały vs T



Przedział ufności dla π; $\lambda = 2$; $n_1 = n_2 = 100$

Długość vs T



Długość przedziału dla π

Długość zależy od

- prawdopodobieństwa drażliwego π
- parametru λ pytania dodatkowego
- wielkości prób n_1 i n_2
- wzajemnego stosunku $n_1 : n_2$

Długość przedziału dla π

Długość zależy od

- prawdopodobieństwa drażliwego π
- parametru λ pytania dodatkowego
- wielkości prób n_1 i n_2
- wzajemnego stosunku $n_1 : n_2$

Zadanie

Jak sterować długością przedziału ufności?

Długość przedziału dla π

Długość zależy od

- prawdopodobieństwa drażliwego π
- parametru λ pytania dodatkowego
- wielkości prób n_1 i n_2
- wzajemnego stosunku $n_1 : n_2$

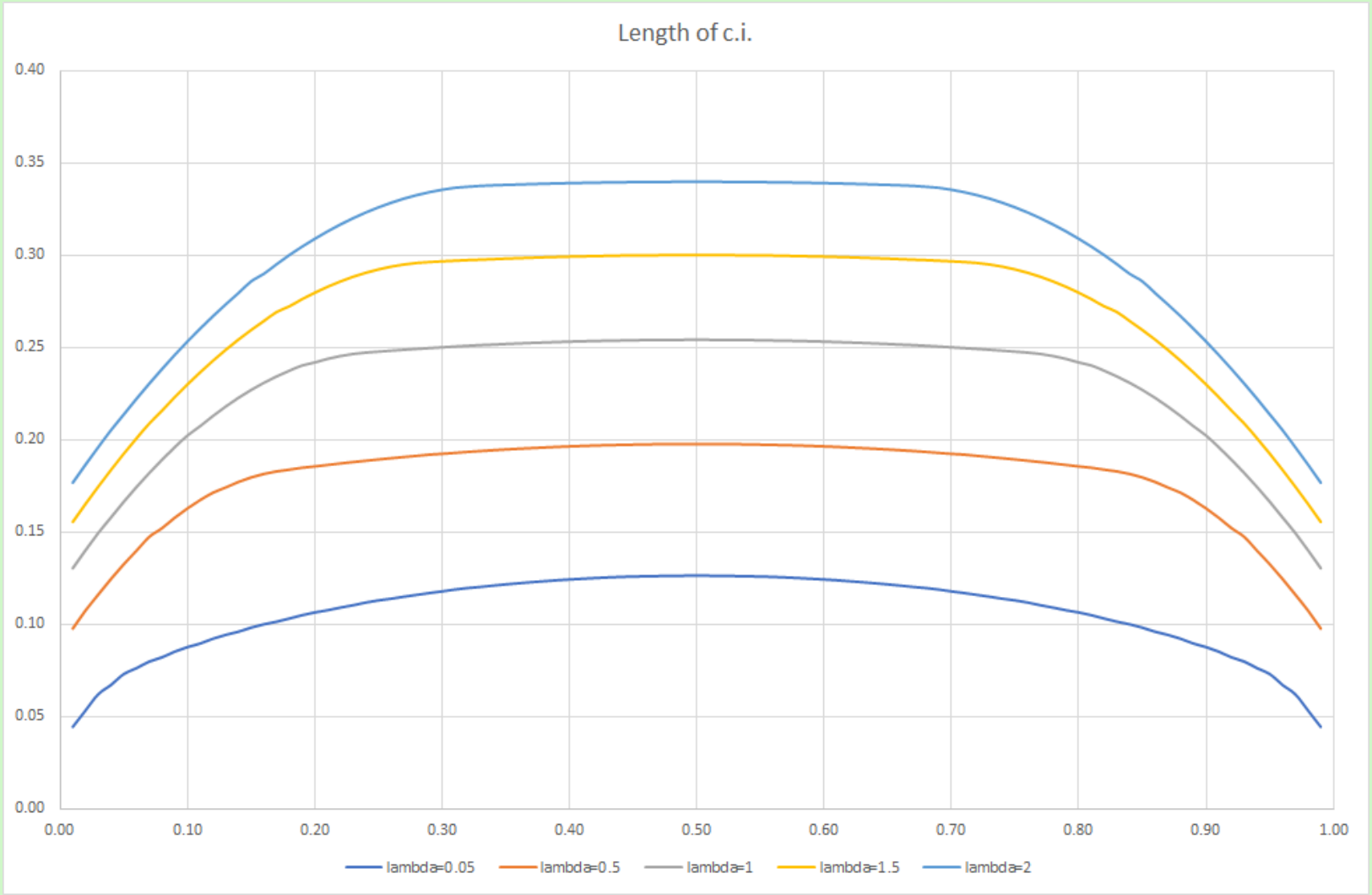
Zadanie

Jak sterować długością przedziału ufności?

Jak można uzyskać krótki przedział?

Analiza długości przedziału ufności dla π

Długość vs π dla różnych λ



Analiza długości przedziału ufności dla π

Długość vs π dla różnych λ

- długość jest największa dla $\pi = 0.5$
- długość rośnie wraz z λ
- długość maleje wraz z licznością prób

Analiza długości przedziału ufności dla π

Długość vs π dla różnych λ

- Długość $l(T; \pi, \gamma, \lambda, n_1, n_2)$ jest zmienną losową.

Analiza długości przedziału ufności dla π

Długość vs π dla różnych λ

- Długość $l(T; \pi, \gamma, \lambda, n_1, n_2)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0

Analiza długości przedziału ufności dla π

Długość vs π dla różnych λ

- Długość $l(T; \pi, \gamma, \lambda, n_1, n_2)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0
- Zadanie: dla jakich λ oraz n_1, n_2 oczekiwana długość przedziału jest najmniejsza

Analiza długości przedziału ufności dla π

Długość vs π dla różnych λ

- Długość $l(T; \pi, \gamma, \lambda, n_1, n_2)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0
- Zadanie: dla jakich λ oraz n_1, n_2 oczekiwana długość przedziału jest najmniejsza
- Najlepszy wynik jest dla $\lambda = 0 \leftarrow$ praktycznie nieużyteczne

Analiza długości przedziału ufności dla π

Długość vs π dla różnych λ

- Długość $l(T; \pi, \gamma, \lambda, n_1, n_2)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0
- Zadanie: dla jakich λ oraz n_1, n_2 oczekiwana długość przedziału jest najmniejsza
- Najlepszy wynik jest dla $\lambda = 0 \leftarrow$ praktycznie nieużyteczne
- Dodatkowe kryterium!

Privacy protection

Privacy protection

Szanse na odgadnięcie odpowiedzi Respondenta na pytanie drażliwe znając całościową odpowiedź na ankietę

Privacy protection

Privacy protection dla grupy pierwszej

$$P^{1st}_{\pi,\lambda}\{Z=1|Y_1\} = \frac{1}{1+\frac{Y_1+1}{\lambda}(\frac{1}{\pi}-1)}$$

Privacy protection dla grupy drugiej

$$P^{2nd}_{\pi,\lambda}\{Z=1|Y_2\} = \frac{1}{1+\frac{\lambda}{Y_2}(\frac{1}{\pi}-1)}$$

Privacy protection

Privacy protection dla grupy pierwszej

$$P^{1st}_{\pi,\lambda}\{Z = 1|Y_1\} = \frac{1}{1 + \frac{Y_1+1}{\lambda}(\frac{1}{\pi} - 1)}$$

Privacy protection dla grupy drugiej

$$P^{2nd}_{\pi,\lambda}\{Z = 1|Y_2\} = \frac{1}{1 + \frac{\lambda}{Y_2}(\frac{1}{\pi} - 1)}$$

Zmienne losowe!

Dobór parametru λ

Ustalamy małe $\alpha > 0$ oraz duże $\delta > 0$

Chcemy dobrać λ tak, by dla wszystkich $\pi \leq \pi_0 < 0.5$

$$\begin{cases} P_{\pi,\lambda}\left\{P^{1st}_{\pi,\lambda}\{Z = 1|Y_1\} \leq \alpha\right\} \geq \delta \\ P_{\pi,\lambda}\left\{P^{2nd}_{\pi,\lambda}\{Z = 1|Y_2\} \leq \alpha\right\} \geq \delta \end{cases}$$

Dobór parametru λ

$$\begin{cases} P_{\pi,\lambda}\left\{\frac{1}{1+\frac{Y_1+1}{\lambda}(\frac{1}{\pi}-1)} \leq \alpha\right\} \geq \delta \\ P_{\pi,\lambda}\left\{\frac{1}{1+\frac{\lambda}{Y_2}(\frac{1}{\pi}-1)} \leq \alpha\right\} \geq \delta \end{cases}$$

Dobór parametru λ

$$\begin{cases} P_{\pi,\lambda}\left\{\frac{1}{1+\frac{Y_1+1}{\lambda}(\frac{1}{\pi}-1)} \leq \alpha\right\} \geq \delta \\ P_{\pi,\lambda}\left\{\frac{1}{1+\frac{\lambda}{Y_2}(\frac{1}{\pi}-1)} \leq \alpha\right\} \geq \delta \end{cases}$$

$$\begin{cases} P_{\pi,\lambda}\left\{Y_1 \geq \lambda \frac{\frac{1}{\alpha}-1}{\frac{1}{\pi}-1} - 1\right\} \geq \delta \\ P_{\pi,\lambda}\left\{Y_2 \leq \lambda \frac{\frac{1}{\pi}-1}{\frac{1}{\alpha}-1}\right\} \geq \delta \end{cases}$$

Dobór parametru λ

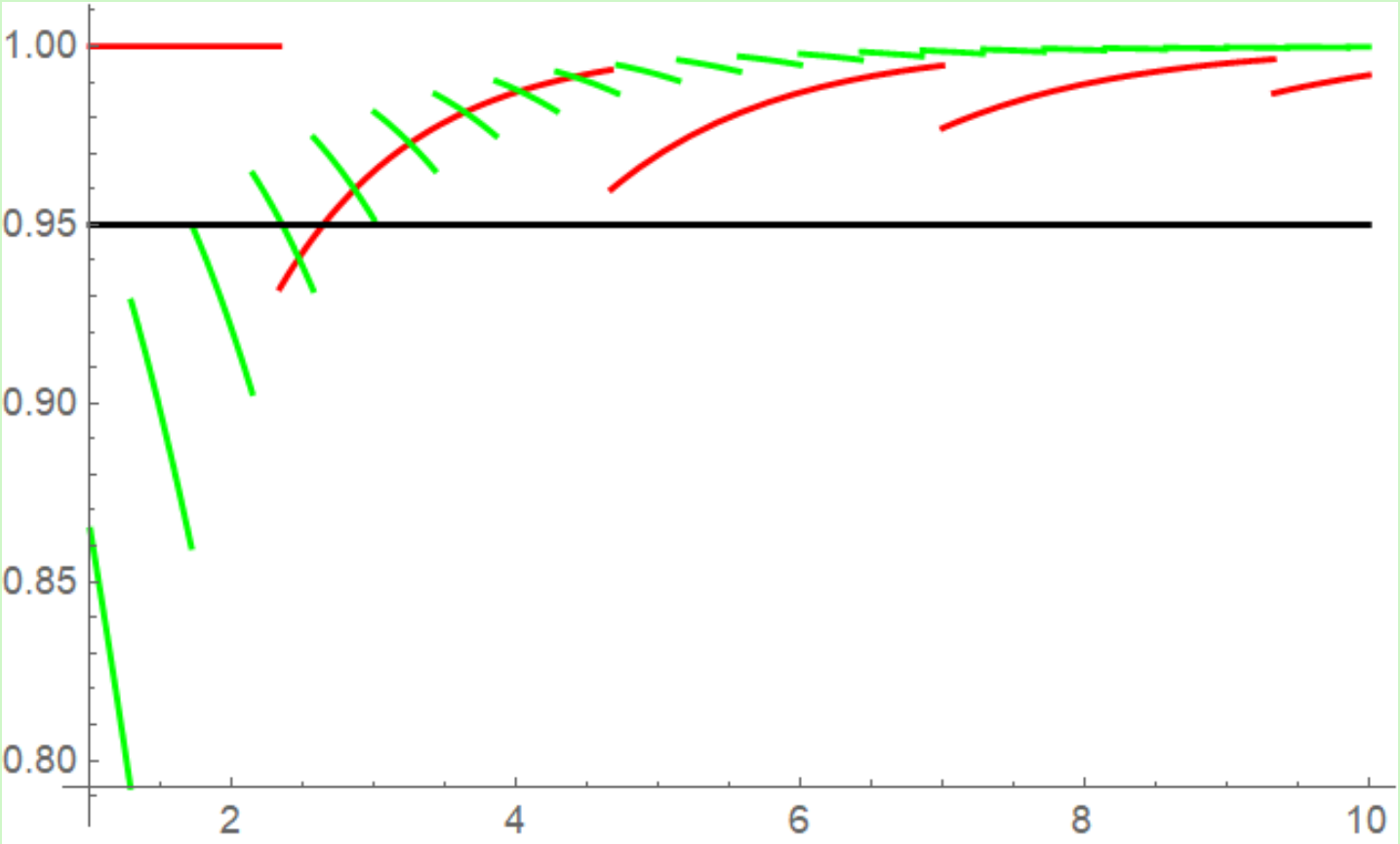
Ustalamy małe $\alpha > 0$ oraz duże $\delta > 0$

Chcemy dobrać λ tak, by dla wszystkich $\pi \leq \pi_0 < 0.5$

$$\begin{cases} P_{\pi,\lambda}\left\{Y_1 \geq \lambda \frac{\frac{1}{\alpha}-1}{\frac{1}{\pi}-1} - 1\right\} \geq \delta \\ P_{\pi,\lambda}\left\{Y_2 \leq \lambda \frac{\frac{1}{\pi}-1}{\frac{1}{\alpha}-1}\right\} \geq \delta \end{cases}$$

Privacy protection vs. λ

$\pi_0 = 0.3, \delta = 0.95$



Przedział ufności

Poziom ufności 2γ

Przedział ufności na poziomie 2γ jest rozwiązaniem dla danego t dwóch równań względem π

$$F_{\pi;\lambda;n,n}(t) = 1 - \gamma \qquad F_{\pi;\lambda;n,n}(t) = \gamma$$

Długość przedziału dla daneqo t

$$l(t;\pi,\gamma,\lambda,n,n) = U(t;\pi,\gamma,\lambda,n,n) - L(t;\pi,\gamma,\lambda,n,n)$$

Średnia długość przedziału

$$\sum_{t=-\infty}^{\infty} l(t;\pi,\gamma,\lambda,n,n) P_{\pi;\lambda;n,n}(t) \mathbf{1}_{(L(t;\pi,\gamma,\lambda,n,n),U(t;\pi,\gamma,\lambda,n,n))}(\pi)$$

Zadanie

Znaleźć

najmniejszą liczność próby taką, że średnia długość przedziału ufności jest **zadaną** liczbą.

Zadanie

Znaleźć

najmniejszą liczność próby taką, że średnia długość przedziału ufności jest **zadaną** liczbą.

Ograniczenia

- wybór parametru λ tak by osiągnąć żądane *privacy protection*
- wybór π_0 określające zdarzenie „rzadkie”

Zadanie

Znaleźć

najmniejszą liczność próby taką, że średnia długość przedziału ufności jest **zadaną** liczbą.

Ograniczenia

- wybór parametru λ tak by osiągnąć żądane *privacy protection*
- wybór π_0 określające zdarzenie „rzadkie”

ROZWIĄZANIE NUMERYCZNE

Przykład numeryczny

Ograniczenia

- poziom ufności $2\gamma = 0.95$
- zdarzenia rzadkie $\pi_0 = 0.1$
- *privacy protection* $1 - \alpha = 0.95$ z prawdopodobieństwem $\delta = 0.95$

Przykład numeryczny

Ograniczenia

- poziom ufności $2\gamma = 0.95$
- zdarzenia rzadkie $\pi_0 = 0.1$
- *privacy protection* $1 - \alpha = 0.95$ z prawdopodobieństwem $\delta = 0.95$
- średnia w rozkładzie Poissona $\lambda = 0.693$

Przykład numeryczny

Ograniczenia

- poziom ufności $2\gamma = 0.95$
- zdarzenia rzadkie $\pi_0 = 0.1$
- *privacy protection* $1 - \alpha = 0.95$ z prawdopodobieństwem $\delta = 0.95$
- średnia w rozkładzie Poissona $\lambda = 0.693$
- maksymalna długość przedziału ufności $d = 0.06$

Przykład numeryczny

Ograniczenia

- poziom ufności $2\gamma = 0.95$
- zdarzenia rzadkie $\pi_0 = 0.1$
- *privacy protection* $1 - \alpha = 0.95$ z prawdopodobieństwem $\delta = 0.95$
- średnia w rozkładzie Poissona $\lambda = 0.693$
- maksymalna długość przedziału ufności $d = 0.06$

minimalna liczność próby 2×1550