

Optymalny dobór próby w modelu krzyżowym oraz trójkątnym pytań drażliwych

Stanisław Jaworski
Wojciech Zieliński

Katedra Ekonometrii i Statystyki
SGGW

V Kongres Statystyki Polskiej
02 lipca 2025

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 .

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 $Y_1, \dots, Y_n$

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

Problem

Kłopot

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

Problem

Kłopot

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

Antidotum

Odpowiedzi na pytanie drażliwe są „ukrywane” przez zadawanie pytania „neutralnego”. Pytanie neutralne ma dwie odpowiedzi *YES* lub *NO*. W ankiecie stawiane są jednocześnie dwa pytania: drażliwe i neutralne. Wynikiem tak skonstruowanej ankiety jest jedna odpowiedź: *YES* lub *NO*, przy czym ankieter ma nie wiedzieć, na które pytanie odpowiada ankietowany.

Modele

Krzyżowy

$$Z = \begin{cases} 1, & \text{gdy obie odpowiedzi są takie same,} \\ 0, & \text{w przeciwnym przypadku.} \end{cases} \quad (1)$$

Modele

Krzyżowy

$$Z = \begin{cases} 1, & \text{gdy obie odpowiedzi są takie same,} \\ 0, & \text{w przeciwnym przypadku.} \end{cases} \quad (1)$$

Trójkątny

$$Z = \begin{cases} 0, & \text{gdy obie odpowiedzi są NIE,} \\ 1, & \text{w przeciwnym przypadku.} \end{cases} \quad (2)$$

Modele

Krzyżowy	NO	YES
NO	1	0
YES	0	1

Trójkątny	NO	YES
NO	0	1
YES	1	1

Model krzyżowy

Model krzyżowy

$$P_{\pi}(Z = 1) = \rho = q\pi + (1 - q)(1 - \pi)$$

Próba niezależnych zmiennych losowych $Z_1, \dots, Z_n$.

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^n Z_i \implies \hat{\pi}_K = \frac{\bar{Z} - (1 - q)}{2q - 1}$$

Model krzyżowy

Model krzyżowy

$$\widehat{Var\pi_K} = \frac{\hat{\pi}_K(1 - \hat{\pi}_K)}{n - 1} + \frac{q(1 - q)}{(n - 1)(2q - 1)^2}$$

Model krzyżowy

Model krzyżowy

$$\widehat{Var\pi_K} = \frac{\hat{\pi}_K(1 - \hat{\pi}_K)}{n - 1} + \frac{q(1 - q)}{(n - 1)(2q - 1)^2}$$

Przedział ufności dla prawdopodobieństwa drażliwego

$$\left(\hat{\pi}_K - u_{(1+\delta)/2} \sqrt{\widehat{Var\pi_K}}, \hat{\pi}_K + u_{(1+\delta)/2} \sqrt{\widehat{Var\pi_K}} \right),$$

gdzie $u_{(1+\delta)/2}$ oznacza kwantyl rozkładu $N(0, 1)$.

Model trójkątny

Model trójkątny

$$P_{\pi}(Z = 1) = \rho = \pi + (1 - \pi)q$$

Próba niezależnych zmiennych losowych $Z_1, \dots, Z_n$.

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^n Z_i \implies \hat{\pi}_T = \frac{\bar{Z} - (1 - q)}{1 - q}$$

Model trójkątny

Model trójkątny

$$\widehat{Var\pi_T} = \frac{\hat{\pi}_T(1 - \hat{\pi}_T)}{n - 1} + \frac{(1 - \hat{\pi}_T)q}{(n - 1)(1 - q)}$$

Model trójkątny

Model trójkątny

$$\widehat{Var\pi_T} = \frac{\hat{\pi}_T(1 - \hat{\pi}_T)}{n - 1} + \frac{(1 - \hat{\pi}_T)q}{(n - 1)(1 - q)}$$

Przedział ufności dla prawdopodobieństwa drażliwego

$$\left(\hat{\pi}_T - u_{(1+\delta)/2} \sqrt{\widehat{Var\pi_T}}, \hat{\pi}_T + u_{(1+\delta)/2} \sqrt{\widehat{Var\pi_T}} \right),$$

gdzie $u_{(1+\delta)/2}$ oznacza kwantyl rozkładu $N(0, 1)$.

Optymalny rozmiar próby

Względna długość

$$l(z; q, n) = \frac{\pi_G(z) - \pi_D(z)}{\pi}$$

Optymalny rozmiar próby

Względna długość

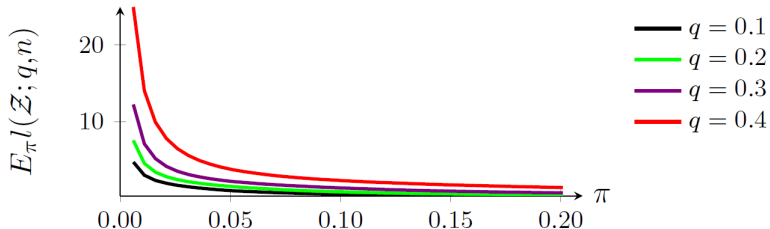
$$l(z; q, n) = \frac{\pi_G(z) - \pi_D(z)}{\pi}$$

$$E_{\pi}l(\mathcal{Z}; q, n) = \sum_{k=0}^n l(k; q, n) \mathbf{1}(\pi \in (\pi_G(k), \pi_D(k))) P_{\pi}\{\mathcal{Z} = k\}$$

Dla każdego $\pi \in (0, 0.5)$ wartość oczekiwana $E_{\pi}l(\mathcal{Z}; q, n)$ rośnie względem q i maleje względem n

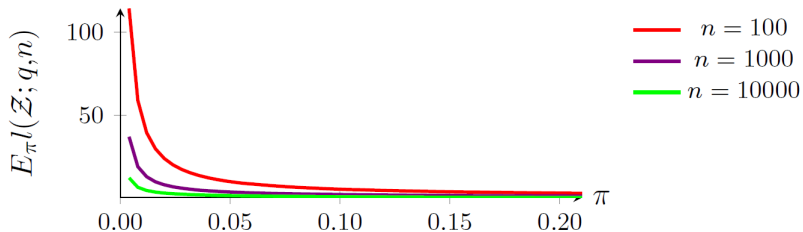
Optymalny rozmiar próby

Przedział ufności



Optymalny rozmiar próby

Przedział ufności



Długość względna przedziału dla π

Długość względna zależy od

- prawdopodobieństwa drażliwego π

Długość względna przedziału dla π

Długość względna zależy od

- prawdopodobieństwa drażliwego π
- parametru q pytania dodatkowego

Długość względna przedziału dla π

Długość względna zależy od

- prawdopodobieństwa drażliwego π
- parametru q pytania dodatkowego
- wielkości próby n

Długość względna przedziału dla π

Długość względna zależy od

- prawdopodobieństwa drażliwego π
- parametru q pytania dodatkowego
- wielkości próby n

Pytania

Jak sterować długością względną przedziału ufności?

Długość względna przedziału dla π

Długość względna zależy od

- prawdopodobieństwa drażliwego π
- parametru q pytania dodatkowego
- wielkości próby n

Pytania

Jak sterować długością względną przedziału ufności?

Jak można uzyskać krótki przedział?

Analiza długości względnej przedziału ufności dla π

Długość względna vs π

- Długość względna $l(\mathcal{Z}; q, n)$ jest zmienną losową.

Analiza długości względnej przedziału ufności dla π

Długość względna vs π

- Długość względna $l(\mathcal{Z}; q, n)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0 .

Analiza długości względnej przedziału ufności dla π

Długość względna vs π

- Długość względna $l(\mathcal{Z}; q, n)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0 .
- Przyjmujemy $\pi \geq \pi_{-1}$ dla zadanego π_{-1} .

Analiza długości względnej przedziału ufności dla π

Długość względna vs π

- Długość względna $l(\mathcal{Z}; q, n)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0 .
- Przyjmujemy $\pi \geq \pi_{-1}$ dla zadanego π_{-1} .
- Pytanie: dla jakich q oraz n oczekiwana długość względna przedziału jest najmniejsza.

Analiza długości względnej przedziału ufności dla π

Długość względna vs π

- Długość względna $l(\mathcal{Z}; q, n)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0 .
- Przyjmujemy $\pi \geq \pi_{-1}$ dla zadanego π_{-1} .
- Pytanie: dla jakich q oraz n oczekiwana długość względna przedziału jest najmniejsza.
- Najlepszy wynik jest dla $q = 0 \leftarrow$ praktycznie nieużyteczne.

Analiza długości względnej przedziału ufności dla π

Długość względna vs π

- Długość względna $l(\mathcal{Z}; q, n)$ jest zmienną losową.
- Przyjmujemy $\pi \leq \pi_0$ dla zadanego π_0 .
- Przyjmujemy $\pi \geq \pi_{-1}$ dla zadanego π_{-1} .
- Pytanie: dla jakich q oraz n oczekiwana długość względna przedziału jest najmniejsza.
- Najlepszy wynik jest dla $q = 0 \leftarrow$ 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

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

$$P_{\pi,q} \{Y = 1|Z\}$$

Privacy protection

Privacy protection w modelu krzyżowym

$$P_{\pi,q} \{Y = 1|Z = 1\} = \frac{\pi q}{\pi q + (1 - \pi)(1 - q)},$$

$$P_{\pi,q} \{Y = 1|Z = 0\} = \frac{\pi(1 - q)}{\pi(1 - q) + (1 - \pi)q}.$$

Privacy protection w modelu trójkątnym

$$P_{\pi,q} \{Y = 1|Z = 1\} = \frac{\pi}{\pi + (1 - \pi)q},$$

$$P_{\pi,q} \{Y = 1|Z = 0\} = 0.$$

Privacy protection

Privacy protection

Im mniejsza jest wartość parametru q , tym słabszy jest poziom ochrony prywatności.

Privacy protection

Privacy protection

Im mniejsza jest wartość parametru q , tym słabszy jest poziom ochrony prywatności.

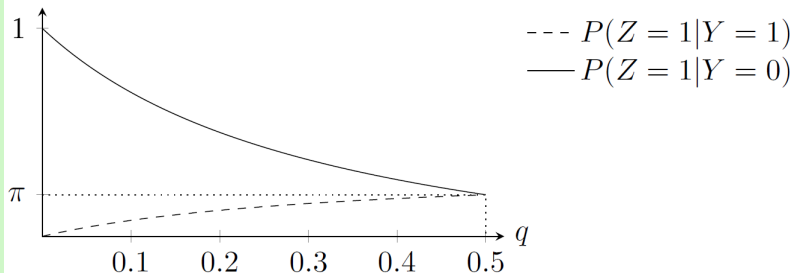
Niech $\gamma \in (0, 1)$.

Chcemy by dla dowolnego $\pi \leq \pi_0$

$$P_{\pi,q} \{Y = 1|Z = 1\} \leq \gamma \text{ oraz } P_{\pi,q} \{Y = 1|Z = 0\} \leq \gamma$$

Privacy protection

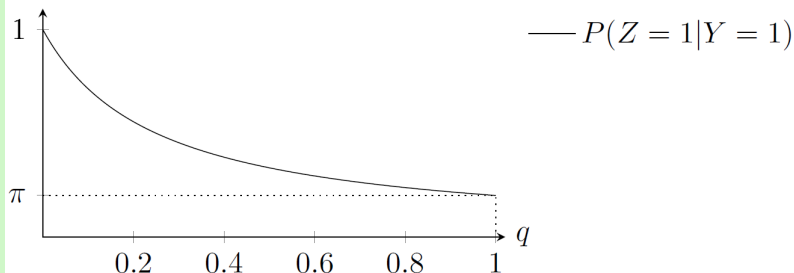
Privacy protection w modelu krzyżowym



$$q(\pi_0; \gamma) = \frac{\pi_0(1 - \gamma)}{\gamma(1 - 2\pi_0) + \pi_0}$$

Privacy protection

Privacy protection w modelu trójkatnym



$$q(\pi_0; \gamma) = \frac{\pi_0(1 - \gamma)}{\gamma(1 - \pi_0)}$$

Optymalny rozmiar próby

Niech $\pi \in \langle \pi_{-1}, \pi_0 \rangle$ ($\pi_{-1} \leq \pi_0 < 0.5$).

Niech $\gamma \leq \pi_0$ oraz d będą ustalone.

Optymalny rozmiar próby: najmniejsze n takie, że

$$E_{\pi} l(\mathcal{Z}; q(\pi_0, \gamma), n) \leq d \quad (\forall \pi \in \langle \pi_{-1}, \pi_0 \rangle)$$

Optymalny rozmiar próby

Niech $\pi \in \langle \pi_{-1}, \pi_0 \rangle$ ($\pi_{-1} \leq \pi_0 < 0.5$).

Niech $\gamma \leq \pi_0$ oraz d będą ustalone.

Optymalny rozmiar próby: najmniejsze n takie, że

$$E_{\pi}l(\mathcal{Z}; q(\pi_0, \gamma), n) \leq d \quad (\forall \pi \in \langle \pi_{-1}, \pi_0 \rangle)$$

Rozwiązanie numeryczne

Optymalny rozmiar próby

Niech $\pi \in \langle \pi_{-1}, \pi_0 \rangle$ ($\pi_{-1} \leq \pi_0 < 0.5$).

Niech $\gamma \leq \pi_0$ oraz d będą ustalone.

Optymalny rozmiar próby: najmniejsze n takie, że

$$E_{\pi}l(\mathcal{Z}; q(\pi_0, \gamma), n) \leq d \quad (\forall \pi \in \langle \pi_{-1}, \pi_0 \rangle)$$

Rozwiązanie numeryczne



Przykład

Optymalny rozmiar próby: $\gamma = 0.5$, $d = 0.5$

Model	π_0	π_{-1}	n^*	π_{-1}	n^*
Krzyżowy	0.1	0.01	83447	0.02	22214
	0.2	0.02	64327	0.04	16726
	0.3	0.03	82676	0.06	21082
	0.4	0.04	209311	0.08	52631
Trójkątny	0.1	0.01	74083	0.02	19690
	0.2	0.02	48016	0.04	12397
	0.3	0.03	46620	0.06	12421
	0.4	0.04	67870	0.08	16593

Na zakończenie

Na zakończenie

Podsumowanie

- Analizowano względną długość przedziału ufności dla frakcji pozytywnych odpowiedzi na pytanie drażliwe.

Na zakończenie

Podsumowanie

- Analizowano względną długość przedziału ufności dla frakcji pozytywnych odpowiedzi na pytanie drażliwe.
- Wyznaczono optymalny rozmiar próby dla modelu krzyżowego i trójkątnego.

Na zakończenie

Podsumowanie

- Analizowano względną długość przedziału ufności dla frakcji pozytywnych odpowiedzi na pytanie drażliwe.
- Wyznaczono optymalny rozmiar próby dla modelu krzyżowego i trójkątnego.
- Zastosowano ważne w tych zagadnieniach kryterium prywatności.

Na zakończenie

Podsumowanie

- Analizowano względną długość przedziału ufności dla frakcji pozytywnych odpowiedzi na pytanie drażliwe.
- Wyznaczono optymalny rozmiar próby dla modelu krzyżowego i trójkątnego.
- Zastosowano ważne w tych zagadnieniach kryterium prywatności.
- Otrzymano bardzo duże liczebności prób.

