Wykład z analizy danych: estymacja punktowa

Marek Kubiak

Instytut Informatyki
Politechnika Poznańska

Plan wykładu

Cel wykładu

Model statystyczny

Pojęcia podstawowe estymacji

Kryteria oceny estymatorów

Przegląd modeli i estymatorów

Cel wykładu

Cel: zaprezentowanie *precyzyjnych* i *praktycznych* sposobów szacowania pewnych *nieznanych* wartości na podstawie badań częściowych (wiedzy niepełnej).

Model

Konstrukcja odwzorowująca dany rodzaj rzeczywistości w sposób uproszczony, sprowadzająca jej cechy do związków najistotniejszych, a mniej istotne pomijająca.

Model statystyczny

W pewnej zbiorowości (*populacji generalnej*) obserwowana jest pewna *cecha* X. Cecha ta ma w tej populacji pewien kształt rozkładu prawdopodobieństwa R (np. $N(\mu, \sigma)$ lub $B_1(p)$) o *nieznanych parametrach* (np. μ, σ, p).

Na podstawie pewnego doświadczenia (badania) zdobyte zostały informacje o wartości cechy X dla pewnej *próbki* elementów populacji w postaci liczb:

$$x_1, x_2, x_3, \ldots, x_n.$$

Model statystyczny

Zazwyczaj powtórzenie takiego doświadczenia daje inne liczby x_i jako wyniki. W takim razie można potraktować te wartości jako *realizacje zmiennych losowych* $X_1, X_2, X_3, \ldots, X_n$. Zmienna losowa X_i określa zachowanie (probabilistyczne) elementu x_i w próbce.

Wektor zmiennych losowych

$$\mathbf{X} = (X_1, X_2, X_3, \ldots, X_n)$$

nazywamy *próbą losową*.

Wektor $\mathbf{X}$ nazywamy *próbą prostą*, jeśli zmienne losowe X_i są niezależne i mają taki sam rozkład R jak badana cecha X populacji.

Wektor wartości $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n)$ nazywamy realizacją próby losowej (*próbką*).

Model statystyczny

Elementy modelu:

- populacja generalna,
- badana (mierzalna) cecha X,
- rozkład R cechy w populacji,
- nieznane parametry θ rozkładu R,
- próba prosta $\mathbf{X}$ (o rozmiarze n) z populacji.

W każdym raporcie z wnioskowania statystycznego trzeba te elementy opisać.

Przykładowe modele sytuacji rzeczywistych

- populacja generalna: wszyscy wyborcy w Polsce
- cecha X: wyborca głosuje na Tuska ($X = 0$) lub na Kaczyńskiego ($X = 1$)
- rozkład R cechy w populacji: $B_1(p)$
- nieznany parametr θ: p
- sens parametru: określa przyszłego prezydenta
- próbka: wylosowana niewielka grupa wyborców (np. $n = 2000$)

Przykładowe modele sytuacji rzeczywistych

- populacja generalna: wszystkie jaja znoszone przez kury pewnej rasy na fermie
- cecha X: waga jaja kury danej rasy
- rozkład R cechy w populacji: $N(\mu, \sigma)$
- nieznane parametry θ: μ, σ
- sens parametrów: określają klasę sprzedaży jaj
- próbka: wylosowana grupa jaj (np. $n = 130$)

Przykładowe modele sytuacji rzeczywistych

- populacja generalna: wszystkie optima lokalne w problemie optymalizacji generowane przez pewien algorytm
- cecha X: wartość funkcji celu dla danego optimum lokalnego
- rozkład R cechy w populacji: nieznany
  (istnieje średnia $E(X) = \mu$)
- nieznany parametr θ: średnia jakość rozwiązania μ
- sens parametru: określa jakość algorytmu
- próbka: wylosowana grupa optimów lokalnych (np. $n = 1000$)

Zadanie estymacji

Zazwyczaj dokładny rozkład R cechy X jest nieznany lub nieznane są jego parametry θ.

Zadanie estymacji: na podstawie pobranej próby losowej oszacować (przybliżyć, ocenić) wartości nieznanych parametrów θ rozkładu R cechy X populacji generalnej z jak największą precyzją i jak najmniejszym kosztem.

Przykład: niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie $N(\mu, \sigma)$, gdzie μ i σ są nieznane; oszacować wartości tych nieznanych parametrów.

Koszt estymacji

Koszt estymacji jest modelowany przez *rozmiar próby*. Im większa próba, tym bardziej kosztowne jest badanie.

Elementy kosztów:

- zniszczenie elementów próby (badania niszczące)
- koszt zakupu lub wynajmu urządzeń pomiarowych
- koszt zatrudnienia i szkolenia ankieterów
- czas procesora
- ...

Wtórny cel estymacji: *minimalizacja wielkości próby*
(przy zachowaniu precyzji oszacowań)

Koszt estymacji

Przykład: badanie wytrzymałości foteli w IKEA

Koszty:

- zniszczenie foteli podlegających badaniu
- zakup urządzeń do automatycznego „siadania”
- eksploatacja tych urządzeń (np. prąd)
- zatrudnienie obsługi urządzeń

Podzadania estymacji

1. Pobranie próby z populacji
2. Obliczenie wartości nieznanych parametrów na podstawie próby

Pobieranie próby prostej z populacji

Postulaty:

- wszystkie elementy populacji mają takie samo prawdopodobieństwo wystąpienia w każdym elemencie X_i próby $\mathbf{X}$
- pobieranie elementów X_i i X_j do próby $\mathbf{X}$ jest niezależne od siebie

Realizacja: schematy losowania z populacji...

Uwaga! Pobieranie próby prostej jest kluczowe w statystyce matematycznej (odróżnia od statystyki opisowej).

Obliczanie wartości nieznanych parametrów z próby

Statystyką nazywamy dowolną funkcję próby losowej postaci:

$$\hat{\theta}(\mathbf{X}) = \hat{\theta}(X_1, X_2, X_3, \ldots, X_n).$$

Przykłady:

$$\hat{\theta}_1(X_1, X_2, X_3, X_4, X_5) = 1$$

$$\hat{\theta}_2(X_1, \ldots, X_{100}) = \frac{X_1 + X_3 + X_5 + \ldots + X_{99}}{45}$$

$$\hat{\theta}_3(X_1, X_2) = X_1 \cdot X_2$$

Obliczanie wartości nieznanych parametrów z próby

Estymatorem parametru θ nazywamy taką statystykę $\hat{\theta}(\mathbf{X})$, którą traktujemy jako oszacowanie parametru θ.

Przykład: dla nieznanego parametru μ w rozkładzie $N(\mu, \sigma)$ mamy estymatory:

$$\hat{\theta}_1(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$$

$$\hat{\theta}_2(X_1, \ldots, X_n) = X_1$$

Kryteria oceny estymatorów

Estymatory można porównywać ze względu na:

- obciążenie,
- zgodność,
- wariancję.

Estymatory – alegoria strzelecka

Estymatory można traktować jak różnych strzelców, którzy próbują trafić w sam środek tarczy.

Równoważne pojęcia:

Estymacja:

- nieznany parametr θ
- wartość estymatora $\hat{\theta}(\mathbf{X})$
- rozmiar (liczność) próby n

Strzelectwo:

- środek tarczy strzeleckiej
- przestrzelone miejsce
- większy wysiłek strzelca (podpórka, skupienie, przymiarki, ocena wiatru)

Estymator nieobciążony

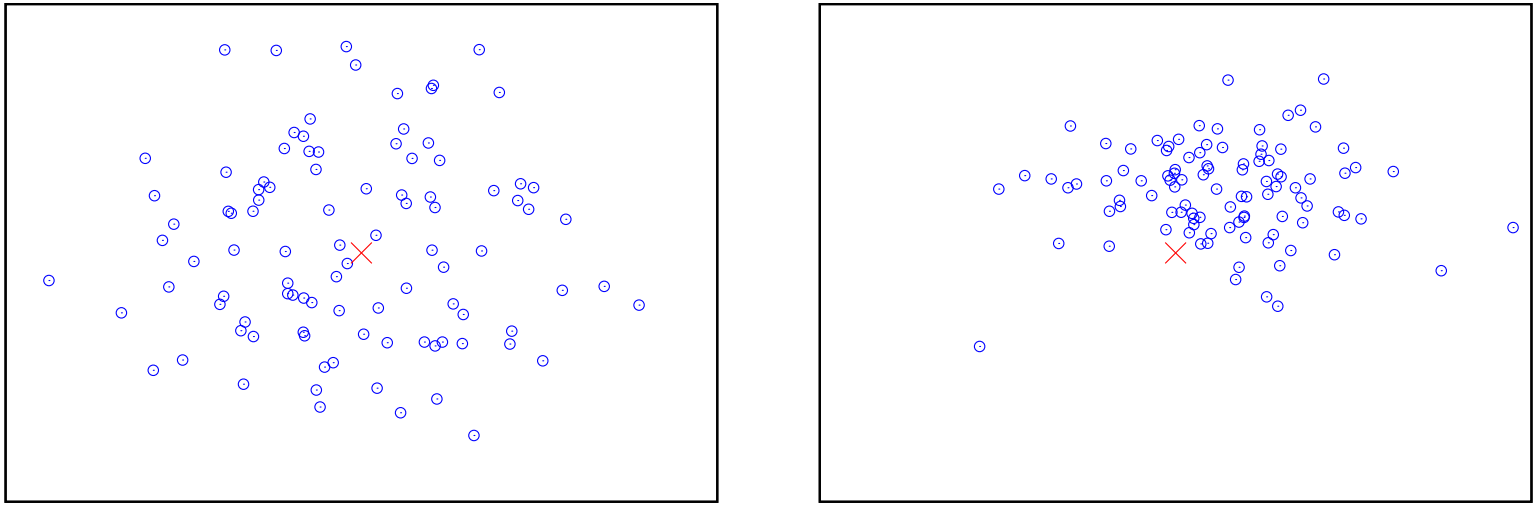
Estymator $\hat{\theta}(\mathbf{X})$ parametru θ nazywamy *nieobciążonym*, gdy spełniony jest warunek:

$$E(\hat{\theta}(\mathbf{X})) = \theta.$$

Komentarz:

- estymator nieobciążony średnio trafia w nieznaną wartość parametru,
- rozsądne wymaganie, chociaż w ogóle nie bierze pod uwagę liczności próby (czyli kosztu zebrania informacji).

Obciążenie w estymacji – alegoria strzelecka



Rysunek: Alegoria estymatora nieobciążonego i obciążonego.

Obciążenie w estymacji

Jeśli estymator jest obciążony (ukierunkowany, skrzywiony), to bez kompensacji obciążenia *zniekształca wnioski*!

Przykłady estymacji obciążonej:

- strzelanie z krzywą lufą,
- ocena momentu południa na podstawie dźwięku dzwonu 1km od kościoła,
- dokładne ustawianie czasu przez Internet na podstawie odległego wzorca.

Przykładowe przyczyny obciążenia estymacji:

- ukierunkowany sposób doboru próby,
- błędy lub przekłamania w zapisie wyników pomiarów,
- uszkodzenie lub dekalibracja urządzeń pomiarowych,
- użycie niewłaściwego estymatora.

Obciążenie w estymacji – przykład C. R. Rao

Tło badania:

- Dehli, Indie, 1947 – rozruchy religijne po odzyskaniu niepodległości,
- konieczność płacenia przez rząd dostawcom żywności do dwu obozów uchodźców religijnych (wysokie wydatki),
- niemożliwość wejścia statystyków do obozów (większość religijna).

Zadanie: oszacować liczbę ludności L w obozach w celu oceny wiarygodności rachunków za żywność.

Dane: rachunki dostawców za ryż, groch, sól.

Obciążenie w estymacji – przykład C. R. Rao

Niech:

- R, G, S: ilości ryżu, grochu i soli zużywane na wyżywienie wszystkich uchodźców przez 1 dzień,
- r, g, s: ilości dziennego zapotrzebowania na osobę wg badań konsumenckich.

Wtedy: $R/r, G/g, S/s$ to estymatory tej samej, nieznanej liczby ludności L.

Spostrzeżenia:

- R/r jest największe, S/s jest najmniejsze, różnica jest znaczna,
- rynkowa cena ryżu jest wysoka, cena soli jest znikoma.

Obciążenie w estymacji – przykład C. R. Rao

Wnioski:

- wartość R jest sztucznie zawyżona przez dostawców (efektywne zawyżanie rachunków),
- dla przyszłych rozliczeń zastosować estymatę $L = S/s$ (*najmniej obciążona*).

Weryfikacja wniosków:

- niezależne oszacowanie liczby uchodźców w mniejszym obozie potwierdziło najmniejszy błąd oceny $L = S/s$.

Obciążenie w estymacji – wybory prezydenckie 1936

Tło badania:

- wybory prezydenckie w USA, rok 1936;
  A. E. Landon vs. F. D. Roosevelt

Zadanie: przewidzieć zwycięzcę (oszacować p w rozkładzie $B_1(p)$)

Dane: badanie ankietowe

Obciążenie w estymacji – wybory prezydenckie 1936

Badanie Literary Digest (dużego czasopisma):

- 2 000 000 ankietowanych
- wynik: Landon 57%, Roosevelt 43%

Badanie George'a Gallupa:

- 300 000 ankietowanych
- wynik: wygra Roosevelt

Wybory wygrywa Roosevelt

Obciążenie w estymacji – wybory prezydenckie 1936

Przyczyny błędu badania Literary Digest:

- pobranie próby wyborców z listy właścicieli telefonów i samochodów
- *obciążenie próby*: preferencja dla klas średniej i wyższej (wyborcy Landona) przed klasą niższą (wyborcy Roosevelta)

Wnioski:

- sposób pobrania próby z populacji może mieć ogromny wpływ na wartości szacowane z próby (obciążenie)
- duży rozmiar próby nie gwarantuje poprawności (jedynie wysokie koszty)

Obciążenie estymacji – przykładowe modele

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie $N(\mu, \sigma)$, gdzie μ jest nieznane, a σ znane.

Estymatory dla μ:

$$\hat{\theta}_1(\mathbf{X}) = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\hat{\theta}_2(\mathbf{X}) = X_1$$

Obydwa estymatory są nieobciążone, choć drugi jest mało zdroworozsądkowy.

Obciążenie estymacji – przykładowe modele

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie $N(\mu, \sigma)$, gdzie μ i σ są nieznane.

Estymatory dla σ^2:

$$\hat{\theta}_1(\mathbf{X}) = S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\hat{\theta}_1(\mathbf{X}) = S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Estymator S^2 jest obciążony (zaniża wartość wariancji)
Estymator S^{*2} jest nieobciążony.

Stopnie swobody (dygresja)

Estymator dla σ^2:

$$S^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

korzysta z $\overline{X}$ (zamiast z nieznanego μ).

Przyjęcie:

$$\mu = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

wprowadza jedną liniową zależność pomiędzy zmiennymi X_i, np.:

$$\frac{X_1 + \ldots + X_{10}}{10} = 5$$

Stopnie swobody (dygresja)

Ta liniowa zależność pozwala na *ustalenie wartości* jednej zmiennej w próbie, np.:

$$X_1 = 50 - (X_2 + \ldots + X_{10})$$

czyli oszacowanie wpierw μ z próby powoduje „stratę" jednej zmiennej niezależnej w próbie!

Stopnie swobody to (intuicyjnie) liczba niezależnych zmiennych w próbie, na podstawie których dokonujemy estymacji

Estymator zgodny

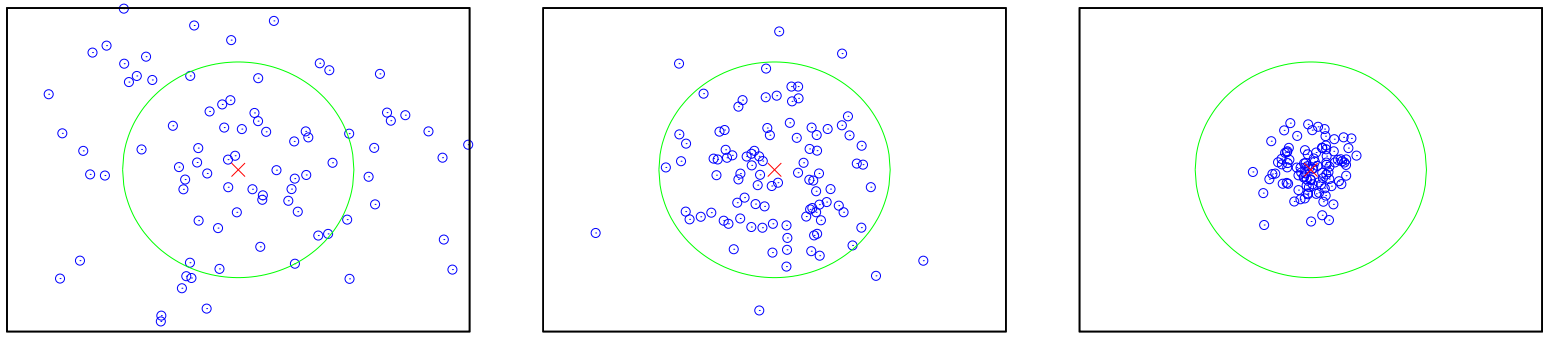
Estymator $\hat{\theta}(\mathbf{X})$ parametru θ nazywamy *zgodnym*, gdy spełniony jest warunek:

$$\forall\ \mathbf{X}, \theta, \epsilon > 0 \qquad \lim_{n \to \infty} P(|\hat{\theta}_n(\mathbf{X}) - \theta| < \epsilon) = 1$$

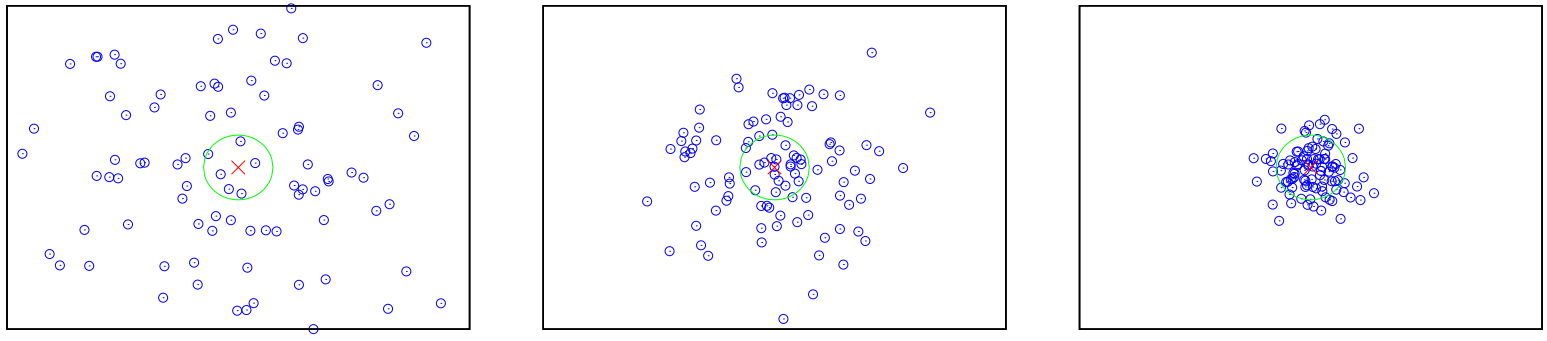
Komentarz:

- w miarę wzrostu liczności próby n prawdopodobieństwo znacznego odchylenia estymatora od parametru maleje do zera,
- rozsądne wymaganie: zwiększanie kosztów badania ma prowadzić do większej dokładności przybliżenia,
- estymator bez tej własności jest *bardzo kiepski*.

Estymator zgodny – alegoria strzelecka



Rysunek: Alegoria zgodności estymatora: $\epsilon = 1$ i $n = 10, 100, 1000$.



Rysunek: Alegoria zgodności estymatora: $\epsilon = 0.3$ i $n = 10, 100, 1000$.

Estymator najefektywniejszy

Estymator nieobciążony $\hat{\theta}_1$ nazywamy *efektywniejszym* od nieobciążonego estymatora $\hat{\theta}_2$, gdy spełniony jest warunek:

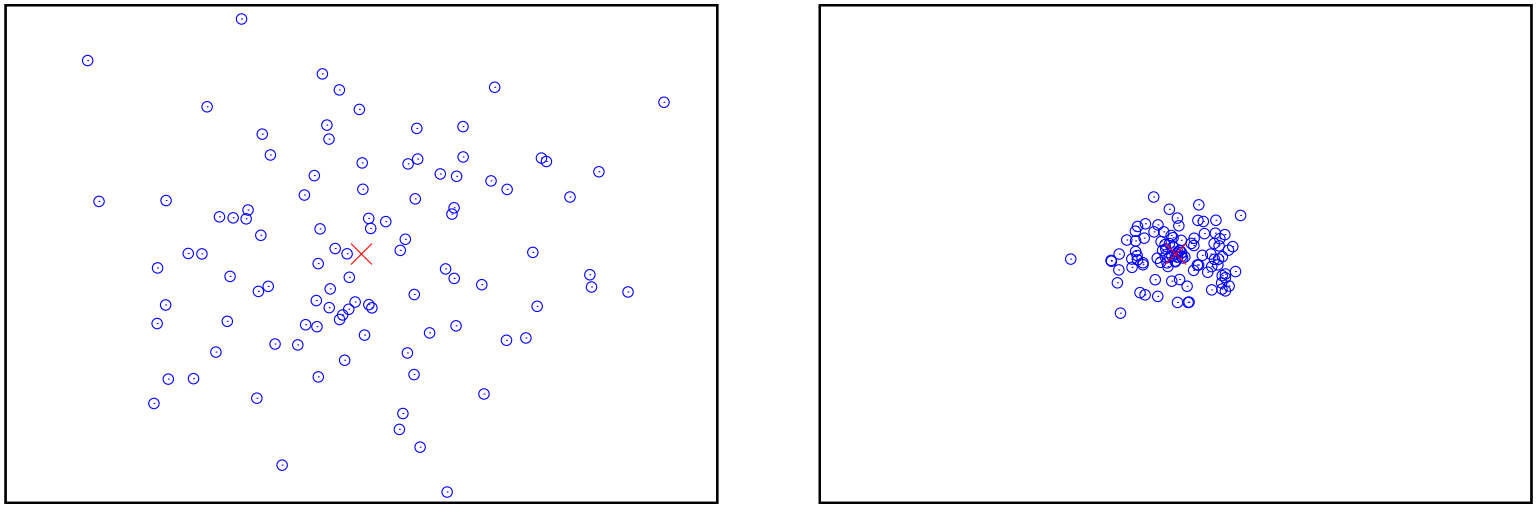
$$\forall\ \theta \quad D^2(\hat{\theta}_1) \leq D^2(\hat{\theta}_2).$$

Estymator nieobciązony parametru θ nazywamy *najefektywniejszym*, gdy ma najmniejszą możliwą wariancję ze wszystkich estymatorów nieobciążonych tego parametru.

Komentarz:

- wariancja estymatora to rozrzut wokół nieznanego parametru,
- rozsądne wymaganie: minimalizacja tego rozrzutu,
- w miarę możliwości używamy estymatora najefektywniejszego.
- wariancja estymatora ma zazwyczaj spory związek z wykorzystaniem przez niego pełnej informacji z próby losowej,
- estymator najefektywniejszy może nie istnieć.

Wariancja estymatorów nieobciążonych – alegoria strzelecka



Rysunek: Alegoria wariancji estymatorów nieobciążonych.

Wariancja estymatorów nieobciążonych – przykładowy model

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie $N(\mu, \sigma)$, gdzie μ jest nieznane, a σ znane.

Estymatory nieobciążone dla μ:

$$\hat{\theta}_1(\mathbf{X}) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\theta}_2(\mathbf{X}) = X_1$$

Można pokazać, że:

$$\sigma^2/n = D^2(\hat{\theta}_1) \leq D^2(\hat{\theta}_2) = \sigma^2.$$

Wnioski:

- zdecydowanie wybieramy estymator $\hat{\theta}_1(\mathbf{X}) = \overline{X}$,
- wiadomo, że jest to estymator najefektywniejszy parametru μ.

Kryteria oceny estymatorów – podsumowanie

- obciążenie $\Leftrightarrow$ systematyczny błąd w ocenie
- zgodność $\Leftrightarrow$ im większa próba, tym lepsza dokładność
- wariancja $\Leftrightarrow$ rozrzut wokół ocenianego parametru

Przegląd estymatorów

Typy rozkładów badanej cechy X:

- rozkład dowolny (nieznany)
- rozkład normalny
- rozkład zero-jedynkowy

Przegląd estymatorów – rozkład dowolny

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie dowolnym R, gdzie $E(X_i) = \mu < \infty$ i $0 < D^2(X_i) = \sigma^2 < \infty$, oraz μ, σ^2 są nieznane.

Estymator parametru μ postaci:

$$\hat{\theta}(\mathbf{X}) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

jest zgodny i nieobciążony.

Estymator parametru σ^2 postaci:

$$\hat{\theta}(\mathbf{X}) = S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

jest zgodny, nieobciążony i najefektywniejszy (asymptotycznie).

Przegląd estymatorów – rozkład dowolny

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie dowolnym R, gdzie $E(X_i) = \mu < \infty$ i $0 < D^2(X_i) = \sigma^2 < \infty$, oraz μ, σ^2 są nieznane.

Estymator parametru σ postaci:

$$\hat{\theta}(\mathbf{X}) = S^* = \sqrt{S^{*2}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

jest zgodny i *obciążony*.

Przegląd estymatorów – rozkład normalny

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie $N(\mu, \sigma)$, gdzie μ, σ^2 są nieznane.

Estymator parametru μ postaci:

$$\hat{\theta}(\mathbf{X}) = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

jest zgodny, nieobciążony i najefektywniejszy.

Estymator parametru σ^2 postaci:

$$\hat{\theta}(\mathbf{X}) = S^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

jest zgodny, nieobciążony i najefektywniejszy (asymptotycznie).

Przegląd estymatorów – rozkład zero-jedynkowy

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie $B_1(p)$, gdzie p jest nieznane.

Estymator parametru p postaci:

$$\hat{\theta}(\mathbf{X}) = \hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

jest zgodny, nieobciążony i najefektywniejszy.

Podstawowe estymatory w MS Excel

Funkcje MS Excel odpowiadające podstawowym estymatorom:

- $\overline{X}$: ŚREDNIA
- S^{*2}: WARIANCJA
- S^2: WARIANCJA.POPUL
- S^*: ODCH.STANDARDOWE
- S: ODCH.STANDARDOWE.POPUL

Estymacja przedziałowa – inny sposób estymacji parametrów

Dwa rodzaje estymacji parametrów rozkładu:

- punktowa: wynikiem jest *punkt*, który *najlepiej przybliża* nieznaną wartość parametru θ
- przedziałowa: wynikiem jest *przedział ufności*, który z zadanym prawdopodobieństwem *pokrywa* nieznaną wartość parametru θ

Alegoria strzelecka: nie strzelamy już kulą (punktem), ale np. *siecią*, która ma owinąć nieznany cel.

Estymacja przedziałowa – przykład

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą prostą z populacji o rozkładzie $N(\mu, \sigma)$, gdzie μ jest nieznane, a σ^2 jest znane.

Przedziałem ufności na poziomie $(1 - \alpha)$ dla parametru μ jest:

$$\left(\overline{X} - u(1 - \frac{\alpha}{2}) \cdot \frac{\sigma}{\sqrt{n}} \ , \ \overline{X} + u(1 - \frac{\alpha}{2}) \cdot \frac{\sigma}{\sqrt{n}} \right)$$

Szerokość przedziału (chcemy minimalizować):

- rośnie, gdy rośnie poziom ufności $(1 - \alpha)$
- maleje, gdy rośnie rozmiar próby n

Wymagana wiedza

- model i model statystyczny
- zadanie estymacji
- koszt estymacji w modelu statystycznym
- próba losowa, różnica w stosunku do zbioru danych
- estymator: pojęcie i alegoria strzelecka
- obciążenie: sens, alegoria i możliwe przyczyny
- zgodność: sens i alegoria
- wariancja, efektywność: sens, alegoria
- pojęcie stopni swobody
- przykładowe estymatory i ich cechy:
  - rozkład dowolny
  - rozkład normalny
  - rozkład zero-jedynkowy
- podstawowe estymatory w MS Excel

Literatura

- W. Starzyńska, *Statystyka praktyczna*, Wydawnictwo Naukowe PWN, 2000.
- J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, 2001.
- G. A. Ferguson, Y. Takane, *Analiza statystyczna w psychologii i pedagogice*, Wydawnictwo Naukowe PWN, Warszawa, 2003.
- W. Krysicki i inni, *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, Część II, Statystyka matematyczna*, Wydawnictwo Naukowe PWN, Warszawa, 1995.
- C. R. Rao, *Statystyka i prawda*, Wydawnictwo Naukowe PWN, 1994.

Dziękuję za uwagę!