Seminarium WFiIS AGH, 25 listopada 2005

STATYSTYKA ODPORNOŚCIOWA

referat dydaktyczny

Plan:

1. Statystyka klasyczna
2. Powstanie statystyki odpornościowej
3. Estymatory statystyki odpornościowej
4. Własności estymatorów
5. Związek estymatorów z funkcjami rozkładu
6. Iteracyjnie ważona metoda najmniejszych kwadratów
7. Dopasowanie prostej & innych funkcji
8. Zastosowanie elementów statystyki odpornościowej w rachunku niepewności pomiaru
9. Podsumowanie

1. STATYSTYKA KLASYCZNA

■ Wyróżniony rozkład: normalny $g(x) \propto \exp\left[-\frac{(x_i-\mu)}{2\sigma^2}\right]$

Powód: centralne twierdzenie graniczne (od XVIII w)

■ Wyróżnione parametry:

położenia – wartość oczekiwana $\mu = \int_{-\infty}^{\infty} x\ g(x)\ dx$

skali – odchylenie standardowe $\sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2\ g(x)\ dx}$

Powód: własność addytywności μ oraz σ^2 dla sumy zmiennych losowych

■ Wyróżnione estymatory:

położenia – średnia $\bar{x} = \frac{\sum x_i}{n}$

skali – odchylenie standardowe $s_x = \sqrt{\frac{\sum (x_i-\bar{x})^2}{n-1}}$

Powody:

- łatwy w obliczeniach wzór analityczny,
- Gauss (początek XIX w): średnia jest – dla rozkładu normalnego – najbardziej efektywnym estymatorem położenia środka rozkładu
- XIX w: dyskusja, co jest lepszą miarą rozrzutu

  Edington i inni: preferencja dla **odchylenia średniego** $= \frac{\sum |x_i-\bar{x}|}{n}$
- Fisher (1920): odchylenie standardowe jest – dla rozkładu normalnego – estymatorem skali 12% bardziej efektywnym od odchylenia średniego

⟶ DOMINACJA STATYSTYKI KLASYCZNEJ

2. POWSTANIE STATYSTYKI ODPORNOŚCIOWEJ

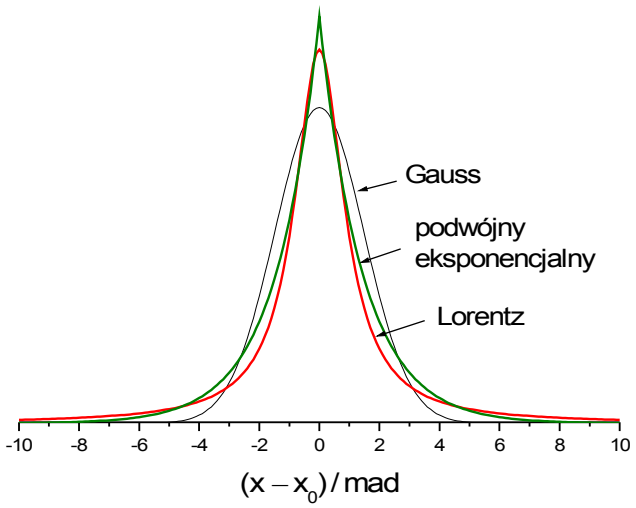
Przesłanki:

- istnieją rozkłady dla których parametry μ i σ nie istnieją:

np. rozkład Lorentza (Cauchy'ego)

$$g(x) \propto \frac{1}{1+[(x_i - x_0)/\Gamma]^2}$$

są to rozkłady o wolno malejących ogonach



- Dla rozkładów takich estymatory klasyczne rozbiegają się:

Przykładowa symulacja MC dla liczb o znormalizowanym rozkładzie Lorentza: zachowanie średniej arytmetycznej, estymatora odchylenia standardowego, mediany i odchylenia medianowego w funkcji liczebności *n* próby losowej (wykonano przy użyciu programu ORIGIN).

n	$\bar{x}$	*s*	med	MAD
5	0,03	0,05	0,49	0,185
10	2,05	7,01	0,49	1,019
20	–2,38	6,18	–0,22	1,235
50	–5,61	31,4	0,15	1,333
100	1,09	43,9	0,093	0,845
200	–0,04	36,2	0,093	0,909
500	0,12	24,0	0,05	0,956
1000	0,33	19,2	0,054	0,993
2000	–3,03	134	0,041	1,016
5000	1,25	218	0,044	1,01
10000	0,87	156	0,007	1,003
20000	4,11	613	0,006	0,998
50000	1,34	391	0,003	1,005
100000	0,40	289	–0,0004	0,9983
200000	0,64	285	–0,0004	1,0006
500000	–2,17	1766	–0,00014	1,00026
1000000	–9,90	9175	0,00005	1,0001

- Statystyka w praktyce: co robić z punktami odstającymi (ang. *outliers*)?
  - subiektywne odrzucanie „błędów grubych"
  - kryteria statystyczne odrzucania (trzeba zadać graniczne P)
  - obcięta średnia (*trimmed mean*): odrzucamy zadaną część punktów

- Tukey (1960): obliczenia (analityczne) dla modelu punktów odstających postaci

$$g(x)=(1-\varepsilon)g_N(x) + \varepsilon\, g_{out}(x)$$

r. normalny r. zakłócający

Dla $g_{out}(x)$ będącej rozkładem też normalnym, ale o trzykrotnie większym σ :
- dla $\varepsilon = 0$ odchylenie standardowe o 12% bardziej efektywne od średniego
- już dla $\varepsilon = 1\%$ odchylenie średnie jest o 44% bardziej efektywne od standardowego !

Uogólnienie: małe odchylenia od wyidealizowanych założeń powodują zaskakująco duże skutki

Tukey: autor nazwy ROBUST STATISTIC

- Lata 60-te: powstanie statystyki odpornościowej jako ścisłego działu teorii prawdopodobieństwa
  Tukey, Huber, Hampel .....

- Obecnie:
  popularność „robust statistics" wśród matematyków (morze problemów),

  niedostateczna „transmisja" do dydaktyki, popularnych podręczników, programów komputerowych i zastosowań

  nazwa polska STATYSTYKA ODPORNOŚCIOWA
  (wg. A. L. Dawidowicz, UJ)

3. ESTYMATORY STATYSTYKI ODPORNOŚCIOWEJ

MEDIANA

Definicja: $$\mathrm{med}\{x_i\} = \begin{cases} x_{n/2} & (n\ parzyste) \\ (x_{(n-1)/2} + x_{(n+1)/2})/2 & (n\ nieparzyste) \end{cases}$$

Przykład: 9 pomiarów okresu wahadła [ms]

p. odstający

1275 1278 1280 1284 1287 1292 1293 1306 1416

↑

$\mathrm{med}\{x_i\} = 1287$ ms

ODCHYLENIE MEDIANOWE

ang. MAD – *median absolute deviation,* lub *median deviation*

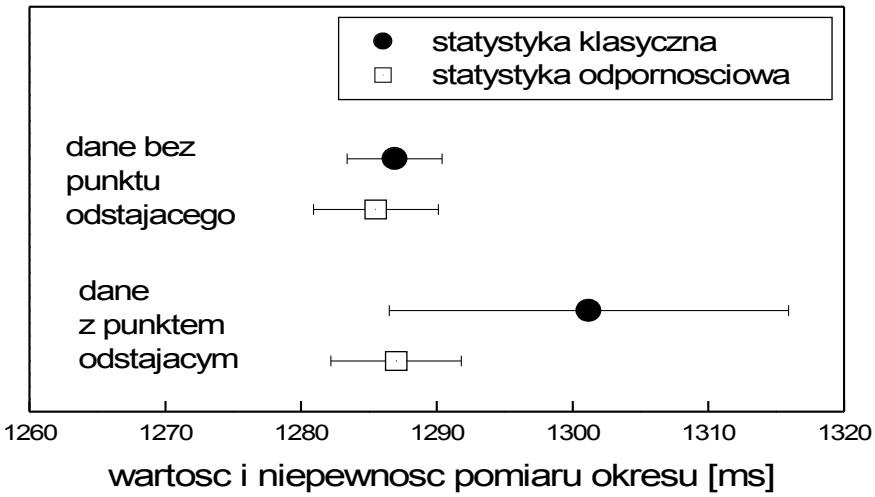
Definicja: $\mathrm{MAD}\{x_i\} = \mathrm{med}\{\,|x_i - \mathrm{med}\{x_i\}|\,\}.$

Przykład: uszeregowany zbiór różnic $|x_i - med\{x_i\}|$:

0 3 4 5 7 9 12 19 129

↑

$\mathrm{MAD}\{T_i\} = 7$ ms.



I WIELE INNYCH ESTYMATORÓW.....

4. WŁASNOŚCI ESTYMATORÓW

(badane w statystyce odpornościowej)

- Odporny – nieodporny
  (*robust – nonrobust*)

Nieodporny – dodanie pojedynczego punktu może zmienić wartość estymatora o dowolną liczbę, niezależnie od liczebności próby

Np. nieodporne są: $\bar{x}$, s_x , odchylenie średnie

- Granica załamania ε^*
  (*breaking point, breaking bound*)

Maksymalna udział „błędów grubych", który nie zniweczy odporności

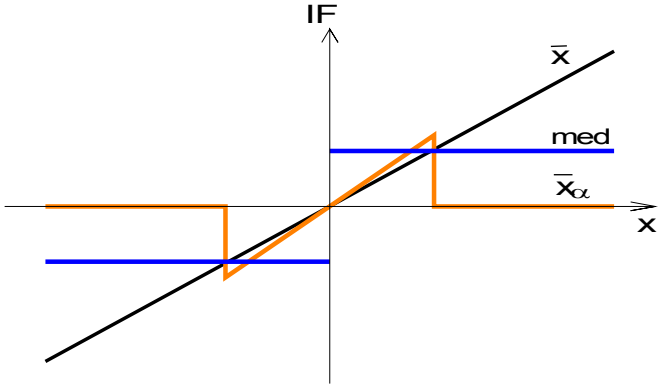
Dla estymatorów nieodpornych $\varepsilon^* = 0$
Mediana, odchylenie medianowe $\varepsilon^* = 0{,}5$

- Funkcja wpływu IF
  ang. influence function (Hampel 1974)

Jakościowo: wpływ dodania liczby x na wartość estymatora

$$IF(x, F, T) = \lim_{\varepsilon \to 0} \frac{T\{(1-\varepsilon)F + \varepsilon \delta_x\} - T\{F\}}{\varepsilon}$$

T –estymator
F – funkcja rozkładu
δ_x – funkcja δ Diraca (Huber: pointmass 1 at x)



Efektywność

$$\text{Efektywność} = \frac{\text{najmniejsza mozliwa wariancja}}{\text{wariancja aktualna}}$$

Efektywność gaussowska = jw., dla rozkładu normalnego

Zwykle da się podać analitycznie efektywność asymptotyczną (dla $n >> 1$)

Estymator	Granica załamania ε *	Efektywność gaussowska (asymptotyczna)
Średnia	0	1
Obcięta średnia	α[1]	
Mediana	0,5	$2/\pi \approx 0{,}64$
Dwukwadrat Tukeya	0,5	0,95
Odchylenie standardowe	0	1
Odchylenie średnie	0	0,88
Odległość 1 – 3 kwantyl	0,25	0,37
Odchylenie medianowe	0,5	0,37

Pożądana cecha odporności estymatora jest uzyskana kosztem niezbyt wielkiej straty efektywności.

(Stwierdzenie o stracie efektywności dotyczy próba losowej o rozkładzie normalnym. Dla rozkładów z wolniej zanikającymi ogonami estymator odporny jest zwykle również bardziej efektywny.)

[1] Parametr α oznacza, że obcinamy od góry i dołu ułamek punktów równy α

5. ESTYMATORY A FUNKCJA ROZKŁADU

●● $\sum_{i=1}^{n} w_i \left[x_i - T \right]^2 = \min$

tj. otrzymujemy metodę najmniejszych kwadratów (NK) z wagami

$$w_i = \frac{1}{\sigma_i^2}$$

●●● $\sum w_i \left[x_i - T \right] = 0 \quad \rightarrow \quad \sum w_i x_i - T \sum w_i = 0 \quad \rightarrow$

●●●● $T = \bar{x} = \frac{\sum w_i x_i}{\sum w_i}$ średnia arytmetyczna (ważona)

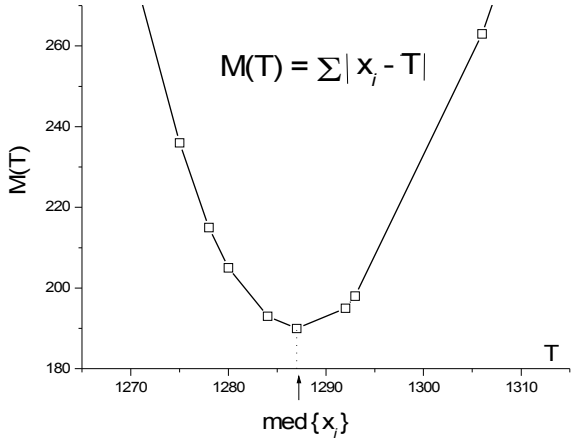
Przykład: podwójny rozkład wykładniczy
[dla prostoty – bez wag]

- $P \propto \prod_{i=1}^{n} \exp\left[-|x_i - T|\right] \rightarrow P \propto \exp\left\{-\sum_{i=1}^{n} |x_i - T|\right\} = \max$

●● $M(T) = \sum |x_i - T| = \min$

(tj. minimalizacja sumy wartości bezwzględnych)

Wykres M(T) dla danych z punktu 3 --->



●●● $\sum \text{sign}(x_i - T) = 0$

●●●● $T = \text{med}\{x_i\}$

(mediana)

Różne sposoby obliczenia estymatora:

(i) numeryczne poszukiwanie minimum funkcji kryterialnej M (●●)

(ii) rozwiązanie r. normalnego ●●● . Uzyskanie zamkniętego wzoru (●●●●) możliwe tylko w wybranych przypadkach (średnia, mediana)

(iii) rozwiązanie równania normalnego metodą iteracyjnie ważonych najmniejszych kwadratów (verte)

6. ITERACYJNIE WAŻONA METODA NAJMNIEJSZYCH KWADRATÓW

(iteratively reweighted least squares method) :

Najważniejsza w praktyce metoda obliczania estymatorów największej wiarygodności vel. typu M (ang. M-estimators)

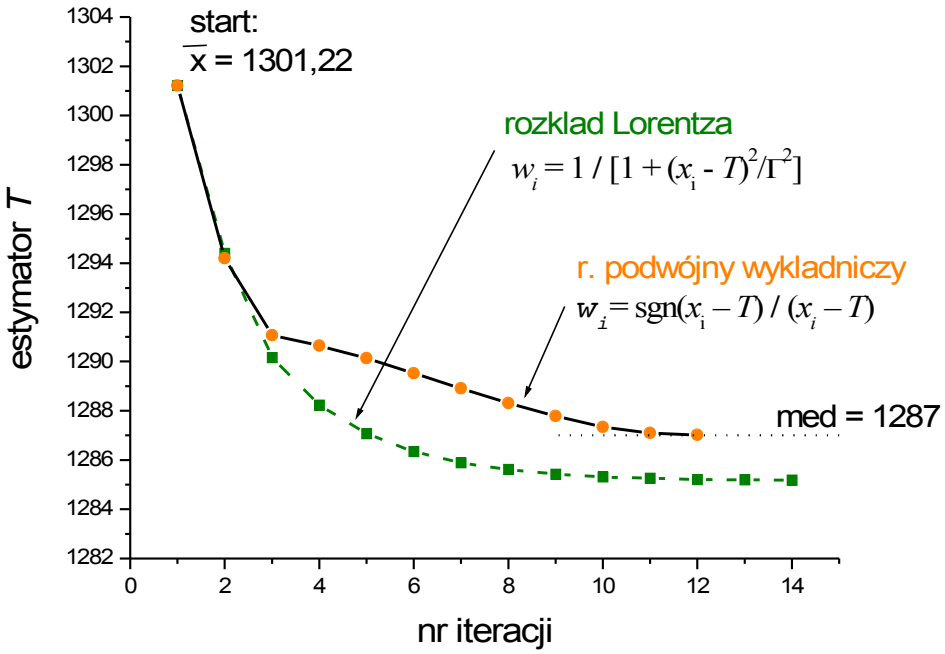
Wyprowadzenie:
równanie normalne (•••) zapisać można jako:

$$\sum \psi(x_i - T) = \sum \frac{\psi(x_i - T)}{x_i - T}(x_i - T) = \sum w_i (x_i - T) = \min$$

formalnie dostaliśmy metodę NK z wagami $w_i = \frac{\psi(x_i - T)}{x_i - T}$,
tyle, że wagi te zależą od szukanego estymatora T

Rozwiązanie iteracyjne: obliczamy cyklicznie wagi w_i i estymator T, aż do uzyskania samouzgodnienia:

Przykład: (dla danych jak w pkt. 3)



7. DOPASOWANIE PROSTEJ & DOWOLNEJ FUNKCJI

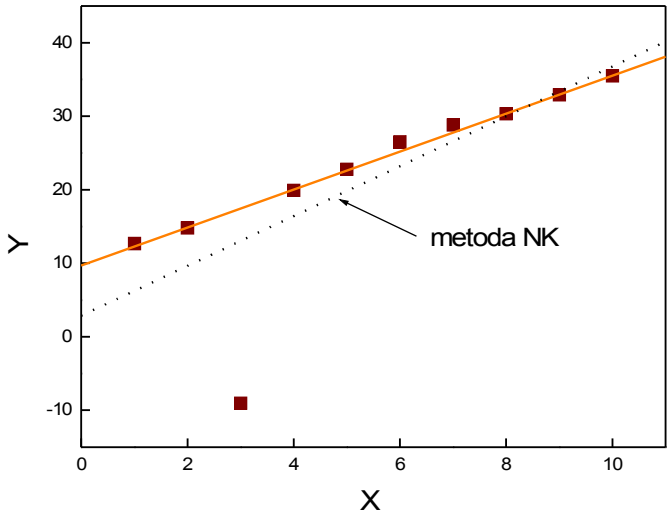
Prawie bezpośrednia adaptacja formalizmu estymatorów największej wiarygodności. Zamiast jednego równania normalnego →

●●● Układ równań normalnych (liczba równań = liczbie parametrów),

Rozwiązanie: najczęściej ITERATYWNIE WAŻONA METODA NAJMNIEJSZYCH KWADRATÓW, czyli

zwykła metody NK (dla prostej lub innej funkcji) + wyrażenia na wagi (jak poprzednio) + rozwiązanie samouzgodnione metodą iteracji.

Przykład dopasowania prostej:



metodą najmniejszych kwadratów (- - - - -)
minimalizacja sumy modułów (———)

Metoda jest ZAIPLEMENTOWANA w wielu programach

np. w dostępnym dla pracowników WFiIS programie MatLab
(info: Z. Stęgowski – dziękuję !)

Są tam do wyboru 3 opcje dopasowania:

LS - metoda najmniejszych kwadratów

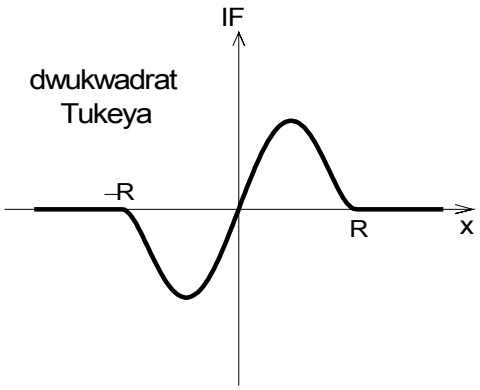
i dwie metody odporne (robust):

LAR – minimalizacja sumy wartości bezwzględnych

Bisquare – iteracyjnie ważona metoda najmniejszych kwadratów
z wagami zaproponowanymi przez Tukeya:

$$w_i = \begin{cases} [1+(r/R)^2]^2 & dla\ |r| \le R \\ 0 & dla\ |r| > R \end{cases}$$

gdzie $\boldsymbol{R = 6{,}946\ \mathrm{MAD}\{r_i\}}$ jest parametrem obcięcia. (Punkty z błędem resztkowym $r_i = y_i - y_i^{(calc)}$ większym od R są całkowicie ignorowane, co pokazuje też wykres funkcji wpływu dla tego estymatora, vide ↓).
Zaletą dwukwadratu Tukeya (bisquare, Tukey's bisquare) jest wysoka efektywność (vide tabela własności estymatorów).



8. ZASTOSOWANIE ELEMENTÓW STATYSTYKI ODPORNOŚCIOWEJ W RACHUNKU NIEPEWNOŚCI POMIARU

Dokumenty konwencji GUM nie przewidują *explicite* stosowania statystyki odpornościowej w rachunku niepewności pomiaru.
Amatorska analiza prawna:

Przewodnik, opis metod typu A:
„W większości przypadków, najlepszym oszacowaniem wartości oczekiwanej ... jest średnia arytmetyczna ...".

→ Możliwość utożsamiania innych estymatorów – np. mediany, z wynikiem pomiaru.

Przewodnik, definicja niepewności:
Niepewność pomiaru jest związanym z rezultatem pomiaru parametrem, charakteryzującym rozrzut wyników, który można w uzasadniony sposób przypisać wartości mierzonej.

→ nie wyklucza stosowania miar innych niż odchylenie standardowe

Propozycja:

Niepewność pojedynczego pomiaru ≡ 1,483 MAD

(Czynnik 1,483 jest stosunkiem σ/MAD dla rozkładu normalnego)

Niepewność pomiaru dla próby losowej o n elementach:

$$u(x) \equiv \frac{1{,}859\ MAD}{\sqrt{n}}$$

(Współczynnik 1,859 = współczynnik poprzedni razy czynnik $(\pi/2)^{1/2}$ uwzględniający stratę efektywności wynikającą z zastąpienia średniej przez medianę (vide tabela na str.7). Obydwa wzory pokryją się z definicją standardową dla liczb o rozkładzie Gaussa i w granicy asymptotycznej $n \gg 1$).

PODSUMOWANIE

Statystyka odpornościowa nie jest alternatywą statystyki klasycznej, lecz jej uzupełnieniem, gdy dane charakteryzują długie „ogony" funkcji rozkładu i/lub obecność punktów odstających.

Zasługuje na szersze stosowanie w badaniach naukowych i aplikacyjnych, niż to ma miejsce obecnie

Opracowanie pliku pdf 19-01-06. Przemysław Gawroński