Ryszard Zieliński, IMPAN Warszawa

O ŚREDNIEJ ARYTMETYCZNEJ I MEDIANIE

XXXIX Ogólnopolska Konferencja Zastosowań Matematyki
Zakopane-Kościelisko 7 - 14 września 2010 r.

Model statystyczny pomiaru: wynik pomiaru $\mathbf{X} = \mu + \varepsilon$

Model statystyczny pomiaru: wynik pomiaru $\mathbf{X} = \mu + \varepsilon$

1. ε jest zmienną losową

Model statystyczny pomiaru: wynik pomiaru $\mathbf{X} = \mu + \varepsilon$

1. ε jest zmienną losową
2. $E(\varepsilon) = 0$ – pomiar nieobciążony, pomiar bez błędu systematycznego

Model statystyczny pomiaru: wynik pomiaru $\mathbf{X} = \mu + \varepsilon$

1. ε jest zmienną losową
2. $E(\varepsilon) = 0$ – pomiar nieobciążony, pomiar bez błędu systematycznego
3. błąd symetryczny względem zera (jednakowo prawdopodobne błędy dodatnie i ujemne)

Model statystyczny pomiaru: wynik pomiaru $\mathbf{X} = \mu + \varepsilon$

1. ε jest zmienną losową
2. $E(\varepsilon) = 0$ – pomiar nieobciążony, pomiar bez błędu systematycznego
3. błąd symetryczny względem zera (jednakowo prawdopodobne błędy dodatnie i ujemne)
4. duże (bezwzględnie) błędy mniej prawdopodobne niż małe

Model statystyczny pomiaru: wynik pomiaru $\mathbf{X} = \mu + \varepsilon$

1. ε jest zmienną losową
2. $E(\varepsilon) = 0$ – pomiar nieobciążony, pomiar bez błędu systematycznego
3. błąd symetryczny względem zera (jednakowo prawdopodobne błędy dodatnie i ujemne)
4. duże (bezwzględnie) błędy mniej prawdopodobne niż małe
5. krzywa Gaussa

Model statystyczny pomiaru: wynik pomiaru $X = \mu + \varepsilon$

1. ε jest zmienną losową
2. $E(\varepsilon) = 0$ – pomiar nieobciążony, pomiar bez błędu systematycznego
3. błąd symetryczny względem zera (jednakowo prawdopodobne błędy dodatnie i ujemne)
4. duże (bezwzględnie) błędy mniej prawdopodobne niż małe
5. krzywa Gaussa
6. $\varepsilon \sim N(0, \sigma)$, $X \sim N(\mu, \sigma)$, σ – dokładność pomiaru, znana lub nieznana

Model statystyczny pomiaru: wynik pomiaru $X = \mu + \varepsilon$

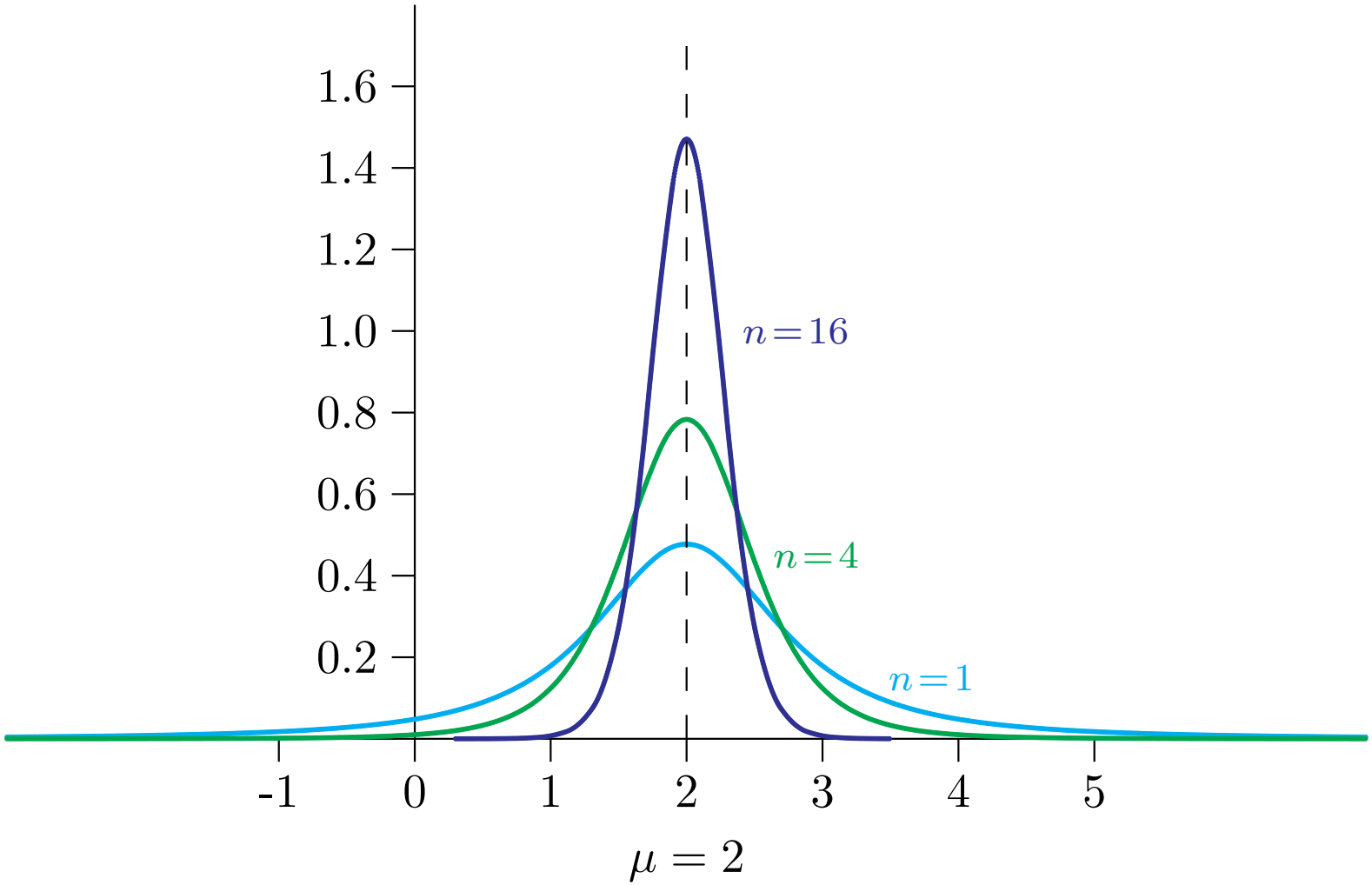
1. ε jest zmienną losową
2. $E(\varepsilon) = 0$ – pomiar nieobciążony, pomiar bez błędu systematycznego
3. błąd symetryczny względem zera (jednakowo prawdopodobne błędy dodatnie i ujemne)
4. duże (bezwzględnie) błędy mniej prawdopodobne niż małe
5. krzywa Gaussa
6. $\varepsilon \sim N(0, \sigma)$, $X \sim N(\mu, \sigma)$, σ – dokładność pomiaru, znana lub nieznana

$\Longrightarrow$ **MODEL BŁĘDU NORMALNEGO**

MODEL BŁĘDU NORMALNEGO:

estymatorem parametru μ jest średnia $\bar{X}$ obserwacji $X_1, X_2, \ldots, X_n$

MODEL BŁĘDU NORMALNEGO:



JAK TO SIĘ DZIEJE?

Funkcja charakterystyczna rozkładu normalnego $N(\mu, \sigma)$:

$$\phi_X(t) = \exp\left\{ i\mu t - \frac{1}{2}\sigma^2 t^2 \right\}$$

Funkcja charakterystyczna średniej $\overline{X} = \sum_{j=1}^{n} X_j/n$:

$$\phi_{\overline{X}}(t) = \exp\left\{ i\mu t - \frac{1}{2}\left(\frac{\sigma^2}{n}\right) t^2 \right\}$$

Ogólny model statystyczny:

wynik obserwacji $\mathbf{X} = \mu + \varepsilon$

1. ε jest zmienną losową
2. $\varepsilon \sim F$, F znane lub nie, $F \in \mathcal{F}$

Ogólny model statystyczny:

wynik obserwacji $\mathbf{X = \mu + \varepsilon}$

1. ε jest zmienną losową
2. $\varepsilon \sim F$, F znane lub nie, $F \in \mathcal{F}$

$$X = \mu + (X - \mu)$$

Ogólny model statystyczny:

wynik obserwacji $\mathbf{X = \mu + \varepsilon}$

1. ε jest zmienną losową
2. $\varepsilon \sim F$, F znane lub nie, $F \in \mathcal{F}$

$$X = \mu + (X - \mu)$$

μ - "poziom odniesienia"

Ogólny model statystyczny:

wynik obserwacji $\mathbf{X = \mu + \varepsilon}$

1. ε jest zmienną losową
2. $\varepsilon \sim F$, F znane lub nie, $F \in \mathcal{F}$

$$X = \mu + (X - \mu)$$

μ - "poziom odniesienia"

- „średnia" cena akcji danej spółki w danym okresie czasu

Ogólny model statystyczny:

wynik obserwacji $\mathbf{X = \mu + \varepsilon}$

1. ε jest zmienną losową
2. $\varepsilon \sim F$, F znane lub nie, $F \in \mathcal{F}$

$$X = \mu + (X - \mu)$$

μ - "poziom odniesienia"

- „średnia" cena akcji danej spółki w danym okresie czasu
- „średni" poziom wskazań wodomierza na rzece

Ogólny model statystyczny:

wynik obserwacji $\mathbf{X = \mu + \varepsilon}$

1. ε jest zmienną losową
2. $\varepsilon \sim F$, F znane lub nie, $F \in \mathcal{F}$

$$X = \mu + (X - \mu)$$

μ - "poziom odniesienia"

- „średnia" cena akcji danej spółki w danym okresie czasu
- „średni" poziom wskazań wodomierza na rzece
- „średnie" roszczenie z polisy

Ogólny model statystyczny:

wynik obserwacji $X = \mu + \varepsilon$

1. ε jest zmienną losową
2. $\varepsilon \sim F$, F znane lub nie, $F \in \mathcal{F}$

$$X = \mu + (X - \mu)$$

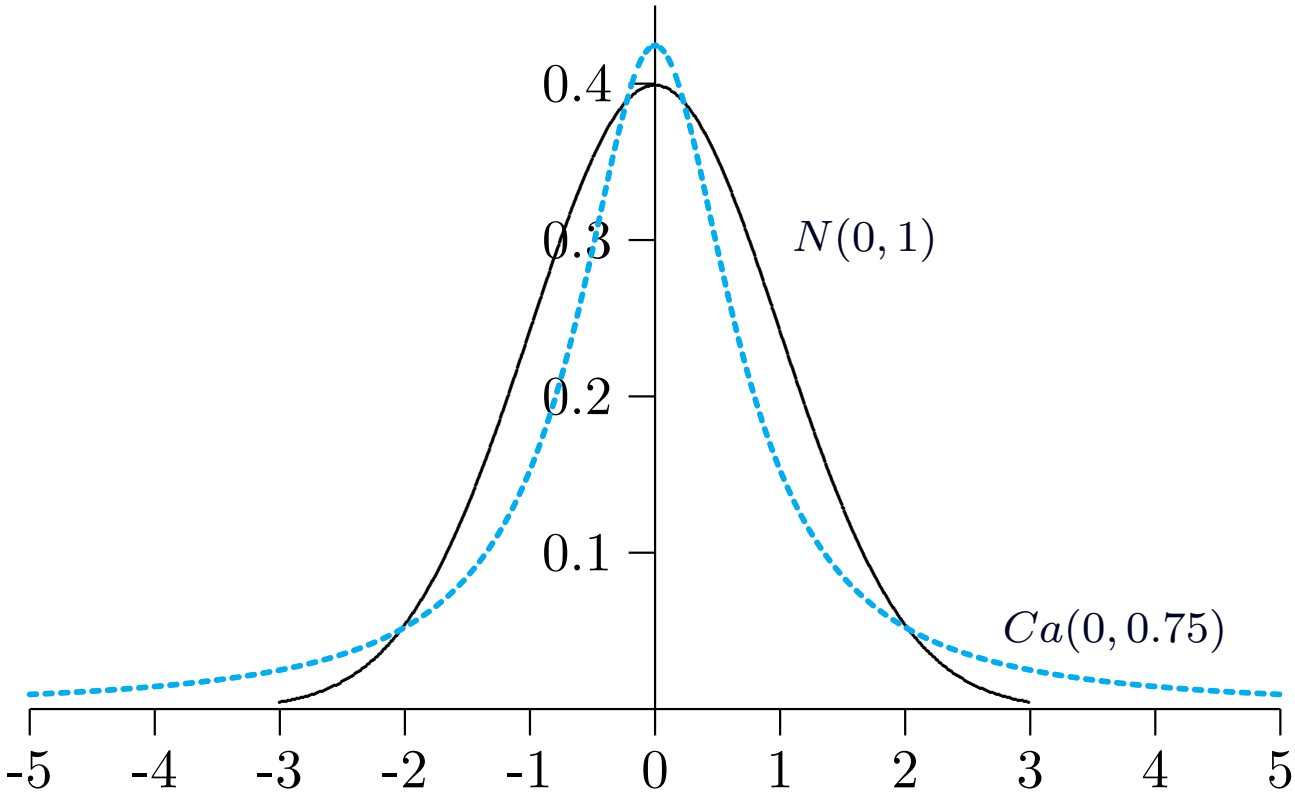
μ - "poziom odniesienia"

- „średnia" cena akcji danej spółki w danym okresie czasu
- „średni" poziom wskazań wodomierza na rzece
- „średnie" roszczenie z polisy

$\varepsilon := X - \mu$ ma rozkład normalny ???

Przypadek rozkładu symetrycznego o tłustych ogonach

Przykład: rozkład Cauchy'ego – *rozkład o trochę tłuściejszych ogonach:*



Rozkład normalny i rozkład Cauchy'ego

Funkcja charakterystyczna rozkładu Cauchy'go

$$\phi_Y(t) = \exp\{i\mu t - |\lambda t|\}$$

Funkcja charakterystyczna rozkładu Cauchy'go

$$\phi_Y(t) = \exp\{i\mu t - |\lambda t|\}$$

.

Funkcja charakterystyczna średniej $\overline{Y} = \sum_{j=1}^{n} Y_j/n$: .

Funkcja charakterystyczna rozkładu Cauchy'go

$$\phi_Y(t) = \exp\{i\mu t - |\lambda t|\}$$

.

Funkcja charakterystyczna średniej $\overline{Y} = \sum_{j=1}^{n} Y_j/n$: .

$$\phi_{\overline{Y}}(t) = \exp\{i\mu t - |\lambda t|\}$$

ROZKŁAD CAUCHY'EGO

ROZKŁAD ŚREDNIEJ ARYTMETYCZNEJ Z PRÓBY JEST TAKI SAM JAK

ROZKŁAD POJEDYNCZEJ OBSERWACJI

Ogólniej:

SYMETRYCZNE ROZKŁADY α-STABILNE

$$exp\{i\mu t - |\lambda t|^{\alpha}\}$$

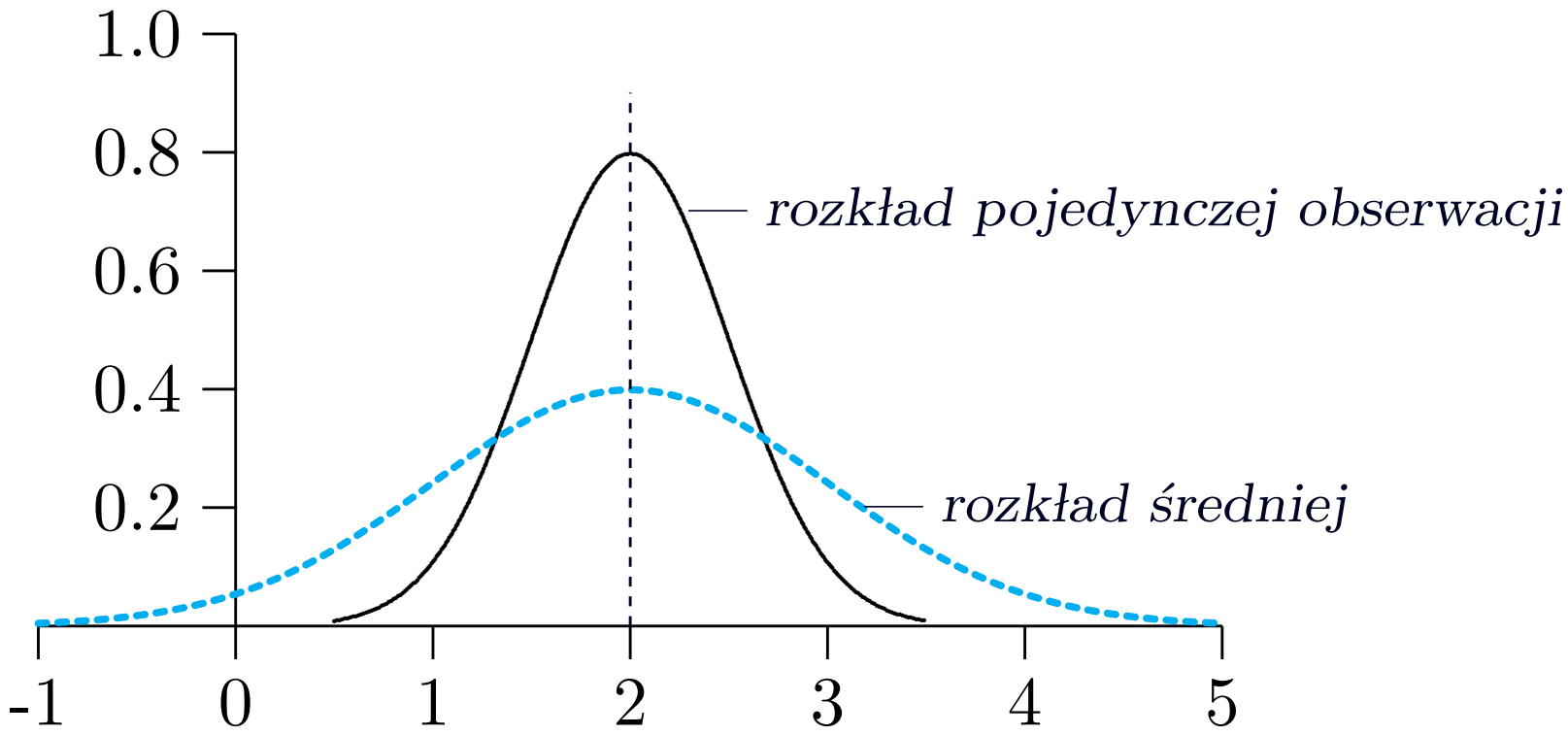
Ogólniej:

SYMETRYCZNE ROZKŁADY α-STABILNE

$$exp\{i\mu t - |\lambda t|^{\alpha}\}$$

$$\left(exp\{i\mu\frac{t}{n} - |\lambda\frac{t}{n}|^{\alpha}\}\right)^{n} = exp\{i\mu t - |n^{1/\alpha-1}\lambda t|^{\alpha}\}$$

$\alpha=2$ – *rozkład normalny;* $\alpha=1$ – *rozkład Cauchy'ego*


1.0
0.8
0.6
0.4
0.2
-1
0
1
2
3
4
5
rozkład pojedynczej obserwacji
rozkład średniej

Teraz średnia z próby traci swoje zalety, bo

1. rozkład może nie mieć wartości oczekiwanej, czyli średnia może nie mieć wartości oczekiwanej

2. rozkład średniej z próby może być „gorszy” do wnioskowania o parametrze położenia niż rozkład pojedynczej obserwacji

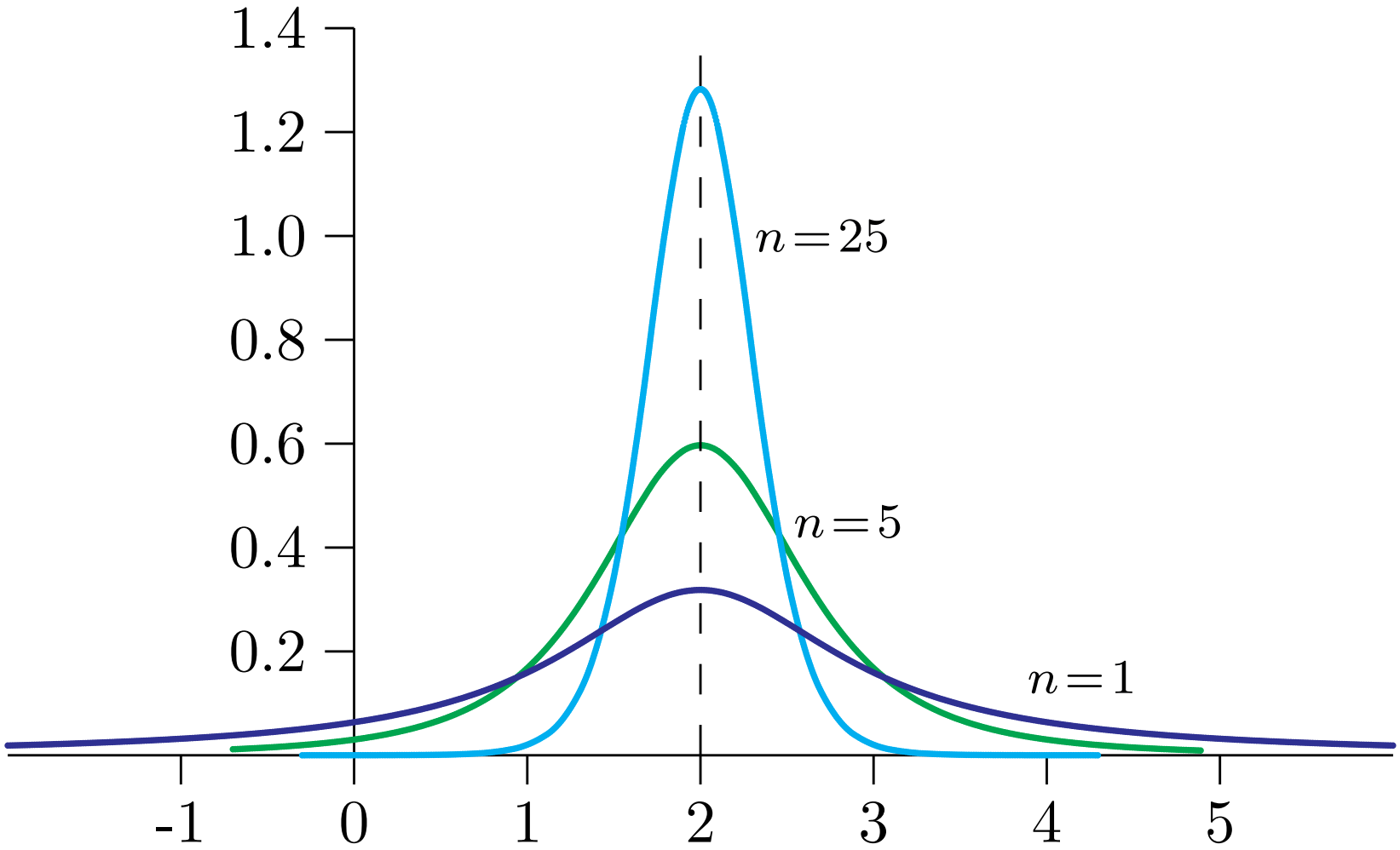
Zamiast średniej - MEDIANA

Model:

Obserwacja $X = \mu + \varepsilon$

$\varepsilon \sim F, \quad F \in \mathcal{F}, \quad F$ - rozkład znany lub nieznany

$Med_F(\varepsilon) = F^{-1}(\frac{1}{2}) = 0.$ Teraz $Med_\mu X = \mu$



Rozkład mediany M_n w modelu z błędem $\varepsilon \sim Ca(0,1)$

MEDIANA Z PRÓBY

Próba: $X_1,\ X_2,\ \ldots,\ X_n$

Statystyki pozycyjne: $X_{1:n},\ X_{2:n},\ \ldots,\ X_{n:n}$

$$X_{1:n} \leqslant\ X_{2:n} \leqslant\ \ldots\ \leqslant X_{n:n}$$

MEDIANA Z PRÓBY

Próba: $X_1,\ X_2,\ \ldots,\ X_n$

Statystyki pozycyjne: $X_{1:n},\ X_{2:n},\ \ldots,\ X_{n:n}$

$$X_{1:n} \leqslant\ X_{2:n} \leqslant\ \ldots\ \leqslant X_{n:n}$$

Mediana M_n z próby $X_1, X_2, \ldots, X_n$

$$M_n = \begin{cases} \frac{1}{2}\left(X_{\frac{n}{2}:n} + X_{\frac{n}{2}+1:n}\right), & \text{jeżeli } n \text{ jest parzyste,} \\ X_{\frac{n+1}{2}:n}, & \text{jeżeli } n \text{ jest nieparzyste} \end{cases}$$

Mediana M_n z próby jako estymator mediany populacji

Mediana M_n z próby jako estymator mediany populacji

Obciążenie ?

Mediana M_n z próby jako estymator mediany populacji

Obciążenie ?

Rozrzut ?

DEFINICJA (przypadek ciągłego rozkładu F_T estymatora T)

T jest estymatorem MEDIANOWO–NIEOBCIĄŻONYM (nieobciążonym w sensie mediany) parametru θ, jeżeli

$$P_\theta\{T \leqslant \theta\} = P_\theta\{T \geqslant \theta\} = 0.5, \quad \text{dla każdego } \theta$$

DEFINICJA (przypadek ciągłego rozkładu F_T estymatora T)

T jest estymatorem MEDIANOWO–NIEOBCIĄŻONYM (nieobciążonym w sensie mediany) parametru θ, jeżeli

$$P_\theta\{T \leqslant \theta\} = P_\theta\{T \geqslant \theta\} = 0.5, \quad \text{dla każdego } \theta$$

ROZRZUT ?

DEFINICJA (przypadek ciągłego rozkładu F_T estymatora T)

T jest estymatorem MEDIANOWO–NIEOBCIĄŻONYM (nieobciążonym w sensie mediany) parametru θ, jeżeli

$$P_\theta\{T \leqslant \theta\} = P_\theta\{T \geqslant \theta\} = 0.5, \quad \text{dla każdego } \theta$$

ROZRZUT ?

ROZSTĘP MIĘDZYKWARTYLOWY

$$F_T^{-1}\left(\frac{3}{4}\right) - F_T^{-1}\left(\frac{1}{4}\right)$$

jest miarą rozrzutu estymatora T

DEFINICJA (przypadek ciągłego rozkładu F_T estymatora T)

T jest estymatorem MEDIANOWO–NIEOBCIĄŻONYM (nieobciążonym w sensie mediany) parametru θ, jeżeli

$$P_\theta\{T \leqslant \theta\} = P_\theta\{T \geqslant \theta\} = 0.5, \quad \text{dla każdego } \theta$$

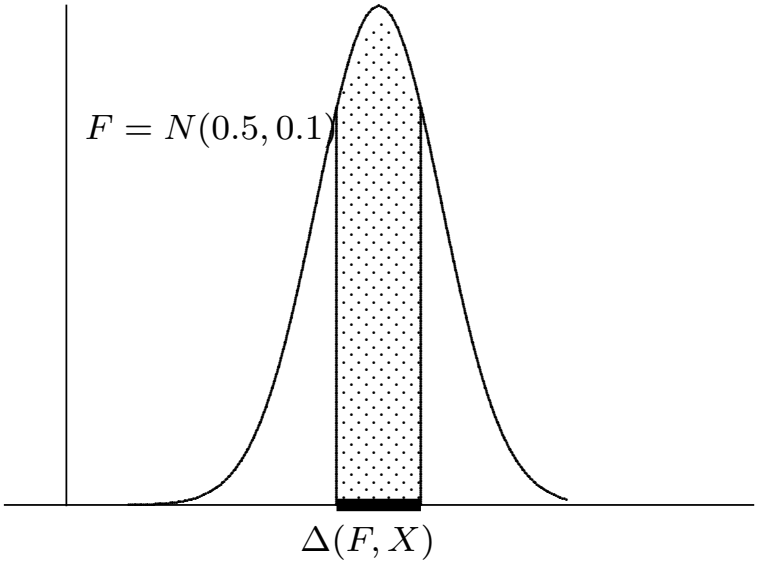
ROZRZUT ?

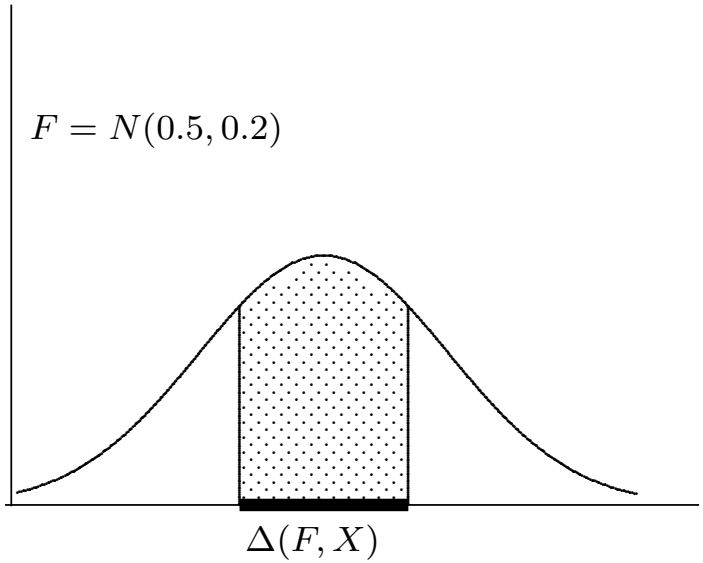
ROZSTĘP MIĘDZYKWARTYLOWY

$$F_T^{-1}\left(\frac{3}{4}\right) - F_T^{-1}\left(\frac{1}{4}\right)$$

jest miarą rozrzutu estymatora T

Ew. $Med(X - Med(X))$


$F = N(0.5, 0.1)$
$\Delta(F, X)$


$F = N(0.5, 0.2)$
$\Delta(F, X)$

ROZSTĘP MIĘDZYKWARTYLOWY Δ_n

n	$Ca(0,1)$	$N(0,1), M_n$	$N(0,1), \bar{X}$
5	0.9455	0.7199	0.6033
15	0.5472	0.4294	0.3483
25	0.4239	0.3348	0.2698
51	0.2968	0.2356	0.1889
101	0.2108	0.1678	0.1342

Narzędzie pomocnicze: rozkład beta

OZNACZENIA:

Gęstość:

$$\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}, \quad x \in (0,1), \quad p,q>0$$

Dystrybuanta w punkcie x: $B(x;p,q)$

Kwantyl rzędu q: $B^{-1}(q;p,q)$

Brak jawnych wzorów. Łatwo dostępne jako funkcje standardowe w pakietach statystycznych

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu znany

Rozkład mediany M_n. Gęstość:

$$\frac{\Gamma(n+1)}{\Gamma^2\left(\frac{n+1}{2}\right)} \Big(F_\mu(x)\Big[1 - F_\mu(x)\Big]\Big)^{(n-1)/2} f_\mu(x)$$

Dystrybuanta:

$$P_\mu\{M_n \leqslant x\} = B\left(F(x-\mu); \frac{n+1}{2}, \frac{n+1}{2}\right)$$

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu znany

OBCIĄŻENIE

Mediana M_n z próby jest medianowo-nieobciążonym estymatorem mediany populacji:

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu znany

ROZSTĘP MIĘDZYKWARTYLOWY Δ_n:

$$\Delta_n = F^{-1}\left(B^{-1}(\frac{3}{4}; \frac{n+1}{2}, \frac{n+1}{2})\right) - F^{-1}\left(B^{-1}(\frac{1}{4}; \frac{n+1}{2}, \frac{n+1}{2})\right)$$

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu znany

Przedział (jednostronny) ufności na poziomie ufności γ:

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

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu znany

Przedział (dwustronny) ufności na poziomie ufności γ:

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

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

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu znany

TESTOWANIE HIPOTEZY $H : \mu = \mu_0,\ K : \mu > \mu_0$

Wartość krytyczna testu:

$$x_{1-\alpha}(M_n) = \mu_0 + F^{-1}\left(B^{-1}(1-\alpha; \frac{n+1}{2}, \frac{n+1}{2})\right)$$

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu znany

Moc tego testu:

$$\beta(\mu) =$$

$$= 1 - B\left(F\left[\mu_0 - \mu + F^{-1}\left(B^{-1}(1-\alpha; \frac{n+1}{2}, \frac{n+1}{2})\right)\right]; \frac{n+1}{2}, \frac{n+1}{2}\right)$$

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

ROZKŁAD F NIEZNANY

Obserwacje

$$X_{1:n}, \ldots, X_{i:n}, \ldots, X_{j:n}, \ldots, X_{n:n}$$

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu nieznany

Nieobciążonym estymatorem parametru μ jest mediana z próby

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste

Rozkład F błędu nieznany

JEDNOSTRONNY PRZEDZIAŁ UFNOŚCI DLA MEDIANY

Przedział ufności postaci $(X_{i:n}, +\infty)$

JEDNOSTRONNY PRZEDZIAŁ UFNOŚCI DLA MEDIANY

Przedział ufności postaci $(X_{i:n}, +\infty)$

Jeżeli $i(n, \gamma)$ jest najmniejszą liczbą taką, że

$$P_F\{X_{i:n} \leqslant F^{-1}(q)\} = \sum_{s=i(n,\gamma)}^{n} \binom{n}{s} q^s (1-q)^{n-s} \geqslant \gamma$$

to $\left(X_{i(n,\gamma):n}, +\infty\right)$ jest przedziałem ufności dla kwantyla rzędu q, na poziomie ufności (co najmniej) γ

Mediana $M_n = X_{\frac{n+1}{2}:n}$ z próby $X_1, X_2, \ldots, X_n$, n nieparzyste
Rozkład F błędu nieznany

DWUSTRONNY PRZEDZIAŁ UFNOŚCI DLA MEDIANY

(dla q-tego kwantyla)

Przedział ufności postaci

$$(X_{i:n}, X_{j:n})$$

Takie przedziały ufności nie zawsze istnieją!

Mediana z próby $X_1, X_2, \ldots, X_n$, n PARZYSTE

$$M_n = \begin{cases} \frac{1}{2}\left(X_{\frac{n}{2}:n} + X_{\frac{n}{2}+1:n}\right), & \text{jeżeli } n \text{ jest parzyste,} \\ X_{\frac{n+1}{2}:n}, & \text{jeżeli } n \text{ jest nieparzyste} \end{cases}$$

n - parzyste

DWA WYNIKI MOGĄCE BUDZIĆ NIEPOKÓJ

Pierwszy wynik.

Efektywność mediany w stosunku do średniej arytmetycznej (średnia arytmetyczna w modelu gaussowskim jest estymatorem nieobciążonym o jednostajnie minimalnej wariancji)

$$e(n) = \frac{Var(\bar{X}_n)}{Var(M_n)}$$

n	$N(0,1)$	$U(0,1)$
1	1.000	1.000
2	1.000	1.000
3	0.743	0.556
4	0.838	0.625
5	0.697	0.467
6	0.776	0.519
7	0.679	0.429
8	0.743	0.469
9	0.669	0.407
10	0.723	0.440

n	$N(0,1)$	$U(0,1)$
1	1.000	1.000
2	1.000	1.000
3	0.743	0.556
4	0.838	0.625
5	0.697	0.467
6	0.776	0.519
7	0.679	0.429
8	0.743	0.469
9	0.669	0.407
10	0.723	0.440

Czyż nie wygląda na paradoks fakt, że zwiększenie liczności próby z $2n$ do $2n+1$ pogarsza efektywność estymatora?

Drugi wynik.

$\mathcal{F}$ - rodzina wszystkich rozkładów o ciągłych i ściśle rosnących dystrybuatach

$Med(F, T)$ - mediana rozkładu statystyki T, gdy próba pochodzi z rozkładu o dystrybuancie F

m_F - mediana rozkładu $F \in \mathcal{F}$.

Okazuje się, że

Drugi wynik.

$\mathcal{F}$ - rodzina wszystkich rozkładów o ciągłych i ściśle rosnących dystrybuatach

$Med(F, T)$ - mediana rozkładu statystyki T, gdy próba pochodzi z rozkładu o dystrybuancie F

m_F - mediana rozkładu $F \in \mathcal{F}$.

Okazuje się, że dla każdej liczby $C > 0$ znajdzie się taki rozkład $F \in \mathcal{F}$, że

$$Med(F, M_{2n}) - m_F > C$$

Drugi wynik.

$\mathcal{F}$ - rodzina wszystkich rozkładów o ciągłych i ściśle rosnących dystrybuatach

$Med(F, T)$ - mediana rozkładu statystyki T, gdy próba pochodzi z rozkładu o dystrybuancie F

m_F - mediana rozkładu $F \in \mathcal{F}$.

Okazuje się, że dla każdej liczby $C > 0$ znajdzie się taki rozkład $F \in \mathcal{F}$, że

$$Med(F, M_{2n}) - m_F > C$$

Praktyczny wniosek jest następujący: unikaj prób o parzystej liczbie elementów, a jeżeli trafi Ci się taka próba, wyrzuć jedną z obserwacji !

Drugi wynik.

$\mathcal{F}$ - rodzina wszystkich rozkładów o ciągłych i ściśle rosnących dystrybuatach

$Med(F, T)$ - mediana rozkładu statystyki T, gdy próba pochodzi z rozkładu o dystrybuancie F

m_F - mediana rozkładu $F \in \mathcal{F}$.

Okazuje się, że dla każdej liczby $C > 0$ znajdzie się taki rozkład $F \in \mathcal{F}$, że

$$Med(F, M_{2n}) - m_F > C$$

Praktyczny wniosek jest następujący: unikaj prób o parzystej liczbie elementów, a jeżeli trafi Ci się taka próba, wyrzuć jedną z obserwacji !

Lekarstwo - RANDOMIZACJA!