STATYSTYKA MATEMATYCZNA

WYKŁAD 2

ESTYMACJA PUNKTOWA

Niech θ - nieznany parametr rozkładu cechy X.
Wartość parametru θ będziemy estymować (przybliżać) na podstawie n elementowej próby.

- wybieramy statystykę U_n o rozkładzie zależnym od θ
- obliczamy na podstawie próby jej wartość u_n
- przyjmujemy, że $\theta \approx u_n$

Statystykę U_n nazywamy **estymatorem** parametru θ.

Klasyfikacja estymatorów.

Estymator U_n jest:

- **zgodny** jeśli $U_n \xrightarrow{n\to\infty} \theta$ *wg prawdopodobieństwa*
- **nieobciążony** jeśli $E(U_n) = \theta$
- **asymptotycznie nieobciążony** jeśli $\lim_{n\to\infty} E(U_n) = \theta$
- **najefektywniejszy** gdy jest nieobciążony i ma najmniejszą wariancję w klasie nieobciążonych estymatorów tego parametru,
- **asymptotycznie najefektywniejszy** gdy jest nieobciążony lub asymptotycznie nieobciążony i jego wariancja dąży do wariancji estymatora najefektywniejszego.

Przykład

Niech $X - N(m;\ \sigma)$. Przyjmijmy, że mamy próbę (X_1, X_2, X_3, X_4). Zakładamy, że $\sigma = 1$ jest znane , szukamy estymatora parametru m .

Rozpatrzmy kilka prostych estymatorów.

$$U_1 = X_1 + X_4$$

$$U_2 = \frac{1}{2}(X_1 + X_3)$$

$$U_3 = \overline{X}_4$$

$$U_4 = \frac{1}{3}(X_1 + X_3)$$

$$U_5 = X_2 + U_2 - X_4$$

$$U_6 = \frac{1}{10}\sum_{i=1}^{4} iX_i$$

Sprawdzimy własności tych estymatorów.

Policzmy wartości oczekiwane tych estymatorów (zbadamy czy są nieobciążone).

$$E(U_1) = 2m$$
$$E(U_2) = m$$
$$E(U_3) = m$$
$$E(U_4) = \frac{2}{3}m$$
$$E(U_5) = m$$
$$E(U_6) = m$$

Zatem estymatory U_1 i U_4 są obciążone, należy je odrzucić.

Policzmy wariancje pozostałych estymatorów.

$D^2(U_2) = 0{,}5$

$D^2(U_3) = 0{,}25$

$D^2(U_5) = 2{,}5$

$D^2(U_6) = 0{,}3$

Zatem estymator U_3 ma najmniejszą wariancję.

Estymatory parametrów rozkładu N(m, σ).

Parametr	Estymator	Własności estymatora
m	$\overline{X}_n$	Zgodny. Nieobciążony. Najefektywniejszy.
σ^2	S_n^2	Zgodny. Asymptot. nieobciążony. Asymptot. najefektywniejszy.
	$\hat{S}_n^2$	Zgodny. Nieobciążony. Asymptot. najefektywniejszy.
	$S_n^{0\,2}$	Zgodny. Nieobciążony. Najefektywniejszy.
σ	$S_n \quad \hat{S}_n$ S_n^0	Zgodne. Asymptot. nieobciążone. Asymptot. najefektywniejsze.

Estymatory innych parametrów.

Parametr	Estymator	Własności estymatora
Wartość oczekiwana (rozkład dowolny)	$\overline{X}_n$	Zgodny. Nieobciążony.
λ (rozkład Poissona)	$\overline{X}_n$	Zgodny. Nieobciążony. Najefektywniejszy.
p (rozkład zero-jedynkowy)	$W = \frac{liczba\ sukcesów}{n}$ = średnia częstość sukcesu	Zgodny. Nieobciążony. Najefektywniejszy.
Wariancja (rozkład dowolny)	S_n^2	Zgodny. Asymptot. nieobciążony.
	$\hat{S}_n^2$	Zgodny. Nieobciążony.

Uwaga

a) w praktyce zgodność estymatora sprawdza się na podstawie praw wielkich liczb lub korzysta się z faktu, że estymator nieobciążony (asymptotycznie nieobciążony), którego wariancja dąży do zera (tzn. $\lim_{n\to\infty} D^2 U_n = 0$) jest estymatorem zgodnym.

b) w praktyce efektywność estymatora bada się na podstawie nierówności Rao-Cramera:

Dla (praktycznie każdego) estymatora nieobciążonego U_n prawdziwa jest nierówność

$$D^2 U_n \geq \frac{1}{n \sum_i \left(\frac{d}{d\theta} \ln p_i(\theta) \right)^2 p_i(\theta)}$$

dla zmiennej losowej skokowej

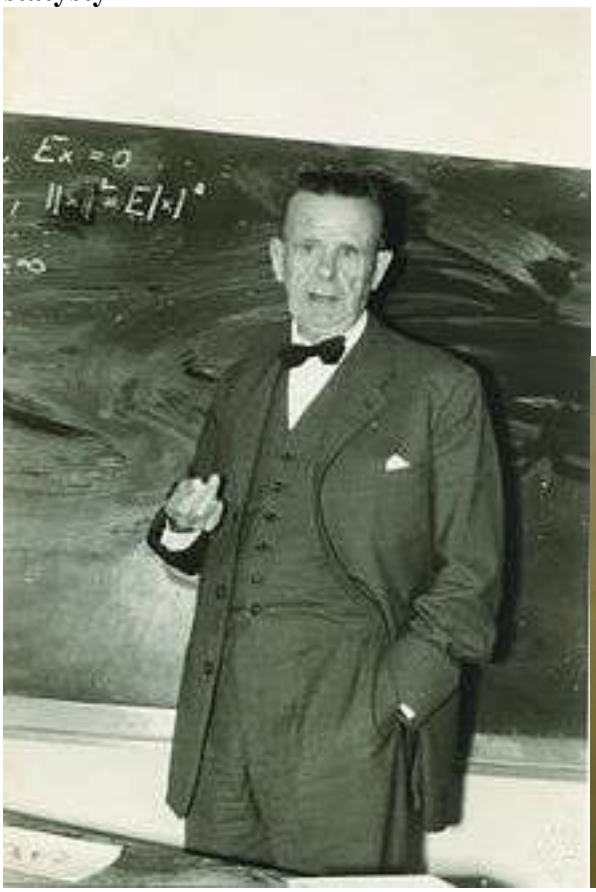
$$D^2 U_n \geq \frac{1}{n \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \theta} ln\, f(x,\theta) \right)^2 f(x,\theta) dx}$$

dla zmiennej losowej ciągłej

Przy czym **dla estymatora najefektywniejszego zachodzi równość** (jeśli istnieje estymator najefektywniejszy to prawe strony powyższych nierówności są równe jego wariancji).

C. R. Rao **(1920 -),**
statystyk

Harald Cramér **(1893-1985),**
matematyk, statystyk,



C. R. Rao and Harald Cramér, 1978

Przykład

Niech $X - N(m; \sigma)$. Przyjmijmy, że estymatorem parametru m jest $\overline{X}_n$.
Sprawdzimy własności tego estymatora.

Rozwiązanie:

$$E\left(\overline{X}_n\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} m = \frac{1}{n} nm = m$$

zatem jest to estymator nieobciążony.

$$D^2\left(\overline{X}_n\right) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

$$\lim_{n\to\infty} D^2\left(\overline{X}_n\right) = \lim_{n\to\infty}\frac{\sigma^2}{n} = 0$$

zatem jest to estymator zgodny.

$$f(x,m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Wyznaczmy prawą stronę nierówności Rao-Cramera:

$$\frac{1}{n\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial m}\ln f(x,m)\right)^2 f(x,m)dx} =$$

$$= \frac{1}{\frac{n}{\sigma^4}\int_{-\infty}^{\infty}(x-m)^2 f(x,m)dx} = \frac{1}{\frac{n}{\sigma^4}\sigma^2} = \frac{\sigma^2}{n}$$

zatem jest to estymator najefektywniejszy.

Przykład

Niech $X - N(m;\ \sigma)$.

Obliczymy $E\left(S_n^2\right)$, $E\left(\hat{S}_n^2\right)$, $E\left(S_n^{02}\right)$.

Rozwiązanie:

$$E\left(S_n^2\right) = E\left(\frac{\sigma^2}{n}\frac{nS_n^2}{\sigma^2}\right) = \frac{\sigma^2}{n}E\left(\frac{nS_n^2}{\sigma^2}\right) =$$

$$= \frac{\sigma^2}{n}E\left(Y_{n-1}\right) = \frac{\sigma^2}{n}(n-1) \neq \sigma^2$$ (estymator obciążony)

bo statystyka $\frac{nS_n^2}{\sigma^2}$ ma rozkład chi kwadrat z $n-1$ stopniami swobody, oraz wartość oczekiwana zmiennej losowej o rozkładzie chi kwadrat jest równa liczbie stopni swobody.

$$E\left(\hat{S}_n^2\right) = E\left(\frac{n}{n-1} S_n^2\right) = \frac{n}{n-1} E\left(S_n^2\right) = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2$$

(estymator nieobciążony)

$$E\left(S_n^{02}\right) = E\left(\frac{\sigma^2}{n}\frac{nS_n^{02}}{\sigma^2}\right) = \frac{\sigma^2}{n}E\left(\frac{nS_n^{02}}{\sigma^2}\right) =$$

$$= \frac{\sigma^2}{n}E\left(Y_n\right) = \frac{\sigma^2}{n}n = \sigma^2$$

(estymator nieobciążony)

Wniosek

S_n^2 jest estymatorem asymptotycznie nieobciążonym parametru σ^2 bowiem:

$$\lim_{n\to\infty} E\left(S_n^2\right) = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$$

$\hat{S}_n^2$ jest estymatorem nieobciążonym parametru σ^2.

S_n^{02} jest estymatorem nieobciążonym parametru σ^2.

Przykład

Niech $X - N(m;\ \sigma)$.

Obliczymy $D^2\left(S_n^2\right)$, $D^2\left(\hat{S}_n^2\right)$, $D^2\left(S_n^{02}\right)$.

Rozwiązanie:

$$D^2 S_n^2 = D^2\left(\frac{\sigma^2}{n}\frac{nS_n^2}{\sigma^2}\right) = \frac{\sigma^4}{n^2} D^2\left(\frac{nS_n^2}{\sigma^2}\right) = \frac{\sigma^4}{n^2} 2(n-1)$$

bo statystyka $\frac{nS_n^2}{\sigma^2}$ ma rozkład chi kwadrat z $n - 1$ stopniami swobody, oraz wariancja zmiennej losowej o rozkładzie chi kwadrat jest równa podwojonej liczbie stopni swobody.

$$D^2\hat{S}_n^2 = D^2\left(\frac{n}{n-1}S_n^2\right) = \frac{n^2}{(n-1)^2}D^2\left(S_n^2\right) =$$

$$= \frac{n^2}{(n-1)^2}\frac{2(n-1)}{n^2}\sigma^4 = \frac{2\sigma^4}{n-1}$$

$$D^2 S_n^{02} = D^2\left(\frac{\sigma^2}{n}\frac{nS_n^{02}}{\sigma^2}\right) = \frac{\sigma^4}{n^2} D^2\left(\frac{nS_n^{02}}{\sigma^2}\right) = \frac{\sigma^4}{n^2} 2n = \frac{2\sigma^4}{n}$$

Wniosek

Wariancje estymatorów S_n^2, $\hat{S}_n^2$, S_n^{02} dążą do zera gdy n dąży do nieskończoności. Zatem

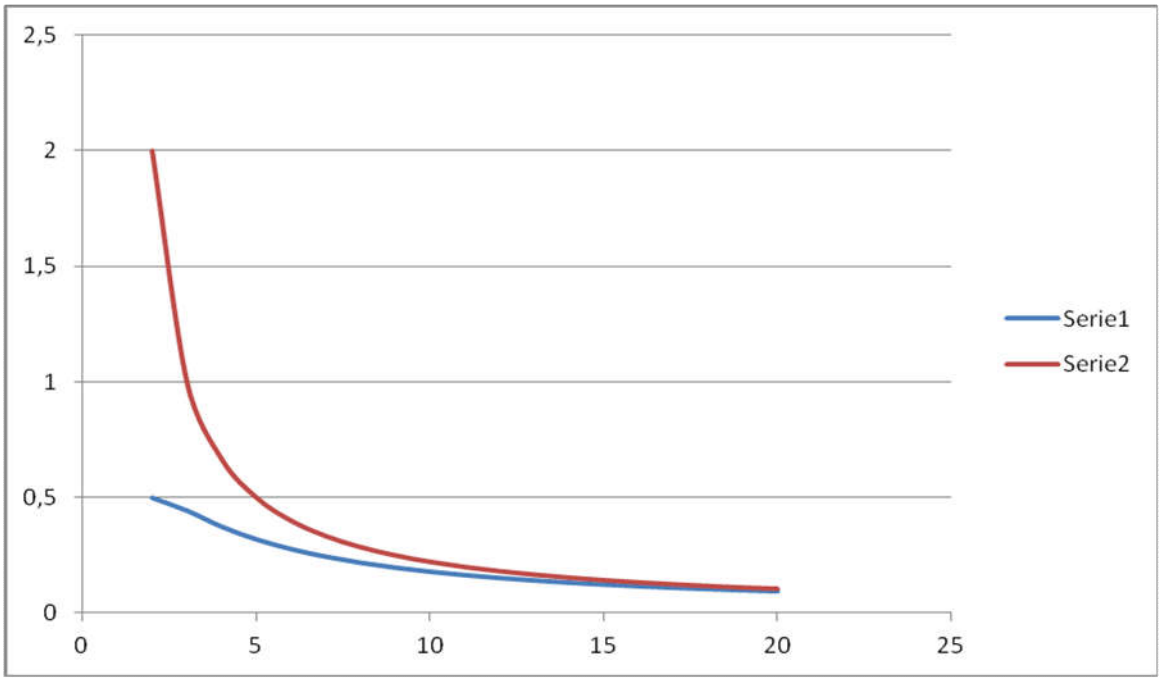
S_n^2 jest estymatorem zgodnym parametru σ^2

$\hat{S}_n^2$ jest estymatorem zgodnym parametru σ^2.

S_n^{02} jest estymatorem zgodnym parametru σ^2.

X-N(m,1) $D^2 S_n^2 = \frac{2(n-1)}{n^2}$ $D^2 \hat{S}_n^2 = \frac{2}{n-1}$

n		
2	0,5	2
3	0,444444444	1
4	0,375	0,666666667
5	0,32	0,5
6	0,277777778	0,4
7	0,244897959	0,333333333
8	0,21875	0,285714286
9	0,197530864	0,25
10	0,18	0,222222222
11	0,165289256	0,2
12	0,152777778	0,181818182
13	0,142011834	0,166666667
14	0,132653061	0,153846154
15	0,124444444	0,142857143
16	0,1171875	0,133333333
17	0,110726644	0,125
18	0,104938272	0,117647059
19	0,099722992	0,111111111
20	0,095	0,105263158



Wyznaczanie estymatorów metodą momentów (K.Pearson)

Nieznane momenty teoretyczne cechy X szacujemy przez momenty empiryczne tego samego rzędu.

Estymatory uzyskane tą metodą są zwykle mało efektywne (zwłaszcza dla rozkładów asymetrycznych).

Momenty teoretyczne:

$m_k = E(X^k)$ – moment rzędu k zmiennej losowej X ($m_1 = EX$).

$m_{kl} = E(X^k Y^l)$ – moment rzędu k, l zmiennej losowej (X, Y).

Momenty empiryczne:

$M_k = \frac{1}{n}\sum x_i^k$ – moment rzędu k cechy X ($M_1 = \overline{X}_n$).

$M_{kl} = \frac{1}{n}\sum x_i^k \cdot y_i^l$ – moment rzędu k, l jednocześnie badanych cech (X, Y).

Zatem przyjmujemy, że:

$m_k \cong M_k$ oraz $m_{kl} \cong M_{kl}$

Parametry będące funkcjami momentów teoretycznych szacuje się przez wartości tych funkcji obliczone dla momentów empirycznych.

Przykład

Dla rozkładu wykładniczego z parametrem a mamy wartość oczekiwaną równą

$EX = m_1 = 1/a$.

Ponieważ przyjmujemy $m_1 \cong M_1$ to $1/a \cong \overline{X}_n$,

zatem estymatorem parametru a jest $\frac{1}{\overline{X}_n}$.

Przykład

Dla rozkładu logarytmiczno-normalnego $LN(m;\ \sigma)$

$$f(x)=\begin{cases}\dfrac{1}{x\sigma\sqrt{2\pi}}e^{\frac{-(\ln x-m)^2}{2\sigma^2}} & dla \quad x>0\\ 0 & dla \quad x\le 0\end{cases}$$

mamy wartość oczekiwaną równą

$$EX=m_1=e^{m+\frac{\sigma^2}{2}}$$

i wariancję $D^2X=e^{2m+\sigma^2}\left(e^{\sigma^2}-1\right)$.

Uwaga.

Jeśli X ma rozkład $LN(m;\ \sigma)$ to zmienna losowa Y = lnX ma rozkład $N(\ln m;\ \sigma)$.

Ponieważ przyjmujemy

$m_1 \cong M_1 = \overline{X}$ i $D^2X \cong S^2$

to rozwiązując układ równań

$$e^{m+\frac{\sigma^2}{2}} = \overline{X}$$

$$e^{2m+\sigma^2}\left(e^{\sigma^2}-1\right) = S^2$$

otrzymamy

$$\sigma^2 = \ln\left(1+\left(\frac{S}{\overline{X}}\right)^2\right) \quad \text{i} \quad m = \ln\left(\overline{X}^2\right) - \frac{1}{2}\sigma^2$$

zatem są to szukane estymatory.

Przykład

Dla zmiennej losowej dwuwymiarowej współczynnik korelacji możemy wyrazić za pomocą momentów

$$\rho = \frac{Cov(X,Y)}{DX \cdot DY} = \frac{m_{11} - m_{10} \cdot m_{01}}{\sqrt{m_{20} - m_{10}^2} \cdot \sqrt{m_{02} - m_{01}^2}}$$

zatem jego estymatorem może być:

$$\rho \approx r = \frac{M_{11} - M_{10} \cdot M_{01}}{\sqrt{M_{20} - M_{10}^2} \cdot \sqrt{M_{02} - M_{01}^2}} =$$

$$= \frac{\frac{1}{n}\sum x_i \cdot y_i - \frac{1}{n}\sum x_i \cdot \frac{1}{n}\sum y_i}{\sqrt{\frac{1}{n}\sum x_i^2 - \left(\frac{1}{n}\sum x_i\right)^2} \cdot \sqrt{\frac{1}{n}\sum y_i^2 - \left(\frac{1}{n}\sum y_i\right)^2}} = \frac{\frac{1}{n}\sum x_i \cdot y_i - \bar{X}\bar{Y}}{S_X \cdot S_Y}$$

Estymatory uzyskane metodą momentów nie zawsze są wyznaczone jednoznacznie.

Przykład

Wyznaczymy metodą momentów estymator parametru λ rozkładu Poissona.

Mamy próbę $(X_1, X_2, X_3, ..., X_n)$.

Skoro $EX = \lambda$, to $\lambda \approx \overline{X}_n$

lecz $D^2X = \lambda$, stąd $\lambda \approx S_n^2$

i mamy dwa różne estymatory tego samego parametru.

Wyznaczanie estymatorów metodą największej wiarygodności (MNW) (R.A.Fisher)

Dla uproszczenia rozpatrujemy przypadek gdy nieznany jest tylko jeden parametr rozkładu.

a) wyznaczamy funkcję wiarygodności

$$L(\theta; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(\theta; x_i)$$

dla zmiennej losowej skokowej

$$L(\theta; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(\theta; x_i)$$

dla zmiennej losowej ciągłej

b) wyznaczamy logarytm funkcji wiarygodności, $l(\theta) = l(\theta; x_1, x_2, ..., x_n) = \ln L(\theta; x_1, x_2, ..., x_n)$

c) wyznaczamy θ dla którego funkcja $l(\theta)$ ma maksimum (w tym celu obliczamy pochodną funkcji $l(\theta)$, wyznaczamy miejsce zerowe pochodnej i sprawdzamy czy w tym punkcie pierwsza pochodna odpowiednio zmienia znak lub druga pochodna jest ujemna),

d) przyjmujemy, że wyznaczony w ten sposób wzór na θ jest poszukiwanym estymatorem.

Uwaga

1) Postać funkcji wiarygodności wynika z wielowymiarowego rozkładu próby (gęstość/funkcja prawdopodobieństwa jest iloczynem gęstości/f.p brzegowych).

2) Logarytmowanie funkcji wiarygodności wynika z potrzeb praktycznych.

3) Jeśli rozpatrujemy przypadek gdy nieznanych jest wiele parametrów rozkładu to postępujemy podobnie stosując rachunek różniczkowy funkcji wielu zmiennych.

Uwaga

Estymatory uzyskane tą metodą są zwykle co najmniej zgodne, asymptotycznie nieobciążone i asymptotycznie najefektywniejsze.
Warto też wiedzieć, że estymatory uzyskane tą metodą mają asymptotyczny rozkład normalny

Uwaga

Niech g będzie funkcją rzeczywistą różnowartościową.
Jeśli u_n jest estymatorem NW parametru θ to estymatorem NW parametru $g(\theta)$ jest $g(u_n)$.
Własność ta jest prawdziwa również dla przypadku wielu parametrów.

Przykład

Wyznaczymy MNW estymator parametru θ rozkładu jednostajnego w [0; θ], $\theta > 0$.

Mamy próbę (X_1, X_2, X_3, ..., X_n).

Wtedy

$$L(\theta) = \frac{1}{\theta^n} \quad dla\ 0 \le x_i \le \theta$$

$$l(\theta) = -\ln\theta$$

$$l'(\theta) = -n/\theta < 0$$

Zauważmy, że $\theta \geq \max_{i=1,2,..n} \{x_i\}$

zatem $L(\theta)$ ma największą wartość dla $\theta = \max_{i=1,2,..n} \{x_i\}$ i jest to szukany estymator NW.

Estymatory uzyskane MNW nie zawsze są wyznaczone jednoznacznie.

Przykład

Wyznaczymy MNW estymator parametru θ rozkładu jednostajnego w $[\theta\,;\ \theta\ + 2]$.

Mamy próbę $(X_1, X_2, X_3, ..., X_n)$.

Wtedy

$$L(\theta) = \frac{1}{2^n} \quad dla\ \theta \le x_i \le \theta + 2$$

jest funkcją stałą względem parametru.

zatem każda wartość

$$\theta \in \left[\max_{i=1,2,..n} \{x_i\} - 2;\ \min_{i=1,2,..n} \{x_i\} \right]$$

może być szukanym estymatorem NW.

Przykład

Wyznaczymy MNW estymator parametru λ rozkładu Poissona.

Mamy próbę $(X_1, X_2, X_3, ..., X_n)$.

Wtedy

$$L(\lambda) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \ldots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = \frac{\lambda^{x_1+..+x_n}}{x_1!\ldots x_n!} e^{-n\lambda}$$

$$l(\lambda) = \ln L(\lambda) = \left(x_1 + .. + x_n\right)\ln\lambda - n\lambda - \ln\left(x_1!\ldots x_n!\right)$$

$$l'(\lambda) = \left(x_1 + .. + x_n\right)/\lambda - n$$

Wyznaczamy punkt krytyczny

$$l'(\lambda) = 0 \quad \Leftrightarrow \quad \left(x_1 + .. + x_n\right)/\lambda - n = 0 \quad \Leftrightarrow$$

$$\Leftrightarrow \lambda = \left(x_1 + .. + x_n\right)/n = \overline{x}_n$$

sprawdzamy istnienie maksimum

$$l''(\lambda) = -\left(x_1 + .. + x_n\right)/\lambda^2 < 0$$

Zatem estymatorem parametru λ jest średnia z próby.

Przykład

Dla rozkładu logarytmiczno-normalnego $LN(m;\ \sigma)$ wyznaczymy estymatory parametrów $m;\ \sigma$.

$$f(x)=\begin{cases}\dfrac{1}{x\sigma\sqrt{2\pi}}e^{\frac{-(\ln x-m)^2}{2\sigma^2}} & dla \quad x>0\\ 0 & dla \quad x\le 0\end{cases}$$

$$L(m,\sigma)=\frac{1}{x_1\sigma\sqrt{2\pi}}e^{\frac{-(\ln x_1-m)^2}{2\sigma^2}}\ldots\frac{1}{x_n\sigma\sqrt{2\pi}}e^{\frac{-(\ln x_n-m)^2}{2\sigma^2}}=$$

$$=\prod_{i=1}^{n}\frac{1}{x_i\sigma\sqrt{2\pi}}e^{\sum_{i=1}^{n}\frac{-(\ln x_i-m)^2}{2\sigma^2}}$$

$$l(m,\sigma)=\ln L(m,\sigma)=$$

$$=-\sum_{i=1}^{n}\ln x_i-n\left(\ln\sigma+\frac{1}{2}\ln(2\pi)\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln x_i-m)^2$$

różniczkując względem m i σ otrzymamy

$\sigma^2=S^2_{\ln X}$ i $m=\overline{\ln X}$

zatem otrzymane estymatory są inne niż w przypadku metody momentów.

Przykład zastosowania estymacji

Chcemy w dyskretny sposób (obawa karalności) ocenić odsetek *k* osób dających łapówki.

Można to zrobić następująco.

Pytana osoba rzuca monetą i wynik rzutu zachowuje do swojej wiadomości.

Przygotowujemy dużą liczbę kart na połowie których jest pytanie: "czy wypadł orzeł?" a na drugiej połowie kart jest pytanie "czy dajesz łapówki?". Karty losujemy. Pytany losuje kartę i odpowiada TAK (T) lub NIE na wylosowane pytanie.

Rozpatrywane doświadczenie ma rozkład zerojedynkowy z nieznanym parametrem *p*.

Niech K1 wylosowanie karty z pytaniem nr 1.

Niech K2 wylosowanie karty z pytaniem nr 2.

Wtedy

$p = P(T) = P(K1)\,P(T|K1) + P(K2)\,P(T|K2) =$

$= 0{,}5 \bullet 0{,}5 + 0{,}5k$

Estymatorem dla p jest średnia w.

Stąd estymatorem k jest $k \cong 2w - 0{,}5$.

L.Kowalski 11.10.2015