Rachunek prawdopodobieństwa i statystyka

Wykład 10

Estymacja - podstawowe pojęcia

Celem pomiarów jest znalezienie rozkładu p-twa leżącego u podstaw badanej cechy (bądź przynajmniej jego najważniejszych parametrów).

Populacja – zbiór wszystkich przedstawicieli posiadających badaną cechę.

Próbka losowa – reprezentatywna próbka całej populacji, tzn. taka, która odzwierciedla wszystkie cechy i związki w niej występujące.

Próbka obciążona – próbka w której brakuje pewnych klas przypadków,
(np. ze względu na skończone rozmiary detektora, skończony czas pomiaru, itd.)

Mówimy, że próbka jest prosta, jeśli wszystkie występujące w niej zmienne losowe są niezależne (schemat losowania niezależnego).
W przeciwnym wypadku próbkę nazywamy złożoną.

Statystyka – dowolna funkcja $f(x_1, x_2, \ldots, x_n)$ próby losowej $x_1, x_2, \ldots, x_n$.

Niech rozkład badanej cechy X populacji zależy od nieznanego parametru θ, który będziemy szacowali na podstawie n elementowej próby prostej $x_1, x_2, \ldots, x_n$ pobranej z tej populacji.

Estymator i estymata

Estymatorem parametru θ nazywamy każdą statystykę $\hat{\theta}_n\left(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\right)$, której wartości przyjmujemy jako oceny tego parametru.

Estymata (oszacowanie, ocena) parametru θ to wartość estymatora $\hat{\theta}_n$ jaką przyjmuje on dla konkretnej próbki losowej.

Przykład: Estymatory parametru θ rozkładu płaskiego: $f\left(x;\theta\right)=\frac{1}{\theta}$ dla $0 \le x \le \theta$

Załóżmy, że dysponujemy n elementową próbką prostą z tego rozkładu $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$

- $\hat{\mathbf{T}}_1 = \mathbf{x}_{n:n}$ - ze wzrostem liczebności próbki coraz bliższy wartości θ, ale zawsze ją zaniża
- $\hat{\mathbf{T}}_2 = \frac{n+1}{n}\mathbf{x}_{n:n}$ - $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ dzielą $\left[0,\theta\right]$ na n+1 przedziałów i są równomiernie rozmieszczone
- $\hat{\mathbf{T}}_3 = \mathbf{x}_{1:n} + \mathbf{x}_{n:n}$ - korzystamy z faktu, że $\mathbf{x}_{1:n}$ i $\mathbf{x}_{n:n}$ są średnio równoodległe od brzegów
- $\hat{\mathbf{T}}_4 = (n+1)\,\mathbf{x}_{1:n}$ - to nie jest dobry estymator ze względu na dużą wariancję
- $\hat{\mathbf{T}}_5 = 2\bar{\mathbf{x}}$ - wartość średnia wypada w środku przedziału

Definicja: Estymator $\hat{\theta}_n$ parametru θ nazywamy zgodnym, jeśli zachodzi:

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| \le \varepsilon\right) = 1 \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0$$

Estymatory zgodne i nieobciążone

Przykład: Czy estymator $\hat{T}_1 = x_{n:n}$ z poprzedniego przykładu jest zgodny?

$$P(x_i \le t) = \frac{t}{\theta} \quad \text{dla} \quad 0 \le t \le \theta \quad \Rightarrow \quad P(\hat{T}_1 \le t) = P(x_1 \le t, ..., x_n \le t) = \left(\frac{t}{\theta}\right)^n$$

$$\text{Dla } 0 < \varepsilon < \theta \text{ mamy } \quad P(|\hat{T}_1 - \theta| \le \varepsilon) = P(\hat{T}_1 \ge \theta - \varepsilon) = 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n \xrightarrow[n\to\infty]{} 1$$

Estymator, którego wariancja dąży do zera dla liczebności próby dążącej do nieskończoności jest estymatorem zgodnym (na podstawie nierówności Czebyszewa):

$$P(|\hat{\theta}_n - \theta| \ge \varepsilon) \le \frac{\sigma^2_{\hat{\theta}_n}}{\varepsilon^2} \quad \Rightarrow \quad \left(\sigma^2_{\hat{\theta}_n} \underset{n\to\infty}{\to} 0 \quad \Rightarrow \quad \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \ge \varepsilon) = 0\right)$$

Definicja: Estymator, którego wartość oczekiwana jest równa wielkości estymowanej, $\mathcal{E}[\hat{\theta}_n] = \theta$ nazywamy estymatorem nieobciążonym.

Jeśli istnieje $\mathcal{E}[\hat{\theta}_n]$, lecz $\mathcal{E}[\hat{\theta}_n] \ne \theta$, to $\hat{\theta}_n$ nazywamy estymatorem obciążonym parametru θ, natomiast różnicę $B_n(\theta) = \mathcal{E}[\hat{\theta}_n] - \theta$ - obciążeniem estymatora.

Estymator nazywamy asymptotycznie nieobciążonym jeśli zachodzi:

$$\lim_{n\to\infty} B_n(\theta) = \lim_{n\to\infty} \mathcal{E}[\hat{\theta}_n] - \theta = 0$$

Estymator wartości oczekiwanej i wariancji

Przykład: Estymatorem nieobciążonym wartości oczekiwanej jest średnia arytmetyczna:

$$\langle \bar{x} \rangle = \left\langle \frac{1}{n}\sum_{i=1}^{n} x_i \right\rangle = \frac{1}{n}\left\langle \sum_{i=1}^{n} x_i \right\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle x_i \rangle = \frac{1}{n} n \langle x \rangle = \langle x \rangle = \mu$$

Sprawdźmy czy estymatorem wariancji jest wielkość: $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$\langle S_x^2 \rangle = \left\langle \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \right\rangle = \frac{1}{n}\left\langle \sum_{i=1}^{n}\left((x_i-\mu)-(\bar{x}-\mu)\right)^2 \right\rangle = \frac{1}{n}\left\langle \sum_{i=1}^{n}\left((x_i-\mu)^2 - 2(x_i-\mu)(\bar{x}-\mu)+(\bar{x}-\mu)^2\right) \right\rangle =$$

$$= \frac{1}{n}\left\langle \sum_{i=1}^{n}(x_i-\mu)^2 - 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu) + n(\bar{x}-\mu)^2 \right\rangle = \frac{1}{n}\left\langle \sum_{i=1}^{n}(x_i-\mu)^2 - 2n(\bar{x}-\mu)(\bar{x}-\mu) + n(\bar{x}-\mu)^2 \right\rangle =$$

$$= \frac{1}{n}\left\langle \sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2 \right\rangle = \frac{1}{n}\left\langle \sum_{i=1}^{n}(x_i-\mu)^2 \right\rangle - \left\langle (\bar{x}-\mu)^2 \right\rangle = \left\langle (x-\mu)^2 \right\rangle - \left\langle (\bar{x}-\mu)^2 \right\rangle = \mathcal{V}[x] - \mathcal{V}[\bar{x}]$$

Wariancja wartości średniej:

$$\mathcal{V}[\bar{x}] = \left\langle (\bar{x}-\mu)^2 \right\rangle = \left\langle \left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right)^2 \right\rangle = \frac{1}{n^2}\left\langle \left(\sum_{i=1}^{n} x_i - n\mu\right)^2 \right\rangle = \frac{1}{n^2}\left\langle \left(\sum_{i=1}^{n}(x_i-\mu)\right)^2 \right\rangle =$$

$$= \frac{1}{n^2}\left\langle \sum_{i,j=1}^{n}(x_i-\mu)(x_j-\mu) \right\rangle = \frac{1}{n^2}\sum_{i,j=1}^{n}\left\langle (x_i-\mu)(x_j-\mu) \right\rangle = \frac{1}{n^2}\sum_{i=1}^{n}\left\langle (x_i-\mu)^2 \right\rangle + \frac{1}{n^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n} \underbrace{\langle x_i-\mu \rangle}_{=0}\underbrace{\langle x_j-\mu \rangle}_{=0} =$$

$$= \frac{1}{n^2} n \mathcal{V}[x] = \frac{1}{n}\mathcal{V}[x]$$

Nieobciążony estymator wariancji: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad \langle s_x^2 \rangle = \mathcal{V}[x]$

Błąd błędu

Wariancja estymatora błędu pojedynczego pomiaru dana jest przez:

$$\mathcal{V}\left[s_x^2\right] = \left\langle s_x^4 \right\rangle - \left\langle s_x^2 \right\rangle^2 = \frac{1}{n}\left\langle (x-\mu)^4 \right\rangle - \frac{n-3}{n(n-1)}\mathcal{V}^2[x]$$

Estymatorem wielkości $\mathcal{V}\left[s_x^2\right]$ jest kwadrat błędu kwadratu błędu:

$$s^2_{s_x^2} = \frac{n}{(n-2)(n-3)}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4 - \frac{n^2-3}{n^2}s_x^4\right) \cong \frac{1}{(n-2)(n-3)}\sum_{i=1}^{n}\left((x_i-\bar{x})^2 - s_x^2\right)^2$$

Gdy próbę pobieramy z rozkładu normalnego wówczas mamy:

$$\mathcal{D}\left[s_x^2\right] = \sqrt{\frac{3}{n}\sigma^4 - \frac{n-3}{n(n-1)}\sigma^4} = \sqrt{\frac{2}{n-1}}\sigma^2$$

$$\left\langle (x-\mu)^{2k} \right\rangle = \frac{(2k)!}{2^k k!}\sigma^{2k}$$

Potraktujmy wielkość s_x^2 jako kwadrat błędu ($u = v^2 \Rightarrow s_u \cong 2\langle v \rangle s_v$):

$$\mathcal{D}\left[s_x^2\right] \cong 2\sigma_x \mathcal{D}\left[s_x\right]$$

W przypadku rozkładu normalnego dostajemy:

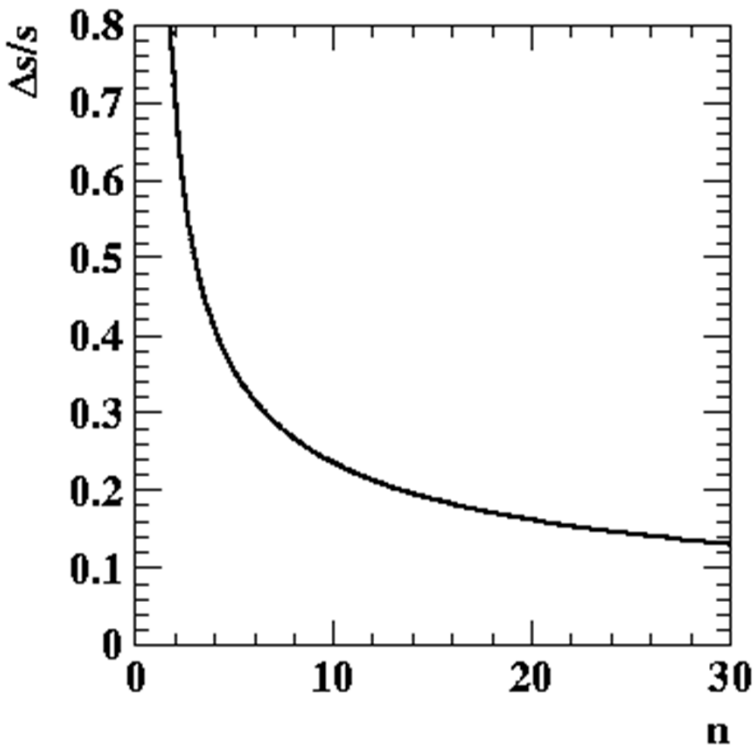
$$\mathcal{D}\left[s_x\right] \cong \frac{\mathcal{D}\left[s_x^2\right]}{2\sigma_x} = \frac{1}{2\sigma_x}\sqrt{\frac{2}{n-1}}\sigma_x^2 = \frac{1}{\sqrt{2(n-1)}}\sigma_x$$

Math Player

Błąd błędu – uwagi praktyczne

Podobnie dla średniej arytmetycznej mamy:

$$\frac{\Delta s}{s} \equiv \frac{\mathcal{D}\left[s_{\bar{x}}\right]}{s_{\bar{x}}} \cong \frac{\mathcal{D}\left[s_{\bar{x}}\right]}{\sigma_{\bar{x}}} = \frac{1}{\sqrt{2(n-1)}}$$



Przykład: Dokładność zapisu wyniku pomiaru.

$$\bar{x} = 9.87654321 \qquad s_{\bar{x}} = 1.23456789$$

Dla n=5 mamy $\Delta s/s \cong 35\%$ czyli $\Delta s \cong 0.4$

dlatego wynik zapisujemy w postaci 10 ± 1

Dla n=50 mamy $\Delta s/s \cong 10\%$ czyli $\Delta s \cong 0.1$

dlatego wynik zapisujemy w postaci 9.8 ± 1.2

Interpretacja zapisu „wynik ± błąd”:

- wartość zmierzona estymuje wartość oczekiwaną,
- błąd jest statystyczny, a jego kwadrat estymuje wariancję,
- rozkład p-twa wielkości mierzonej jest symetryczny.

Estymator kowariancji i wsp. korelacji

Nieobciążony estymator kowariancji to $\mathbf{R \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}$

$$\langle R\rangle = \frac{1}{n-1}\left\langle \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right\rangle = \frac{1}{n-1}\left\langle \sum_{i=1}^{n}\left(x_i-\mu_x-(\bar{x}-\mu_x)\right)\left(y_i-\mu_y-(\bar{y}-\mu_y)\right)\right\rangle =$$

$$= \frac{1}{n-1}\left\langle \sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)-(\bar{y}-\mu_y)\sum_{i=1}^{n}(x_i-\mu_x)-(\bar{x}-\mu_x)\sum_{i=1}^{n}(y_i-\mu_y)+\sum_{i=1}^{n}(\bar{x}-\mu_x)(\bar{y}-\mu_y)\right\rangle =$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left\langle (x_i-\mu_x)(y_i-\mu_y)\right\rangle - n\left\langle (\bar{x}-\mu_x)(\bar{y}-\mu_y)\right\rangle\right) =$$

$$= \frac{1}{n-1}\left(n\left\langle (x-\mu_x)(y-\mu_y)\right\rangle - \frac{1}{n}\sum_{i,j=1}^{n}\left\langle (x_i-\mu_x)(y_j-\mu_y)\right\rangle\right) =$$

$$= \frac{1}{n-1}\left(n\,\mathrm{cov}[x,y]-\frac{1}{n}\left(\sum_{i=1}^{n}\left\langle (x_i-\mu_x)(y_i-\mu_y)\right\rangle + \sum_{i\neq j=1}^{n}\left\langle (x_i-\mu_x)(y_j-\mu_y)\right\rangle\right)\right) =$$

$$= \frac{n\,\mathrm{cov}[x,y]-\mathrm{cov}[x,y]}{n-1} = \mathrm{cov}[x,y]$$

Asymptotycznie nieobciążony estymator współczynnika korelacji: $\mathbf{r = \frac{R}{s_x s_y}}$

Estymatory - rozkład dwumianowy

Estymator parametru p: $$\langle \hat{\mathrm{p}} \rangle = \left\langle \frac{\mathrm{k}}{n} \right\rangle = \frac{1}{n} \langle \mathrm{k} \rangle = \frac{1}{n} np = p$$

Wariancja estymatora parametru p:

$$\mathcal{V}[\hat{\mathrm{p}}] = \langle \hat{\mathrm{p}}^2 \rangle - \langle \hat{\mathrm{p}} \rangle^2 = \left\langle \left(\frac{\mathrm{k}}{n} \right)^2 \right\rangle - p^2 = \frac{1}{n^2} \langle \mathrm{k}^2 \rangle - p^2 = \frac{1}{n^2} np\big((n-1)p+1\big) - p^2 = \frac{1}{n} p(1-p)$$

Estymator wariancji estymatora parametru p:

$$\frac{1}{n} \langle \hat{p} \rangle - \frac{1}{n} \langle \hat{p}^2 \rangle = \frac{1}{n} p - \frac{1}{n} \frac{1}{n^2} np\big((n-1)p+1\big) = \frac{1}{n} p - \frac{1}{n^2} \left(np^2 - p^2 + p \right) =$$

$$= \frac{1}{n} p - \frac{1}{n} p^2 + \frac{1}{n^2} p^2 - \frac{1}{n^2} p = \frac{1}{n} p(1-p) - \frac{1}{n^2} p(1-p) = \frac{n-1}{n^2} p(1-p)$$

Nieobciążonym estymatorem wariancji jest: $$\hat{\mathcal{V}}[\hat{\mathrm{p}}] \equiv \mathrm{s}^2_{\hat{\mathrm{p}}} = \frac{1}{n-1} \hat{\mathrm{p}}(1-\hat{\mathrm{p}})$$

Estymator wariancji zmiennej losowej k:

$$\langle n\hat{\mathrm{p}}(1-\hat{\mathrm{p}}) \rangle = n \left\langle \frac{\mathrm{k}}{n} \left(1 - \frac{\mathrm{k}}{n} \right) \right\rangle = \langle \mathrm{k} \rangle - \frac{1}{n} \langle \mathrm{k}^2 \rangle = np - \frac{1}{n} np\big((n-1)p+1\big) = (n-1)pq$$

Nieobciążonym estymatorem wariancji jest: $$\hat{\mathcal{V}}[\mathrm{k}] \equiv \mathrm{s}^2_{\mathrm{k}} = \frac{n}{n-1} n\hat{\mathrm{p}}(1-\hat{\mathrm{p}})$$

Estymatory - rozkład wykładniczy

Estymator parametru τ: $$\left\langle \hat{\tau} \right\rangle = \left\langle \bar{t} \right\rangle = \left\langle \frac{1}{n}\sum_{i=1}^{n} t_i \right\rangle = \frac{1}{n}\sum_{i=1}^{n} \left\langle t_i \right\rangle = \frac{1}{n}\sum_{i=1}^{n} \tau = \tau$$

Wariancja estymatora parametru τ:

$$\nu[\hat{\tau}] = \nu\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\nu\left[\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\nu[t_i] = \frac{1}{n^2}\sum_{i=1}^{n}\tau^2 = \frac{1}{n}\tau^2$$

Estymator wariancji estymatora parametru τ:

$$\left\langle \hat{\tau}^2 \right\rangle = \left\langle \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^2 \right\rangle = \frac{1}{n^2}\left\langle \sum_{i=1}^{n} t_i^2 + \sum_{i\neq j=1}^{n} t_i\, t_j \right\rangle = \frac{1}{n^2}\left(\sum_{i=1}^{n}\left\langle t_i^2 \right\rangle + \sum_{i\neq j=1}^{n}\left\langle t_i \right\rangle\left\langle t_j \right\rangle \right) =$$

$$= \frac{1}{n^2}\left(n2\tau^2 + n(n-1)\tau^2\right) = \frac{n+1}{n}\tau^2$$

Nieobciążonym estymatorem wariancji jest: $$\hat{\nu}[\hat{\tau}] \equiv s_{\hat{\tau}}^2 = \frac{1}{n+1}\hat{\tau}^2$$

Estymator parametru λ:

$$\left\langle \frac{1}{\bar{t}} \right\rangle = \left\| t = \sum_{i=1}^{n} t_i \right\| = \left\langle \frac{n}{t} \right\rangle = n\lambda \int_0^\infty \frac{1}{t}\frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}\, dt = \left\| x = \lambda t \right\| = \frac{n}{n-1}\lambda \int_0^\infty \frac{x^{n-2}}{(n-2)!} e^{-x}\, dx = \frac{n}{n-1}\lambda$$

Estymatory - rozkład wykładniczy

Nieobciążonym estymatorem parametru λ jest: $\hat{\lambda} = \frac{n-1}{t} = \frac{n-1}{\sum_{i=1}^{n} t_i}$

Wariancja estymatora parametru λ:

$$\varepsilon[\hat{\lambda}^2] = \lambda (n-1)^2 \int_0^\infty \frac{1}{t^2} \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}\, dt = \lambda^2 \frac{n-1}{n-2}$$

$$\nu[\hat{\lambda}] = \frac{n-1}{n-2}\lambda^2 - \lambda^2 = \frac{1}{n-2}\lambda^2$$

Nieobciążonym estymatorem wariancji jest: $\hat{\nu}[\hat{\lambda}] \equiv s^2_{\hat{\lambda}} = \frac{1}{n-1}\hat{\lambda}^2$

Uwaga: Jeśli $\hat{\varphi}$ jest estymatorem parametru φ rozkładu, a $\theta = h(\varphi)$ jest funkcją tego estymatora to w ogólności $\hat{\theta} \neq h(\hat{\varphi})$.

Estymator wariancji zmiennej losowej t z rozkładu wykładniczego:

$$\langle \hat{\tau}^2 \rangle = \langle \overline{t}^2 \rangle = \left\langle \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \sum_{i=1}^{n} t_i^2 + \sum_{i \neq j=1}^{n} t_i\, t_j \right\rangle = \frac{1}{n^2} \left(\sum_{i=1}^{n} \langle t_i^2 \rangle + \sum_{i \neq j=1}^{n} \langle t_i \rangle \langle t_j \rangle \right) =$$

$$= \frac{1}{n^2}\left(n 2\tau^2 + n(n-1)\tau^2\right) = \frac{n+1}{n}\tau^2 \quad \Rightarrow \quad \hat{\nu}[t] \equiv s_t^2 = \frac{1}{n+1}\left(\sum_{i=1}^{n} t_i \right)^2$$

Metoda momentów

Przykład: Rozważmy rozkład liniowy z nieznanym parametrem θ:

$$f(x;\theta)=\frac{1}{2}(1+\theta x) \qquad -1\le x\le 1, \qquad -1\le\theta\le 1$$

Wartość oczekiwana zmiennej x: $\mathcal{E}[\mathrm{x}]=\int_{-1}^{1} xf(x;\theta)\,dx=\frac{1}{2}\int_{-1}^{1} x(1+\theta x)\,dx=\frac{1}{3}\theta$

Estymator parametru θ: $\bar{\mathrm{x}}=\frac{1}{3}\hat{\theta} \quad\Rightarrow\quad \hat{\theta}=3\bar{\mathrm{x}}$

Jego wariancja: $\mathcal{V}[\hat{\theta}]=9\mathcal{V}[\bar{\mathrm{x}}]=\frac{9}{n}\mathcal{V}[\mathrm{x}]=\frac{1}{n}(3-\theta^2)$

Estymator parametru θ na podstawie wariancji:

$$\mathcal{V}[\mathrm{x}]=\mathcal{E}[\mathrm{x}^2]-\mathcal{E}^2[\mathrm{x}]=\frac{1}{2}\int_{-1}^{1} x^2(1+\theta x)\,dx-\frac{1}{9}\theta^2=\frac{1}{3}-\frac{1}{9}\theta^2$$

$$\mathrm{s}_x^2=\frac{1}{3}-\frac{1}{9}\hat{\theta}^2 \quad\Rightarrow\quad \hat{\theta}=\sqrt{3-9\mathrm{s}_x^2}$$

W ogólnym przypadku do znalezienia estymatorów dla *n* parametrów rozkładu potrzebujemy *n* momentów (najłatwiej wykorzystać najniższe):

$$\left\langle\frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i^k\right\rangle=\frac{1}{n}\left\langle\sum_{i=1}^{n}\mathrm{x}_i^k\right\rangle=\frac{1}{n}\sum_{i=1}^{n}\left\langle\mathrm{x}_i^k\right\rangle=\frac{1}{n}n\left\langle\mathrm{x}^k\right\rangle=\left\langle\mathrm{x}^k\right\rangle=m_k$$