

Rachunek prawdopodobo- bieństwa i statystyka

Wykład 9

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 $f(x_1, x_2, \dots, x_n)$ próby losowej x_1, x_2, \dots, 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, \dots, x_n pobranej z tej populacji.

Estymator i estymata

Estymatorem parametru θ nazywamy każdą statystykę $\hat{\theta}_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, 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.

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 $\mathcal{E}[\hat{\theta}_n] \neq \theta$, 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:

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

Estymator nazywamy **zgodnym**, jeśli wariancja tego estymatora dąży do zera dla liczebności próby dążącej do nieskończoności.

Przykład: Estymatorem nieobciążonym wartości

oczekiwanej jest średnia arytmetyczna: $\langle \bar{x} \rangle = \left\langle \frac{1}{n} \sum_{i=1}^n x_i \right\rangle = \frac{1}{n} \left\langle \sum_{i=1}^n x_i \right\rangle = \frac{1}{n} \sum_{i=1}^n \langle x_i \rangle = \frac{1}{n} n \langle x \rangle = \langle x \rangle = \mu$

Estymator wariancji

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

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

Wariancja wartości średniej:

$$\begin{aligned}\mathcal{V}[\bar{x}] &= \langle (\bar{x} - \mu)^2 \rangle = \left\langle \left(\frac{1}{n} \sum_{i=1}^n x_i - \mu \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \left(\sum_{i=1}^n x_i - n\mu \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \left(\sum_{i=1}^n (x_i - \mu) \right)^2 \right\rangle = \\ &= \frac{1}{n^2} \left\langle \sum_{i,j=1}^n (x_i - \mu)(x_j - \mu) \right\rangle = \frac{1}{n^2} \sum_{i,j=1}^n \langle (x_i - \mu)(x_j - \mu) \rangle = \frac{1}{n^2} \sum_{i=1}^n \langle (x_i - \mu)^2 \rangle + \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^n \underbrace{\langle x_i - \mu \rangle}_{=0} \underbrace{\langle x_j - \mu \rangle}_{=0} = \\ &= \frac{1}{n^2} n \mathcal{V}[x] = \frac{1}{n} \mathcal{V}[x]\end{aligned}$$

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

Błąd błędu

Wariancja estymatora błędu pojedynczego pomiaru dana jest przez:

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

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

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

Gdy próbę pobieramy z rozkładu normalnego wówczas mamy:

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

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

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

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

W przypadku rozkładu normalnego dostajemy:

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

Math
Player

Błąd błędu - uwagi praktyczne

Podobnie dla średniej arytmetycznej mamy:

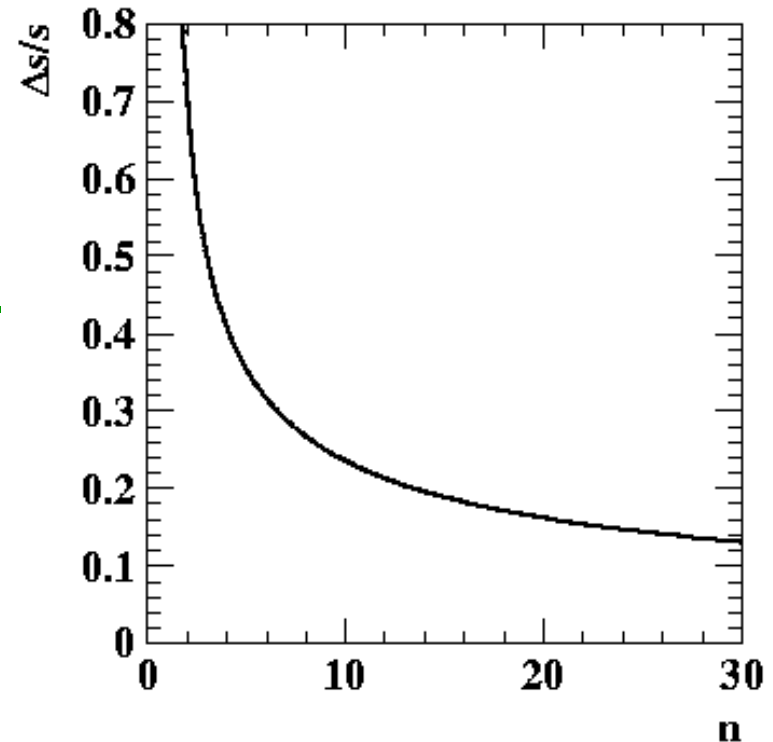
$$\frac{\Delta s}{s} \equiv \frac{\mathcal{D}[s_{\bar{x}}]}{s_{\bar{x}}} \approx \frac{\mathcal{D}[s_{\bar{x}}]}{\sigma_{\bar{x}}} = \frac{1}{\sqrt{2(n-1)}}$$

Przykład: Dokładność zapisu wyniku pomiaru.

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

Dla $n=5$ mamy $\Delta\sigma/\sigma \approx 35\%$ czyli $\Delta\sigma \approx 0.4$
dlatego wynik zapisujemy w postaci 10 ± 1

Dla $n=50$ mamy $\Delta\sigma/\sigma \approx 10\%$ czyli $\Delta\sigma \approx 0.1$
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 $\mathbf{R} \equiv \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_i - \bar{\mathbf{y}})$

$$\begin{aligned}\langle \mathbf{R} \rangle &= \frac{1}{n-1} \left\langle \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_i - \bar{\mathbf{y}}) \right\rangle = \frac{1}{n-1} \left\langle \sum_{i=1}^n (\mathbf{x}_i - \mu_x - (\bar{\mathbf{x}} - \mu_x))(\mathbf{y}_i - \mu_y - (\bar{\mathbf{y}} - \mu_y)) \right\rangle = \\ &= \frac{1}{n-1} \left\langle \sum_{i=1}^n (\mathbf{x}_i - \mu_x)(\mathbf{y}_i - \mu_y) - (\bar{\mathbf{y}} - \mu_y) \sum_{i=1}^n (\mathbf{x}_i - \mu_x) - (\bar{\mathbf{x}} - \mu_x) \sum_{i=1}^n (\mathbf{y}_i - \mu_y) + \sum_{i=1}^n (\bar{\mathbf{x}} - \mu_x)(\bar{\mathbf{y}} - \mu_y) \right\rangle = \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n \langle (\mathbf{x}_i - \mu_x)(\mathbf{y}_i - \mu_y) \rangle - n \langle (\bar{\mathbf{x}} - \mu_x)(\bar{\mathbf{y}} - \mu_y) \rangle \right) = \\ &= \frac{1}{n-1} \left(n \langle (\mathbf{x} - \mu_x)(\mathbf{y} - \mu_y) \rangle - \frac{1}{n} \sum_{i,j=1}^n \langle (\mathbf{x}_i - \mu_x)(\mathbf{y}_j - \mu_y) \rangle \right) = \\ &= \frac{1}{n-1} \left(n \text{cov}[\mathbf{x}, \mathbf{y}] - \frac{1}{n} \left(\sum_{i=1}^n \langle (\mathbf{x}_i - \mu_x)(\mathbf{y}_i - \mu_y) \rangle + \sum_{i \neq j=1}^n \langle (\mathbf{x}_i - \mu_x)(\mathbf{y}_j - \mu_y) \rangle \right) \right) = \\ &= \frac{n \text{cov}[\mathbf{x}, \mathbf{y}] - \text{cov}[\mathbf{x}, \mathbf{y}]}{n-1} = \text{cov}[\mathbf{x}, \mathbf{y}]\end{aligned}$$

Asymptotycznie nieobciążony estymator
współczynnika korelacji:

$$\mathbf{r} = \frac{\mathbf{R}}{\mathbf{S}_x \mathbf{S}_y}$$

Estymatory - rozkład dwumianowy

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

Wariancja estymatora parametru p :

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

Estymator wariancji estymatora parametru p :

$$\begin{aligned} \frac{1}{n} \langle \hat{p} \rangle - \frac{1}{n} \langle \hat{p}^2 \rangle &= \frac{1}{n} p - \frac{1}{n} \frac{1}{n^2} np((n-1)p+1) = \frac{1}{n} p - \frac{1}{n^2} (np^2 - p^2 + p) = \\ &= \frac{1}{n} p - \frac{1}{n} p^2 + \frac{1}{n^2} p^2 - \frac{1}{n^2} p = \frac{1}{n} p(1-p) - \frac{1}{n^2} p(1-p) = \frac{n-1}{n^2} p(1-p) \end{aligned}$$

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

Estymator wariancji zmiennej losowej k :

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

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

Estymatory - rozkład wykładniczy

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

Wariancja estymatora parametru τ :

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

Estymator wariancji estymatora parametru τ :

$$\begin{aligned} \langle \hat{\tau}^2 \rangle &= \left\langle \left(\frac{1}{n} \sum_{i=1}^n t_i \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \sum_{i=1}^n t_i^2 + \sum_{i \neq j=1}^n t_i t_j \right\rangle = \frac{1}{n^2} \left(\sum_{i=1}^n \langle t_i^2 \rangle + \sum_{i \neq j=1}^n \langle t_i \rangle \langle t_j \rangle \right) = \\ &= \frac{1}{n^2} (n2\tau^2 + n(n-1)\tau^2) = \frac{n+1}{n} \tau^2 \end{aligned}$$

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

Estymator parametru λ :

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

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 :

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

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

Nieobciążonym estymatorem wariancji jest: $\hat{\mathcal{V}}[\hat{\lambda}] \equiv s_{\hat{\lambda}}^2 = \frac{1}{n-1} \hat{\lambda}^2$

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:

$$\begin{aligned} \langle \hat{t}^2 \rangle &= \langle \bar{t}^2 \rangle = \left\langle \left(\frac{1}{n} \sum_{i=1}^n t_i \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \sum_{i=1}^n t_i^2 + \sum_{i \neq j=1}^n t_i t_j \right\rangle = \frac{1}{n^2} \left(\sum_{i=1}^n \langle t_i^2 \rangle + \sum_{i \neq j=1}^n \langle t_i \rangle \langle t_j \rangle \right) = \\ &= \frac{1}{n^2} (n2\tau^2 + n(n-1)\tau^2) = \frac{n+1}{n} \tau^2 \quad \Rightarrow \quad \hat{\mathcal{V}}[t] \equiv s_t^2 = \frac{1}{n+1} \left(\sum_{i=1}^n t_i \right)^2 \end{aligned}$$

Metoda momentów

Przykład: Rozważmy rozkład liniowy z nieznanym parametrem θ :

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

Wartość oczekiwana zmiennej x :

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

Estymator parametru θ :

$$\bar{x} = \frac{1}{3} \hat{\theta} \quad \Rightarrow \quad \hat{\theta} = 3\bar{x}$$

Jego wariancja:

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

Estymator parametru θ na podstawie wariancji:

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

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

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