Štatistická inferencia

Asymptotické vlastnosti odhadov

Stanislav Katina[1]

[1]Ústav matematiky a štatistiky, Masarykova univezita
Honorary Research Fellow, The University of Glasgow

11. decembra 2018

Asymptotické vlastnosti odhadov

Konvergencie

Štatistická inferencia o θ_* sa vykonáva na základe odhadu $\widehat{\theta} = \widehat{\theta}_n$. Podobne štatistická inferencia o $g(\theta_*)$ sa vykonáva na základe odhadu $g(\widehat{\theta}_n)$. **Bodový odhad** parametrickej funkcie $g(\theta)$ je štatistika $T_n = T(X_1, X_2, \ldots, X_n)$ nadobúdajúca hodnoty blízko $g(\theta)$. Nech $T_n = g(\widehat{\theta}_n)$ je nejaký odhad $g(\theta)$, podobne $T_n = \widehat{\theta}_n$ je nejaký odhad θ.

Potom môžeme definovať nasledovné typy konvergencií pre T_n (čo platí aj pre $g(\theta) = \theta$):

Asymptotické vlastnosti odhadov

Konvergencie

1. **konvergencia skoro všade**
$$\Pr\left(\lim_{n\to\infty} T_n = g(\theta)\right) = 1, \text{ pre } \forall\theta\in\Theta,$$
2. **konvergencia v kvadratickom strede**
$$T_n \overset{n\to\infty}{\rightrightarrows} g(\theta) \Leftrightarrow E_\theta\left[(T_n - g(\theta))^2\right] \overset{n\to\infty}{\rightrightarrows} 0, \text{ pre } \forall\theta\in\Theta,$$
3. **konvergencia podľa pravdepodobnosti** (ozn. $\xrightarrow{\mathcal{P}}$)
$$\lim_{n\to\infty}\left(\Pr\left(|T_n - g(\theta)| \geq \varepsilon\right)\right) = 0, \text{ pre } \varepsilon > 0, \text{ pre } \forall\theta\in\Theta,$$
4. **konvergencia v distribúcii** (ozn. $\overset{\mathcal{D}}{\sim}$)
$$\lim_{n\to\infty} F_n(x) = F_X(x)$$
vo všetkých bodoch x, kde $F_n(x)$ je empirická distribučná funkcia a $F_X(x)$ je spojitá distribučná funkcia.

Asymptotické vlastnosti odhadov

Vlastnosti odhadov

Nech $X_1, X_2, \ldots, X_n$ je náhodný výber z rozdelenia F_*, $g(\theta)$ je parametrická funkcia a T_n, T_1 a T_2 sú štatistiky, $E[T_n]$ je stredná hodnota a $Var[T_n]$ rozptyl štatistiky T_n.

- hovoríme, že štatistika T_n je **nevychýlený odhad** parametrickej funkcie $g(\theta)$, ak $E[T_n] = g(\theta), \forall\theta \in \Theta$, t.j. odhad T_n nesmie paramatrickú funkciu *systematicky nadhodnocovať* alebo *podhodnocovať* – ak táto podmienka nie je splnená, hovoríme o **vychýlenom odhade**,
- ak T_1, T_2 sú dva nevychýlené odhady tej istej parametrickej funkcie $g(\theta)$, potom T_1 je **lepší odhad** $g(\theta)$ ako T_2, ak $Var[T_1] < Var[T_2], \forall\theta \in \Theta$, t.j. ak existuje odhad s menším rozptylom, je potrebné ho v štatistickej inferencii použiť namiesto odhadu s väčším rozptylom (to môžeme docieliť optimalizáciou dizajnu experimentu vhodnou voľbou bodov, v ktorých budeme merať),

- T_n sa nazýva **asymptoticky nevychýlený odhad** $g(\theta)$, ak platí $\lim_{n\to\infty} E[T_n] = g(\theta)$, t.j. s rastúcim rozsahom náhodného výberu n, klesá výchylka odhadu (to môžeme docieliť optimalizáciou dizajnu experimentu vhodnou voľbou n),
- T_n sa nazýva **konzistentný odhad** $g(\theta)$, ak platí $\lim_{n\to\infty} \Pr(|T_n - g(\theta)| \geq \epsilon) = 0$, t.j. s rastúcim rozsahom náhodného výberu n klesá pravdepodobnosť, že odhad sa bude realizovať ďaleko od $g(\theta)$ (to môžeme docieliť optimalizáciou dizajnu experimentu vhodnou voľbou n),

- T_n sa nazýva **konzistentný** a **asymptoticky efektívny (eficientný, výdatný) odhad** $g(\theta)$, ak platí $(T_n - g(\theta)) \overset{\mathcal{D}}{\sim} N(0, C(g(\theta)))$, kde $\mathcal{D}$ znamená v distribúcii, $C(g(\theta))$ je Raova-Cramerova spodná hranica definovaná ako $C(g(\theta)) = -\frac{[g'(\theta)]^2}{E_\theta[\frac{\partial^2}{\partial\theta^2} \ln L(\theta|\mathbf{x})]} = \frac{[g'(\theta)]^2}{ni(\theta)}$, kde $i(\theta)$ je Fisherova miera informácie jedného pozorovania, $L(\theta|\mathbf{x})$ je funkcia vierohodnosti (táto vlastnosť sa používa pri testovaní hypotéz ako predpoklad asymptotického rozdelenia testovacej štatistiky za predpokladu normality rozdelenia X),
- T_n sa nazýva **asymptoticky normálny odhad** $g(\theta)$, ak platí $\frac{T_n - g(\theta)}{\sqrt{C(g(\theta))}} \overset{\mathcal{D}}{\sim} N(0, 1)$ (táto vlastnosť sa tiež používa pri testovaní hypotéz ako predpoklad asymptotického rozdelenia testovacej štatistiky za predpokladu normality rozdelenia X).

Z nevychýlenosti odhadu vyplýva jeho asymptotická nevychýlenosť a z asymptotickej nevychýlenosti vyplýva konzistencia, ak $\lim_{n\to\infty} Var[T_n] = 0$. Implikáciou posleného bodu je

$$T_n \overset{\mathcal{D}}{\sim} N(g(\theta), \frac{[g'(\theta)]^2}{ni(g(\theta))}),$$

kde $ni(g(\theta)) = \mathcal{I}(g(\theta))$. Ak $g(\theta) = \theta$, potom

$$T_n \overset{\mathcal{D}}{\sim} N(\theta, \frac{1}{ni(\theta)}),$$

kde $ni(\theta) = \mathcal{I}(\theta)$. Ak $k > 1$, potom je $\boldsymbol{\theta}$ vektor a môžeme písať

$$\sqrt{n}\left(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right) \overset{\mathcal{D}}{\sim} N_k\left(\mathbf{0}, \Delta^T (i(\boldsymbol{\theta}))^{-1} \Delta\right),$$

kde $\Delta = \frac{\partial}{\partial\boldsymbol{\theta}} g(\boldsymbol{\theta})$. Ak $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$, potom

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \overset{\mathcal{D}}{\sim} N_k\left(\mathbf{0}, (i(\boldsymbol{\theta}))^{-1}\right).$$

Nech náhodná premenná X pochádza z **normálneho rozdelenia** s parametrami μ a σ^2, $X \sim N(\mu, \sigma^2)$, kde $E[X] = \mu$ je stredná hodnota a $Var[X] = \sigma^2$ je rozptyl náhodnej premennej X. Potom výberový aritmetický priemer $\overline{X}_n = \overline{X}$ je **nevychýleným odhadom** μ a výberový rozptyl $S^2_{n-1} = S^2$ je **nevychýleným odhadom** σ^2 ($n-1$ v dolnom indexe S^2_{n-1} znamená počet stupňov voľnosti). Zároveň platí, že $\overline{X}$ a S^2 sú **asymptoticky nevychýlené**, **konzistentné**, **asymptoticky efektívne** a **asymptoticky normálne**. Preto môžeme písať nasledovné

$$\sqrt{n}(\overline{X}_n - \mu) \overset{\mathcal{D}}{\sim} N(0, \sigma^2), \text{ čo je ekvivalentné } \overline{X}_n \overset{\mathcal{D}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

a

$$\sqrt{n}(S^2_n - \sigma^2) \overset{\mathcal{D}}{\sim} N(0, 2\sigma^4), \text{ kde } S^2_n = \frac{n-1}{n} S^2_{n-1}.$$

Tiež môžeme písať

$$\sqrt{n}(S_{n-1}^2-\sigma^2) \overset{\mathcal{D}}{\sim} \sqrt{n}\sigma^2\left(\frac{\chi_{n-1}^2}{n-1}-1\right) \text{ a } \sqrt{n}(S_n^2-\sigma^2) \overset{\mathcal{D}}{\sim} \sqrt{n}\sigma^2\left(\frac{\chi_{n-1}^2}{n}-1\right),$$

čo je ekvivalentné $(n-1)\frac{S_{n-1}^2}{\sigma^2} \overset{\mathcal{D}}{\sim} \chi_{n-1}^2$ a $n\frac{S_n^2}{\sigma^2} \overset{\mathcal{D}}{\sim} \chi_{n-1}^2$. Potom môžeme písať nasledovné

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}) \overset{\mathcal{D}}{\sim} N_2(\mathbf{0},\mathbf{\Sigma}),$$

kde $\widehat{\boldsymbol{\theta}}_n=(\overline{X}_n,S_n^2)^T$, $\boldsymbol{\theta}=(\mu,\sigma^2)^T$, $\mathbf{0}=(0,0)^T$ a $\mathbf{\Sigma}=\mathrm{diag}(\sigma^2,2\sigma^4)$. Na tomto mieste ale treba zdôrazniť, že n použité pri odhadovaní musí byť dostatočne veľké, zvyčajne sa uvádza $n>30$. Realizáciou $\overline{X}$ je $\overline{x}$ a realizáciou S^2 je s^2.

Nech dvojrozmerná náhodná premenná $(X_1,X_2)^T$ pochádza z **dvojrozmerného normálneho rozdelenia** s parametrami $\boldsymbol{\mu}$ a $\mathbf{\Sigma}$, $(X_1,X_2)^T\sim N_2(\boldsymbol{\mu},\mathbf{\Sigma})$, kde $\boldsymbol{\mu}=(\mu_1,\mu_2)^T$, $\mathbf{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{pmatrix}$, $\sigma_{12}=\sigma_{21}=\rho\sigma_1\sigma_2$, $E[X_1]=\mu_1$ je stredná hodnota a $Var[X_1]=\sigma_1^2$ je rozptyl náhodnej premennej X_1; $E[X_2]=\mu_2$ je stredná hodnota a $Var[X_2]=\sigma_2^2$ je rozptyl náhodnej premennej X_2. Potom výberové aritmetické priemery $\overline{X}_j$ ($j=1,2$) sú **nevychýlené odhady** μ_j a výberové rozptyly $S_{n-1,j}^2=S_j^2$ sú **nevychýlenými odhadmi** σ_j^2. Zároveň platí, že $\overline{X}_j$ a $S_{n-1,j}^2$ sú **asymptoticky nevychýlené**, **konzistentné**, **asymptoticky efektívne** a **asymptoticky normálne**.

Ďalej platí, že výberová kovariancia S_{12} je **nevychýleným odhadom** σ_{12}, ako aj, že je **asymptoticky nevychýlený**, **konzistentný**, **asymptoticky efektívny** a **asymptoticky normálny** odhad. Na tomto mieste ale treba zdôrazniť, že n použité pri odhadovaní musí byť dostatočne veľké, zvyčajne sa uvádza $n>30$.

Avšak $E[R_{12}]$ sa rovná ρ len približne (zhoda nastane až pre $n>30$). Naviac píšeme, že R_{12} je **nevychýleným odhadom** ρ. Zároveň platí, že R_{12} je **asymptoticky nevychýlený** a **konzistentný**.

Realizáciou S_{12} je s_{12} a realizáciou R_{12} je r.

Nech náhodná premenná X pochádza z normálneho rozdelenia s parametrami μ a σ^2, $X\sim N(\mu,\sigma^2)$, kde $E[X]=\mu$ je stredná hodnota a $Var[X]=\sigma^2$ je rozptyl náhodnej premennej X. Nech $g(\boldsymbol{\theta})=\sigma/\mu$, kde $\boldsymbol{\theta}=(\mu,\sigma^2)^T$, $g(\widehat{\boldsymbol{\theta}}_n)=\frac{S_n}{\overline{X}_n}$ a $\Delta=\frac{\partial}{\partial\boldsymbol{\theta}}g(\boldsymbol{\theta})=\left(-\frac{\sigma}{\mu^2},\frac{1}{2\sigma\mu}\right)^T$. Potom

$$\sqrt{n}\left(\frac{S_n}{\overline{X}_n}-\frac{\sigma}{\mu}\right) \overset{\mathcal{D}}{\sim} N_k\left(0,\Delta^T(i(\boldsymbol{\theta}))^{-1}\Delta\right),$$

kde

$$\Delta^T(i(\boldsymbol{\theta}))^{-1}\Delta=\left(-\frac{\sigma}{\mu^2},\frac{1}{2\sigma\mu}\right)\begin{pmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{pmatrix}\left(-\frac{\sigma}{\mu^2},\frac{1}{2\sigma\mu}\right)^T=\frac{\sigma^2}{\mu^2}\left(\frac{\sigma^2}{\mu^2}+\frac{1}{2}\right).$$