Becsléselméleti alapfogalmak

A becslésnél az $X_1, X_2, \ldots, X_n$ mintából szeretnénk az X háttérváltozó egy jellemzőjét (tipikusan: paraméterét vagy várható értékét vagy szórását) - jelölje most ezt θ - megbecsülni. Ehhez kiszámoljuk a becslést:

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$$

Azt mondjuk, hogy $\widehat{\theta}$ a θ <u>torzítatlan becslése</u>, ha

$$E(\widehat{\theta}) = \theta$$

Például az $\overline{X}$ az $E(X)$ egy torzítatlan becslése, mert

$$E(\overline{X}) = E(X)$$

De az $E(X)$ egy másik torzítatlan becslése X_1, mert

$$E(X_1) = E(X)$$

Melyik a jobb becslés?

Célszerű azt választani, amelyiknek kisebb a szórása.

Az X háttérváltozó egy θ jellemzőjének a

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$$

torzítatlan becslés hatásos becslése, ha neki a torzítatlan becslések közül legkisebb a szórása.

Eloszlás	Maximum-likelihood becslés	Hatásos becslés
(N, p) binomilis	$\widehat{p} = \overline{X}$	$\overline{X}$
λ Poisson	$\widehat{\lambda} = \overline{X}$	$\overline{X}$
λ exponenciális	$\widehat{\lambda} = \frac{1}{\overline{X}}$	$\frac{n-1}{n}\frac{1}{\overline{x}}$
(m, σ) normális	$\widehat{m} = \overline{X}$, $\widehat{\sigma} = S$	$m = \overline{X}$
$(0, \theta)$ egyenletes	$\widehat{\theta} = X_n^*$	$\frac{n+1}{n} X_n^*$

Belátható, hogy a maximum-likelihood becslés majdnem torzítatlan és majdnem hatásos lesz, tehát jól használható.

Az X háttérváltozó egy θ jellemzőjének a

$$\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, \ldots, X_n)$$

torzítatlan becslés konzisztens becslése, ha

$$\widehat{\theta}_n \to \theta$$

majdnem biztosan, amint $n \to \infty$.

Megmutatható, hogy ha az X háttérváltozónak véges a szórása, akkor $\overline{X}$ konzisztens becslése $E(X)$ várható értéknek.

Intervallumbecslések

Az alapfeladat az, hogy az X háttérváltozó egy θ jellemzője esetén adott $0 < \epsilon < 1$ valós számhoz az $X_1, X_2, \ldots, X_n$ mintából készítsünk olyan

$$\widehat{\theta}_1 = \widehat{\theta}_1(X_1, X_2, \ldots, X_n)$$

és

$$\widehat{\theta}_2 = \widehat{\theta}_2(X_1, X_2, \ldots, X_n)$$

értékeket, hogy

$$P_\theta(\widehat{\theta}_1 < \theta < \widehat{\theta}_2) \geq 1 - \epsilon$$

teljesüljön minden szóba jöhető θ-ra.

Általában $\epsilon = 0,05$ vagy $\epsilon = 0,01$.

Az $(\widehat{\theta}_1, \widehat{\theta}_2)$ intervallumot $1 - \epsilon$-szintű konfidencia intervallumnak hívják.

Konfidenciaintervallum szerkesztése (m, σ_0) paraméterű normális eloszlásra, ahol σ_0 ismert

Az $0 < \epsilon < 1$ legyen rögzítve. Adott $X_1, X_2, \ldots, X_n$ minta esetén az $\overline{X}$-ból indulunk ki. Célszerű olyan r_ϵ-t keresni, amire igaz, hogy

$$P(\overline{X} - r_\epsilon < \mu < \overline{X} + r_\epsilon) = 1 - \epsilon,$$

azaz

$$P(-r_\epsilon < \overline{X} - \mu < r_\epsilon) = 1 - \epsilon$$

Tudjuk, hogy $\overline{X} - \mu$ olyan normális eloszlsú v.v., aminek a várható értéke 0 és szórása $\frac{\sigma_0}{\sqrt{n}}$. Ekkor

$$\frac{\overline{X} - \mu}{\frac{\sigma_0}{\sqrt{n}}}$$

standard normális eloszlású v.v. lesz.

Olyan r_ϵ számra van szükségünk, amire

$$P(-\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}} < \frac{\overline{X}-\mu}{\frac{\sigma_0}{\sqrt{n}}} < \frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}}) = 1-\epsilon.$$

Tudjuk, hogy

$$P(-\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}} < \frac{\overline{X}-\mu}{\frac{\sigma_0}{\sqrt{n}}} < \frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}}) = \Phi(\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}}) - \Phi(-\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}}) =$$

$$\Phi(\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}}) - (1-\Phi(\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}})) = 2\Phi(\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}}) - 1 = 1-\epsilon,$$

azaz

$$\Phi(\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}}) = 1-\frac{\epsilon}{2}.$$

Jelölje $u_{\frac{\epsilon}{2}}$ azt az értéket, amelyre

$$\Phi(u_{\frac{\epsilon}{2}}) = 1 - \frac{\epsilon}{2}$$

Ekkor

$$1 - \frac{\epsilon}{2} = \Phi(\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}}) = \Phi(u_{\frac{\epsilon}{2}}),$$

ezért

$$\frac{r_\epsilon}{\frac{\sigma_0}{\sqrt{n}}} = u_{\frac{\epsilon}{2}} \quad \Rightarrow \quad r_\epsilon = \frac{u_{\frac{\epsilon}{2}}\sigma_0}{\sqrt{n}}.$$

Tehát az $1 - \epsilon$-szintű intervallum:

$$(\overline{X} - \frac{u_{\frac{\epsilon}{2}}\sigma_0}{\sqrt{n}}, \overline{X} + \frac{u_{\frac{\epsilon}{2}}\sigma_0}{\sqrt{n}})$$

Megjegyzések:

1. Például $\epsilon = 0,05$ esetén $u_{\frac{\epsilon}{2}} = 1,96$ így a konfidencia intervallum:

$$\left(\overline{X} - 1,96\frac{\sigma_0}{\sqrt{n}}, \overline{X} + 1,96\frac{\sigma_0}{\sqrt{n}}\right)$$

   és $\epsilon = 0,01$ esetén $u_{\frac{\epsilon}{2}} = 2,58$, így a konfidencia intervllum:

$$\left(\overline{X} - 2,58\frac{\sigma_0}{\sqrt{n}}, \overline{X} + 2,58\frac{\sigma_0}{\sqrt{n}}\right)$$

2. Több $1 - \epsilon$-szintű konfidenciaintarvallum szerkeszthető, de ezek közül a legrövidebbet az előbb meghatározott intervallum adja, ezért ez optimális megoldása a feladatnak.
3. Látjuk, hogy az intervallum hossza a mintaelemszám n növelésével csökken, 0-hoz tart.
4. Azt is látjuk, hogy az ϵ csökkentésével az intervallum hossza nő, tehát ha biztosabbak akarunk lenne, akkor nagyobb intervallomot kell választanunk.

Példák: Az 50 kg-os cementeszsákok készítésénél a töltőtömeget lehet beállítani, a szórás adott. Tegyük fel, hogy cementeszsákok súlya normális eloszlást követ és a szórás: 0,2.

1. Tegyük fel, hogy 16 zsákot veszünk, melyek súlyai:

$$49,9 \quad 50,2 \quad 50,1 \quad 49,7 \quad 49,8 \quad 49,9 \quad 50 \quad 49,8$$

$$49,8 \quad 50,3 \quad 50 \quad 49,9 \quad 49,8 \quad 49,9 \quad 50,1 \quad 49,8$$

   Adja meg az $\epsilon = 0,05$ értékhez tartozó konfidencia intervallumot, azaz adja meg azt az intervallumot, ahová a tényleges átlagérték 95% biztonsággal esik

2. Hány zsákot kell lemérnem ahhoz, hogy 99% biztonsággal $0,2$ kg intervallumba ismerjem a tényleges átlagértéket?

Megoldás: Az átlagérték:

$$\overline{X} = \frac{49,9 + 50,2 + \cdots + 49,8}{16} = 49,9375$$

Az $\epsilon = 0,05$ esetén $u_{0,025} = 1,96$, ezért

$$\left(\overline{X} - 1,96\frac{\sigma_0}{\sqrt{n}}, \overline{X} + 1,96\frac{\sigma_0}{\sqrt{n}}\right) =$$

$$\left(49,9375 - 1,96\frac{0,2}{\sqrt{16}}, \quad 49,9375 + 1,96\frac{0,2}{\sqrt{16}}\right) =$$

$$(49,8395; \quad 50,0355)$$

A második kérdés megválaszolásához olyan n-et kell keresnünk, hogy arra már az

$$(\overline{X} - \frac{u_{\frac{\epsilon}{2}}\sigma_0}{\sqrt{n}}, \overline{X} + \frac{u_{\frac{\epsilon}{2}}\sigma_0}{\sqrt{n}})$$

intervallum hossza legfeljebb 0,2. Ennek az intervallumnak a hossza:

$$\overline{X} + \frac{u_{\frac{\epsilon}{2}}\sigma_0}{\sqrt{n}} - (\overline{X} - \frac{u_{\frac{\epsilon}{2}}\sigma_0}{\sqrt{n}}) = 2\frac{u_{\frac{\epsilon}{2}}\sigma_0}{\sqrt{n}} < 0,2$$

Most $1 - \epsilon = 0,99$, azaz $\epsilon = 0,01$, így $u_{\epsilon/2} = u_{0,005} = 2,58$. Innen

$$2 * 2,58\frac{0,2}{\sqrt{n}} \leq 0,2$$

$$5,16\frac{0,2}{\sqrt{n}} \leq 0,2 \quad \Rightarrow \quad n \geq 26,7$$

azaz 27 cementeszsákot kell megvizsgálnunk.

Konfidenciaintervallum szerkesztése (m, σ) paraméterű normális eloszlásra, ahol σ nem ismert

Az $0 < \epsilon < 1$ legyen rögzítve. Adott $X_1, X_2, \ldots, X_n$ minta esetén az $\overline{X}$-ból indulunk ki. Célszerű olyan r_ϵ-t keresni, amire igaz, hogy

$$P(\overline{X} - r_\epsilon < \mu < \overline{X} + r_\epsilon) = 1 - \epsilon,$$

azaz

$$P(-r_\epsilon < \overline{X} - \mu < r_\epsilon) = 1 - \epsilon$$

Mivel σ nem ismert, ezért nehezebb megoldani a feladatot. Bevezetünk egy új eloszlást az ún. t- vagy Student-eloszlást:

Legyen $Z_1, Z_2, \ldots, Z_n$ független standard normális eloszlású v. v.-k. Jelölje

$$Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

Ekkor Y-t n-szabadságfokú χ^2-eloszlású v.v-nak hívjuk.

Legyen Z szinén egy standard normális eloszlású v.v változó. Ekkor

$$\frac{Z}{\sqrt{Y/n}}$$

v.v-t n-szabadságfokú Student eloszlásnak nevezzük és $t(n)$-nel jelöljük, az ő eloszásfüggvénye: $G_n(x)$.

Visszatérve az eredeti feladaunkra: megmutatható, hogy $\overline{X}$-sal jelölve a mintaátlagot és S_n^*-gal a korrigált empirikus szórást

$$\frac{\overline{X}-\mu}{S_n^*}\sqrt{n}$$

eloszlása $t(n-1)$ eloszlás lesz, azaz $(n-1)$-szabadságfokú Student-eloszlás.

A feladat megoldásához válasszunk egy olyan $t_{\epsilon/2}(n-1)$ számot, melyre

$$G_{n-1}(t_{\epsilon/2}(n-1)) = 1 - \frac{\epsilon}{2},$$

amit táblázatból lehet meghatározni.

Ekkor

$$r_\epsilon = \frac{t_{\epsilon/2}(n-1)S_n^*}{\sqrt{n}},$$

azaz a keresett $(1-\epsilon)$-szintű konfidencia intervallum:

$$\left(\overline{X} - \frac{t_{\epsilon/2}(n-1)S_n^*}{\sqrt{n}}, \overline{X} + \frac{t_{\epsilon/2}(n-1)S_n^*}{\sqrt{n}}\right)$$

Megjegyzés: Ha n nagy (ezt általában $n \geq 30$ esetén mondják), akkor a $t(n)$ eloszlású v.v. jól közelíthető normális eloszlású v.v-val. ezért ekkor a

$$\left(\overline{X} - \frac{u_{\epsilon/2} S_n^*}{\sqrt{n}}, \overline{X} + \frac{u_{\epsilon/2} S_n^*}{\sqrt{n}}\right)$$

képlettel számolnak, míg az előző oldalon szereplő képletet kis n (azaz $n < 30$) esetén használják.

Feladat: Az 50 kg-os cementeszsákok tömege véletlen mennyiség, amiről tudjuk, hogy normális eloszlást követ, de az (m, σ) paramétereket nem ismerjük. Megmérjük 5 db cementeszsák tömegét és ezekre a következő értékeket kapjuk kg-ban:

$$49,7 \quad 50,1 \quad 49,8 \quad 50 \quad 49,9$$

Adja meg 95% valószínűséggel, hogy az m várható érték melyik intervallumba esik!

Megoldás: A Student-eloszlást fogjuk használni. Először kiszámítjuk az átlagot és a korrigált tapasztalati szórást:

$$\overline{X} = \frac{49,7 + 50,1 + 49,8 + 50 + 49,9}{5} = 49,9$$

$$(49,7 - 49,9)^2 + (50,1 - 49,9)^2 + (49,8 - 49,9)^2 +$$
$$(50 - 49,9)^2 + (49,9 - 49,9)^2 = 0,1$$

$$(S^*)^2 = \frac{0,1}{4} = 0,025$$

Ezért

$$S^* = 0,158$$

Tudjuk, hogy $n = 5$, $1 - \epsilon = 0,95$, azaz $\epsilon = 0,05$. A táblázatból látjuk, hogy

$$t_{0,025}(n-1) = t_{0,025}(4) = 2,77,$$

innen a konfidenciaintervallum:

$$\left(\overline{X} - \frac{t_{\epsilon/2}(n-1)S_n^*}{\sqrt{n}}, \overline{X} + \frac{t_{\epsilon/2}(n-1)S_n^*}{\sqrt{n}}\right)$$

$$\left(49,9 - \frac{2,77 * 0,158}{\sqrt{5}}; 49,9 + \frac{2,77 * 0,158}{\sqrt{5}}\right)$$

azaz

$$(49,704; 50,086)$$