Kako se lotiš: Statistika

Patrik Žnidaršič

Prevedeno 28. junij 2024

1 Centralni limitni izrek

Izrek (centralni limitni izrek)**.** *Naj bodo $X_1, X_2, \ldots$ neodvisne in enako porazdeljene slučajne spremenljivke s končnim drugim momentom. Označimo $\mu_1 = E(X_1)$, $\sigma_1^2 = \operatorname{var}(X_1)$ in $S_n = X_1 + \cdots + X_n$. Za $W_n = \frac{S_n - \mu_1 n}{\sigma_1 \sqrt{n}}$ potem velja*

$$\lim_{n\to\infty} P(W_n \leq w) = \phi(w)$$

enakomerno za $w \in \mathbb{R}$, torej

$$\lim_{n\to\infty} \sup_{w\in\mathbb{R}} |P(W_n \leq w) - \phi(w)| = 0.$$

Izrek lahko uporabimo za ocenjevanje porazdelitve vsote veliko IID slučajnih spremenljivk. Izračunamo $\mu_1 = E(X_1)$ in $\sigma_1^2 = \operatorname{var}(X_1)$ in ocenimo verjetnost kot

$$P(S_n \leq x) = \phi\left(\frac{x - n\mu_1}{\sigma_1\sqrt{n}}\right).$$

Za x vzamemo mejo, ki nas zanima, pri čemer za vsote z vrednostmi na mreži $a\mathbb{Z} + b$ za $a, b \in \mathbb{N}$ po dogovoru vzamemo srednjo vrednost; torej za ocenjevanje verjetnosti, da je padlo manj kot M pik, npr. vzamemo $x = M - \frac{1}{2}$, za več kot M pa $x = M + \frac{1}{2}$.

Povemo lahko tudi nekaj o napaki te ocene. Če so $X_1, X_2, \ldots$ neodvisne, za $\mu_n = E(S_n)$ in $\sigma_n^2 = \operatorname{var}(S_n)$ velja

$$\sup_{x\in\mathbb{R}} \left| P(S_n \leq x) - \phi\left(\frac{x - \mu_n}{\sigma_n}\right) \right| \leq \frac{0.5583}{\sigma_n^3} \sum_{k=1}^{n} E\Big(|X_k - E(X_k)|^3\Big).$$

Pri tem nismo predpostavili, da so spremenljivke enako porazdeljene, torej lahko to oceno uporabimo v več primerih. Če npr. pokažemo, da desna stran konvergira k 0 za $n \to \infty$, lahko pokažemo rezultat CLI tudi za različno porazdeljene slučajne spremenljivke.

Za računanje ϕ glej tabelo, in se spomni $\phi(-x) = 1 - \phi(x)$.

2 Konvergenca slučajnih spremenljivk

Zaporedje $X_1, X_2, \ldots$ konvergira proti X

- ŠIBKO, če je za vsako zvezno in omejeno h
$$\lim_{n\to\infty} E(h(X_n)) = E(h(X)),$$
- V VERJETNOSTI, če je za vsak $\varepsilon > 0$
$$\lim_{n\to\infty} P(d(X_n, X) > \varepsilon) = 0,$$
- SKORAJ GOTOVO, če je
$$P\Big(\{\lim_{n\to\infty} X_n = X\}\Big) = 1.$$

Tretja točka implicira drugo, druga pa prvo. Če imajo spremenljivke vrednosti v $\mathbb{R}$, je prva točka ekvivalentna pogoju, da

$$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$$

za vsak x z $P(X = x) = 0$.

Glede konvergence slučajnih spremenljivk lahko povemo marsikaj. Vzemimo take slučajne spremenljivke $X_1, X_2, X_3, \ldots, X$, da velja $X_n \xrightarrow[n\to\infty]{d} X$. Če je g zvezna funkcija, velja tudi $g(X_n) \xrightarrow[n\to\infty]{d} g(X)$.

Če so X_i neodvisne in enako porazdeljene z $E(X_i) = \mu$, velja ZAKON VELIKIH ŠTEVIL

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{d} \mu.$$

Za konvergenco parov lahko povemo manj kot bi pričakovali; če $X_n \xrightarrow[n\to\infty]{d} X$ in $Y_n \xrightarrow[n\to\infty]{d} c$, kjer je c konstanta, potem konvergirajo tudi pari $(X_n, Y_n) \xrightarrow[n\to\infty]{d} (X, c)$. Kot posledico dobimo IZREKE SLUCKEGA, ki pravijo, da $X_n + Y_n \xrightarrow[n\to\infty]{d} X + c$ in $X_nY_n \xrightarrow[n\to\infty]{d} cX$. Podobno velja tudi za deljenje, kjer moramo izrek formulirati malce drugače:

Trditev. *Naj bodo $X_1, X_2, \ldots, X, Y_1, Y_2, \ldots$ in c kot prej. Privzamemo še $c \neq 0$.*

- *Če so $Z_1, Z_2, \ldots$ taki slučajni vektorji, da je $Z_n = {}^{X_n}/_{Y_n}$ za $Y_n \neq 0$, potem gre $Z_n \xrightarrow[n\to\infty]{d} {}^{X}/_{c}$.*
- *Za vsak $a \in \mathbb{R}$, za katerega je $P({}^{X}/_{c} = a) = 0$, velja*
$$\lim_{n\to\infty} P\left(Y_n \neq 0, \frac{X_n}{Y_n} \leq a\right) = P\left(\frac{X}{c} \leq a\right)$$
ter
$$\lim_{n\to\infty} P\left(Y_n \neq 0, \frac{X_n}{Y_n} \geq a\right) = P\left(\frac{X}{c} \geq a\right).$$

3 Cenilke

Recimo, da imamo model nekega dogajanja in želimo preveriti, če drži vodo. Karakteristika modela y je neka lastnost (parameter, ...), ki je za ta model značilna, mi pa je ne poznamo. Iz opažanj lahko ocenimo vrednost te karakteristike $\hat{y}$, oceni pravimo CENILKA. Za cenilko definiramo PRIČAKOVANO ali SREDNJO KVADRATIČNO NAPAKO

$$\mathrm{MSE}(\hat{y} \mid y) = E((y - \hat{y})^2)$$

($\hat{y}$ je slučajna spremenljivka, pisana z malo). Poleg tega definiramo PRISTRANSKOST

$$\mathrm{Bias}(\hat{y} \mid y) = E(\hat{y}) - y.$$

Cenilka je NEPRISTRANSKA, če je $\mathrm{Bias}(\hat{y} \mid y) = 0$. Pri takih cenilkah je MSE enaka varianci, in lahko definiramo

$$\mathrm{SE}(\hat{y}) = \mathrm{RMSE}(\hat{y} \mid y) = \sqrt{\mathrm{var}(\hat{y})},$$

v splošnem pa dobimo

$$\mathrm{var}(\hat{y}) = \mathrm{MSE}(\hat{y} \mid y) - \left(\mathrm{Bias}(\hat{y} \mid y)\right)^2.$$

Če podatke dobivamo postopoma, dobimo zaporedje cenilk $(\hat{y}_i)_i$. Za tako zaporedje pravimo, da je ŠIBKO DOSLEDNO, če

$$\hat{y}_n \xrightarrow[n\to\infty]{d} y.$$

Zadosten pogoj je, da srednja kvadratna napaka konvergira k 0, čemur pravimo DOSLEDNOST.

Pogosta tema je iskanje NAJBOLJŠE NEPRISTRANSKE LINEARNE CENILKE (NNLC). To je preprosto cenilka, ki ima izmed vseh linearnih cenilk najmanjšo srednjo kvadratno napako. Če so $X_1, \ldots, X_n$ nekorelirane z enakimi pričakovanimi vrednostmi in variancami, je NNLC kar povprečje.

4 Slučajni vektorji

Za slučajni vektor $\underline{X}$ definiramo PRIČAKOVANO VREDNOST

$$E(\underline{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix},$$

za par $\underline{X}, \underline{Y}$ pa KOVARIANČNO MATRIKO

$$\mathrm{Cov}(\underline{X}, \underline{Y}) = \begin{bmatrix} \mathrm{cov}(X_1, Y_1) & \cdots & \mathrm{cov}(X_1, Y_n) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, Y_1) & \cdots & \mathrm{cov}(X_n, Y_n) \end{bmatrix} = E(\underline{X}\underline{Y}^T) - E(\underline{X})E(\underline{Y})^T.$$

Za deterministično matriko A je $E(A\underline{X}) = AE(\underline{X})$, podobno za slučajno matriko M velja $E(AM) = AE(M)$. Analogni enakosti dobimo za množenje z desne. Za deterministični matriki A, B in slučajna vektorja $\underline{X}, \underline{Y}$ je

$$\operatorname{Cov}(A\underline{X}, B\underline{Y}) = A \operatorname{Cov}(\underline{X}, \underline{Y}) B^T.$$

Dva omembe vredna trika pri obračanju slučajnih matrik sta s sledjo. Ker je sled linearna preslikava, se lepo obnaša s pričakovano vrednostjo, in je $E(\operatorname{sl}(M)) = \operatorname{sl}(E(M))$. Spomnimo se, da je skalar enak svoji sledi, da je sled linearna, $\operatorname{sl}(A + B) = \operatorname{sl} A + \operatorname{sl} B$, in da lahko ciklično zamenjujemo argumente:

$$\operatorname{sl}(ABC) = \operatorname{sl}(BCA) = \operatorname{sl}(CAB).$$

5 Pridobivanje cenilk

Pri enostavnem slučajnem vzorčenju imamo populacijo velikosti N in vzorec velikosti n. Iz populacije izberemo enote $K_1, \ldots, K_n$, $K_i \neq K_j$, tako, da so vse n-terice enako verjetne. Označimo $X_i = X_{K_i}$. Imamo nepristranske cenilke

$$\hat{\mu} = \frac{1}{n}(X_1 + \ldots + X_n),$$

za $S = \sum_i (X_i - \hat{\mu})^2$

$$\widehat{\sigma^2} = \frac{N-1}{N} \frac{S}{n-1},$$
$$\widehat{\mu^2} = \hat{\mu}^2 \frac{N-n}{N} \frac{S}{n(n-1)},$$

in napaka

$$\widehat{\mathrm{SE}^2} = \frac{N-n}{N-1} \frac{\widehat{\sigma^2}}{n}.$$

Pri stratificiranem vzorčenju je populacija razdeljena na k stratumov z velikostmi $N_1, \ldots, N_k$. Označimo $w_i = N_i/N$. Na vsakem stratumu vzorčimo enostavno slučajno.

Poznamo dve pogosti metodi za pridobivanje cenilk. Prva je metoda momentov, kjer cenilke izpeljemo iz predpisov za momente (potrebujemo toliko momentov, kolikor je argumentov v porazdelitvi). Potem poiščemo inverzno preslikavo, s katero izrazimo cenilke z momenti;

$$(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m) = F^{-1}(\overline{X}, \overline{X^2}, \ldots, \overline{X^m}).$$

Druga metoda je metoda največjega verjetja. Pri tej metodi iz opažanja $X = (X_1, \ldots, X_n)$ in gostote $f_X(x \mid \theta)$ pridobimo cenilko za θ kot

$$\hat{\theta} = \arg\max L(\theta \mid x)$$

kjer je $L(\theta \,|\, x) = f_X(x \,|\, \theta)$ verjetje. V diskretnem primeru zamenjamo f_X z verjetnostjo $P(X = x \,|\, \theta)$. Pogosto se splača namesto verjetja maksimizirati njegov logaritem, torej rešiti $\partial_\theta l = \partial_\theta \log L = 0$.

Ko imamo cenilko po MNV, morda tudi želimo oceniti, za koliko smo se zmotili. To lahko naredimo s pomočjo Fisherjeve informacije

$$\mathrm{FI}(\theta) = E\left(\left(\frac{\partial l(\theta(X))}{\partial \theta}\right)^2\right) = -E\left(\frac{\partial^2 l(\theta(X))}{\partial \theta^2}\right).$$

Če je cenilka $\hat{\theta}_{\mathrm{MNV}}$ nepristranska, namreč velja Rao-Cramerjeva ocena

$$\mathrm{SE}^2(\hat{\theta}) = \operatorname{var} \theta \geq \frac{1}{\mathrm{FI}(\theta)}.$$

Če so opažanja neodvisna in enako porazdeljena, lahko izračunamo Fisherjevo informacijo enega opažanja

$$\mathrm{FI}_1(\theta) = E\left(\left(\frac{\partial l_1(\theta(X))}{\partial \theta}\right)^2\right) = -E\left(\frac{\partial^2 l_1(\theta(X))}{\partial \theta^2}\right),$$

potem je

$$\mathrm{FI}(\theta) = n\,\mathrm{FI}_1(\theta).$$

Če je model dovolj regularen (kar predpostavljamo, da vedno je), potem je cenilka po MNV asimptotsko nepristranska, in za $n \to \infty$ velja

$$\mathrm{SE}^2(\hat{\theta}) \sim \frac{1}{n\,\mathrm{FI}_1(\theta)}.$$

Če je model večparametričen, torej če je $\underline{\theta}$ vektor, je Fisherjeva informacija matrika z elementi

$$[\mathrm{FI}(\theta)]_{ij} = E\left(\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right) = -E\left(\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right).$$

Če je $\xi = h(\theta)$ neka karakteristika modela in $\hat{\xi} = h(\hat{\theta})$ nepristranska cenilka za ξ, velja

$$\operatorname{var}(\xi) \geq \vec{\nabla}.h(\theta)^T\,\mathrm{FI}^{-1}(\theta)\vec{\nabla}.h(\theta).$$

Tudi tu lahko računamo po opažanjih, če so le ta IID.

6 Večrazsežna normalna porazdelitev

Standardna p-razsežna normalna porazdelitev je porazdelitev vektorja $Z = (Z_1, \ldots, Z_p)$, kjer so $Z_i \sim N(0,1)$ neodvisne. Potem je splošna n-razsežna normalna porazdelitev vektorja $X = AZ + \mu$, kjer sta $\mu \in \mathbb{R}^n$ in $A \in \mathbb{R}^{n \times p}$ deterministični. Pišemo $X \sim N(b, \Sigma)$, kjer je b pričakovana vrednost in Σ variančna matrika.

Če sta $X_1 \sim N(\mu_1, \Sigma_1)$ in $X_2 \sim N(\mu_2, \Sigma_2)$ neodvisna, je vektor

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \right).$$

Velja tudi obratno. Komponente normalnega vektorja so neodvisne natanko tedaj, ko so nekorelirane.

Poznamo tudi porazdelitev $\chi^2(p)$, kar je porazdelitev $Z_1^2 + \cdots + Z_p^2$, kjer so $Z_i \sim N(0,1)$ neodvisni. Velja $\chi^2(p) = \Gamma(\frac{p}{2}, \frac{1}{2})$.

7 Intervali zaupanja

Če imamo cenilko $\hat{\theta}$ parametra porazdelitve, želimo poiskati zgornjo in spodnjo mejo $(L(\hat{\theta}), U(\hat{\theta}))$ za θ, da bo parameter v tem intervalu verjetnostjo $1 - \alpha$. Običajno poiščemo cenilko, ki jo lahko z determinističnimi operacijami transformiramo v slučajni vektor $W = g(\hat{\theta}(X))$, katerega porazdelitev poznamo vsaj asimptotično. Potem poiščemo meji c_1, c_2, da bo $P(c_1 \leq W) = 1 - \frac{\alpha}{2}$ in $P(W \leq c_2) \leq 1 - \frac{\alpha}{2}$ (če je zahtevan najmanjši interval zaupanja, se morda drugače odločiš za desno stran). Kumulativna porazdelitvena funkcija F_W nam potem da c_1 in c_2, nato pa naš interval transformiramo nazaj v

$$g^{-1}(F_W^{-1}(\frac{\alpha}{2})) \leq \theta \leq g^{-1}(F_W^{-1}(1 - \frac{\alpha}{2})).$$

Za to je dobro poznati porazdelitve raznih stvari. Vedno si lahko pomagaš s CLI, ki ti bo dal, da je W porazdeljen normalno. Če poznaš pričakovano vrednost in varianco, je to to, sicer pa moraš malo kreativno obračati stvari.

Če so $X_1, \ldots, X_n$ neodvisni in porazdeljeni $N(\mu, \sigma^2)$, potem je

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1).$$

Pogosto ne poznaš σ, tedaj lahko uporabiš dejstvo

$$\frac{\overline{X} - \mu}{\hat{\mathrm{SE}}_+} \sim \mathrm{Student}(n-1)$$

za cenilko

$$\hat{\mathrm{SE}}_+ = \sqrt{\frac{1}{n-1} \sum_i (X_i - \overline{X})^2}.$$

Poleg tega velja

$$n \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n)$$

za običajen $\hat{\sigma}$ (po MNV). Če vzameš $\hat{\sigma}_+$, potem dobiš

$$(n-1)\frac{\hat{\sigma}_+^2}{\sigma^2} \sim \chi^2(n-1).$$

Uporabiš lahko tudi dejstvo, da je cenilka $\hat{\theta}$, dobljena po MNV, pri dovolj lepih modelih in dovolj velikih n porazdeljena približno $N(\theta, \hat{\mathrm{SE}}^2)$, kjer je $\hat{\mathrm{SE}}^2 = \mathrm{FI}(\hat{\theta})^{-1}$.

8 Preizkusi domnev

Dano imamo neko domnevo H_0, za katero želimo preizkusiti veljavnost. Poleg tega imamo alternativno domnevo H_1, ki bo veljala, če H_0 završemo; nikoli pa H_0 ne sprejmemo. Preizkus je postopek odločanja, in ima stopnjo tveganja α. Preizkus vedno naredimo tako, da bo za vsako verjetnostno mero θ, ki ustreza H_0, velja $P_\theta(H_0 \text{ zavrnemo}) \leq \alpha$. Gledamo lahko tudi moč preizkusa $P_{H_1}(H_0 \text{ zavrnemo}) =: 1-\beta$. Količinam α in β pravimo napaka prve oz. druge vrste.

Pri konstrukciji preizkusa običajno vzamemo neko cenilko parametra θ, katere porazdelitev poznamo pod predpostavko, da H_0 drži (ali pa $g(\theta)$ za nek g, če poznamo porazdelitev tega). To je enak postopek kot pri iskanju intervalov zaupanja, z eno malo razliko; če je H_1 le enostranska negacija (npr. $H_0 : \theta = \theta_0$, $H_1 : \theta < \theta_0$), potem poiščemo enostranski interval zaupanja.

Na voljo imamo tudi preizkus na podlagi razmerja verjetij, ki deluje drugače. Definiramo

$$\Lambda = \frac{\sup_{\theta \in I} L(\theta \mid X)}{\sup_{\theta \in I_0} L(\theta \mid X)}$$

in domnevo zavrnemo, če je $\Lambda \gg 1$. Podobno lahko gledamo Λ', ki je definirana enako, le da je supremum v števcu po I_1, ne po I. Če znamo izračunati porazdelitev Λ (ali Λ'), potem preprosto določimo tak c, da je $P(\Lambda > c) = \alpha$, in zavrnemo, če je $\Lambda > c$. Če pa ne poznamo porazdelitve, pa si lahko pomagamo z Wilksovim izrekom:

$$2 \ln \Lambda \xrightarrow[n\to\infty]{d} \chi^2(\dim I - \dim I_0),$$

kjer gledamo dimenzijo v smislu mnogoterosti. Zavrnemo, če je $2\ln\Lambda > F^{-1}(1-\alpha)$, kjer je F CDF za $\chi^2(\dim I - \dim I_0)$.

Če H_0 pravi, da vzorec prihaja iz multinomske porazdelitve za neke koeficiente $(p_1, \ldots, p_n) \in I_0$, potem lahko uporabimo Pearsonov χ^2-test. Označimo s T_i število pojavitev i-tega razreda, N velikost vzorca in $E_i = Np_i$. Pearsonova statistika pravi

$$T = \sum_{i=1}^{n} \frac{(T_i - E_i)^2}{E_i} \sim \chi^2(\dim I - \dim I_0)$$

(tu $\sim$ pomeni, da je porazdelitev približno enaka, ne enaka). Zavrnemo, če je T prevelik.

Poseben primer Pearsonovega testa je preizkus neodvisnosti; če imamo tabelo $[T_{ij}]_{ij}$ pojavitev v ij-tem razredu, potem za $E_{ij} = Np_{i,\cdot}p_{\cdot,j}$, kjer je $p_{i,\cdot} = \frac{T_{i,\cdot}}{N}$, pika pa pomeni vsoto po tej komponenti, dobimo

$$T = \sum_{ij} \frac{(T_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((r-1)(c-1)).$$

Porazdelitev je spet približno enaka, r označuje število vrstic, c pa število stolpcev. Spet zavrnemo, je če T prevelik.

9 Linearna regresija

Uporabljamo linearen model $Y = X\beta + \varepsilon$, kjer je Y slučajen n-razsežni vektor, X poznana deterministična $n \times p$ matrika, β determinističen, a nepoznan p-razsežen vektor, in ε šum z $E(\varepsilon) = 0$ in $\mathrm{var}(\varepsilon) = \sigma^2 I_n$. Cilj je poiskati β.

Po izreku Gauss-Markova je $\hat{\beta} = (X^T X)^{-1} X^T Y$ NNLC za β, v primeru $\varepsilon \sim N(0, \sigma^2 I_n)$ pa je tudi cenilka po MNV. Za ta $\hat{\beta}$ velja $\mathrm{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. Če potrebuješ oceno variance, lahko σ nadomestiš z nepristransko cenilko

$$\hat{\sigma}^2 = \frac{1}{n-p} \left\| Y - X\hat{\beta} \right\|^2,$$

v primeru normalnega modela je potem $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$ in

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

V primeru $p > 1$ je za konstanten $c \in \mathbb{R}^p$ NNLC za $c^T\beta$ kar enaka $c^T\hat{\beta}$. Potem je za $\hat{\mathrm{SE}}(c^T\hat{\beta}) = \hat{\sigma}\sqrt{c^T (X^T X)^{-1} c}$ in normalni model

$$\frac{c^T\hat{\beta} - c^T\beta}{\hat{\mathrm{SE}}(c^T\hat{\beta})} \sim \mathrm{Student}(n-p).$$