

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \rightarrow E(\underline{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix} \rightarrow \text{cov}(\underline{X}) = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{var}(X_n) \end{pmatrix}$$

kovariančna matrika (simetrična, pozitivno semi-definitna)

$$\underline{u} \otimes \underline{v} = \underline{u} \underline{v}^T$$

$$\underline{X} \underline{X}^T = \begin{pmatrix} X_1^2 & X_1 X_2 & \cdots & X_1 X_n \\ X_2 X_1 & X_2^2 & \cdots & X_2 X_n \\ \vdots & \vdots & \ddots & \vdots \\ X_n X_1 & X_n X_2 & \cdots & X_n^2 \end{pmatrix}$$

$$\text{cov}(\underline{X}) = E[(\underline{X} - E(\underline{X}))(\underline{X} - E(\underline{X}))^T] = E(\underline{X} \underline{X}^T) - E(\underline{X})E(\underline{X})^T$$

$$\text{korelacijski koeficient: } \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1) \text{var}(X_2)}}$$

Če je A deterministična matrika (konstantna), velja: $E(A\underline{X}) = AE(\underline{X})$, $\text{cov}(A\underline{X}) = A \text{cov}(\underline{X}) A^T$

$$\text{cov}(\langle \underline{X}, \underline{u} \rangle, \langle \underline{X}, \underline{v} \rangle) = \text{cov}(\underline{X} \underline{u}, \underline{X} \underline{v}), \text{cov}(\underline{u}^T \underline{X}, \underline{v}^T \underline{X}) = \underline{v}^T \text{cov}(\underline{X}) \underline{u}$$

Standardna p -razsežna normalna porazdelitev je porazdelitev slučajnega vektorja (Z_1, Z_2, \dots, Z_n) ,

kjer so $Z_1, \dots, Z_p \sim N(0, 1)$ in neodvisne.

Če je Q ortogonalna matrika in \underline{Z} standarden normalen vektor, potem je $\underline{W} = Q\underline{Z}$ tudi standarden normalen.

Splošna n -razsežna normalna porazdelitev je vsaka porazdelitev slučajnega vektorja $\underline{W} = A\underline{Z} + \underline{u}$, kjer je \underline{Z} standarden p -razsežni normalni vektor, A matrika $n \times p$ polnega ranga in $\underline{u} \in \mathbb{R}^n$.

$$E(\underline{Z}) = 0, \text{cov}(\underline{Z}) = I, E(\underline{W}) = \underline{u}, \text{cov}(\underline{Z}) = AA^T$$

Če $A \in \mathbb{R}^{n \times p}$ polnega ranga, je AA^T polnega ranga (in obrnljiva).

$$\sigma > 0, X \sim N(\mu, \sigma^2) \implies P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$X_1, \dots, X_n \sim N(\mu, \sigma^2), \text{potem } X_1 + \dots + X_n \sim N(n\mu, \frac{\sigma^2}{n}), \bar{X} \sim N(\mu, \sigma^2), \bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$$

$$\text{Pogojna gostota: } f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \implies X_2|X_1 \sim N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

$$\|\underline{X}\|^2 = \underline{X}^T \underline{X} = \text{sl}(\underline{X} \underline{X}^T)$$

Če poznamo porazdelitev slučajne spremenljivke Y in $f_{X|Y}$, potem velja $f_X(x) = E[f_{X|Y}(x)]$.

Pogojne pričakovane vrednosti: $E(X) = E[E(X|Y)]$, $E[Xg(Y)|Y] = E(X|Y)g(X)$, v abstraktnem smislu definiramo $E[Y|X]$ kot funkcijo $\Psi(x)$, za katero za vsako omejeno zvezno funkcijo g velja $E[Yg(x)] = E[\Psi(x)g(x)]$.

$$\text{cov}(X, Y) = \text{cov}(E(X|Z), E(Y|Z)) + E(\text{cov}(X, Y|Z)), \text{med drugim } \text{var}(X) = \text{var}(E(X|Z)) + E(\text{var}(X|Z))$$

Če so X_1, \dots, X_n neodvisne med seboj in tudi od Y_1, \dots, Y_n , potem so X_1, \dots, X_n neodvisne tudi pogojno na Y .

CENTRALNI LIMITNI IZREK

Izrek: Naj bodo X_1, X_2, \dots neodvisne, enako porazdeljene z $E(X_i^2) < \infty$ in $E(X_i) = \mu_1$ ter $\text{var}(X_i) = \sigma_1^2$. $S_n = X_1 + X_2 + \dots + X_n$. Tedaj:

$$\frac{S_n - n\mu_1}{\sigma_1 \sqrt{n}} \xrightarrow[n \rightarrow \infty]{\text{šibko}} N(0, 1),$$

kjer $n\mu_1 = E(S_n)$ in $\sigma_1 \sqrt{n} = \sigma(S_n)$.

Bolj ohlapno: n velik $\implies S_n \sim N(n\mu_1, n\sigma^2)$

$$P(a \leq S_n \leq b) \approx \Phi\left(\frac{b-n\mu_1}{\sigma_1 \sqrt{n}}\right) - \Phi\left(\frac{a-n\mu_1}{\sigma_1 \sqrt{n}}\right)$$

Če slučajna spremenljivka živi v celih številih, lahko namesto \leq vzamemo $<$ in mejo povečamo za 1, ali pa vzamemo sredino.

$$\text{Natančnost sredine je odvisna od asimetrije, ki jo meri } A(X) = \frac{E[(X-E(X))^3]}{(\text{var}(X))^{\frac{3}{2}}}.$$

Naj bodo X_1, \dots, X_n neodvisne in identično porazdeljene, $\sigma_1 = \sqrt{\text{var}(X_i)}$, $\gamma_1 = E[|X_i - E(X_i)|^3]^{\frac{1}{3}}$, $S_n = X_1 + \dots + X_n$. Ko $n \rightarrow \infty$, $P(a_n \leq S_n \leq b_n)$ aproksimiramo z ustreznimi normalnimi. Zadosten pogoj, da gre:

- absolutna napaka $\rightarrow 0$: $n \gg \frac{\gamma_1^6}{\sigma_1^6}$
- relativna napaka $\rightarrow 0$: $n \gg \frac{\gamma_1^6}{\sigma_1^6}$ in $\min\{|a_n - E(S_n)|, |b_n - E(S_n)|\} \ll \frac{n^{\frac{2}{3}} \sigma_1^2}{\gamma_1}$

$$\mu_1 = E(X_i), \sup_{x \in \mathbb{R}} |P(S_n \leq x) - \Phi\left(\frac{x-n\mu_1}{\sigma_1 \sqrt{n}}\right)| \leq \frac{0.4774}{\sqrt{n}} \frac{\gamma_1^3}{\sigma_1^3}$$

Porazdelitev χ^2 :

Če so Z_1, \dots, Z_n neodvisne standardno normalne, potem je $Z_1^2 + \dots + Z_n^2 \sim \chi^2(n)$, $\chi^2(n) \xrightarrow[n \rightarrow \infty]{} N(0, 1)$.

$$\chi^2(n) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

Če $U \sim \Gamma(a, \lambda)$ in $V \sim \Gamma(b, \lambda)$, potem $U + V \sim \Gamma(a + b, \lambda)$

Če $U_1, \dots, U_m \sim \Gamma\left(\frac{n}{2m}, \frac{1}{2}\right)$ neodvisne, potem $U_1 + \dots + U_m \sim \chi^2(n)$.

Razmerje Ljapunova:

$S = X_1 + \dots + X_n$, $\mu = E(S)$, $\sigma^2 = \text{var}(S)$, X_1, \dots, X_n neodvisne. $P(a \leq S \leq b) \approx \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$

$\sup_{x \in \mathbb{R}} |P(S \leq x) - \Phi(\frac{x-\mu}{\sigma})| \leq \frac{0.5591}{\sigma^3} \sum_{k=1}^n E[|X_k - E(X_k)|^3]$

Če desna stran konvergira k 0, imamo konvergenco k $N(0, 1)$.

DEJANSKA STATISTIKA – VZROČENJE

\hat{a} je nepristranska cenilka za a , če je $E(\hat{a}) = a$, srednja kvadratična napaka: $q(\hat{a}) = E[(\hat{a} - a)^2]$, standardna napaka: $\sqrt{q(\hat{a})} = se(\hat{a})$.

Slučajne spremenljivke X_1, \dots, X_n so izmenljive, če velja: $(X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(n)}) \stackrel{d}{=} (X_1, X_2, \dots, X_n) \quad \forall \pi \in S_n$. Za izmenljive sl. spr. X_1, \dots, X_n s pričakovano vrednosti $E(X_i) = \mu$, varianco $\text{var}(X_i) = \sigma^2$, korelacijo $\text{corr}(X_i, X_j) = \rho$, za $i \neq j$ je vzorčno povprečje $\bar{X} = \frac{X_1 + \dots + X_n}{n}$ nepristranska cenilka za μ , $\text{var}(\bar{X}) = \frac{\sigma^2}{n}(1 + \rho(n-1))$, nepristranska cenilka za σ^2 pa je $\hat{\sigma}^2 = \frac{1}{(n-1)(1-\rho)} \sum_{i=1}^n (X_i - \bar{X})^2$

Enostavno slučajno vzorčenje

Populacija: $1, 2, \dots, N$, vzorec: K_1, K_2, \dots, K_n . Vrednosti spremenljivk na populaciji x_1, x_2, \dots, x_N (ne poznamo vseh). Poznamo vrednosti na vzorcu: $X_i = x_{K_i}$ (izmenljive, ker je vsaka n -terica enako verjetna)

Stratificirano vzorčenje:

Populacijo velikosti N razdelimo na k stratumov velikosti N_1, \dots, N_k , kjer w_1, w_2, \dots, w_k ($w_i = \frac{N_i}{N}$ in $w_1 + \dots + w_k = 1$) predstavljajo delež populacije v stratumih, μ_1, \dots, μ_k povprečja stratificiranih spremenljivk, $\sigma_1, \dots, \sigma_k$ standardne odklone. Povprečje na celotni populaciji je $\mu = w_1\mu_1 + \dots + w_k\mu_k$.

Varianca na celi populaciji: $\sigma^2 = \sigma_B^2 + \sigma_w^2$, kjer $\sigma_B^2 = \sum_{i=1}^k w_i(\mu_i - \mu)^2$ in $\sigma_w^2 = \sum_{i=1}^k w_i\sigma_i^2$.

Enostavni slučajni vzorci po stratumih:

$X_{11}, \dots, X_{1n_1} \quad \bar{X}_1$
 $X_{21}, \dots, X_{2n_2} \quad \bar{X}_2$
 $\vdots \quad \vdots$, kjer so \bar{X}_i vzorčna povprečja po stratumih. $\bar{X} = \sum_{i=1}^k w_i \bar{X}_i$ je nepristranska cenilka za μ ,
 $X_{k1}, \dots, X_{kn_k} \quad \bar{X}_k$
 $\sigma_i^2 = \frac{1}{n_i-1} \frac{N_i-1}{N_i} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$ nepristranska cenilka za σ_i^2 , $\hat{\sigma}_w^2 = \sum_{i=1}^k \frac{w_i}{n_i-1} \frac{N_i-1}{N_i} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$ nepristranska cenilka za σ_w^2 in $\hat{\sigma}_B^2 = \sum_{i=1}^k w_i(\bar{X}_i - \bar{X})^2 - \sum_{i=1}^k (w_i - w_i^2) \frac{N_i-1}{N_i} \frac{1}{n_i} \hat{\sigma}_i^2$ nepristranska cenilka za σ_B^2 .

INTERVALI ZAUPANJA

a bi radi ocenili: $a_{\min} < a < a_{\max}$, kjer je (a_{\min}, a_{\max}) interval zaupanja. $P(a_{\min} < a < a_{\max}) \geq 1 - \alpha$, kjer je $1 - \alpha$ stopnja zaupanja (95%, 99%) in α stopnja tveganja (5%, 1%).

Določanje IZ: pivotna funkcija $T(\underline{X}, a)$, kjer je \underline{X} opažanje in a ocenjevani parameter.

IZ za μ , kjer je σ znan: $P(|\bar{X} - \mu| < M_\alpha) = 1 - \alpha$, $M_\alpha = \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \frac{\alpha}{2})$

Če tudi σ ne poznamo in če so $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, potem $T(\underline{X}, \mu) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sqrt{n}$, kjer $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Za velike n je $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sqrt{n} \sim N(0, 1)$, v splošnem pa je $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sqrt{n} \sim \text{Student}(n-1)$.

OCENJEVANJE PARAMETROV

Metoda momentov:

Vzorec $= (X_1, \dots, X_n)^T$ je iz neke porazdelitve $f_X(x|\vartheta)$. Nastavimo parametre ϑ , da bo povprečje enako vzorčnemu povprečju, varianca vzorčni varianci. ... Rešimo sistem enačb: $E[X] = \bar{X}$, $\text{var}(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Metoda največjega verjetja: Postavimo se na stališče, da je vzorec $\underline{x} = (x_1, \dots, x_n)^T$ neke porazdelitve $f_X(x|\vartheta)$, ki smo ga dobili verjetno eden izmed bolj verjetnih, in nastavimo parametre ϑ tako, da bo najbolj verjeten. Kako verjeten je vzorec je podano s funkcijo verjetja L . Če so vzorci neodvisni, je $L(\vartheta|\underline{x}) = \prod_{i=1}^n f_X(x_i|\vartheta)$. V diskretnem primeru je to $L(\vartheta|\underline{x}) = \prod_{i=1}^n P_{\vartheta}(X = x_i)$.

Definiramo tudi logaritemsko funkcijo verjetja $\ell(\vartheta|\underline{x}) = \log(L(\vartheta|\underline{x}))$.

Cenilka po metodi največjega verjetja (MLE) je vrednost $\hat{\vartheta}$, kjer je dosežen maksimum L , ali ekvivalentno, ℓ . Cenilka je nepristranska. Ponavadi jo dobimo tako da rešimo sistem $\frac{\partial \ell}{\partial \vartheta}(\vartheta|\underline{x}) = 0$ za ϑ .

Za MLE velja: $\sqrt{n}(\vartheta - \hat{\vartheta}) \sim N(0, I^{-1}(\vartheta))$.

Fisherjeva matrika informacije:

Varianca MLE je dana s Fisherjevo matriko informacije.

Matrike je dana z $I_{kl}(\vartheta) = -E \left[\frac{\partial^2}{\partial \vartheta_k \partial \vartheta_l} (\log(f_X(x|\vartheta))) \right]$, $I(\vartheta) = -E[\text{hesse}(\log(f_X(x|\vartheta)))]$.

Velja: $\text{var}(\hat{\vartheta}_k) = \frac{(I^{-1})_{kk}}{n}$, $se(\hat{\vartheta}_k) = \sqrt{\frac{(I^{-1})_{kk}}{n}}$.

Za primer, ko je ϑ skalar, se formule poenostavijo: $I(\vartheta) = E[\frac{\partial^2}{\partial \vartheta^2} \log(f_X(x|\vartheta))]$,

$se(\hat{\vartheta}) = \frac{1}{\sqrt{nI(\vartheta)}}$. Aproximativni interval zaupanja: $\vartheta \in \left(\hat{\vartheta} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{nI(\hat{\vartheta})}} \right)$.

TESTIRANJE HIPOTEZ

LINEARNA REGRESIJA

Predpostavljamo, da so opaženi slučajni podatki $\underline{Y} = (Y_1, \dots, Y_n)^T$ nastali kot $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, kjer je X znana deterministična $n \times m$ matrika, $\underline{\beta}$ neznan determinističen m vektor, $\underline{\varepsilon}$ neznan slučajen vektor napak, za katerega predpostavimo standardni regresijski model, ki pravi: $E[\underline{\varepsilon}] = 0$ in $\text{var}(\underline{\varepsilon}) = \sigma^2 I$. Drugače povedano, napake ε_i so nekorelirane, varianca vsake posamezne pa je σ^2 .

Za $\underline{\beta} = (\alpha, \beta)^T$ in $X = (1, x_1; 1, x_2; \dots; 1, x_n)$ preidemo na standardno ocenjevanje s premico.

Nepistranska cenilka za $\underline{\beta}$ je cenilka po metodi najmanjših kvadratov $\hat{\underline{\beta}} = (X^T X)^{-1} X^T Y$.

Varianca cenilke: $\text{var}(\underline{\hat{\beta}}) = \sigma^2 (X^T X)^{-1} = \sigma^2 C$.

Nepistranska cenilka za σ^2 je $\hat{\sigma}^2 = \frac{1}{n-m} \sum_{i=1}^n \hat{\varepsilon}_i^2$, kjer je $\hat{\varepsilon} = \underline{Y} - X\hat{\underline{\beta}}$.

Če gledamo samo posamezne komponente je cenilka za β_i enaka $\hat{\beta}_i = (\hat{\underline{\beta}})_i$ in $\text{var}(\hat{\beta}_i) = \sigma^2 C_{ii}$, se $(\hat{\beta}_i) = \sigma \sqrt{C_{ii}}$.

Za lin. kombinacijo komponent je cenilka enaka $a^T \hat{\underline{\beta}}$ in njena varianca je $\text{var}(a^T \hat{\underline{\beta}}) = a^T C a$, za poljuben vektor a .

Izrek (Gauss-Markov): Cenilka po metodi najmanjših kvadratov je najboljša med vsemi linearnimi nepristranskimi cenilkami (ima najmanjšo varianco). Za vsako linearno cenilko $\tilde{\underline{\beta}} = L\underline{Y}$ mora veljati $LX = I$ in posledično $\text{var}(\tilde{\underline{\beta}}) = \text{var}(\tilde{\underline{\beta}} - \hat{\underline{\beta}}) + \text{var}(\hat{\underline{\beta}})$, saj je $\text{cov}(\tilde{\underline{\beta}} - \hat{\underline{\beta}}, \hat{\underline{\beta}}) = 0$. Enako velja tudi za cenilko $a^T \tilde{\underline{\beta}}$ za vsak vektor a , v posebnem za $a = e_i$ tudi za cenilke po komponentah.

Če za $\underline{\varepsilon}$ predpostavljamo $\text{var}(\underline{\varepsilon}) = \sigma^2 \Sigma$ za neko pd. matriko Σ , potem to prevedemo na standardni model z množenjem z leve s $(\Sigma)^{-\frac{1}{2}}$. Cenilka za $\underline{\beta}$ postane $\hat{\underline{\beta}} = (X^T \Sigma X)^{-1} X^T \Sigma^{-1} \underline{Y}$.

Za testiranje hipoteze $\beta_i = 0$ proti $\beta_i \neq 0$ uporabimo testno statistiko $t = \frac{\hat{\beta}_i}{\hat{\sigma} \sqrt{C_{ii}}} \sim t_{n-m}$.