Inferencijalna statistika: točkovne i intervalne procjene

ak. god. 2021./2022.

Inferencijalna statistika – uvod

Neka je X slučajna varijabla.

Cilj inferencijalne statistike je na osnovu uzorka izvesti određene zaključke o varijabli X, tj. njezinoj raspodjeli.

Sjetimo se da je slučajni pokus matematički opisan vjerojatnosnim prostorom $(\Omega, \mathcal{F}, \mathbb{P})$, gdje vjerojatnost $\mathbb{P} : \mathcal{F} \to [0, 1]$ u pravilu ne znamo.

Metodama inferencijalne statistike možemo procijeniti vjerojatnost $\mathbb{P}$ jer na nju možemo gledati kao na raspodjelu od X.

Slučajni uzorak i statistika

Neka je X kvantitativna varijabla sa slikom $R(X)$ koja opisuje ishode eksperimenta na elementima populacije.

Slučajni uzorak duljine n iz raspodjele od X je n-dimenzionalni slučajni vektor $(X_1, \ldots, X_n)$ čije su komponente nezavisne i jednako distribuirane slučajne varijable s raspodjelom od X.

Realizaciju slučajnog uzorka $(X_1, \ldots, X_n)$ označavamo s $(x_1, \ldots, x_n)$.
Statistika je funkcija slučajnog uzorka, ona je i sama slučajna varijabla.
Primjerice,

$$\overline{x} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{i} \qquad s = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

su dvije statistike od $(X_1, \ldots, X_n)$.

Procjenitelji parametara

Želimo procijeniti očekivanje $\mu = \mathbb{E}(X)$ i varijancu $\sigma^2 = \mathrm{Var}(X)$ na temelju prikupljenih podataka (informacija koje dobivamo iz realizacije slučajnog uzorka).

Procjenitelj parametra kojeg promatramo je funkcija slučajnog uzorka, dakle statistika. Primjerice, procjenitelji za μ i σ^2 su, redom, $\overline{X}_n$ i S_n^2.

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{i} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Dobri procjenitelji su **nepristrani**. Konkretno, ovdje to znači da vrijedi $\mathbb{E}(\overline{X}_n) = \mu$ i $\mathbb{E}(S_n^2) = \sigma^2$.

Dva pristupa u procjeni parametara

U procjeni parametara μ i σ^2 od X imamo dva pristupa:

(a) točkovne procjene

(b) intervalne procjene.

Na idućem predavanju u tu svrhu diskutirat ćemo i postupak testiranja statističkih hipoteza, što je općenitiji pristup koji ne služi samo za procjenu parametara raspodjele.

Točkovne procjene

Definicija

Točkovna procjena je procjena kod koje se vrijednost (realizacija) procjenitelja uzima kao procjena parametra promatrane raspodjele.

Ranije smo naveli da je $\overline{X}_n$ "dobar" izbor za procjenitelja od μ, a S_n^2 od σ^2.

S vjerojatnošću 1 vrijedi

$$\lim_{n\to\infty} \overline{X}_n = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \mathbb{E}(X) = \mu$$

Dakle, $\overline{X}_n$ konvergira prema očekivanju μ.

Slično, S_n^2 konvergira prema varijanci populacije ($\lim_{n\to\infty} S_n^2 = \sigma^2$).

Primjer

U nekom gradu je 10000 glasača. Ispitivanjem slučajnog uzorka od 1000 njih želimo procijeniti postotak glasača stranke S. U uzorku ima 700 glasača S-a. Odredimo točkovnu procjenu populacijske proporcije (postotka) glasača stranke S.

Uočimo prvo da je na populaciji od 10000 glasača raspodjela slučajne varijable kojom modeliramo glasačku preferenciju prema S-u dana s

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},$$

gdje 0 znači da glasač nije glasao za S, a 1 znači da je glasao za S. Populacijska proporcija glasača stranke S koju trebamo procijeniti je $\mathbb{E}(X) = p$. Kako je realizacija od $X_1 + \cdots + X_{1000}$ jednaka 700, procjena od p je $700/1000 = 0.7$.

Intervalne procjene

t-distribucija

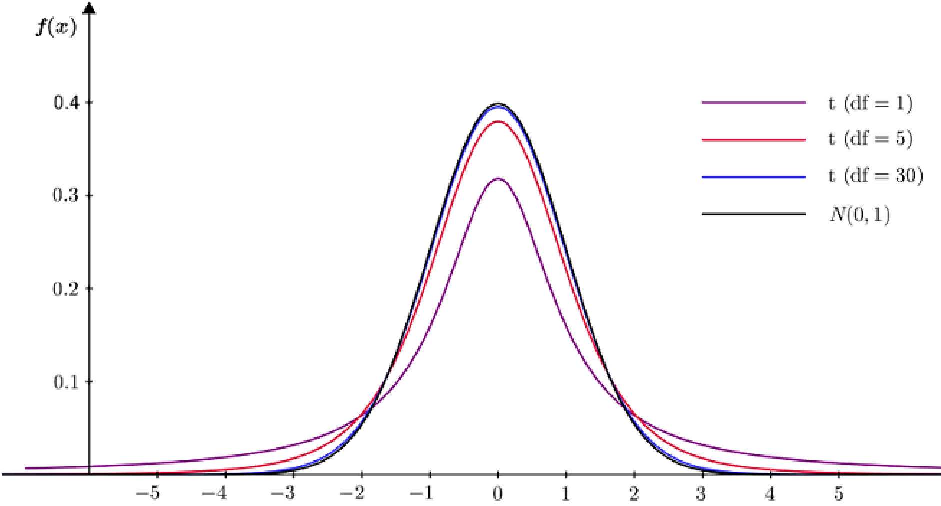
Za neprekidnu slučajnu varijablu X kažemo da ima t-**raspodjelu** (ili **Studentovu** raspodjelu) s n stupnjeva slobode (engl. degrees of freedom), $n \in \mathbb{N}$, u oznaci $X \sim t(n)$, ako je $R(X) = \mathbb{R}$ i funkcija gustoće od X je dana s

$$f(x) = C_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

gdje je C_n konstanta (ovisi samo o n).

Za $n \geqslant 3$ vrijedi $\mathbb{E}(X) = 0$ i $\text{Var}(X) = \dfrac{n}{n-2}$.

t-distribucija za različite stupnjeve slobode



Pojašnjenje stupnjeva slobode

Ako je aritmetička sredina pet vrijednosti 6, četiri od njih možemo odabrati po volji, a peta je onda jedinstveno određena aritmetičkom sredinom. Npr. ako odaberemo brojeve 4, 6, 7 i 10, peta vrijednost mora biti jednaka 3. U ovom primjeru imamo četiri stupnja slobode.

Tipično ćemo u uzorku imati n opažanja i $n-1$ stupanj slobode.

Za niz nezavisnih i jednako distribuiranih slučajnih varijabli $X_1, \ldots, X_n$ s raspodjelom $N(\mu, \sigma^2)$ vrijedi da slučajna varijabla

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{S_n^2}{n}}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n}$$

ima $t(n-1)$ raspodjelu. Stoga se t-raspodjela prirodno javlja u procjeni očekivanja.

Ideja intervalne procjene

Ideja intervalne procjene je konstruirati interval oko točkovne procjene.

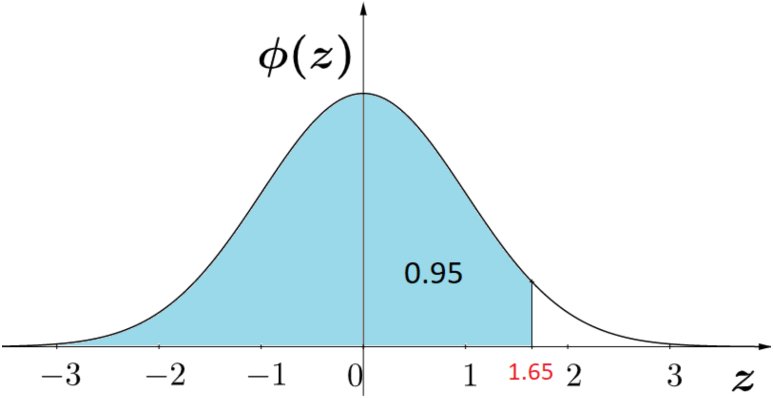
Koeficijent pouzdanosti je vjerojatnost da promatrani interval sadrži vrijednost nepozatog parametra promatrane raspodjele i zato taj interval konstruiramo kao slučajni interval, tj. interval čije su granice slučajne varijable.

Za danu pouzdanost $1-\alpha \in (0,1)$, slučajni uzorak $(X_1, \dots, X_n)$ iz X i statistike $L_n = f(X_1, \dots, X_n)$ i $D_n = g(X_1, \dots, X_n)$, kažemo da je $[L_n, D_n]$ **interval pouzdanosti** (pouzdanosti $1-\alpha$) za parametar τ (μ ili σ^2) ako $\mathbb{P}(L_n \leq \tau \leq D_n) \geqslant 1-\alpha$.

Dakle, za barem $(1-\alpha) \cdot 100\%$ realizacija od $(X_1, \dots, X_n)$ interval pouzdanosti $[L_n, D_n]$ sadrži stvarnu vrijednost τ nepoznatog parametra.

Kvantili

X je slučajna varijabla s funkcijom raspodjele $F(x)$ i neka je $q \in (0, 1)$. Broj $x_q \in \mathbb{R}$ zove se q-**ti kvantil** od X ako vrijedi $F(x_q) = q$.



Slika: 1.65 je 95-postotni kvantil jedinične normalne raspodjele

Intervali pouzdanosti za očekivanje μ

Neka je $(X_1, \ldots, X_n)$ slučajni uzorak iz $N(\mu, \sigma^2)$. Na početku odredimo željenu pouzdanost (najčešće 95%, 99%, a ponekad se koristi 90%). Razlikujemo dva slučaja: **prvi**, u kojem je varijanca populacije σ^2 poznata.

Interval pouzdanosti $1-\alpha$ za μ je dan s

$$\left[\overline{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right],$$

gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $N(0,1)$.

Drugi, češći slučaj je kad varijanca populacije σ^2 nije poznata. Interval pouzdanosti $1-\alpha$ za μ je dan s

$$\left[\overline{X}_n - t_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}\right],$$

gdje je $t_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $t(n-1)$ raspodjele.

Intervali pouzdanosti za očekivanje μ

Ako varijabla koja nas zanima nije normalno distribuirana u populaciji, možemo koristiti iste intervale pouzdanosti. Preporuka je koristiti uzorke veličine barem 30, a u tom slučaju t-distribuciju (s bar 30 stupnjeva slobode) možemo aproksimirati s $N(0,1)$.

Interval pouzdanosti $1-\alpha$ je tada

$$\left[\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right],$$

gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $N(0,1)$.

Primjer intervala pouzdanosti za očekivanje μ

U istraživanju kvalitete boca na uzorku veličine 25 izmjerene su debljine stijenki boca. Uzoračka aritmetička sredina je 4.05 milimetara, a uzoračka standardna devijacija 0.08 milimetara. Poznato je da debljina stijenki boca ima normalnu raspodjelu s nepoznatim očekivanjem i nepoznatom varijancom. Pronađimo 95%-tni pouzdani interval za očekivanu debljinu stijenke.

Budući da nam varijanca populacije nije poznata, tražimo interval oblika

$$\left[\overline{X}_{25} - t_{0.975}\frac{S_{25}}{\sqrt{25}}, \overline{X}_{25} + t_{0.975}\frac{S_{25}}{\sqrt{25}}\right].$$

Pritom, realizacije od $\overline{X}_{25}$ i S_{25} su, redom, $\overline{x}_{25} = 4.05$ i $s_{25} = 0.08$. Naredbom u Excelu iščitavamo da je za $n-1=24$ stupnja slobode $t_{0.975} = 2.06$. Dobivamo interval

$$\left[4.05 - 2.06 \cdot \frac{0.08}{\sqrt{25}}, 4.05 + 2.06 \cdot \frac{0.08}{\sqrt{25}}\right] = [4.01704, 4.08296].$$