



INTERVALI ZAUPANJA

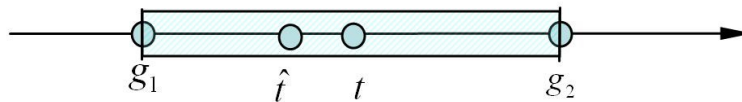
(seminarska naloga pri predmetu STATISTIČNE METODE)

Mentor: izred. prof. dr. Blaž Zmazek
Avtor: Marko Leskovšek
Številka indeksa: 95030615
Študijsko leto: 2007 / 2008

Šentjur, april 2008

INTERVALI ZAUPANJA

Ocene statističnih parametrov delimo na točkaste ocene in intervalne ocene. Natančneje si bomo tukaj ogledali intervalne ocene neznanega statističnega parametra. Za niz podatkov x_1, \dots, x_n je potrebno izbrati interval (g_1, g_2) , ki naj ima lastnost, da z veliko verjetnostjo pokriva neznan parameter t .



Skica intervala ocene parametra

Privzamemo, da je vrednost $\hat{t} \in (g_2 - g_1)$ in velja, da je $\hat{t} \approx t$. Z veliko verjetnostjo lahko jamčimo, da je absolutna napaka pri približevanju neznanega parametra t vrednosti \hat{t} ni večja od $\delta = |g_2 - g_1|$. Meje g_1 in g_2 intervala $\{g_1, g_2\}$ so odvisne od niza podatkov x_1, \dots, x_n , tako da lahko zapišemo:

$$(1) \quad g_1 = h_1(x_1, \dots, x_n), \quad g_2 = h_2(x_1, \dots, x_n),$$

kjer sta h_1 in h_2 izbrani funkciji n realnih spremenljivk, tako da velja:

$$h_1(x_1, \dots, x_n) < h_2(x_1, \dots, x_n), \quad (x_1, \dots, x_n) \in \mathbb{R}^n$$

Niz x_1, \dots, x_n predstavimo kot vrednost naključnega vektorja (X_1, \dots, X_n) , kjer so X_1 in X_2 neodvisne slučajne spremenljivke s skupno verjetnostjo porazdelitve $P_i \in P$, (P je razred dopustnih porazdelitev verjetnosti), tako da lahko g_1 in g_2 razumeta kot vrednosti statistike.

$$(2) \quad G_1 = h_1(x_1, \dots, x_n), \quad G_2 = h_2(x_1, \dots, x_n),$$

Statistike G_1 in G_2 so izbrane naključne spremenljivke, katerima v odvisnosti od parametra t , pripadajo odgovarjajoče porazdelitve verjetnosti. Postavimo zahtevo:

$$(3) \quad P(G_1 < t < G_2) \geq \gamma, \quad t \in \theta,$$

kjer je $\gamma(0 < \gamma < 1)$ izbrana realna vrednost.

Če so meje g_1 in g_2 intervala $\{g_1, g_2\}$ dobljeni kot vrednosti statistike G_1 in G_2 , pri čemer je zadoščeno pogoju (3), tedaj lahko rečemo da je $\{g_1, g_2\}$ interval zaupanja verjetnosti γ za neznan parameter t .

Vrednost γ imenujemo tudi stopnja zaupanja, običajno se za parameter γ privzema vrednost $\gamma=0,95$ oziroma $\gamma=0,99$, govorimo o 95% oziroma 98% zanesljivosti izvedene ocene intervala, odvisno od 95 odstotne ali 98 odstotne stopnje zaupanja.

Če, na primer vzamemo 95% interval zaupanja, lahko pričakujemo da se bo ob večkratnem jemanju naključnega vzorca vsaj v 95% primerih dobili interval zaupanja ki bo vseboval neznan parameter, oziroma bo največ 5% premerov, ko interval zaupanja tega neznanega

parametra ne bo vseboval. Seveda imamo z 99% intervalom zaupanja bistveno večjo verjetnost, da bomo v interval zajeli neznan parameter, vendar bo ta interval širši od 95% intervala zaupanja.

Vprašanje kateri interval (95% ali 99% ali kakšno tretjo vrednost) zaupanja izbrati ni vprašanje, ki bi ga obravnavala statistika oziroma matematika, temveč je izbor povsem stvar ocene in izbora strokovnjaka o vplivu posameznih faktorjev (zanesljivost, velikost napake...itd) na končno odločitev in na praktične posledice posledice.

Teoretično gledano je izbor intervala zaupanja verjetnosti γ za neznan parameter t , ki obstaja v dani množici dopustnih porazdelitev $P = \{P_t : t \in \theta\}$ statistika (2), ki izpolnjuje pogoj (3). V praksi se interval zaupanja izbira tako, da se za vrsto n neodvisnih merjenj x_1, \dots, x_n vrednosti naključne spremenljivke X izračunajo vrednosti g_1 in g_2 in statistike G_1 in G_2 . Statistike G_1 in G_2 niso enoznačno določene glede pogoja (3). V kolikor obstajajo statistike, ki zadovoljujejo pogoj

$$(4) \quad P(G_1 < t < G_2) = \gamma, \quad t \in \theta,$$

potem te statistike zadovoljujejo tudi pogoj (3), očitno pa obstajajo tudi statistike, G_1 in G_2 ki zadovoljujejo pogoj (4), ne pa tudi pogoja (3).

Če statistike G_1 in G_2 zadovoljujejo pogoj (4), dobimo tako imenovani najožji interval zaupanja verjetnosti γ .

Ne obstaja neka univerzalna metoda za iskanje intervala zaupanja, vendar veliko konkretnih primerov rešujemo na naslednji način: Predpostavimo da je \hat{T} izbrana cenika za neznan parameter t , tako je \hat{T} naključna spremenljivka, kateri pripada odgovarjajoča porazdelitev verjetnosti, odvisna od parametra t . Postavimo zahtevo:

$$(5) \quad P(c_1(t) < \hat{T} < c_2(t)) = \gamma, \quad t \in \theta,$$

pri čemer so $t \mapsto c_1(t)$ in $t \mapsto c_2(t)$ monotone funkcije za katere velja, da je

$c_1(t) \leq c_2(t) \quad (t \in \theta)$, sledi da obstajajo odgovarjajoče inverzne funkcije h_1 in h_2 , tako da lahko (5) zapišemo v obliko:

$$(6) \quad P(h_1(\hat{T}) < t < h_2(\hat{T})) = \gamma$$

Predpostavimo da so $h_1(\hat{T})$ in $h_2(\hat{T})$ kot funkcije naključne spremenljivke \hat{T} , prav tako naključne spremenljivke, če primerjamo z (4) in (6) vidimo, da je $G_1 = h_1(\hat{T})$ in $G_2 = h_2(\hat{T})$. S tem je rešeno tudi problem skrajnih točk intervala zaupanja verjetnosti γ za neznan parameter t . Problem ostane v tem, da je potrebno izbrati porazdelitev verjetnostne cenilke \hat{T} in potrebno je najti funkcije c_1 in c_2 , ki izpolnjujejo pogoj (5).

PRIMER 1

Predpostavimo, da je za $P = \{N(t, \sigma^2); t \in \mathbb{R}\}$ potrebno izbrati interval zaupanja verjetnosti γ za neznano verjetnost $\mu = t$ za normalno porazdelitev, pri čemer je znana varianca σ^2 .

V praksi takšne primere srečujemo na primer, kadar proučujemo proces proizvodnje določenega izdelka (vijak, ploščica...itd) na nekem stroju, ki je nastavljen tako, da ima proizvod določeno predpisano dimenzijo (npr dolžino). Zaradi različni vzrokov je dimenzija X naključna spremenljivka, za katero se predpostavlja, da ima normalno porazdelitev $N(\mu, \sigma^2)$, kjer je μ odvisen od nastavitve stroja, medtem ko je σ^2 odvisna od natančnosti in tovarniške kvalitete stroja, kar je znano v naprej. Na podlagi merjenj x_1, \dots, x_n dimenzije X na izdelanih proizvodih, želimo preveriti ali je stroj pravilno nastavljen; oziroma ali se parameter μ nahaja v mejah tolerančnega polja, ki je vnaprej predpisano.

Teoretični del problema lahko rešimo tako: Pojavi se spoznanje, da je

$\hat{T} = \bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$ cenilka na neznan parameter $t \approx \mu$ in da je $\bar{X} \sim N(t, \frac{1}{n}\sigma^2)$. Potrebno je v skladu s (5) najti funkciji c_1 in c_2 .

Predpostavimo da je

$$(7) \quad P\left(t - \lambda \frac{\sigma}{\sqrt{n}} < \bar{X} < t + \lambda \frac{\sigma}{\sqrt{n}}\right) = 2\Phi(\lambda) - 1, \quad \lambda > 0$$

in vzamemo

$$(8) \quad c_1(t) = t - \lambda \frac{\sigma}{\sqrt{n}}, \quad c_2(t) = t + \lambda \frac{\sigma}{\sqrt{n}}$$

vidimo, da so to strogo naraščajoče funkcije parametra $t \in \mathbb{R}$ in lahko (7) zapišemo v obliko:

$$(9) \quad P\left(\bar{X} - \lambda \frac{\sigma}{\sqrt{n}} < t < \bar{X} + \lambda \frac{\sigma}{\sqrt{n}}\right) = 2\Phi(\lambda) - 1.$$

Primerjamo (9) in (6), vidimo, da λ določena iz pogoja $2\Phi(\lambda) - 1 = \lambda$, čemur sledi, da je

$$(10) \quad \lambda = z_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

Vidimo, da so iskane statistike G_1 in G_2 , kot naključne meje intervala zaupanja verjetnosti γ podane s formulami:

$$(11) \quad G_1 = \bar{X} - z_\gamma \frac{\sigma}{\sqrt{n}}, \quad G_2 = \bar{X} + z_\gamma \frac{\sigma}{\sqrt{n}}$$

Opazimo, da so statistike G_1 in G_2 , poleg tega da sta funkciji cenilke \bar{X} , tudi odvisni od verjetnosti γ in velikosti vzorca n , medtem ko je širina intervala zaupanja:

$$(12) \quad \delta = |G_2 - G_1| = 2z_\gamma \frac{\sigma}{\sqrt{n}}$$

odvisna le od verjetnosti γ in ni naključna spremenljivka.

Kot je znano se funkcija ϕ in njena inverzna funkcija ϕ^{-1} nanašajo na standardno normalno porazdelitev $N(0,1)$, tako, da lahko vrednosti z_γ izračunamo iz tabel. Za najpogosteje uporabljene vrednosti verjetnosti γ so odgovarjajoče vrednosti z_γ izračunane po formuli (10) so navedene v tabeli:

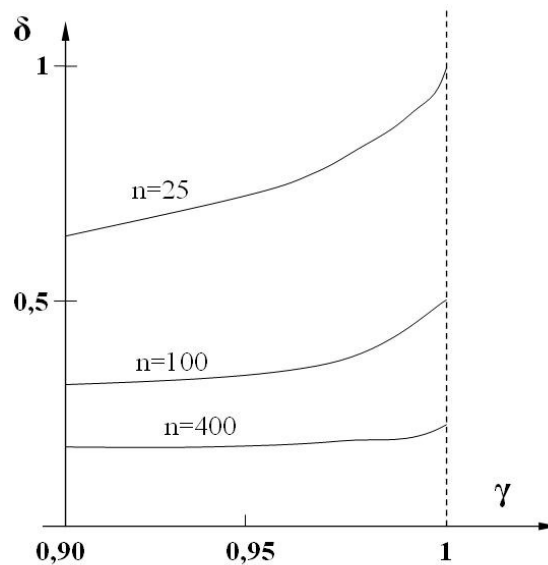
Tabela 1

γ	0,90	0,95	0,99
z_γ	1,65	1,96	2,58

$\gamma \backslash n$	0,90	0,95	0,99
25	0,66	0,78	1,03
100	0,33	0,39	0,52
400	0,16	0,20	0,26

Tabela 2

Glede na (12) in tabelo 1 opazimo odvisnost med širino intervala zaupanja, pripadajoče verjetnosti in velikosti vzorca, kot je prikazano v tabeli 2. Tam so za izbrane velikosti vzorcev n in verjetnosti γ navedene vrednosti veličine $\delta = \frac{2}{\sqrt{n}} z_\gamma$, kar prikazuje širino intervala za $\sigma = 1$. Grafični prikaz je na sliki spodaj.



Slika odvisnosti veličin n , γ in δ

Sedaj lahko navedemo konkreten primer izbire intervala zaupanja. Recimo da je $\sigma = 2$, in da smo na vzorcu velikosti $n=25$ izračunali povprečno vrednost aritmetične sredine $\bar{x} = 12$, nato pa bo glede na (11) in tabelo 1 interval zaupanja pri verjetnosti $\gamma = 0,95$ imel naslednje meje:

$$g_1 = \bar{x} - z_\gamma \frac{\sigma}{\sqrt{n}} = 12 - 1,96 \cdot \frac{2}{5} = 11,22$$

$$g_2 = \bar{x} + z_\gamma \frac{\sigma}{\sqrt{n}} = 12 + 1,96 \cdot \frac{2}{5} = 12,78$$

Rezultat je, da se neznano pričakovanje μ nahaja v intervalu $\{11,22; 12,87\}$, vendar ta rezultat ni absolutno natančen, saj ima verjetnost 95%, kar pomeni, da z uporabo omenjenega postopka lahko pričakujemo 5% napačnih zaključkov (stopnja tveganja); če vzamemo verjetnosti $\gamma = 0,99$, se pravi da toleriramo samo 1% napačnih rezultatov dobimo vrednosti za meje intervala zaupanja naslednje:

$$g_1 = 12 - 2,58 \cdot \frac{2}{5} = 10,97, \quad g_2 = 12 + 2,58 \cdot \frac{2}{5} = 13,03.$$

Glede na formulo (12) je mogoče, da za v naprej določeno napako δ_0 pri aproksimaciji neznanega parametra t na vrednost \hat{t} iz intervala zaupanja verjetnosti γ določimo potrebo število n naključnega vzorca. Vstavimo $\delta = \delta_0$ in iz (12) dobimo:

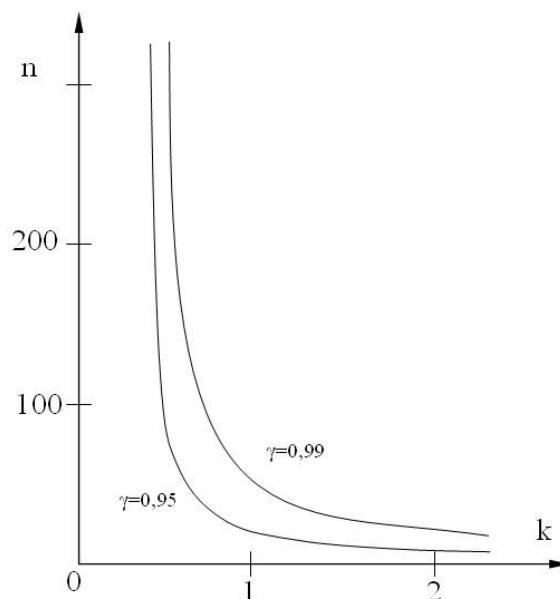
$$(13) \quad n = z_y^2 \left(\frac{\sigma}{\delta_0} \right)^2$$

In če za mero napake vzamemo standardni odklon δ , oziroma vstavimo $\delta_0 = k\sigma$ se izraz (13) spremeni v obliko:

$$(14) \quad n = \frac{4z_y^2}{k^2},$$

vidimo, da je število n , potrebnih merenj za izbor intervala zaupanja verjetnost γ obratno sorazmeren s kvadratom dovoljene napake. Na primer, če za verjetnost γ vzamemo vrednost 0,95, se tolerira napaka za eno polovico standardnega odklona, $k=0,5$, in lahko na podlagi formule (14) izračunamo, da je potrebno vsaj 62 merenj preizkušane naključne spremenljivke X .

Odvisnost širine intervala zaupanja od dopuščene napake δ_0 in velikosti vzorca n je prikazana na sliki.



Odvisnost napake (k) od velikosti vzorca (n)

2. Primer

Zastavimo si nalogo, da izberemo interval zaupanja verjetnosti γ za neznani parameter t enakomerne porazdelitve $U(0, t)$ ($t > 0$). ML-cenilki $\hat{T}_2 = \max(x_1, \dots, x_n)$ za neznani parameter t pripada verjetnostna porazdelitev, prikazana s formulo (19) iz poglavja 6.1. Z apliciranjem relacije (5) na cenilko \hat{T}_2 dobimo:

$$P(c_1(t) < \hat{T}_2 < c_2(t)) = \left[\frac{c_2(t)}{t} \right]^n - \left[\frac{c_1(t)}{t} \right]^n = \gamma,$$

Potrebno je najti monotone funkcije c_1 in c_2 , katere zadovoljijo pogoje zgornje enačbe. Glede na to, da je pri enakomerni porazdelitvi $U(0, t)$ vsaka izmerjena vrednosti $x_i \leq t$ ($i=1, \dots, n$), bo tudi $\hat{t}_2 = \max(x_1, \dots, x_n) \leq t$. Vzamemo, da je $c_2(t) = t$, $c_1(t) = kt$, in postavimo zahtevo k ($0 < k < 1$), tako da velja:

$$\left[\frac{c_2(t)}{t} \right]^n - \left[\frac{c_1(t)}{t} \right]^n = 1 - k^n = \gamma$$

Iz česar opazimo, da je $k = \sqrt[n]{1 - \gamma}$ in lahko zapišemo:

$$P(t \sqrt[n]{1 - \gamma} < \hat{T}_2 < t) = \gamma, \text{ in } P(\hat{T}_2 < t < \frac{\hat{T}_2}{\sqrt[n]{1 - \gamma}}) = \gamma$$

Primerjanje te relacije z relacijo (4) nam pokaže, da so naključni robovi intervala zaupanja verjetnosti γ za parameter t enakomerne porazdelitve $U(0, t)$ podani z formulami:

$$G_1 = \hat{T}_2, \quad G_2 = \frac{\hat{T}_2}{\sqrt[n]{1 - \gamma}}$$

Na primer če vzamemo $\gamma = 0,95$ in $n=25$, dobimo vrednost $\sqrt[n]{1 - \gamma} = \sqrt[25]{0,05} \approx 0,89$, in je \hat{t}_2 maksimalna vrednost v nizu merjenj (x_1, \dots, x_{25}), potem so odgovarjajoči robovi intervala zaupanja za parameter t enaki:

$$g_1 = \hat{t}_2, \quad g_2 = \frac{\hat{t}_2}{0,89} = 1,12\hat{t}_2$$

V primeru je širina intervala enaka:

$$\Delta = |G_2 - G_1| = \hat{T}_2 \left(\frac{1}{\sqrt[n]{1 - \gamma}} - 1 \right)$$

Vrednost širine intervala zaupanja:

$$\delta = \hat{t}_2 \left(\frac{1}{\sqrt[n]{1-\gamma}} - 1 \right)$$

je sorazmerna vrednosti \hat{t}_2 cenilke \hat{T}_2 s faktorjem sorazmernosti d

$$d = \frac{1}{\sqrt[n]{1-\gamma}} - 1$$

pri čemer bolj konkreten pregled odvisnosti med veličinami n, d in γ najdemo v tabeli.

n	d		
	$\gamma = 0,90$	$\gamma = 0,95$	$\gamma = 0,99$
2	2,16	3,47	9
5	0,58	0,82	1,51
10	0,26	0,35	0,58
29	0,12	0,16	0,26
50	0,05	0,06	0,10
100	0,02	0,03	0,05

Viri

**Pauše, Željko: Uvod u matematičku statistiku
Zapiski predmeta STATISTIKA**