Uporabna statistika

Gregor Dolinar

Fakulteta za elektrotehniko
Univerza v Ljubljani

12. november 2013

Na podlagi statistične analize vzorca želimo sklepati o lastnostih populacije in sprejemati različne odločitve.
Slučajno spremenljivko X oziroma populacijo, ki jo ta spremenljivka predstavlja, opišemo z njeno slučajno porazdelitvijo.

Slučajna porazdelitev je točno določena z nekaj parametri, ki jih običajno ne poznamo:

- povprečna vrednost μ
- varianca σ^2
- delež "dobrih" elementov populacije p
- razlika povprečnih vrednosti dveh populacij $\mu_1 - \mu_2$
- razlika deležev v dveh populacijah $p_1 - p_2$

S pomočjo vzorca dobimo oceno vrednosti parametrov. Slučajna porazdelitev je potem določena in s tem dobimo opis celotne populacije.

Primer

Zanima nas življenjska doba procesorja. Primeren model za življensko dobo je Weibullova porazdelitev

$$f(x) = \frac{\beta}{\delta}\left(\frac{x}{\delta}\right)^{\beta-1} e^{\left(\frac{x}{\delta}\right)^{\beta-1}}.$$

Ne poznamo parametrov β in δ, vemo pa da se $E(X)$ in $V(X)$ izražata s pomočjo β in δ.
Na podlagi vzorca naredimo oceno za $E(X)$ in $V(x)$.

Ločimo dve vrsti ocen parametrov:

- točkovna ocena
- intervalna ocena

Točkovna ocena parametrov

Cenilka Θ parametra θ je funkcija slučajnih spremenljivk $X_1, \ldots, X_n$ slučajnega vzorca velikosti n.

Primer

Ena izmed cenilk za $E(X)$ je na primer

$$\Theta = h(X_1, \ldots, X_n) = \frac{X_1 + \ldots + X_n}{n} = \overline{X}.$$

Cenilka za $V(X)$ je na primer

$$\Theta = h(X_1, \ldots, X_n) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

Definicija

Cenilka za parameter θ je nepristranska (unbiased estimator), če je $E(\Theta) = \theta$.
Če cenilka ni nepristranska, potem razliko $E(\Theta) - \theta$ imenujemo pristranskost (bias).

Primer

$X_1, \ldots, X_n$ normalno porazdeljene. Zanima nas točkovna ocena za μ. Možne cenilke:

- X_1
- $\overline{X}$
- 10 % največjih in najmanjših vrednosti vzorca ne upoštevamo
- $\frac{2X_1 - X_2 + X_3}{2}$

Katera cenilka je najboljša?

Cenilka naj bi bila nepristranska.
Vse cenilke v prejšnjem primeru so nepristranske.

Trditev

Najboljša je nepristranska cenilka z najmanjšo varianco - minimal variance unbiased estimator (MVUE).

Opomba

Najbolj verjetno je, da bo napoved $\hat{\theta}$ nepristranske cenilke z najmanjšo varianco najbližje pravi vrednosti parametra θ od vseh nepristranskih cenilk.

Izrek

$X_1, \ldots, X_n$ normalno porazdeljene s parametroma μ, σ^2, potem je vzorčna sredina $\overline{X}$ nepristranska cenilka z najmanjšo varianco (MVUE) za parameter μ.

Opomba

Opomba. Včasih je pristranska cenilka z majhno varianco boljša od nepristranske z veliko varianco.

Določanje cenilk

Kako poiskati cenilko, s katero bomo na podlagi vzorca ocenjevali vrednost iskanega parametra?

Cenilke lahko določamo z:

- metodo momentov
- metodo največjega verjetja

Metoda momentov

Definicija

Naj bodo $X_1, \ldots, X_n$ vzorčne slučajne spremenljivke s slučajno porazdelitvijo $f(x)$. Populacijski moment r-tega reda je potem $E(X^r)$, vzorčni moment r-tega reda pa $\frac{1}{n}\sum_{i=1}^{n} X_i^r$

Vrednosti parametrov določimo tako, da momente izbrane porazdelitve z neznanimi parametri izenačimo z momenti, ki jih izračunamo s pomočjo vzorca.

Primer.
Gama porazdelitev:

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)},$$

$$E(X) = \frac{r}{\lambda},$$

$$E(X^2) = \frac{r(r+1)}{\lambda^2}.$$

Izenačimo

$$\frac{r}{\lambda} = \overline{X},$$

$$\frac{r(r+1)}{\lambda^2} = \sum_{i=1}^{n} \frac{1}{n} X_i^2.$$

Življenjska doba izdelka je modelirana z Gamma porazdelitvijo. Za 6 izdelkov so bile življenjske dobe $5, 12, 15, 22, 33$ in 62 dni. Koliko odstotkov izdelkov ima pričakovano življenjsko dobo krajšo od 30 dni?