

1 Vzorčenje - neskončna populacija

Oceniti želimo znižanje vrednosti pritiska pri pacientih z esencialno hipertenzijo po treh mesecih jemanja nekega zdravila. V ta namen smo zbrali vzorec 25 bolnikov, naj bo X_i vrednost razlike pri i -tem bolniku našega vzorca. Predpostavimo, da so slučajne spremenljivke X_i neodvisne in enako porazdeljene.

- Pokažite, da je povprečje našega naključnega vzorca nepristranska ocena povprečnega znižanja v populaciji bolnikov (to označimo z μ).

Povprečje vzorca je $\frac{1}{n} \sum_{i=1}^n X_i$, povprečje populacije označimo z μ . Predpostavljamo, da je vzorec naključen, torej da so vrednosti X_i enako porazdeljene, za vse velja $E(X_i) = \mu$.

$$E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} n\mu = \mu$$

- Kaj lahko rečemo o $cov(X_i, X_j)$, če je $i \neq j$?

Ker so vrednosti bolnikov med seboj neodvisne, je kovarianca enaka 0

- Naj bo varianca v populaciji enaka σ^2 . Kakšna je standardna napaka naše ocene?

$$\begin{aligned} var[\bar{X}] &= cov\left[\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} n \cdot cov\left[X_i, \frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \{cov[X_i, X_i] + (n-1)cov[X_i, X_j]\} = \frac{1}{n} \{cov[X_i, X_i]\} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

- Na podlagi našega vzorca želimo oceniti σ^2 . Naj bo naša cenilka $\hat{\sigma}^2 = c \sum_{i=1}^n (X_i - \bar{X})^2$. Kakšna mora biti vrednost konstante c , da bo naša ocena nepristranska?

Vemo, da velja $\sigma^2 = E(X^2) - E(X)^2$, torej $E(X^2) = \sigma^2 + \mu^2$ in podobno tudi $SE^2 = \text{var}(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{X}^2) - E(\bar{X})^2$, torej $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}$:

$$\begin{aligned} E\left[c \sum_{i=1}^n (X_i - \bar{X})^2\right] &= cE\left[\sum_{i=1}^n X_i^2 - \bar{X}^2\right] \\ &= c \sum_{i=1}^n [E(X_i^2) - E(\bar{X}^2)] \\ &= c \sum_{i=1}^n \left[(\mu + \sigma^2) - \left(\mu^2 + \frac{\sigma^2}{n}\right) \right] \\ &= cn \left[\sigma^2 \left(1 - \frac{1}{n}\right) \right] \\ &= \sigma^2(n-1)c \end{aligned}$$

Ker želimo, da velja $E(\hat{\sigma}^2) = \sigma^2$, mora biti $c = \frac{1}{n-1}$.

- Na vzorcu smo dobili naslednje rezultate: $\bar{x} = 4$, $s = 20$ (s označuje vzorčno oceno standardnega odklona). Izračunajte standardno napako. Ali boste na podlagi podatkov lahko trdili, da se tlak zniža tudi v populaciji?

Ocena standardne napake je enaka $\widehat{SE} = \frac{s}{\sqrt{n}} = \frac{20}{5} = 4$. Znižanje pritiska je enako standardni napaki - prave razlike ne moremo razločiti od naključne variabilnosti. V ta namen bi potrebovali precej večji vzorec.

Povzemimo rezultate naloge: če je populacija neskončna in enote v vzorcu neodvisne, velja:

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}; \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

2 Vzorčenje - končna populacija

Oceniti želimo povprečno število zaposlenih v podjetjih neke panoge ob začetku letošnjega leta. Panoga je razdeljena na podskupine, v eni izmed skupin je le 11 podjetij. Uspeli smo pridobiti podatke za naključen vzorec šestih izmed teh podjetij. Naj bo X_i število zaposlenih v i -tem podjetju našega vzorca.

- Naj X_1 in X_2 označujeta vrednosti prvih dveh naključno izbranih podjetij. Kaj lahko rečemo o $cov(X_1, X_2)$? Kaj pa za splošen $i \neq j$?

Populacija končna, označimo vrednosti v populaciji z x_k , $i = k, \dots, 11$. Izberimo po nekem vrstnem redu vseh 11 podjetij, prvih 6 naj jih predstavlja vzorec, X_i označuje število zaposlenih v i -tem izbranem podjetju. Ker je vsak izmed X_i ena od vrednosti x_k in imajo vse enako verjetnost, je $cov(X_1, X_2) = cov(X_i, X_j)$ za poljubna različna i in j . Vendar pa sedaj X_i in X_j nista neodvisni slučajni spremenljivki - če je $X_i = x_k$, X_j ne more zavzeti te vrednosti.

$$cov(X_i, \sum_{j=1}^N X_j) = cov(X_i, \sum_{k=1}^N x_k) = 0$$

Ker je vsota vseh vrednosti konstanta, je zgornji izraz enak 0, torej

$$cov(X_i, \sum_{j=1}^N X_j) = cov(X_i, X_i) + (N-1)cov(X_i, X_j) = 0$$

In zato (za $i \neq j$)

$$cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

Spremenljivki sta torej negativno povezani. Ta povezanost je seveda šibka, korelacija je enaka

$$cor(X_i, X_j) = -\frac{\sigma^2}{(N-1)\sigma^2} = -\frac{1}{N-1}$$

Korelacija je torej odvisna izključno od velikosti populacije.

- Izračunajte standardno napako našega vzorca (privzemite da poznate σ^2). V drugi skupini imamo 100 podjetij. Kako velik vzorec moramo vzeti iz te skupine, da bomo dobili približno enako veliko standardno napako (privzemimo da je varianca tudi v tej skupini enaka σ^2)?

$$\begin{aligned} \text{var}[\bar{X}] &= \text{cov}\left[\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} n \cdot \text{cov}\left[X_i, \frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \{ \text{cov}[X_i, X_i] + (n-1) \text{cov}[X_i, X_j] \} \\ &= \frac{1}{n} \left\{ \sigma^2 - (n-1) \frac{\sigma^2}{N-1} \right\} = \frac{\sigma^2}{n} \frac{N-n}{N-1} \end{aligned}$$

Za $N = 11$ in $n = 6$ dobimo $SE^2 = \frac{\sigma^2}{6} \frac{11-6}{10} = \frac{\sigma^2}{12}$. Če je populacija večja, bomo za enako standardno napako potrebovali več enot. Pri $N = 100$, dobimo željeno velikost napake z vzorcem velikosti 11.

- Kakšna mora biti vrednost konstante c , da bo cenilka $\hat{\sigma}^2 = c \sum_{i=1}^n (X_i - \bar{X})^2$ nepristransko ocenila vrednost σ^2 ?

Ponovimo izračun iz zadnje točke prejšnje naloge, upoštevamo, da je $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n} \frac{N-n}{N-1}$:

$$\begin{aligned} E\left[c \sum_{i=1}^n (X_i - \bar{X})^2\right] &= cE\left[\sum_{i=1}^n X_i^2 - \bar{X}^2\right] \\ &= c \sum_{i=1}^n \left[(\mu + \sigma^2) - \left(\mu^2 + \frac{\sigma^2}{n} \frac{N-n}{N-1}\right) \right] \\ &= cn \left[\sigma^2 \left(1 - \frac{N-n}{n(N-1)}\right) \right] \\ &= \sigma^2 c \frac{N(n-1)}{N-1} \end{aligned}$$

Ker želimo, da velja $E(\hat{\sigma}^2) = \sigma^2$, mora biti $c = \frac{1}{n-1} \frac{N-1}{N}$, torej

$$\hat{\sigma}^2 = \frac{1}{n-1} \frac{N-1}{N} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Zapišite še nepristransko cenilko za standardno napako

Združimo dosedanje rezultate in dobimo

$$\begin{aligned}\widehat{SE}^2 &= \frac{\widehat{\sigma}^2(N-n)}{N-1} = \frac{1}{n-1} \frac{N-1}{N} \sum_{i=1}^n (X_i - \bar{X})^2 \frac{(N-n)}{N-1} \\ &= \frac{s^2}{n} \frac{N-n}{N},\end{aligned}$$

kjer je $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Povzemimo rezultate naloge: če je populacija končna, velja:

Vrednosti izbrane v vzorec navkljub naključnemu izbiranju med seboj niso neodvisne, kovarianca med enotami je enaka:

$$\text{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

varianca povprečja in nepristranska cenilka za varianco v populaciji pa sta enaki

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{(N-1)}; \widehat{\sigma}^2 = \frac{1}{n-1} \frac{N-1}{N} \sum_{i=1}^n (X_i - \bar{X})^2$$