2 Vzorčenje

2.1 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 $E(X_i) = \mu$.

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

- 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 varianca (oz. standardna napaka) naše ocene?

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

Uporabimo, da so vrednosti med seboj neodvisne, torej da je $\mathrm{cov}[X_i, X_j] = 0$ za vsak $i \neq j$.

$$\begin{aligned}
\mathrm{var}[\bar{X}] &= \frac{1}{n^2}\sum_{i=1}^{n}\{\mathrm{cov}[X_i, X_i]\} \\
&= \frac{1}{n^2}\cdot n\mathrm{cov}[X_i, X_i] = \frac{1}{n}\mathrm{var}[X_i] \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Standardna napaka je enaka $SE = \frac{\sigma}{\sqrt{n}}$.

- Na podlagi našega vzorca želimo oceniti σ^2. Naj bo naša cenilka $\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$. 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 $SE^2 = \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{X}^2) - E(\bar{X})^2$, torej $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}$. Upoštevamo še, da velja $\sum_i X_i\bar{X} = \bar{X}\sum_i X_i = n\bar{X}^2 = \sum_i \bar{X}^2$:

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

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

- Na vzorcu smo dobili naslednje rezultate: $\bar{x} = 4$, $\hat{\sigma} = 20$. Ocenite standardno napako (torej standardni odklon vzorčnega povprečja). Ali boste na podlagi podatkov lahko trdili, da se tlak zniža tudi v populaciji?

  Ocena standardne napake je enaka $\widehat{SE} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{20}{5} = 4$. 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:

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

Predlogi za vaje v R-u:

- Generirajte vzorce velikosti 30 iz enakomerne porazdelitve, vsakič izračunajte povprečje ter si oglejte porazdelitev tega povprečja. Ocenite pričakovano vrednost za povprečje in standardno napako ter oceni primerjajte s teoretičnimi vrednostmi
- Postopek ponovite še za druge (morda bolj asimetrične) porazdelitve.

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, μ naj označuje njihovo povprečje, σ pa standardni odklon.

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

  Populacija je končna, označimo vrednosti v populaciji z x_k, $k = 1, \ldots, 11$. 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 $\mathrm{cov}(X_1, X_2) = \mathrm{cov}(X_i, X_j)$ 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 k-te vrednosti.

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

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

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

In zato (za $i \neq j$)

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

Spremenljivki sta torej negativno povezani.

- Izračuanjte še korelacijo $\mathrm{cor}(X_i, X_j)$ za neka $i \neq j$. Od česa je odvisna?

  Korelacija med spremenljivkama je enaka

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

  Korelacija je torej odvisna izključno od velikosti populacije in z večanjem populacije hitro postane zelo majhna.

- Izračunajte standardno napako našega vzorca (privzemite da poznate σ^2).

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

  Standardna napaka je torej enaka $SE = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{(N-1)}}$.

- 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)? Kaj pa če bi imeli skupino z zelo velikim številom podjetij?

  Za $N = 11$ in $n = 6$ dobimo $SE^2 = \frac{\sigma^2}{6}\frac{11-6}{(10)} = \frac{\sigma^2}{12}$. Pri populaciji velikosti 100 nam vzorec velikosti 10 da standardno napako $SE^2 = \frac{\sigma^2}{11}$, vzorec velikosti 11 pa standardno napako $SE^2 = \frac{\sigma^2}{12,2}$.
  Če je populacija večja, bomo za enako standardno napako torej potrebovali več enot. Če bi imeli skupino z zelo velikim številom podjetij, bi bil izraz $\frac{N-n}{(N-1)}$ približno enak 1, zato bi potrebovali 12 podjetij. Velikost potrebnega vzorca se z večanjem populacije veča, vendar pa to večanje kmalu ni več bistveno.

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

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

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

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

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

- Zapišite še nepristransko cenilko za varianco povprečja

Združimo dosedanje rezultate in dobimo

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

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

Povzemimo rezultate naloge: če je populacija končna, velja:
Vrednosti izbrane v vzorec navkljub naključnemu izbiranju med seboj niso neodvisne, kovarianca med enotami je enaka:

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

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

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

Predlogi za vaje v R-u:

- Izmislite si 11 vrednosti ter nato generirajte vzorce velikosti 6. Oglejte si porazdelitev vzorčnih povprečij. Pokažite, da gornja formula predstavlja nepristransko oceno populacijske variance.