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).
- Kaj lahko rečemo o $cov(X_i, X_j)$, če je $i \neq j$?
- Naj bo varianca v populaciji enaka σ^2. Kakšna je varianca (oz. standardna napaka) naše ocene?
- 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?
- 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?

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 $i \neq j$?
- Izračuanjte še korelacijo $\text{cor}(X_i, X_j)$ za neka $i \neq j$. Od česa je odvisna?
- 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)? Kaj pa če bi imeli skupino z zelo velikim številom podjetij?
- 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?
- Zapišite še nepristransko cenilko za varianco povprečja

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.

2.3 Določitev načrta vzorčenja

Zanima nas povprečna teža bolnikov (μ) s hipertenzijo v starostni skupini 60 do 80 let. Težo bi radi ocenili na podlagi vzorca, jasno je, da bo teža precej različna pri moških (μ_1) kot pri ženskah (μ_2). Čas in denar, ki ju imamo na voljo za raziskavo, nam dopuščata vzorec velikosti 100. Vemo, da se v populaciji deleža moških in žensk s hipertenzijo razlikujeta, delež moških označimo z d. Zanima nas, kakšen delež moških in kakšen delež žensk naj naberemo v vzorec, da bo standardna napaka naše ocene najmanjša možna. Pri tem predpostavimo, da je standardni odklon teže moških k-krat večji od standardnega odklona teže žensk.

- Zapišite nepristransko cenilko za populacijsko povprečje

- Standardno napako ocene izrazite z velikostima podvzorcev (n_1 je število moških, n_2 število žensk).
- Naj velja $\sigma_1 = k\sigma_2$. Pri kakšni razdelitvi vzorca je standardna napaka najmanjša? Izračunajte n_1 za primera $k = 1$ in $k = 2$, vzemite, da je delež moških enak 0,7.

Predlogi za vaje v R-u:

- Izmislite si smiselne vrednosti za parametre v nalogi ter generirajte podatke. Grafično prikažite, kako se pri različnih vrednosti d in izbirah velikosti vzorcev spreminja kvaliteta vaše ocene.