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.

2.4 Enostavni vzorec iz končne populacije, še enkrat

Vzemimo še enkrat enostavni slučajni vzorec velikosti n iz populacije N, vrednosti v populaciji označimo z x_i; $i = 1, \ldots, N$, populacijsko vrednost povprečja označimo z μ, variance pa z σ^2. Definirajmo slučajno spremenljivko $I_i = I_{[i \text{ je izbran v vzorec}]}$ in zapišimo cenilko populacijskega povprečja μ kot $C = \frac{1}{n}\sum_{i=1}^{N} I_i x_i$.

- Koliko je vsota $\sum_{i=1}^{N} I_i$? Kakšna je verjetnost $P(I_i = 1)$?
- Pokažite, da je cenilka nepristranska.
- Izračunajte $\text{var}(I_i)$ in $cov(I_i, I_j)$.
- Pokažite še, da je varianca tako zapisane cenilke enaka $\text{var}(C) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

2.5 Vzorčenje po skupinah

Oceniti želimo dosežek ljubljanskih sedmošolcev na nekem testu znanja, ki ga izvajajo v večih državah. Populacijo $N = 2800$ učencev te starosti bomo vzorčili po šolah ($K = 46$). V vzorec bomo najprej slučajno (in neodvisno od števila N_i sedmošolcev na šoli i) vzorčili $k = 10$ šol, nato pa bomo na vsaki šoli izbrali vzorec $n = 15$ učencev. Naj μ označuje populacijsko povprečje dosežka na testu, μ_i pa naj bo povprečje za vsako šolo posebej. Vzorčenje znotraj šol je neodvisno od vzorčenja na prvem koraku.

- Zapišite nepristransko cenilko za μ.
- Kako bi ocenili populacijsko povprečje, če bi imele vse šole enako število učencev L?
- Ali je za nepristranskost pomembno, koliko učencev z vsake šole vzamete?
- Označimo varianco znotraj vsake šole z $\sigma_{wi}^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(x_{ij} - \mu_i)^2$. Kaj je $\mathrm{var}(I_i\bar{X}_i)$ in kaj $\mathrm{cov}(I_i\bar{X}_i, I_j\bar{X}_j)$?
- Izpeljite formulo za varianco cenilke v primeru, ko so vse vrednosti N_i enake L in je varianca znotraj šole enaka za vse šole, varianco med šolami označite z σ_b^2.

Predlogi za vaje v R-u:

- Preverite rezultate naloge z R-om.

2.6 Ocena kovariance

V nekem podjetju velikosti N so izvedli izobraževanje za naključen vzorec n zaposlenih. Ob koncu izobraževanja so novo znanje preverili s testom. Podjetje se želi odločiti, ali je smiselno uvesti izobraževanje za vse zaposlene, zato jih zanima povezanost med starostjo zaposlenega (X_i) in rezultatom na testu (Y_i).
Za vsakega posameznika iz vzorca imamo torej par slučajnih spremenljivk (X_i, Y_i), $i = 1 \ldots n$.

- Utemeljite, da je količina $\mathrm{cov}(X_i, Y_j)$ za poljubna $i \neq j$ enaka.

- Naj bo $\gamma = \mathrm{cov}(X_i, Y_i)$. Izračunajte kovarianco $\mathrm{cov}(X_i, Y_j)$ za $i \neq j$.
- Kako bi ocenili korelacijo? Kaj vemo o nepristranskosti te ocene?

Predlogi za vaje v R-u:

- Ker ne vemo, ali je ocena pristranska ali ne, pristranskost preverimo s simulacijo:

  Vzamemo populacijo velikosti $N = 300$, vzorci naj bodo velikosti $n = 10$. Naj bo X starost porazdeljena enakomerno med 25 in 65, uspeh na testu pa negativno povezan s starostjo, tako, da je v povprečju enak $100 - starost$ (predpostavimo, da so odstopanja od tega povprečja razpršena s standardnim odklonom 20 in normalno porazdeljena)

```
> set.seed(1)
> xi <- runif(300)*40+25                #300 posameznikov, starosti 25-65 let
> yi <- 100 - xi + rnorm(300)*20        #rezultat na testu za populacijo
> cov(xi,yi)                            #kovarianca v populaciji
[1] -136.8110
> cor(xi,yi)                            #korelacija v populaciji
[1] -0.5207052

> runs <- 10000                         #stevilo korakov simulacije
> cova <- cora <- rep(NA,runs)          #sem bomo zapisali rezultate simulacije
> for(it in 1:runs){                    #simulacija po korakih
+ inx <- sample(1:length(xi),size=10,replace=F)     #izberemo vzorec 10-ih
+ xa <- xi[inx]                         #pogledamo njihove starosti
+ ya <- yi[inx]                         #pogledamo njihove rezultate
+ cova[it] <- 1/9*299/300*
+   sum( (xa-mean(xa))*(ya-mean(ya)))    #izracunamo kovarianco
+ cora[it] <-  sum( (xa-mean(xa))*(ya-mean(ya)))/
+   sqrt(sum( (xa-mean(xa))^2)*sum((ya-mean(ya))^2))   #izracunamo korelacijo
+ }

> mean(cova)                            #povprecna kovarianca
[1] -135.4745
> mean(cora)                            #povprecna korelacija
[1] -0.5034081
```

- Vidimo, da sta obe vrednosti nekoliko manjši od populacijskih, preverimo ali je odstopanje veliko glede na standardno napako, ki jo lahko

pričakujemo pri takem številu simulacij:

Zanima nas ali povprečna kovarianca (`mean(cova)`) bistveno odstopa od prave vrednosti (`cov(xi,yi)`). Povprečna kovarianca je slučajna spremenljivka, če bomo vnovič pognali simulacijo (vseh 10000 korakov), bomo dobili drugo vrednost. Predpostavimo, da je približno normalno porazdeljena, ocenimo njeno varianco (varianca povprečja n i.i.d spremenljivk je varianca spremenljivk deljeno z n, pri nas je n število korakov simulacije). Ničelna domneva, ki jo preverjamo, je: H_0 : povprečna kovarianca je enaka populacijski vrednosti. Odstopanje od te ničelne domneve preverjamo s testom t.

```
> (mean(cova)-cov(xi,yi))/sqrt(var(cova)/runs)
[1] 1.509540
```

Ta rezultat je v okviru pričakovanj, saj smo teoretično pokazali, da je ocena kovariance nepristranska. Enako ponovimo za korelacijo:

```
> (mean(cora)-cor(xi,yi))/sqrt(var(cora)/runs)
[1] 6.66459
```

Odstopanje pri korelaciji je bistveno večje, verjamemo, da se v naši simulaciji ni zgodilo po naključju, temveč je ocena dejansko pristranska.