

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

Populacijsko povprečje μ lahko zapišemo kot $\mu = d\mu_1 + (1 - d)\mu_2$. Ker velja $E(\bar{X}_1) = \mu_1$ in $E(\bar{X}_2) = \mu_2$, je nepristranska cenilka enaka

$$M = d\bar{X}_1 + (1 - d)\bar{X}_2.$$

- Standardno napako ocene izrazite z velikostima podvzorcev (n_1 je število moških, n_2 število žensk).

Ker je povprečje pri moških neodvisno od povprečja pri ženskah, velja

$$\text{var}(M) = d^2 \text{var}(\bar{X}_1) + (1 - d)^2 \text{var}(\bar{X}_2) = d^2 \frac{\sigma_1^2}{n_1} + (1 - d)^2 \frac{\sigma_2^2}{n_2}.$$

- 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.

Varianco vzorčnega povprečja zapišemo kot

$$\text{var}(M) = \sigma^2 \left(\frac{d^2 k^2}{n_1} + \frac{(1 - d)^2}{n - n_1} \right),$$

minimizirati moramo torej izraz v oklepaju. Odvajamo po n_1 in izenačimo z 0

$$\begin{aligned} -\frac{d^2 k^2}{n_1^2} + \frac{(1 - d)^2}{(n - n_1)^2} &= 0 \\ -(n - n_1)^2 d^2 k^2 + n_1^2 (1 - d)^2 &= 0 \\ (d^2 k^2 - (1 - d)^2) n_1^2 - 2n d^2 k^2 n_1 + n^2 d^2 k^2 &= 0 \end{aligned}$$

Rešimo kvadratno enačbo in dobimo rešitev (druga rešitev je večja od n in tako nesmiselna):

$$n_1 = \frac{ndk}{dk + 1 - d}$$

Za $k = 1$ dobimo rešitev $n_1 = nd$, v konkretnem primeru torej $n_1 = 70$ in $n_2 = 30$. Če sta deleža moških in žensk na vzorcu enaka deležema v populaciji, je standardna napaka najmanjša možna.

Če je standardni odklon teže moških 2x večji od standardnega odklona žensk, dobimo $n_1 = 82,4$, vzeti moramo torej ustrezno več posameznikov iz bolj variabilne skupine.

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 ; $i = 1, \dots, N$, 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 $P(I_i = 1)$?

Vsota $\sum_{i=1}^N I_i = n$, saj smo vzeli vzorec velikosti n . Izračunajmo še verjetnost, da bo izbran element i :

Jemljem vzorce velikosti n in iz populacije velikosti N . Vseh možnih kombinacij je $\binom{N}{n}$, kakšno je število tistih vzorcev, v katerih je element i ? Pri teh vzorcih en element že poznamo, izmed ostalih $N - 1$ smo jih izbrali $n - 1$. Torej je iskana verjetnost enaka:

$$P(I_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

- Pokažite, da je cenilka nepristranska.

Radi bi ocenili $\mu = \frac{1}{N} \sum_{i=1}^N x_i$

$$E(C) = \frac{1}{n} \sum_{i=1}^N E(I_i)x_i$$

Ker lahko I_i zavzame le vrednosti 0 in 1, je $E(I_i) = P(I_i = 1) = \frac{n}{N}$ (vzorec je slučajen, zato so verjetnosti za vse i enake), zato dobimo

$$E(C) = \frac{1}{n} \sum_{i=1}^N \frac{n}{N} x_i = \frac{1}{N} \sum_{i=1}^N x_i = \mu$$

- Izračunajte $\text{var}(I_i)$ in $\text{cov}(I_i, I_j)$.

Spremenljivka I_i je Bernoullijeva, z verjetnostjo $P(I_i = 1) = \frac{n}{N}$. Njena varianca je zato enaka

$$\text{var}(I_i) = \frac{n}{N} \left(1 - \frac{n}{N}\right) = \frac{n}{N} \frac{N-n}{N}$$

Kovarianco izračunamo tako, da upoštevamo $\text{cov}(I_1, I_1 + \dots + I_N) = \text{cov}(I_1, n) = 0$ in $\text{cov}(I_i, I_j)$ je enaka za vsak $i \neq j$:

$$\text{cov}(I_i, I_j) = -\frac{\frac{n}{N} \frac{N-n}{N}}{N-1} = -\frac{n(N-n)}{N^2(N-1)}$$

- Pokažite še, da je varianca tako zapisane cenilke enaka $\text{var}(C) = \frac{\sigma^2}{n} \frac{N-n}{N-1}$

$$\begin{aligned} \text{var}(C) &= \frac{1}{n^2} \text{cov}\left(\sum_{i=1}^N I_i x_i, \sum_{j=1}^N I_j x_j\right) \\ &= \frac{1}{n^2} \sum_{i=1}^N \text{cov}\left(I_i x_i, \sum_{j=1}^N I_j x_j\right) \\ &= \frac{1}{n^2} \sum_{i=1}^N \left[\text{cov}(I_i x_i, I_i x_i) + \sum_{j=1, j \neq i}^N \text{cov}(x_i I_i, I_j x_j) \right] \\ &= \frac{1}{n^2} \sum_{i=1}^N \left[x_i^2 \text{cov}(I_i, I_i) + \sum_{j=1, j \neq i}^N x_i x_j \text{cov}(I_i, I_j) \right] \end{aligned}$$

Populacijska varianca definirana kot:

$$\begin{aligned}\sigma^2 &= \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \\ &= \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2 \\ \sum_{i=1}^N x_i^2 &= N(\sigma^2 + \mu^2)\end{aligned}$$

Izpeljimo varianco:

$$\begin{aligned}\text{var}(C) &= \frac{1}{n^2} \sum_{i=1}^N \left[x_i^2 \text{var}(I_i) - \sum_{j=1, j \neq i}^N x_i x_j \frac{\text{var}(I_i)}{N-1} \right] \\ &= \frac{\text{var}(I_i)}{n^2(N-1)} \sum_{i=1}^N \left[(N-1)x_i^2 - x_i \sum_{j=1, j \neq i}^N x_j \right] \\ &= \frac{\text{var}(I_i)}{n^2(N-1)} \sum_{i=1}^N \left[(N-1)x_i^2 - x_i \left(\sum_{j=1}^N x_j - x_i \right) \right] \\ &= \frac{\text{var}(I_i)}{n^2(N-1)} \sum_{i=1}^N [(N-1)x_i^2 - x_i(N\mu - x_i)] \\ &= \frac{\text{var}(I_i)}{n^2(N-1)} \sum_{i=1}^N [Nx_i^2 - x_i N\mu] \\ &= \frac{\text{var}(I_i)}{n^2(N-1)} \left[N \sum_{i=1}^N x_i^2 - N\mu \sum_{i=1}^N x_i \right] \\ &= \frac{\text{var}(I_i)}{n^2(N-1)} [N^2(\sigma^2 + \mu^2) - N^2\mu^2] \\ &= \frac{\text{var}(I_i)}{n^2(N-1)} [N^2\sigma^2] \\ &= \frac{N^2\sigma^2}{n^2(N-1)} \frac{n}{N} \frac{N-n}{N} \\ &= \frac{\sigma^2}{n} \frac{N-n}{N-1}\end{aligned}$$

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 .

Najprej izrazimo μ s povprečji šol, torej μ_i . Naj bo x_{ij} vrednost j -tega učenca na i -ti šoli. Velja

$$\mu = \frac{1}{N} \sum_{i=1}^K \sum_{j=1}^{N_i} x_{ij} = \frac{1}{N} \sum_{i=1}^K N_i \cdot \mu_i \quad (1)$$

Označimo ocenjeno povprečje vsake šole z \bar{X}_i , I_i pa naj bo indikatorska spremenljivka, ki je enaka 1, če je šola izbrana v vzorec. Naša cenilka naj bo enaka

$$\bar{X} = \sum_{i=1}^K c_i I_i \bar{X}_i$$

Določiti moramo vrednost konstante c_i , tako da bo cenilka nepristranska. Upoštevamo, da smo na vsaki šoli vzeli naključni vzorec in zato velja $E(\bar{X}_i) = \mu_i$. Ker je vzorčenje na drugem koraku neodvisno od vzorčenja na prvem, velja $E(I_i \bar{X}_i) = E(I_i)E(\bar{X}_i)$. Ker smo na prvem koraku vzorčili vse šole z enako verjetnostjo, je $E(I_i) = \frac{k}{K}$ za vsak i . Uporabimo vse naštetu in dobimo

$$\begin{aligned} E(\bar{X}) &= \sum_{i=1}^K c_i E(I_i \bar{X}_i) = \sum_{i=1}^K c_i E(I_i) E(\bar{X}_i) \\ &= \sum_{i=1}^K c_i \frac{k}{K} \mu_i \end{aligned}$$

Zaradi (1) mora veljati $c_i \frac{k}{K} = \frac{N_i}{N}$, zato je naša cenilka enaka

$$\bar{X} = \frac{K}{N} \frac{1}{k} \sum_{i=1}^K N_i I_i \bar{X}_i.$$

- Kako bi ocenili populacijsko povprečje, če bi imele vse šole enako število učencev L ?

Ker velja $N = \sum_{i=1}^K N_i$, za enake $N_i = L$ velja $N = KL$ in zato

$$\bar{X} = \frac{1}{L} \frac{1}{k} \sum_{i=1}^K L I_i \bar{X}_i = \frac{1}{k} \sum_{i=1}^K I_i \bar{X}_i$$

- Ali je za nepristranskost pomembno, koliko učencev z vsake šole vzamete?

Ne, \bar{X}_i je nepristranska cenilka μ_i ne glede na velikost vzorca. Seveda pa velikost vzorca vpliva na standardno napako te cenilke.

- Zapišite varianco cenilke s pomočjo varianc in kovarianc

$$\begin{aligned} \text{var}(\bar{X}) &= \text{var}\left(\frac{K}{N} \frac{1}{k} \sum_{i=1}^K N_i I_i \bar{X}_i\right) \\ &= \left(\frac{K}{Nk}\right)^2 \sum_{i=1}^K \left[N_i^2 \text{var}(I_i \bar{X}_i) + \sum_{j=1, i \neq j}^K N_i N_j \text{cov}(I_i \bar{X}_i, I_j \bar{X}_j) \right] \end{aligned}$$

- 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 $\text{var}(I_i \bar{X}_i)$ in kaj $\text{cov}(I_i \bar{X}_i, I_j \bar{X}_j)$?

Uporabimo, da je vzorčenje na drugem koraku neodvisno od vzorčenja na prvem in da je $I_i^2 = I_i$ ($1^2 = 1$, $0^2 = 0$):

$$\begin{aligned} \text{var}(I_i \bar{X}_i) &= E(I_i^2 \bar{X}_i^2) - E(I_i \bar{X}_i)^2 = E(I_i)E(\bar{X}_i^2) - E(I_i)^2 E(\bar{X}_i)^2 \\ &= \frac{k}{K} E(\bar{X}_i^2) - \left(\frac{k}{K}\right)^2 \mu_i^2 \end{aligned}$$

Upoštevamo še, da je $E(\bar{X}_i^2) = \text{var}(\bar{X}_i) + E(\bar{X}_i)^2 = \frac{\sigma_{wi}^2}{n} \frac{N_i - n}{N_i - 1} + \mu_i^2$ in dobimo

$$\text{var}(I_i \bar{X}_i) = \frac{k}{K} \left(\frac{\sigma_{wi}^2}{n} \frac{N_i - n}{N_i - 1} + \mu_i^2 \right) - \left(\frac{k}{K} \right)^2 \mu_i^2 = \mu_i^2 \frac{k(K - k)}{K^2} + \frac{k}{K} \frac{\sigma_{wi}^2}{n} \frac{N_i - n}{N_i - 1}$$

Sedaj izrazimo še kovarianco:

$$\text{cov}(I_i \bar{X}_i, I_j \bar{X}_j) = E(I_i I_j \bar{X}_i \bar{X}_j) - E(I_i \bar{X}_i) E(I_j \bar{X}_j)$$

Upoštevamo neodvisnost vzorčenja na prvem in drugem koraku in dejstvo, da je povprečje na eni šoli neodvisno od povprečja druge šole:

$$\begin{aligned} \text{cov}(I_i \bar{X}_i, I_j \bar{X}_j) &= E(I_i I_j) \mu_i \mu_j - E(I_i) E(I_j) \mu_i \mu_j \\ &= \mu_i \mu_j \text{cov}(I_i, I_j) = -\mu_i \mu_j \frac{k(K - k)}{K^2(K - 1)} \end{aligned}$$

- 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 .

$$\begin{aligned} \text{var}(\bar{X}) &= \left(\frac{1}{Lk} \right)^2 \sum_{i=1}^K \left[L^2 \text{var}(I_i \bar{X}_i) + \sum_{i=1, i \neq j}^K L^2 \text{cov}(I_i \bar{X}_i, I_j \bar{X}_j) \right] \\ &= \left(\frac{1}{k} \right)^2 \sum_{i=1}^K \left[\mu_i^2 \frac{k(K - k)}{K^2} + \frac{k}{K} \frac{\sigma_w^2}{n} \frac{L - n}{L - 1} \right. \\ &\quad \left. - \sum_{j=1, i \neq j}^K \mu_i \mu_j \frac{k(K - k)}{K^2(K - 1)} \right] \end{aligned}$$

Uporabimo izraz za varianco med šolami:

$$\begin{aligned} \sigma_b^2 &= \frac{1}{K} \sum_{i=1}^K [\mu_i - \mu]^2 \\ &= \frac{1}{K} \sum_{i=1}^K [\mu_i^2 - 2\mu\mu_i + \mu^2] \\ &= \frac{1}{K} \sum_{i=1}^K \mu_i^2 - \mu^2 \end{aligned}$$

Velja:

$$\begin{aligned}
& \sum_{i=1}^K \mu_i^2 \frac{k(K-k)}{K^2} - \sum_{i=1}^K \sum_{j=1, i \neq j}^K \mu_i \mu_j \frac{k(K-k)}{K^2(K-1)} \\
&= \frac{k(K-k)}{K^2(K-1)} \sum_{i=1}^K \left[(K-1)\mu_i^2 - \mu_i \left(\sum_{j=1}^K \mu_j - \mu_i \right) \right] \\
&= \frac{k(K-k)}{K^2(K-1)} \sum_{i=1}^K [(K-1)\mu_i^2 - \mu_i(K\mu - \mu_i)] \\
&= \frac{k(K-k)}{K^2(K-1)} \sum_{i=1}^K [K\mu_i^2 - K\mu\mu_i] \\
&= \frac{k(K-k)K}{K^2(K-1)} \left[\sum_{i=1}^K \mu_i^2 - \mu \sum_{i=1}^K \mu_i \right] \\
&= \frac{k(K-k)K}{K^2(K-1)} [K(\sigma_b^2 + \mu^2) - K\mu^2] \\
&= \frac{k(K-k)K}{K^2(K-1)} K\sigma_b^2 \\
&= \frac{k(K-k)}{(K-1)} \sigma_b^2
\end{aligned}$$

in zato

$$\begin{aligned}
\text{var}(\bar{X}) &= \frac{1}{k^2} \left[\frac{k(K-k)}{(K-1)} \sigma_b^2 + \sum_{i=1}^K \left(\frac{k}{K} \frac{\sigma_w^2}{n} \frac{L-n}{L-1} \right) \right] \\
&= \frac{1}{k^2} \frac{k(K-k)}{(K-1)} \sigma_b^2 + \frac{K}{k^2} \frac{k}{K} \frac{\sigma_w^2}{n} \frac{L-n}{L-1} \\
&= \frac{1}{k} \frac{K-k}{(K-1)} \sigma_b^2 + \frac{1}{k} \frac{\sigma_w^2}{n} \frac{L-n}{L-1}
\end{aligned}$$

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) , $i = 1 \dots n$.

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

Vzorčenje si lahko predstavljamo tako, da smo populacijo naključno uredili, nato pa v vzorec zajeli prvih n posameznikov. Ker imajo vsi vrstni redi enako verjetnost, bo na i -tem mestu z enako verjetnostjo katerikoli posameznik. Vsi pari (X_i, Y_i) imajo tako enako porazdelitev in zato je enaka tudi kovarianca X_i in Y_j .

- Naj bo $\gamma = \text{cov}(X_i, Y_i)$. Izračunajte kovarianco $\text{cov}(X_i, Y_j)$ za $i \neq j$.

Uporabimo isto idejo ko pri prejšnji nalogi - vsota vseh vrednosti iz populacije je konstanta, zato velja

$$\text{cov}(X_i, \sum_{j=1}^N Y_j) = \text{cov}(X_i, Y_i) + (N-1)\text{cov}(X_i, Y_j) = 0.$$

Velja torej (za $i \neq j$)

$$\text{cov}(X_i, Y_j) = -\frac{\gamma}{N-1}.$$

- Kovarianca med spremenljivkama X in Y je definirana kot

$$\text{cov}(X, Y) = \text{cov}(X_1, Y_1) = \frac{1}{N} \sum_{i=1}^N [(x_i - \mu)(y_i - \nu)] = \frac{1}{N} \sum_{i=1}^N x_i y_i - \mu \nu,$$

kjer smo z μ in ν označili povprečji $\mu = \frac{1}{N} \sum_{i=1}^N x_i$ in $\nu = \frac{1}{N} \sum_{i=1}^N y_i$.

S pomočjo vzorca bi kovarianco radi ocenili s cenilko $\hat{\gamma} = c [\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}]$.

Določite vrednost konstante c , da bo cenilka nepristranska.

Pričakovana vrednost cenilke je

$$E(\hat{\gamma}) = c \left[\sum_{i=1}^n E(X_i Y_i) - nE(\bar{X}\bar{Y}) \right] \quad (2)$$

Zaradi simetrije je $E(X_i Y_i) = E(X_j Y_j)$ za poljubna i in j . Vemo, da velja

$$\text{cov}(X_i, Y_i) = E(X_i Y_i) - E(X_i)E(Y_i) = E(X_i Y_i) - \mu\nu$$

Torej je $E(X_i Y_i) = \mu\nu + \gamma$. Oglejmo si še drugi člen na desni strani (2):

$$\begin{aligned} E(\bar{X}\bar{Y}) &= E \left[\frac{1}{n} \sum_{i=1}^n X_i \frac{1}{n} \sum_{j=1}^n Y_j \right] \\ &= \frac{1}{n^2} E \sum_{i=1}^n \left[X_i Y_i + X_i \sum_{j=1, i \neq j}^n Y_j \right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \left[E(X_i Y_i) + \sum_{j=1, i \neq j}^n E(X_i Y_j) \right] \\ &= \frac{1}{n^2} \sum_{i=1}^n [E(X_i Y_i) + (n-1)E(X_i Y_j)] \end{aligned}$$

Uporabimo rezultat

$$\begin{aligned} \text{cov}(X_i, Y_j) &= E(X_i Y_j) - E(X_i)E(Y_j) \\ E(X_i Y_j) &= \mu\nu - \frac{\gamma}{N-1} \end{aligned}$$

in zato

$$\begin{aligned} E(\bar{X}\bar{Y}) &= \frac{1}{n^2} n \left[\mu\nu + \gamma + (n-1) \left(\mu\nu + \frac{-\gamma}{N-1} \right) \right] \\ &= \frac{1}{n} \left[n\mu\nu + \gamma \left(1 - \frac{(n-1)}{N-1} \right) \right] \\ &= \frac{1}{n} \left[n\mu\nu + \gamma \frac{N-n}{N-1} \right] \end{aligned}$$

To vstavimo v enačbo (2)

$$\begin{aligned}
 E(\hat{\gamma}) &= c \left[\sum_{i=1}^n (\mu\nu + \gamma) - n \frac{1}{n} \left[n\mu\nu + \gamma \frac{N-n}{N-1} \right] \right] \\
 &= c \left[n\mu\nu + n\gamma - n\mu\nu - \gamma \frac{N-n}{N-1} \right] \\
 &= c \left[n\gamma - \gamma \frac{N-n}{N-1} \right] \\
 &= c\gamma \left[\frac{nN-n}{N-1} - \frac{N-n}{N-1} \right] \\
 &= c\gamma \frac{N(n-1)}{N-1}
 \end{aligned}$$

c mora biti torej enak $\frac{1}{n-1} \frac{N-1}{N}$.

- Kako bi ocenili korelacijo? Kaj vemo o nepristranskosti te ocene?

Uporabimo izpeljane formule za oceno kovariance in varianc:

$$\begin{aligned}
 \hat{\rho} &= \frac{\widehat{\text{cov}}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y} \\
 &= \frac{\frac{1}{n-1} \frac{N-1}{N} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1} \frac{N-1}{N} \sum_{i=1}^n (X_i - \bar{X})^2 \frac{1}{n-1} \frac{N-1}{N} \sum_{i=1}^n (Y_i - \bar{Y})^2}} \\
 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}
 \end{aligned}$$

Nepriustranskosti te ocene še nismo dokazali, saj pričakovana vrednost kvocienta ni enaka kvocientu pričakovanih vrednosti.

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 $100 - \text{starost}$ (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 ($\text{mean}(cova)$) bistveno odstopa od prave vrednosti ($\text{cov}(xi,yi)$). 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.