1 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_j) in rezultatom na testu (Y_j).
Za vsakega posameznika iz vzorca imamo torej par slučajnih spremenljivk (X_i, Y_i), $i = 1 \dots n$.

- Utemeljite, da je količina $cov(X_i, Y_j)$ 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 $\gamma = cov(X_i, Y_i)$. Izračunajte kovarianco $cov(X_i, Y_j)$ za $i \neq j$.

  Vsota vseh vrednosti iz populacije je konstanta, zato velja

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

  Velja torej (za $i \neq j$)

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

- Kovarianca med spremenljivkama X in Y je definirana kot

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

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

  Na vzorcu bi kovarianco radi ocenili s cenilko $\widehat{\gamma} = c\left[\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right]$. Določite vrednost konstante c, da bo cenilka nepristranska.

Pričakovana vrednost cenilke je

$$E(\widehat{\gamma}) = c\left[\sum_{i=1}^{n} E(X_iY_i) - nE(\bar{X}\bar{Y})\right] \tag{1}$$

Zaradi simetrije je $E(X_iY_i) = E(X_jY_j)$ za poljubna i in j. Vemo, da velja

$$cov(X_i, Y_i) = E(X_iY_i) - E(X_i)E(Y_i) = E(X_iY_i) - \mu\nu$$

Torej je $E(X_iY_i) = \mu\nu + \gamma$. Oglejmo si še drugi člen na desni strani (1):

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

Uporabimo rezultat

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

in zato

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

To vstavimo v enačbo (1)

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

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

- Kako bi ocenili korelacijo? Ali je taka ocena korelacije nepristranska? Preverite s simulacijo.

Uporabimo izpeljane formule za oceno kovariance in varianc:

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

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

```
> 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:

```
> (mean(cova)-cov(xi,yi))/sqrt(var(cova)/runs)
[1] 1.509540
>
> (mean(cora)-cor(xi,yi))/sqrt(var(cora)/runs)
[1] 6.66459
```

Odstopanje pri korelaciji je bistveno večje, medtem ko je odstopanje pri kovarianci v okviru naključne variabilnosti.

2 Enostavni slučajni vzorec, 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 $P(I_i = 1)$?

  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 $var(I_i)$ in $cov(I_i, I_i)$.

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

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

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

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

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

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

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:

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

3 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) 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

$$\mu = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{K} N_i \cdot \mu_i \tag{2}$$

Označimo 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 našteto 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 (2) 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

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

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

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

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

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

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

  Sedaj izrazimo še kovarianco:

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

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

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

- 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.

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

Velja:

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

in zato

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

- Kaj bi se razlikovalo v naših izračunih če bi šole vzorčili proporcionalno glede na njihovo velikost, tako da bi bila verjetnost, da je izbrana šola

i enaka $\frac{kN_i}{N}$?

Če bi bile vse šole enako velike, bi se spremenil le izračun kovariance. Ker ne vemo, kakšna bo velikost vzorca, vsota I_i ni več konstanta, zato ne moremo kovariance izračunati z istim "trikom" - potrebno jo bo izračunati po definiciji.