Končna populacija, CLI

Nataša Kejžar

Povzetek

Končna populacija

- N – število enot v populaciji
- vzorčimo brez ponavljanja (brez vračanja)
- $cov(X_i, X_j) = -\frac{1}{N-1}$
- $P(X_2 = x_j|X_1) \neq P(X_2 = x_j)$; pogojna in robna porazdelitev nista enaki
- $SE = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$; popravek za končnost
- $\hat{SE} = \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}}$
- popravek za končnost je pomemben, ko je N majhen in n predstavlja nezanemarljiv delež N

Centralni limitni izrek

- $\lim_{n\to\infty} P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) = \Phi(x)$
- X_i so iid. – neodvisne, enako porazdeljene spremenljivke (angl. independent identically distributed)
- varianca spremenljivke mora biti končna
  - X ima varianco σ^2
  - $\bar{X}$ ima varianco $\frac{\sigma^2}{n}$
- SE – standardna napaka (angl. standard error) je standardni odklon vzorčnega povprečja spremenljivke
\end{itemize}

Naloge

1. Končnost populacije vpliva na velikost standardne napake (SE), in sicer se ta zmanjša za koeficient $\sqrt{\frac{N-n}{N-1}}$.
   - Prikažite to razmerje med SE_{koncna} in $SE_{\text{neskoncna}}$ na grafu za različne velikosti končne populacije (in enako velikost vzorca).
2. V Sloveniji je 175 srednjih šol. Ministrstvo za izobraževanje, znanost in šport (MIZŠ) zanima, kakšno je povprečno število podpornih delavcev na srednjih šolah. Radi bi oceno, ki ne bi variirala več kot za 4 delavce. Najmanj kako velik vzorec naj raziskovalec izbere, da bo njegova ocena dovolj natančna?
   a. Katero dodatno populacijsko količino mora MIZŠ še povedati, da raziskovalec lahko naredi približen izračun?
   b. Naredite najprej izračun za primer neskončne populacije.
   c. Naredite pravi izračun (za primer končne populacije).
   d. Na grafu prikažite velikost vzorca v odvisnosti od standardnega odklona za neskončno populacijo. (`curve`)
   e. Na istem grafu prikažite z rdečo tudi rezultate, ki jih dobite, če upoštevate končnost populacije. Komentirajte rezultate.
   f. Kaj bi se v vaših izračunih spremenilo, če veste, da je spremenljivka normalno porazdeljena, MIZŠ pa vam pove le oceno standardnega odklona?
3. Pred volitvami naj velja naslednje: 55 % volilnih upravičencev bo volilo demokratsko stranko, ostali pa republikansko. Naredimo anketo in v vzorec izberemo 10 ljudi.
   a. Zapišite cenilko za delež, ki ga dobijo demokrati.
   b. Naj bo ta cenilka naša nova spremenljivka Y. Izpeljite njeno pričakovano vrednost in varianco.
   c. Izpeljite formulo za nepristransko cenilko za varianco Bernoullijeve spremenljivke. Kako iz nje dobimo nepristransko cenilko za standardno napako za delež?
   d. Kakšna je verjetnost, da bo na vzorcu 10 ljudi večina volila za republikansko stranko?
   e. Preverite to tudi s simulacijami.
4. Janez odgovarja na vprašanja o zgodovini. Trdi, da obvlada, predvidevamo, da zna odgovoriti na približno 90 % vprašanj. Na vsako vprašanje odgovori bodisi pravilno, ali pa nepravilno. Delež pravilnih odgovorov bi radi ocenili na 10 % natančno (širina 95 % intervala zaupanja). V nalogi upoštevajte, da delež lahko zapišemo kot povprečje.
   a. Zapišite cenilko za delež pravilnih odgovorov. Kaj veste o njeni porazdelitvi (eksaktni in aproksimativni)?
   b. Zapišite nepristransko cenilko za varianco cenilke. Pri tem uporabite, da je izraz
   $$\widehat{\sigma}^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$$
   nepristranska cenilka variance za spremenljivko iz $Bernoulli(p)$.
   c. Koliko vprašanj mu moramo zastaviti za želeno natančnost?
   d. Kaj se bo zgodilo z intervalom zaupanja, če v resnici zna manj kot predvidevamo?
5. Zapišite ekzaktno porazdelitev za delež in utemeljite, zakaj lahko pri dovolj velikih vzorcih rečemo, da je delež približno normalno porazdeljen.
6. Pokažite s simulacijami, da je korelacija med spremenljivkama X_i in X_j, kjer i in j označujeta $i$to in $j$to vrednost v **končni populaciji** $cor(X_i, X_j) = -1/(N-1)$.
   - Zapišite premislek, s pomočjo katerega ste se odločili za **velikost** in za **vrednosti** vaše končne populacije. *Pomoč:* Za izračun ene ocene korelacije je potrebno generirati veliko vzorcev. Iz vsakega vzorca shranite i-to in j-to vrednost in nato izračunajte korelacijo med i-timi in j-timi vrednostmi. Da boste videli, ali je *pričakovana vrednost* korelacije prava, morate ta postopek ponoviti velikokrat.

7. Predpostavite, da zdravilo ne naredi nikakršne razlike pri zmanjšanju sistoličnega tlaka pacientov ($\mu = 0$), standardni odklon razlike naj bo $\sigma = 50$, v vzorec vzamemo $n = 100$ bolnikov.

   a. Izračunajte pričakovani delež primerov, ko na vzorcu opazimo povprečno zmanjšanje za 10 mmHg, čeprav v populaciji ni sprememb po jemanju zdravila za primer, ko je populacija porazdeljena normalno. Preverite to tudi s simulacijami.
   b. Preverite s simulacijami pričakovani delež za primer, ko je populacija porazdeljena enakomerno. *Pomoč:* Najprej izpeljite, pri kakšnih vrednostih a in b ima enakomerna porazdelitev na intervalu $[a, b]$ povprečje 0 in standardni odklon 50.
   c. Ali bo/je delež za enakomerno porazdeljeno populacijo precej drugačen od normalno porazdeljene populacije? Zakaj?

8. Število decilitrov mineralne vode, ki jo posameznik popije v službi, je porazdeljeno po normalni porazdelitvi s povprečjem 6 dl in standardnim odklonom 2 dl. Podjetje oskrbuje svoje zaposlene dnevno s 650 dl mineralne vode. Podjetje ima 100 zaposlenih.

   a. Kakšna je verjetnost, da je količina mineralne vode, ki jo podjetje naroči za en dan, premalo za potrebe vseh zaposlenih?
   b. Kakšna je verjetnost, da količine mineralne vode vsaj enkrat v naslednjih 4 dneh ne bo dovolj? Predpostavite, da je količina mineralne vode, ki jo popijejo zaposleni, med dnevi neodvisna.
   c. Kakšna je verjetnost, da v naslednjem letu (365 dnevih) podjetje ne bo zadostilo porabi mineralne vode za zaposlene v več kot 2 dneh?

9. Spremenljivka X v populaciji naj bo porazdeljena po porazdelitvi, ki je mešanica dveh neodvisnih enakomernih porazdelitev: z verjetnostjo 0,9 $U(0, 1)$, z verjetnostjo 0,1 pa $U(0, 10)$. Spremenljivko zapišete kot

$$X = W \cdot Y + (1 - W) \cdot Z.$$

   a. Skicirajte porazdelitev spremenljivke X.
   b. Povejte, kako so porazdeljene spremenljivke W, Y in Z in kaj veste o njihovi odvisnosti in korelaciji.
   c. Izračunajte pričakovano vrednost in varianco za tako porazdelitev X.
   d. Naredite funkcijo, ki grafično pokaže, da centralni limitni izrek za vzorce iz take populacije velja.
   e. Komentirajte hitrost konvergence CLI.

   Pomoč: Za $U \sim Unif(a, b)$ velja $E(U) = \frac{b-a}{2}$ in $var(U) = \frac{(b-a)^2}{12}$.

10. Pokažite grafično, da se binomsko porazdeljena spremenljivka $X \sim Bin(n, \pi)$ za velike vrednosti n porazdeljuje približno normalno:

    a. Narišite histograme za nekaj različnih n-jev in $\pi = 0, 85$.
    b. Kakšni sta vrednosti μ in σ (parametra prilegajoče se normalne porazdelitve) in zakaj?
    c. Narišite čez histogram z največjim n še gostoto te normalne porazdelitve.

11. Za naslednje trditve povejte, ali so pravilne ali ne in svoj odgovor **dobro** utemeljite.

    a. Verjetnost, da je povprečje 20 vrednosti znotraj 0,4 standardnega odklona od populacijskega povprečja je večje kot verjetnost, da je povprečje 40 vrednosti znotraj 0,4 standardnega odklona od populacijskega povprečja.
    b. $P(\bar{X} > 4)$ je večje kot $P(X > 4)$, če je $X \sim N(8, \sigma^2)$.
    c. Če je $\bar{X}$ povprečje n vrednosti iz normalne porazdelitve s povprečjem μ in če je je c neko pozitivno število, potem se $P(\mu - c \leq \bar{X} \leq \mu + c)$ zmanjšuje, ko večamo n.

12. CLI govori o tem, da se vzorčno povprečje iid. spremenljivk porazdeljuje po normalni porazdelitvi s povprečjem μ in standardnim odklonom $\sigma/\sqrt{n}$. Imate spremenljivko, ki lahko zavzame vrednosti 1, 2 ali 3 (vsaka vrednost ima enako verjetnost).

    a. Kakšno je povprečje take spremenljivke in kakšen je standardni odklon?
    b. Opišite po korakih algoritem, ki bi grafično pokazal, da centralni limitni izrek velja za tako spremenljivko.
    c. Kako bi grafično ugotovili, pri kako velikem n je aproksimacija s CLI že dovolj dobra?

d. Spremenite spodnjo funkcijo tako, da boste lahko spreminjali število enot v vzorcu in vrednosti spremenljivke.
e. Ugotovite, pri katerem n je (na oko) aproksimacija z normalno porazdelitvijo že dovolj dobra.
f. Ugotovite s simulacijami, pri katerem n je aproksimacija z normalno porazdelitvijo dovolj dobra za spremenljivko, ki ima 20 različnih vrednosti, ki so enako verjetne (npr. vrednosti 1:20).

```
cli = function(){
  povprecja = NULL
  ponovi=10000
  vrednosti = 1:3
  for(i in 1:ponovi){
    vzorec=sample(vrednosti,size=20,replace=TRUE)
    povprecja = c(povprecja,mean(vzorec))
  }
  hist(povprecja,freq=FALSE,breaks=50)
  povp=sum(vrednosti)/length(vrednosti)
  vari=sum((vrednosti-povp)^2)/length(vrednosti)
  curve(dnorm(x,mean=povp,sd=sqrt(vari/20)),from=1,to=3,
        n=1000,add=TRUE,col="red",lwd=3)
}
```

13. Naredite 100 enot veliko populacijo iz normalne porazdelitve.
    a. Naredite funkcijo, ki grafično pokaže, da centralni limitni izrek za tako populacijo velja. Funkcija naj omogoča enostavno spreminjanje vhodne populacije.
    b. Poženite funkcijo za vzorce velikosti 4, 10, 30 in 50. Zapišite ugotovitve, komentirajte grafe.
    c. Kaj lahko rečete o konvergenci v tem primeru?
14. V vrtcu je 105 otrok. Vemo, da je verjetnost, da otrok zboli (in s tem ne pride v vrtec) običajno 1/15. Izračunajte, pri katerem številu manjkajočih otrok v vrtcu bi moralo vodstvo postati pozorno oziroma razglasiti epidemijo trenutne bolezni. V to številko bi radi imeli 99 % zaupanje.
    a. Izračunajte to z in brez aproksimacije z normalno porazdelitvijo.
    b. Komentirajte razlike in razlago ponazorite z grafom.
        - Poiščite informacije o zveznostnem popravku (angl. *continuity correction*) in razložite namen. Ali v vašem primeru pride prav?
15. Z naključnim vzorčenjem želimo oceniti delež volilcev ZA referendum. Recimo, da je v populaciji 55 % volilcev ZA referendum, 45 % volilcev pa PROTI. Pri naključnem vzorčenju vzemimo najprej odgovore 1500 volilcev.
    a. Simulirajte 1000 vzorcev, za vsakega izračunajte ocenjen delež volilcev ZA referendum in iz teh ocen pridobite 95% interval zaupanja za delež.
    b. Izpeljite formulo za nepristransko *oceno* populacijske variance.
    c. Izračunajte 95 % interval zaupanja za delež iz enega vzorca s pomočjo formule za IZ (vemo, da binomsko porazdelitev, ko je n velik, lahko aproksimiramo z $N(\mu_{\text{binom}}, \sigma_{\text{binom}})$).
    d. Primerjajte in komentirajte oba rezultata (95 % IZ dobljen s simulacijami in s formulo):
        - Kaj lahko rečete o širini obeh intervalov?
        - Komentirajte sredini obeh intervalov.
    e. Zanima nas tudi ocena za razliko v deležu ZA–PROTI (torej v populaciji srednjih 10%).
        - Zapišite izraz za vašo 'sestavljeno' spremenljivko in izpeljite pričakovano vrednost in varianco.
        - Izračunajte IZ za delež iz enega vzorca s pomočjo formule za IZ (uporabite aproksimacijo).
        - Simulirajte 95% interval zaupanja za razliko. Izračunajte IZ za število volilcev in za delež (uporabite aproksimacijo).

- Interpretirajte interval zaupanja.
- Izračunajte (izpeljite), najmanj kako velik vzorec bi morali vzeti, da bi pri dejanskem rezultatu 55 % ZA, s 95 % zaupanjem pravilno trdili, da bo referendum uspel?
  - Na podlagi prejšnje točke narišite in komentirajte graf (uporabite npr. `curve`) za velikost vzorca (y os) v odvisnosti od deleža volilcev ZA referendum (x os)?