Cenilke, nepristranskost, doslednost

Nataša Kejžar

Povzetek

Nepristranskost

- $E(\hat{\theta}) = \theta$
- dokazujemo za majhne vzorce, pri velikih prevlada doslednost (če velja)
- simulacije: gledamo, ali se povprečje ocen približuje pravi vrednosti (simuliramo *veliko* število ocen)

Doslednost

- $\forall \epsilon > 0 : \lim_{n \to \infty} P\left(|\hat{\theta} - \theta| > \epsilon\right) = 0$
- variabilnost ocen se manjša, ko se veča velikost vzorca, ocene konvergirajo proti pravi populacijski vrednosti
- simulacije: gledamo, ali se variabilnost povprečne ocene zmanjšuje, ko večamo velikost vzorca in, ali se približuje pravi populacijski vrednosti

Naloge

1. Za genotipe AA, Aa in aa v genih velja t.i. ravnotežje Hardy-Weinberg (Rice, str. 273), in sicer so frekvence pojavljanj genotipov kot v tabeli pod nalogo.

   a. Iz frekvence za genotip aa (n_{aa}) želimo oceniti parameter θ. Zapišite najenostavnejšo (intuitivno) cenilko.
   b. Izračunajte pričakovano vrednost cenilke za θ iz prvega vprašanja. Ali je cenilka nepristranska?
   c. Izračunajte varianco za spremenljivko X. (*Pomoč:* Izberite možne vrednosti x tako, da boste lahko izračunali $E(X)$. Ali ima $E(X)$ smiselno interpretacijo?)

X	AA	Aa	aa
p(X)	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2

2. Cenilka variance na vzorcu naj bo $\tilde{\sigma}^2 = 1/n \sum_{i=1}^{n}(x_i - \bar{x})^2$.

   a. Ali ta cenilka preceni/podceni pravo varianco? (Nekajkrat izračunajte oceno variance za vzorec velikosti 10 iz $N(120, 30^2)$.)

   b. Na predavanjih ste/boste na primeru pokazali, da ta cenilka ni nepristranska. Pokažite to še s simulacijami na velikosti vzorca 10 iz $N(120, 30^2)$.

      - definirajte funkcijo za varianco
      - definirajte funkcijo za simulacijo n ocen
      - narišite histogram $n = 1000$ ocen, na graf dodajte populacijsko vrednost (rdeče) in povprečno vrednost iz vzorca (modro)

   c. Simulacija je natančnejša, če je število ocen večje. Narišite graf povprečne ocene variance za različno število simulacij (npr. od 50 do 5000 v korakih po 50). Komentirajte.

3. Cenilka za standardni odklon je

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

   Zanjo ste/boste na predavanjih pokazali, da ni nepristranska. Pokažite to še s simulacijami. V ta namen uporabite vzorce velikosti 10 iz porazdelitve $N(120, 30^2)$.

   a. Narišite histogram ocen, populacijske vrednosti in povprečja za simulacijo 1000 vzorcev.
   b. Narišite graf, kjer prikažete zmanjševanje variabilnosti povprečnih ocen, ko se število vzorcev povečuje. Dorišite populacijsko vrednost in komentirajte rezultate.
   c. Kaj se zgodi, če namesto vzorcev velikosti 10 uporabljate vzorce velikosti 100? Komentirajte. Pokažite to tudi s pomočjo simulacij, grafov.
   d. Pokažite (npr. s pomočjo Jensenove neenakosti), v katero smer je cenilka za standardni odklon pristranska (precenjuje ali podcenjuje populacijsko vrednost).

4. Generirajte vzorec velikosti 30 enot s katerim bi lahko preverjali ocenjevanje pri univariatni linearni regresiji. Naj bo populacijski regresijski koeficient enak 2, porazdelitev neodvisne spremenljivke naj bo $N(10, 2)$, odvisna spremenljivka pa naj variira okoli neodvisne s standardnim odklonom 3.

   a. Zapišite model linearne regresije z znanimi količinami.
   b. Narišite razsevni diagram iz generiranih podatkov.
   c. Ocenite regresijsko konstanto in regresijski koeficient za te podatke (funkcija `lm`).

5. Cenilka za oceno regresijskega koeficienta po metodi najmanjših kvadratov je

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})(Y_i - \bar{Y})\right]}{\sum_{i=1}^{n}\left[(x_i - \bar{x})^2\right]}$$

a. Naredite funkcijo za izračun ocene za β_1. Preverite s funkcijo `lm`, da se oceni za β_1 iz vaše funkcije in funkcije `lm` ne razlikujeta.
b. Pokažite s simulacijami, da je ta cenilka nepristranska (vzorec naj bo velik 30 enot, gl. prejšnjo nalogo).

6. Eksponentna porazdelitev se uporablja v analizi preživetja (modeliranje časov od diagnoze do smrti), pri teoriji čakalnih vrst (čakalni časi) ipd. Gostota tako porazdeljene spremenljivke X ima obliko

$$f(x) = \lambda e^{-\lambda x} \qquad ; x \geq 0, \quad \lambda > 0$$

Pričakovana vrednost tako porazdeljene spremenljivke je $E(X) = \lambda^{-1}$. (Za vajo jo izpeljite po definiciji.)

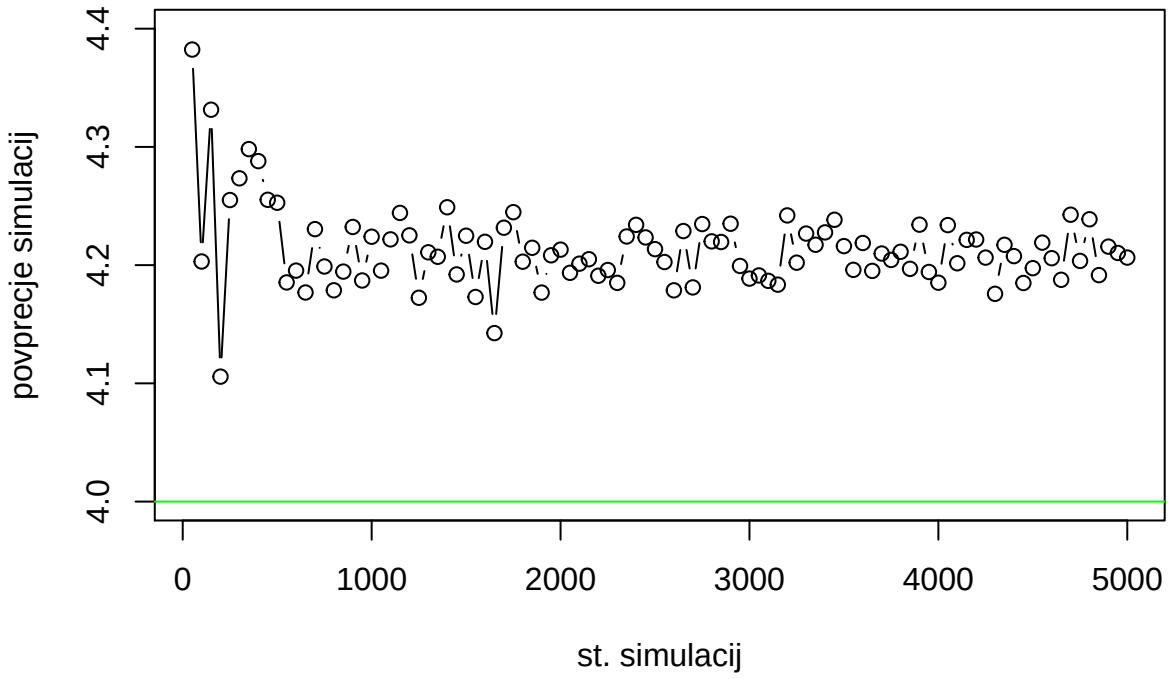
a. Za cenilko za λ izberemo funkcijo

$$h(x) = \frac{n}{\sum_{i=1}^{n} x_i}.$$

Utemeljite, zakaj je to smiselna cenilka.
b. Teoretično preverite, ali je cenilka nepristranska.
c. S simulacijami preverite, ali je cenilka nepristranska. V ta namen uporabite velikost vzorca 20, $X \sim Exp(4)$. Komentirajte ugotovitve glede na različno število ponovitev.
d. S simulacijami preverite, ali je cenilka dosledna.
e. Spodaj napisano simulacijo za (ne)pristranskost cenilke spremenite tako, da boste lahko preverili doslednost. *Pomoč:* Zanima vas konvergenca povprečnih ocen za različne velikosti vzorca (število simulacij naj bo dovolj veliko).
f. Kodo spremenite tako, da boste poleg povprečnih vrednosti cenilke za vsako velikost vzorca, v graf narisali tudi vse ocene.

```
#funkcija za izracun cenilke
cenilka = function(x){
  return(length(x)/sum(x))}

# definiramo funkcijo za simulacijo N vzorcev in izracun N ocen
simN = function(N){
  ocene = NULL # inicializiramo vektor simuliranih ocen
  for(i in 1:N){
    vzorec = rexp(20,rate=4)
    ocene = c(ocene,cenilka(vzorec))} # izracunamo cenilko
  return(ocene) #vrnemo vektor ocen
}
# (ne)pristranskost
stVzorcev = seq(50,5000,by = 50)
povpr = NULL
for(i in stVzorcev){
  povpr = c(povpr,mean(simN(i)))}
plot(stVzorcev,povpr,xlab='st. simulacij',
ylab="povprecje simulacij",main="",type="b",ylim=c(4,4.4))
abline(h=4,col="green") # populacijska vrednost
```



7. Pokažite s simulacijami, da cenilka za μ, ki je definirana kot povprečje prvih 5 vrednosti iz vzorca, ni dosledna, je pa nepristranska.

8. Iz normalne porazdelitve $N(120, 30^2)$ simulirajte:

   a. en vzorec velikosti $n_a = 10^4$
      - zanj izračunajte varianco, ki je podana s funkcijo `varianca` (gl. spodaj); naj bo to količina a

   b. vzorec iz točke (a) razdelite na 1000 vzorcev velikosti $n_b = 10$
      - za vsak vzorec izračunajte varianco, ki je podana s funkcijo `varianca` (gl. spodaj)
      - izračunajte povprečje teh ocen (to naj bo količina b)

   Dobro utemeljite, zakaj se količini a in b razlikujeta in kaj vsaka zase pove.

Koda za varianco:

```
varianca = function(x){1/length(x) * sum((x-mean(x))^2)}
```

9. Teoretično izračunajte povprečje in varianco za uniformno porazdeljeno spremenljivko $X \sim U(a, b)$.

   a. Generirajte velik vzorec ($n = 10^6$) in preverite, da to res velja.
   b. Zakaj je velik vzorec *dovolj* za potrditev teoretičnih izračunov?

10. Cenilka za standardni odklon je

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

    Zanjo ste na predavanjih pokazali, da ni nepristranska. Pokažite s simulacijami, da je cenilka dosledna. V ta namen uporabite vzorce iz porazdelitve $N(120, 30^2)$, ki jim velikosti večate od 10 do 500 po skokih za 10 (uporabite lahko funkcijo `seq`). Narišite najprej prazen razsevni diagram z mejami

```
plot(c(-1,-1),ylim=c(15,45),xlim=c(0,500),xlab="velikost vzorca",
     ylab="ocenjeni standardni odklon")
```

   a. Za vsako velikost vzorca $n_i = \{10, 20, \ldots, 500\}$ izračunajte oceno standardnega odklona 100-krat.

   b. Na graf narišite vseh 100 ocen za velikosti vzorca n_i (`points`) in dodajte kot rdečo točko še povprečje za teh 100 ocen.

c. Na koncu dorišite teoretično vrednost standardnega odklona (zelena vodoravna črta).

d. Pri kateri velikosti vzorca lahko rečemo (približno), da pristranskost zaradi asimptotskih lastnosti cenilke nima več bistvenega vpliva na oceno?

e. V katero smer je cenilka za standardni odklon pristranska? Podcenjuje ali precenjuje populacijsko vrednost? Zakaj? Dokažite s formulami.