Maruša Oražem

SEMINARSKA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

2019/20

Pred vami je seminarska naloga iz statistike, ki je sestavni del obveznosti pri tem predmetu. Predavatelj in asistent sva vam na voljo, če potrebujete nasvet. Naloge so večinoma iz učbenika:

John Rice: *Mathematical Statistics & Data Analysis*, Duxbury, 2007,

a morda so malo modificirane. V primeru težav z dostopom do knjige se oglasite pri asistentu.

Pri določenih nalogah si boste morali pomagati z računalnikom. Pri teh prosim priložite tako program ali datoteko kot tudi izhod (numerične rezultate, grafikone ...). Vsaj izhode programov prosim sproti prilagajte k rešitvam posameznih nalog: vse skupaj sestavite v enotno PDF datoteko ali pa preprosto natisnite. Prosim tudi, da izvozite izhod (še zlasti grafikone) iz programov za obdelavo preglednic (recimo excel, če ga boste že uporabili). Datoteke z besedili nalog ne pošiljajte nazaj.

Če stopnja značilnosti pri testu ni navedena, morate testirati tako pri $\alpha = 0{\cdot}01$ kot tudi pri $\alpha = 0{\cdot}05$.

Veliko uspeha pri reševanju!

1. V datoteki `Kibergrad` se nahajajo informacije o 43.886 družinah, ki stanujejo v mestu *Kibergrad*. Mesto ima štiri četrti: v severni četrti stanuje 10.149 družin, v vzhodni 10.390, v južni 13.457 in v zahodni 9.890. Za vsako družino so zabeleženi naslednji podatki (ne boste potrebovali vseh):

   - Tip družine (od 1 do 3)
   - Število članov družine
   - Število otrok v družini
   - Skupni dohodek družine
   - Četrt, v kateri stanuje družina:

     1: Severna
     2: Vzhodna
     3: Južna
     4: Zahodna
   - Stopnja izobrazbe vodje gospodinjstva (od 31 do 46)

   Vzemite enostavni slučajni vzorec 400 enot.

   a) Na podlagi vzorca ocenite povprečni dohodek v Kibergradu. Ocenite še standardno napako vaše ocene in postavite 95% interval zaupanja.

   b) Ali pri oceni povprečnega dohodka pomaga, če stratificiramo po četrtih? Izvedite prejšnjo točko na stratificiranem vzorcu s proporcionalno alokacijo. Primerjajte!

   c) Za cel Kibergrad izračunajte varianco dohodka med četrtmi in varianco znotraj četrti. Kako se to ujema z opažanji iz prejšnje točke?

2. Narediti želimo raziskavo na populaciji, ki ima K stratumov z velikostmi $N_1, N_2, \ldots, N_K$.

   a) Recimo, da so stroški raziskave enaki $C = C_0 + nC_1$, kjer je n število enot v vzorcu (C_0 je torej začetni strošek, C_1 pa je nadaljnji strošek na enoto). Pri danih sredstvih za raziskavo v višini C poiščite velikosti podvzorcev $n_1, n_2, \ldots, n_K$, pri katerih je varianca standardne cenilke populacijskega povprečja minimalna.

   b) Recimo sedaj, da se stroški opažanja lahko spreminjajo od stratuma do stratuma. Če je n_k število enot iz k-tega stratuma, ki so zajete v vzorec, naj bodo stroški raziskave enaki:

   $$C = C_0 + \sum_{k=1}^{K} n_k C_k \,.$$

   Spet pri danih sredstvih za raziskavo v višini C poiščite tiste velikosti podvzorcev, pri katerih je varianca cenilke populacijskega povprečja minimalna.

c) Naj se stroški raziskave izražajo na enak način kot v prejšnji točki, predpisano pa imamo natančnost raziskave, torej varianco cenilke. Poiščite tiste velikosti podvzorcev, pri katerih bodo stroški najmanjši.

Privzamete lahko naslednje:

- Da poznamo variance na celotnih stratumih. V praksi te variance ocenimo bodisi iz preteklih raziskav bodisi iz manjših *pilotnih vzorcev*.
- Da so deli vzorca na vseh stratumih dovolj veliki, tako da lahko zanemarite popravke zaradi celoštevilskosti (natančneje, sprememba velikosti za fiksno število je zanemarljiva). Ni pa nujno, da so deli vzorca na vseh stratumih precej manjši od samih stratumov.

3. Opazimo n neodvisnih realizacij zvezne porazdelitve z gostoto:

$$f(x \mid \sigma) = \begin{cases} \dfrac{\Gamma(3\alpha)}{\Gamma(\alpha)\,\Gamma(2\alpha)}\, x^{\alpha-1}(1-x)^{2\alpha-1} & ;\ 0 < x < 1 \\ 0 & ;\ \text{sicer,} \end{cases}$$

kjer je $\alpha > 0$ neznan parameter. Če je X slučajna spremenljivka s to gostoto, se da izračunati:

$$E(X) = \frac{1}{3}\,, \qquad \operatorname{var}(X) = \frac{2}{9(3\alpha+1)}\,.$$

a) Ocenite α po metodi momentov.

b) Poiščite enačbo, ki določa cenilko po metodi največjega verjetja. Kdaj ta cenilka sploh obstaja?

c) Poiščite asimptotično varianco cenilke po metodi največjega verjetja.

Namig: splača se izraziti s funkcijo *digama*, ki je logaritemski odvod funkcije gama. Preberite kaj o njej recimo na wikipediji.

4. V eni od klasičnih genetskih študij so preučevali podatke o rojstvih na Saškem v Nemčiji. V tabeli na desni je prikazano število otrok moškega spola pri 6115 družinah z 12 otroki. Če so spoli otrok neodvisni, bi moralo biti število moških potomcev, ki se rodijo družini z 12 otroki, porazdeljeno binomsko $\mathrm{Bin}(12, p)$; če sta verjetnosti za posamezen spol konstantni, so torej dani podatki realizacija 6115 enako porazdeljenih binomskih slučajnih spremenljivk. Preizkusite, ali so podatki v skladu s tem modelom (glejte razdelek 9.5 v knjigi). Bodite pozorni, da je preizkus zgolj aproksimativen in da morajo biti za sprejemljivo natančnost pričakovane frekvence vsaj 5 (razmislite, kako boste to dosegli). Zakaj bi bil lahko ta model napačen?

št. moških potomcev	št. družin
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3

Vir podatkov: A. Geissler: Beiträge zur Frage des Geschlechtsverhältnisses des Gebornen. *Z. K. Sachs. Stat. Bur.* **35** (1889), 1–24.

5. V modelu enostavne linearne regresije:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,; \quad i = 1, 2, \ldots, n \,,$$

kjer so napake ε_i neodvisne in porazdeljene normalno z matematičnimi upanji nič in enakimi variancami, ocenimo β_0 in β_1 po metodi največjega verjetja. Pokažite, da dobimo isto oceno kot po metodi najmanjših kvadratov.