Sara Bizjak

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*. 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
   - Mestna četrt, v kateri stanuje družina (od 1 do 4)
   - Stopnja izobrazbe vodje gospodinjstva:

     31: Brez šolske izobrazbe
     32: Dokončan prvi, drugi, tretji ali četrti razred osnovne šole
     33: Nedokončana osnovna šola, a končanih vsaj pet razredov
     34: Dokončana osnovna šola
     35: Dokončan prvi letnik srednje šole
     36: Dokončan drugi letnik srednje šole
     37: Dokončan tretji letnik srednje šole
     38: Dokončan četrti letnik srednje šole, a brez mature
     39: Poklicna matura
     40: Splošna matura
     41: Dokončan višji strokovni študij
     42: Dokončan visoki strokovni študij
     43: Dokončan univerzitetni študij prve stopnje
     44: Dokončan univerzitetni študij druge stopnje (magisterij)
     45: Magisterij po starem programu
     46: Doktorat znanosti

   a) Vzemite enostavni slučajni vzorec 200 družin in na njegovi podlagi ocenite delež družin v Kibergradu, v katerih vodja gospodinjstva nima srednješolske izobrazbe, t. j. niti poklicne niti splošne mature.

   b) Ocenite standardno napako in postavite 95% interval zaupanja.

   c) Vzorčni delež in ocenjeno standardno napako primerjajte s populacijskim deležem in pravo standardno napako. Ali interval zaupanja pokrije populacijski delež?

   d) Vzemite še 99 enostavnih slučajnih vzorcev in prav tako za vsakega določite 95% interval zaupanja. Narišite intervale zaupanja, ki pripadajo tem 100 vzorcem. Koliko jih pokrije populacijski delež?

   e) Izračunajte standardni odklon vzorčnih deležev za 100 prej dobljenih vzorcev. Primerjajte s pravo standardno napako za vzorec velikosti 200.

   f) Izvedite prejšnji dve točki še na 100 vzorcih po 800 družin. Primerjajte in razložite razlike s teorijo vzorčenja.

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:
   $$f(x \mid \sigma) = \begin{cases} \dfrac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2} \big[x(1-x)\big]^{\alpha-1} & ;\ 0 < x < 1 \\ 0 & ;\ \text{sicer} \end{cases} ,$$
   kjer je $\alpha > 0$ neznan parameter. Če je X slučajna spremenljivka s to gostoto, se da izračunati:
   $$E(X) = \frac{1}{2} \,, \qquad \operatorname{var}(X) = \frac{1}{4(2\alpha+1)} \,.$$

   a) Določite obliko porazdelitve v odvisnosti od α.

   b) Ocenite α po metodi momentov.

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

   d) 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 spodnji tabeli (prav tako pa tudi v datoteki `Samomori`) so podatki Ameriškega nacionalnega centra za statistiko zdravja o številu samomorov v ZDA v letu 1970 po mesecih.

Mesec	Število samomorov	Število dni
Januar	1867	31
Februar	1789	28
Marec	1944	31
April	2094	30
Maj	2097	31
Junij	1981	30
Julij	1887	31
Avgust	2024	31
September	1928	30
Oktober	2032	31
November	1978	30
December	1859	31

   a) Narišite histogram, pri katerem bodo širine stolpcev sorazmerne dolžinam mesecev. Premislite, čemu naj bodo sorazmerne višine stolpcev, da bo histogram verodostojen.

   b) Raziščite, ali podatki kažejo, da samomorilnost iz meseca v mesec variira, oziroma ali so skladni s predpostavko, da je samomorilnost skozi leto konstantna: uporabite primeren preizkus, glejte razdelek 9.5 v knjigi.

5. Naj bosta X in Y slučajni spremenljivki z:

$$\begin{aligned} E(X) = \mu_x\,, \quad & E(Y) = \mu_y\,, \\ \operatorname{var}(X) = \sigma_x^2\,, \quad & \operatorname{var}(Y) = \sigma_y^2\,, \end{aligned}$$

$$\operatorname{cov}(X, Y) = \sigma_{x,y}\,.$$

   Denimo, da opazimo X in želimo napovedati Y.

   a) Poiščite napoved oblike $\hat{Y} = \alpha + \beta X$, kjer α in β izberemo tako, da je srednja kvadratična napaka $E\big[(Y - \hat{Y})^2\big]$ minimalna. Matematični upanji, varianci in kovarianco poznamo.
   Namig: velja $E\big[(Y - \hat{Y})^2\big] = \big[E(Y) - E(\hat{Y})\big]^2 + \operatorname{var}(Y - \hat{Y})$.

   b) Pokažite, da se pri tako izbranih koeficientih *determinacijski koeficient* (kvadrat korelacijskega koeficienta) izraža v obliki:

$$r_{x,y}^2 = 1 - \frac{\operatorname{var}(Y - \hat{Y})}{\operatorname{var}(Y)}\,.$$