Dejan Perić

SEMINARSKA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

2020/21

Pred vami je seminarska naloga iz statistike, ki je sestavni del obveznosti pri tem predmetu. Predavatelj vam je na voljo, če potrebujete nasvet. Morda boste morali uporabiti kakšno različico statistične metode, ki je na predavanjih ali vajah nismo omenili. Lahko si pomagate z učbenikom:

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

ali katero drugo knjigo. V primeru težav z dostopom do učbenika se oglasite pri predavatelju.

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 tveganja pri preizkusu ni navedena, morate preizkusiti 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 (od 31 do 46)

   a) Vzemite enostavni slučajni vzorec 200 družin in na njegovi podlagi ocenite povprečno število otrok na družino v Kibergradu.

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

   c) Vzorčno povprečje in ocenjeno standardno napako primerjajte s populacijskim povprečjem in pravo standardno napako. Ali interval zaupanja iz prejšnje točke pokrije populacijsko povprečje?

   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 populacijsko povprečje?

   e) Izračunajte standardni odklon vzorčnih povprečij 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. V datoteki `TempPulz` se nahajajo odčitki telesnih temperatur (v Fahrenheitovih stopinjah) ter pulzov 65 moških (kodiranih z 1) in 65 žensk (kodiranih z 2). Privzemite, da sta telesna temperatura in pulz tako pri moških kot pri ženskah porazdeljena normalno.

   a) Ocenite povprečje in standardni odklon za telesno temperaturo posebej pri moških in posebej pri ženskah.

   b) Za povprečji iz prejšnje točke določite 95% intervala zaupanja.

   c) Preizkusite domnevo, da imajo moški in ženske v povprečju enako telesno temperaturo.

   Pretvornik med Fahrenheitovimi in Celzijevimi stopinjami: $x°\mathrm{F} = y°\mathrm{C}$, če je $y = 5(x - 32)/9$.

   Vir podatkov: A. L. Shoemaker: What's normal? Temperature, gender, and heart rate. *J. Stat. Edu.* **3**, št. 2 (1996).

3. V datoteki `Temp_LJ` se nahajajo izmerjene mesečne temperature v Ljubljani v letih od 1986 do 2020. Postavimo naslednja dva modela spreminjanja temperature s časom:

   - **Model A:** vključuje linearni trend in sinusno nihanje s periodo eno leto.
   - **Model B:** vključuje linearni trend in spreminjanje temperature za vsak mesec posebej.

   Očitno je model B širši od modela A.

   a) Preizkusite model A znotraj modela B.

   b) Pri modeliranju je nevarno privzeti preširok model: lahko bi recimo postavili model, po katerem je temperatura vsak mesec drugačna, neidvisno od ostalih mesecev, a tak model bi bil neuporaben za napovedovanje. *Akaikejeva informacija* nam pomaga poiskati optimalni model – izberemo tistega, za katerega je le-ta najmanjša. Akaikejeva informacija je sicer definirana z verjetjem, a pri linearni regresiji in Gaussovem modelu je le-ta ekvivalentna naslednji modifikaciji:

   $$\mathrm{AIC} := 2m + n \ln \mathrm{RSS}\,,$$

   kjer je m število parametrov, n pa je število opažanj. Kateri od zgornjih dveh modelov ima manjšo Akaikejevo informacijo?