

Andraž Čepič

PROJEKTNA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

2021/22

Pred vami je projektna 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.

Rešeno nalogo prosim oddajte v ustrezno rubriko na Učilnici v formatu PDF pod imenom `Projektna_naloga.pdf`.

Pri določenih nalogah si boste morali pomagati z računalnikom. Pri teh prosim priložite tako program ali datoteko kot tudi njegov izhod (numerične rezultate, grafikone ...). Vsaj izhode programov pa prosim še **sproti** prilagajte k rešitvam posameznih nalog v glavni datoteki. Na ta način prosim tudi priložite da izvozite izhod (še zlasti grafikone) programov za obdelavo preglednic (recimo excel, če ga boste že uporabili). Datoteke z besedili nalog ne oddajajte.

Če stopnja tveganja pri preizkusu ni navedena, morate preizkusiti tako pri $\alpha = 0.01$ kot tudi pri $\alpha = 0.05$.

Veliko uspeha pri reševanju!

NEKAJ NAPOTKOV ZA STAVLJENJE V T_EX-u oz. L^AT_EX-u

- Spremenljivke se dosledno stavijo ležeče, v T_EX-u torej med dolarji. Tako morate staviti, tudi če formula vsebuje en sam znak.
- Operatorji se stavijo pokončno, kar pa ne pomeni, da jih v T_EX-u postavimo kar izven dolarjev. Za najpogostejše operatorje so že naprogramirani ukazi.
- Če operator še ni definiran, ga sicer lahko stavimo recimo kot `\mathop{\mathrm{var}}` (ukaz `\mathop` je pomemben zaradi presledkov), a bistveno lažje je, če definiramo ukaz, recimo v preambuli:
`\usepackage{amsmath}`
`\DeclareMathOperator{\var}{var}`
- Dele formul je dostikrat smiselno ločiti z dodatnimi presledki. Temu so namenjeni ukazi `\,`, `\;`, `\>`, `\quad` in `\qquad`.
- Formule, ki so predolge za eno vrstico, je treba razlomiti. Najpogosteje se to naredi z uporabo okolij `array`, `align`, `align*`, `gather`, `gather*` in `split` (slednje znotraj okolja `equation` ali `equation*`). Za vse razen prvega potrebujemo knjižnico `amsmath`.
- Grafikone postavite **natančno** na mesto, kamor sodijo. Za to recimo v okolju `figure` uporabite določilo H (ne h), pri tem pa je treba v preambulo dati `\usepackage{float}`.
- Če boste decimalno vejico stavili kot običajno vejico, recimo `23,6`, vam bo T_EX naredil presledek, torej `23, 6`, ker bo mislil, da gre za naštevanje. Rešitev: `23{,}6`.

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. Dana je diskretna porazdelitev:

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ 2\theta/3 & \theta/3 & 2(1-\theta)/3 & (1-\theta)/3 \end{pmatrix},$$

kjer je $0 \leq \theta \leq 1$ neznan parameter. Opazimo naslednjih 10 neodvisnih realizacij te porazdelitve:

3, 0, 2, 1, 3, 2, 1, 0, 2, 2.

- a) Po metodi momentov ocenite θ .
- b) Ocenite standardno napako te ocene.
- c) Ocenite θ po metodi največjega verjetja.
- d) Ocenite standardno napako te ocene.
- e) V *bayesovski statistiki*, ki temelji na Bayesovi formuli, tudi parametre modeliramo kot slučajne spremenljivke in jim pripišemo določeno *apriorno porazdelitev*. Nato izračunamo pogojno porazdelitev glede na opažene vrednosti – *aposteriorno porazdelitev*.

Predpišimo, naj bo parameter θ apriorno porazdeljen enakomerno na intervalu $[0, 1]$. Poiščite aposteriorno porazdelitev, narišite jo ter določite njen modus in pričakovano vrednost; oboje lahko služi kot ocena za θ . Kaj opazite?

- f) Zdaj pa vzemimo malo drugačno parametrizacijo porazdelitve – uvedimo $\theta = \sin^2 \varphi$. Predpišimo, naj bo zdaj φ porazdeljen enakomerno na intervalu od 0 do $\pi/2$. Naredite enako kot v prejšnji točki še za to parametrizacijo: najprej na oba načina ocenite φ , nato pa izračunajte še odgovarjajoči vrednosti za θ . Primerjajte oceni z ocenama, dobljenima v prejšnji točki!

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:

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

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