Jimmy Zakeršnik

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 TEX-u oz. LATEX-u

- Spremenljivke se dosledno stavijo ležeče, v TEX-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 TEX-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 TEX 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:
     1: Družina z zakonskima ali zunajzakonskima partnerjema
     2: Enostarševska družina z očetom
     3: Enostarševska družina z materjo
   - 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 velikosti 500 in primerjajte dohodke glede na tip družine, tako da narišete vzporedne škatle z brki (glejte razdelek 10.6 v knjigi). Ali določen tip družine izstopa?

   b) Vzemite še štiri enostavne slučajne vzorce velikosti 500. Za vseh pet vzorcev spet narišite vzporedne škatle z brki, ki pripadajo dohodkom družin z dvema partnerjema. Komentirajte!

   c) Za celotni Kibergrad izračunajte varianco dohodka, pojasnjeno s tipom družine, in preostalo (rezidualno) varianco. Kako se to ujema z opažanji od prej?

2. *Model slučajnega sprehoda za kromatin.* Človeški kromosom je zelo velika molekula, dolga od 2 do 3 centimetre, ki vsebuje okoli 100 milijonov bazičnih parov *nukleotidov*, črk genetskega zapisa. Po drugi strani pa ima celično jedro, kjer se kromosomi nahajajo, premer komaj kakšno stotinko milimetra. Zato je neizogibno, da so kromosomi zaviti. Skupaj s posebnimi beljakovinami, imenovanimi *histoni*, so zapakirani v tako imenovani *kromatin*. Poenostavljeno povedano je to mešanica DNK in pomožnih beljakovin.

   V celični biologiji je zelo pomemben študij oblike kromatina pri različnih velikostnih merilih: to je med drugim ključnega pomena pri razumevanju celične delitve. Za ta namen so v seriji poskusov (Sachs et al., 1995; Yokota et al., 1995) na večjem številu celic fluorescenčno označili po dve krajši zaporedji nukleotidov (dolžine približno 40.000) na znanih lokacijah, nakar so s fluorescenčno mikroskopijo izmerili razdaljo med njunima projekcijama na določeno ravnino. Empirična porazdelitev teh razdalj nam da informacijo o obliki kromatina.

   V kemiji je dolga tradicija modeliranja oblike polimerov s slučajnimi sprehodi. Po tovrstnem modelu ima dvorazsežna razdalja med dvema točkama *Rayleighovo porazdelitev*:

$$f(r \mid \theta) = \begin{cases} \dfrac{r}{\theta^2} \exp\left(-\dfrac{r^2}{2\theta^2}\right) & ;\ r > 0 \\ 0 & ;\ \text{sicer} \end{cases} .$$

Razlog za to je v osnovi ta, da je po tem modelu vektor, ki pove premik od ene do druge točke, porazdeljen dvorazsežno normalno; krajši račun pokaže, da ima dvorazsežna razdalja potem Rayleighovo porazdelitev.

V datotekah `Kromatin_kratki`, `Kromatin_srednji` in `Kromatin_dolgi` so podane razdalje med pari zaporedij v treh različnih eksperimentih.

a) Poiščite splošno formulo za cenilko za θ po metodi največjega verjetja.

b) Poiščite splošno formulo za cenilko za θ po metodi momentov.

c) Izračunajte asimptotični srednji kvadratični napaki cenilk iz prejšnjih dveh točk. Je katera od cenilk nepristranska? Katera od cenilk je vsaj asimptotično boljša?

d) Za vsakega od treh danih eksperimentov izračunajte številsko oceno za θ po metodi največjega verjetja in ocenite standardno napako (zakaj je slednja relevantna?). Rezultate še grafično prikažite: narišite graf funkcije verjetja, na njem pa označite oceno za θ in standardno napako. Kaj opazite? Zakaj je tako?

e) Za vsakega od treh danih eksperimentov ocenite θ še po metodi momentov in spet ocenite standardno napako vaše ocene (spet zakaj je slednja relevantna?).

f) Za vsakega od eksperimentov narišite histogram meritev ter ustrezno dorišite ocenjeni gostoti po metodi največjega verjetja in po metodi momentov. Je videti razumno? Je kakšna omembe vredna razlika med obema ocenjenima gostotama?

Pri histogramu združite razdalje v enako široke razrede. Širino posameznega razreda določite v skladu z modificiranim Freedman–Diaconisovim pravilom.

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?