Nik Globočnik

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 (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) Narišite histogram dohodkov vseh družin v Kibergradu. Pri tem dohodke razdelite v enako široke razrede. Širino posameznega razreda določite v skladu s *Freedman–Diaconisovim pravilom*, po katerem le-ta znaša približno:

   $$l = \frac{2(q_{3/4} - q_{1/4})}{\sqrt[3]{n}}\,, \qquad (*)$$

   kjer sta $q_{1/4}$ in $q_{3/4}$ prvi in tretji kvartil, n pa je število enot. To vrednost nato smiselno zaokrožite na število oblike $k \cdot 10^r$, kjer je $k \in \{1, 2, 5\}$ in $r \in \mathbb{Z}$.

   b) Dorišite normalno gostoto, katere pričakovana vrednost in standardni odklon se ujemata s povprečjem in standardnim odklonom dohodka družine v Kibergradu. Kako dobro se prilega?

   c) Narišite kumulativno porazdelitveno funkcijo porazdelitve dohodkov družin v Kibergradu in primerjajte s kumulativno porazdelitveno funkcijo ustrezne normalne porazdelitve. Spet komentirajte, kako dobro se prilega.

   d) Narišite še primerjalni kvantilni (Q–Q) grafikon, ki porazdelitev dohodkov družin v Kibergradu primerja z normalno porazdelitvijo (glejte razdelek 9.8 v knjigi).

   e) Vzemite 1000 enostavnih slučajnih vzorcev velikosti 400 in narišite histogram vzorčnih povprečij dohodkov družin.

   f) Dorišite normalno gostoto, katere pričakovana vrednost se ujema s povprečnim dohodkom na družino v Kibergradu, standardni odklon pa s standardno napako za enostavni slučajni vzorec velikosti 400. Komentirajte, kako dobro se prilega.

   g) Za vzorčna povprečja podobno kot prej narišite še kumulativno porazdelitveno funkcijo in primerjalni kvantilni grafikon ter primerjajte z normalno porazdelitvijo. Komentirajte prileganje.

2. V datotekah `Delnice_Haliburton` in `Delnice_McDonalds` se nahajajo podatki o relativnih mesečnih donosih delnic teh dveh družb v letih od 1975 do 1999.

   a) Za vsako od delnic naredite histogram donosov in dorišite odgovarjajočo normalno porazdelitev. Komentirajte prileganje. Katera od delnic je bolj volatilna?

b) Za vsako od delnic naredite še primerjalni kvantilni (Q–Q) grafikon (glejte razdelek 9.8 v knjigi) in ponovno komentirajte prileganje.

Pri histogramu z dorisano normalno gostoto združite donose v enako široke razrede. Širino posameznega razreda določite v skladu z modificiranim Freedman–Diaconisovim pravilom.

3. V datoteki `Zobje` se nahajajo podatki o dolžini zob morskih prašičkov, ki so jim dodajali vitamin C v različnih količinah na dva različna načina: bodisi neposredno (kar je zakodirano z VC) bodisi s pomarančnim sokom (kar je zakodirano z OJ).

   a) Preizkusite, ali dodajanje vitamina C vpliva na rast zob.

   b) Kateri način dodajanja je učinkovitejši? Preizkusite, ali je razlika statistično značilna.