

**Timotej Stibilj**

## PROJEKTNA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

2022/23

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 ZIP. Tam naj bo zapakirana datoteka z imenom `Projektna_naloga.pdf`, v mapi **Priloge** pa naj bodo pomožne datoteke, npr. programi, s katerimi ste dobili rezultate. Toda v glavni datoteki morajo biti sproti vključeni vsi rezultati in grafikoni: imejte v mislih, naj, če je vse prav, pomožne datoteke ne bodo potrebne. Datoteke z besedili nalog ne oddajajte.

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

Rok oddaje je **ponedeljek, 4. september 2023**. Veliko uspeha pri reševanju!

## NEKAJ NAPOTKOV ZA STAVLJENJE V T<sub>E</sub>X-u oz. L<sup>A</sup>T<sub>E</sub>X-u

- Spremenljivke se dosledno stavijo ležeče, v T<sub>E</sub>X-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<sub>E</sub>X-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<sub>E</sub>X naredil presledek, torej `23, 6`, ker bo mislil, da gre za naštevaje. Rešitev: `23{,}6`.

1. V datoteki *Kibergrad* se nahajajo informacije o 43.886 družinah, ki stanujejo v mestu *Kibergrad*. Mesto ima štiri četrti: v severni četrti stanuje 10.149 družin, v vzhodni 10.390, v južni 13.457 in v zahodni 9.890. Za vsako družino so zabeleženi naslednji podatki (ne boste potrebovali vseh):

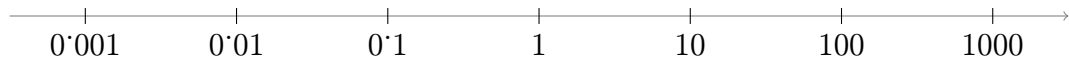
- Tip družine (od 1 do 3)
- Število članov družine
- Število otrok v družini
- Skupni dohodek družine
- Četrt, v kateri stanuje družina:
  - 1: Severna
  - 2: Vzhodna
  - 3: Južna
  - 4: Zahodna
- Stopnja izobrazbe vodje gospodinjstva (od 31 do 46)

Iz vsake četrti vzemite enostavni slučajni vzorec velikosti 100.

- (a) Primerjajte dohodke po četrtih, tako da narišete vzporedne škatle z brki (glejte razdelek 10.6 v knjigi). Je videti, da so določenih četrtih dohodki višji?
  - (b) Iz severne četrti vzemite še štiri enostavne slučajne vzorce velikosti 100. Za vseh pet vzorcev iz severne četrti spet narišite vzporedne škatle z brki. Komentirajte!
  - (c) Za celotni Kibergrad izračunajte varianco dohodka, pojasnjeno s četrtmi, in preostalo (rezidualno) varianco. Kako se to ujema z opažanji od prej?
2. V datoteki *ZarkiGama* se nahajajo podatki o časovnih razmikih med 3.935 zaznanimi fotoni, torej medprihodni časi (v sekundah).
- (a) Naredite histogram medprihodnih časov. Se vam zdi, da je model s porazdelitvijo gama plavzibilen?
  - (b) Ocenite parametra porazdelitve gama po metodi momentov in po metodi največjega verjetja. Primerjajte!  
*Namig:* potrebovali boste funkcijo *digama*, ki je logaritemski odvod funkcije gama. Preberite kaj o njej recimo na wikipediji.  
*Namig:* sistema enačb ne boste mogli rešiti eksaktno, to boste morali narediti numerično. Ena od učinkovitih možnosti je večrazsežna Newtonova metoda.
  - (c) Ocenjeni porazdelitvi dorišite na histogram. Je videti razumno?
  - (d) Histogram z dorisanimi gostotama narišite še na logaritemski lestvici. Lestvico transformirajte le na abscisni osi, vendar pa ustrezno transformirajte tudi dorisani gostoti.
  - (e) Je porazdelitev medprihodnih časov videti konsistentna s Poissonovim modelom, po katerem so ti časi porazdeljeni eksponentno?

Pri histogramih združite čase oz. njihove logaritme v enako široke razrede. Širino posameznega razreda določite v skladu z modificiranim Freedman–Diaconisovim pravilom.

Logaritemska lestvica pomeni, da položaj ustreza logaritmu, oznaka pa izvirni vrednosti, npr.:



3. V datoteki **Temp\_LJ** se nahajajo izmerjene mesečne temperature v Ljubljani v letih od 1986 do 2020. Z linearno regresijo preizkusite, ali januarske in julijske temperature naraščajo enako hitro ali z različno stopnjo.