Luka Ljevar

PROJEKTNA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

2024/25

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 $\alpha = 0{\cdot}01$ kot tudi pri $\alpha = 0{\cdot}05$.

Rok oddaje je **ponedeljek, 8. september 2025**. 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. Torej: slučajna spremenljivka X, ne slučajna spremenljivka X.

- Operatorji se stavijo pokončno, kar pa ne pomeni, da jih v TEX-u postavimo kar izven dolarjev. Za najpogostejše operatorje so že naprogramirani ukazi. Torej $\mathrm{var}(X)$, ne $var(X)$.

- 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}
```

- Levo in desno od formule v besedilu mora biti vedno beseda ali pa ločilo. Med drugim se torej povedi na začenja s formulo. Narobe je torej recimo: "X ima pričakovano vrednost 0, saj ima po trditvi 1 Y in z njim X simetrično porazdelitev." Pravilno: "Slučajna spremenljivka X ima pričakovano vrednost 0, saj ima Y in z njim X po trditvi 1 simetrično porazdelitev."

- Dele formul je dostikrat smiselno ločiti z dodatnimi presledki. Temu so namenjeni ukazi `\,` , `\;` , `\>` , `\quad` in `\qquad` . Med drugim to storite tudi, kadar je faktor v produktu ulomek. Primer:

$$\frac{1}{\sqrt{2\pi}}\, e^{-z^2/2},$$

kar je bilo stavljeno kot `\[ \frac{1}{\sqrt{2 \pi}} \, e^{- z^2/2} \]` .

- Za pogojevanje priporočam ukaz `\mid`, ki okoli navpičnice naredi ustrezen presledek. Če mora biti navpičnica višja, priporočam `\bigm|` , `\Bigm|` itd.

- Če pika ne označuje konca povedi, ji mora slediti ubežni ali pa trdi presledek, da TEX ne naredi prevelikega presledka. Stavite torej npr.
`Smolčki, tj.\ ljudje, rojeni 29.~februarja, naj bi rojstni dan praznovali le vsaka štiri leta.`

- Za tri pike (. . .) uporabljamo ukaz `\ldots` . Toda paketa `xelatex` in `lualatex` te tri pike, kadar so v besedilu (ne v formuli), naredita zelo stisnjene (…). Če želimo tri pike vselej staviti narazen, lahko ukažemo
`\renewcommand{\textellipsis}{$ \mathellipsis $}` ali

```
\renewcommand{\textellipsis}{%
  .\kern\fontdimen3\font
  .\kern\fontdimen3\font
  .\kern\fontdimen3\font
}
```

- Če poved zaključimo s tremi pikami, ne naredimo dodatne pike (tudi če so tiste tri pike del formule). Pač pa z ukazom `\spacefactor=3000{}` TEX-u povemo, naj naredi presledek, primeren za zaključek povedi.
- 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` .
- 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` .
- Za opombe, trditve, izreke, leme, dokaze in podobno priporočam okolje `amsthm` .
- Za spletne povezave priporočam ukaza `\url` in `\href` iz knjižnice `hyperref` .
- 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}` .

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:

     31: Brez šolske izobrazbe
     32: Dokončan prvi, drugi, tretji ali četrti razred osnovne šole
     33: Nedokončana osnovna šola, a končanih vsaj pet razredov
     34: Dokončana osnovna šola
     35: Dokončan prvi letnik srednje šole
     36: Dokončan drugi letnik srednje šole
     37: Dokončan tretji letnik srednje šole
     38: Dokončan četrti letnik srednje šole, a brez mature
     39: Poklicna matura
     40: Splošna matura
     41: Dokončan višji strokovni študij
     42: Dokončan visoki strokovni študij
     43: Dokončan univerzitetni študij prve stopnje
     44: Dokončan univerzitetni študij druge stopnje (magisterij)
     45: Magisterij po starem programu
     46: Doktorat znanosti

   (a) Vzemite enostavni slučajni vzorec 200 družin in na njegovi podlagi ocenite delež družin v Kibergradu, v katerih vodja gospodinjstva nima srednješolske izobrazbe, tj. niti poklicne niti splošne mature.

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

   (c) Vzorčni delež in ocenjeno standardno napako primerjajte s populacijskim deležem in pravo standardno napako. Ali interval zaupanja pokrije populacijski delež?

   (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 populacijski delež?

   (e) Izračunajte standardni odklon vzorčnih deležev 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.

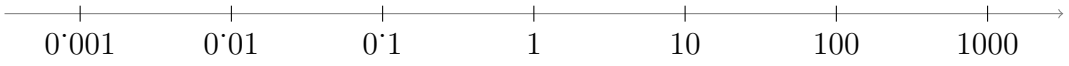
2. Slučajna spremenljivka X ima *log-normalno* porazdelitev, če je $\ln X$ porazdeljena normalno. Ta porazdelitev včasih služi kot model za asimetrične porazdelitve s težkimi repi. Preučite, ali so podatki iz datoteke `Hobotnice`, ki prikazujejo hrbtne dolžine različnih vrst hobotnic, vsaj približno v skladu z log-normalnim modelom, tako da narišete:

   - histogram z dorisano ustrezno log-normalno gostoto;
   - histogram na logaritemski lestvici z dorisano ustrezno normalno gostoto;
   - primerjalni kvantilni (Q–Q) grafikon na logaritemski lestvici.

   Komentirajte!

   Pri histogramih združite hrbtne dolžine oz. njihove desetiške 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.:

   Pri histogramu transformirajte lestvico le na abscisni osi, vendar pa ustrezno transformirajte tudi dorisano gostoto.

   Za primerjalni kvantilni grafikon glejte razdelek 9.8 v knjigi. Ko ga narišete na logaritemski lestvici, transformirajte obe osi.

3. V datoteki `Temp_MN` se nahajajo podatki o temperaturi na nekem kraju v Minnesoti, ZDA, v obdobju od leta 2015 do 2018, ob določeni uri dneva. Določeni dnevi manjkajo.

   Izračunajte 95% napovedni interval za temperaturo na dan, ki sledi zadnjemu, pri več različnih modelih:

   (a) Privzamemo, da temperature tvorijo kar Gaussov beli šum.

   (b) Privzamemo model, po katerem se pričakovana temperatura sinusno spreminja z dnevom v letu, šumi, ki pripadajo dnevnim temperaturam, pa so neodvisni. Učinek globalnega segrevanja zanemarimo.

   (c) Privzamemo model, po katerem imajo neodvisne šume *spremembe* temperatur med zaporednima dnevoma, pričakovane vrednosti teh sprememb pa se sinusno spreminjajo z dnevom v letu, pri čemer je sinusoida centrirana (torej tudi tu zanemarimo učinek globalnega segrevanja). Poskusite pravilno vključiti vse podatke, če ne gre, pa se omejite le na dneve, ki si neposredno sledijo.

   Opomba. Učinek globalnega segrevanja bi imelo smisel vključiti, če bi imeli podatke za daljše obdobje ali pa če bi bilo časovno obdobje večkratnik leta.