Luka Andrenšek

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*. 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)

   Vzemite enostavni slučajni vzorec 400 enot.

   (a) Na podlagi vzorca ocenite število otrok na družino. Ocenite še standardno napako vaše ocene in postavite 95% interval zaupanja.

   (b) Ali pri oceni povprečnega števila otrok pomaga, če stratificiramo po četrtih? Izvedite prejšnjo točko na stratificiranem vzorcu s proporcionalno alokacijo. Primerjajte!

2. Eksponentna porazdelitev često služi kot model za porazdelitev življenjske dobe, precej tudi zaradi matematične preprostosti. V datoteki `Kevlar` se nahajajo podatki za 101 vlakno materiala Kevlar 49/epoksi, ki se uporablja tudi pri Space Shuttlu. Prikazana so trajanja vlaken pri 90% obremenitvi, preden je prišlo do porušitve.

   (a) Narišite primerjalni kvantilni (Q–Q) grafikon, ki empirično porazdelitev primerja s standardno eksponentno porazdelitvijo (glejte razdelek 9.8 v knjigi). Kako je videti?

   (b) Kako so videti primerjalni kvantilni grafikoni, ki primerjajo dve eksponentni porazdelitvi?

   (c) Po metodi največjega verjetja poiščite tisto eksponentno porazdelitev, ki se najbolje prilega danim podatkom.

   (d) Na primerjalni kvantilni grafikon iz točke (a) dorišite ustrezno črto. Je vtis zdaj kaj drugačen?

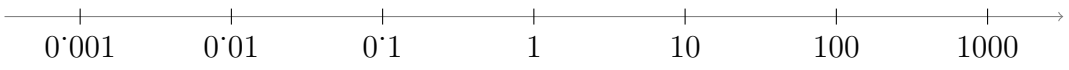
   (e) Narišite viseči histogram iz razlik korenov frekvenc (glejte razdelek 9.7 v knjigi). Kako je videti prileganje?

   (f) Točke (a), (b), (d) in (e) naredite še na logaritemski lestvici in ponovno komentirajte prileganje. Pri primerjalnem kvantilnem grafikonu transformirajte obe osi, pri visečem histogramu pa le abscisno os.

Pri visečih histogramih združite trajanja vlaken oz. njihove desetiške logaritme v enako široke razrede. Širino posameznega razreda določite v skladu z modificiranim Freedman–Diaconisovim pravilom. Nadalje premislite naslednje:

- Približno kakšno porazdelitev imajo frekvence, če so razredi fiksni, podatki neodvisne in enako porazdeljene slučajne spremenljivke, pričakovane frekvence dovolj velike, razredi pa vendar še dovolj ozki?
- Približno kakšno porazdelitev imajo kvadratni koreni teh frekvenc? Pri tem korensko funkcijo v glavnini dogajanja aproksimirajte z ustrezno linearno funkcijo. Utemejite, da je napaka, ki jo naredimo, majhna v primerjavi z variacijo funkcije v glavnini dogajanja.
- Kako torej interpretiramo odstopanja v visečem histogramu?

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



Vir podatkov: R. E. Barlow, R. H. Toland, T. Freeman: A Bayesian analysis of stress-rupture life of Kevlar/epoxy spherical pressure vessels. V *Proceedings of the Canadian Conference in Applied Statistics*, urednik T. D. Dwivedi. Marcel-Dekker, New York, 1984.

3. V datoteki `Temp_LJ` se nahajajo izmerjene mesečne temperature v Ljubljani v letih od 1994 do 2023.
   (a) Modelirajte spreminjanje temperature z enostavno linearno regresijo. Kako hitro se podnebje segreva? Je linearni trend spreminjanja temperature s časom statistično značilen? Vzemite standardno stopnjo tveganja 5%.
   (b) Postavite model, ki vključuje tudi letno nihanje temperature. Kako je zdaj z linearnim trendom spreminjanja temperature s časom? Primerjajte p-vrednosti obeh preizkusov.
   (c) Napovejte januarsko in povprečno letno temperaturo leta 2044. Za oboje konstruirajte tudi 95% napovedni interval.