Pantner Jan

PROJEKTNA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

2023/24

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, 9. september 2024**. 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` .

- 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)

   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. Pri najlonskih palicah so preizkušali lomljivost (Bennett in Franklin, 1954). V podobnih okoliščinah so ulili 280 palic in vsako od njih preizkusili na petih mestih. Rezultati poskusa so prikazani v tabeli na desni.

št. lomov	št. palic
0	157
1	69
2	35
3	17
4	1
5	1

   Če ima palica enakomerno strukturo, bi moralo biti število mest, na katerih se je zlomila, porazdeljeno binomsko Bin$(5, p)$ za določen neznan p. To naj bo naša osnovna ničelna domneva. Privzamemo tudi, da so palice med seboj neodvisne.

   (a) Ob predpostavki osnovne ničelne domneve ocenite p po metodi največjega verjetja.

   (b) Združite zadnje tri vrednosti in s posplošenim Pearsonovim preizkusom hi kvadrat preizkusite osnovno ničelno domnevo proti alternativni domnevi, da ima število lomov katero drugo porazdelitev (glejte razdelek 9.5 v knjigi). Še vedno privzamemo, da ima število lomov na vseh palicah enako porazdelitev.

(c) Za $i = 1, 2, \ldots, n$ naj bodo dana neodvisna opažanja $X_i \sim \mathrm{Bin}(m_i, p_i)$, kjer so parametri m_i znani, parametri p_i pa neznani. Razvijte preizkus na podlagi razmerja verjetij, ki bo preizkusil ničelno domnevo, da so vsi parametri p_i enaki, proti alternativni domnevi, da temu ni tako.

(d) Uporabite preizkus iz prejšnje točke na danih podatkih, vedite pa, da Wilksovega izreka ne morete uporabiti (premislite, zakaj pogoji niso izpolnjeni). Namesto tega uporabite metodo *bootstrap*: simulirajte 10.000 vrednosti preizkusne statistike pri ničelni domnevi, pri čemer za p vzemite oceno iz točke (a). Nato poglejte, koliko teh vrednosti presega vrednost preizkusne statistike, izračunano na konkretnih podatkih. Na podlagi tega ustrezno sklepajte, kaj storiti z ničelno domnevo.

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 v letu 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, pri katerem je lahko pričakovana temperatura prav vsak mesec poljubna (torej npr. pričakovane januarske temperature za različna leta ne bi bile med seboj povezane). A tak model bi bil za za napovedovanje popolnoma neuporaben. *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:

   $$\mathrm{AIC} := 2p + n \ln \mathrm{RSS}\,,$$

   kjer je p število parametrov, n pa je število opažanj. Kateri od zgornjih dveh modelov ima manjšo Akaikejevo informacijo?