Vito Založnik

SEMINARSKA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

2020/21

Pred vami je seminarska 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.

Pri določenih nalogah si boste morali pomagati z računalnikom. Pri teh prosim priložite tako program ali datoteko kot tudi izhod (numerične rezultate, grafikone ...). Vsaj izhode programov prosim sproti prilagajte k rešitvam posameznih nalog: vse skupaj sestavite v enotno PDF datoteko ali pa preprosto natisnite. Prosim tudi, da izvozite izhod (še zlasti grafikone) iz programov za obdelavo preglednic (recimo excel, če ga boste že uporabili). Datoteke z besedili nalog ne pošiljajte nazaj.

Če stopnja tveganja pri preizkusu ni navedena, morate preizkusiti tako pri $\alpha = 0{\cdot}01$ kot tudi pri $\alpha = 0{\cdot}05$.

Veliko uspeha pri reševanju!

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 datoteki `TempPulz` se nahajajo odčitki telesnih temperatur (v Fahrenheitovih stopinjah) ter pulzov 65 moških (kodiranih z 1) in 65 žensk (kodiranih z 2). Privzemite, da sta telesna temperatura in pulz tako pri moških kot pri ženskah porazdeljena normalno.

a) Ocenite povprečje in standardni odklon za telesno temperaturo posebej pri moških in posebej pri ženskah.

b) Za povprečji iz prejšnje točke določite 95% intervala zaupanja.

c) Preizkusite domnevo, da imajo moški in ženske v povprečju enako telesno temperaturo.

Pretvornik med Fahrenheitovimi in Celzijevimi stopinjami: $x°\mathrm{F} = y°\mathrm{C}$, če je $y = 5(x - 32)/9$.

Vir podatkov: A. L. Shoemaker: What's normal? Temperature, gender, and heart rate. *J. Stat. Edu.* **3**, št. 2 (1996).

3. V neki raziskavi:

http://www.statsci.org/data/oz/ms212.html

so študentom merili pulz. Vsakemu študentu so pulz izmerili dvakrat. Določeni so imeli med obema meritvama fizično obremenitev (tek na mestu), določeni ne. Podatki so zbrani v tabeli `Pulz`, pri čemer imajo stolpci naslednje pomene:

`VISINA`	telesna višina
`TEZA`	telesna teža
`STAROST`	starost v letih
`SPOL`	1=moški, 2=ženski
`KADI`	1=kadilec, 2=nekadilec
`ALKOHOL`	1=pije, 2=ne pije
`VADBA`	1=vadi veliko, 2=vadi zmerno, 3=vadi malo ali pa sploh ne
`OBREMENITEV`	1=obremenitev, 2=brez obremenitve
`PULZ1`	prva meritev pulza
`PULZ2`	druga meritev pulza
`LETO`	leto meritve (1993–1998)

a) Raziščite, katere od dejavnikov višina, teža, spol in vadba je smiselno vključiti v linearni model, ki bo opisoval odvisnost prve meritve pulza od teh dejavnikov: poiščite model z najmanjšo *Akaikejevo informacijo*. Slednja nam pomaga izbrati le bistvene dejavnike. Akaikejeva informacija je sicer definirana z verjetjem, a pri linearni regresiji in Gaussovem modelu je le-ta ekvivalentna naslednji modifikaciji:

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

kjer je m število parametrov, n pa je število opažanj.

b) Za vsakega od dejavnikov kajenje in alkohol vaš model preizkusite proti širšemu modelu, ki poleg dejavnikov, izbranih v prejšnji točki, vključi tudi še dodatni dejavnik.