Klara Golob

SEMINARSKA NALOGA IZ STATISTIKE

UL FMF, Matematika — univerzitetni študij

2019/20

Pred vami je seminarska naloga iz statistike, ki je sestavni del obveznosti pri tem predmetu. Predavatelj in asistent sva vam na voljo, če potrebujete nasvet. Naloge so večinoma iz učbenika:

John Rice: *Mathematical Statistics & Data Analysis*, Duxbury, 2007,

a morda so malo modificirane. V primeru težav z dostopom do knjige se oglasite pri asistentu.

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 značilnosti pri testu ni navedena, morate testirati 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) Vzemite enostavni slučajni vzorec 200 družin in na njegovi podlagi ocenite povprečno število otrok na družino v Kibergradu.

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

   c) Vzorčno povprečje in ocenjeno standardno napako primerjajte s populacijskim povprečjem in pravo standardno napako. Ali interval zaupanja iz prejšnje točke pokrije populacijsko povprečje?

   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 populacijsko povprečje?

   e) Izračunajte standardni odklon vzorčnih povprečij 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. Populacijo sestavljajo trije stratumi, prva dva imata 1000, tretji pa ima 500 enot. Iz vsakega stratuma vzamemo enostavni slučajni vzorec desetih enot in vrednosti spremenljivke pridejo:

1. stratum:	94	99	106	106	101	102	122	104	97	97
2. stratum:	183	183	179	211	178	179	192	192	201	177
3. stratum:	343	302	286	317	289	284	357	288	314	276

   Ocenite populacijsko povprečje in standardno napako vaše cenilke ter poiščite aproksimativni 95% interval zaupanja.

3. 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 dorisanima gostotama narišite še na logaritemski lestvici. Lestvico transformirajte le na abscisni osi, vendar pa ustrezno transformirajte tudi dorisani gostoti.

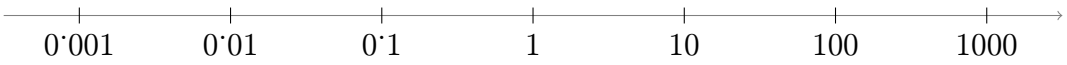
e) Je porazdelitev medprihodnih časov videti konsistentna s Poissonovim modelom, po katerem so ti časi porazdeljeni eksponentno?

Pri histogramih združite čase oz. njihove logaritme v enako široke razrede. Širino posameznega razreda določite v skladu 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 $r \in \mathbb{Z}$.

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



4. Recimo, da opazimo eno vrednost statistične spremenljivke X, porazdeljene enakomerno na intervalu $[0, \theta]$. Preizkusimo ničelno domnevo $H_0\colon\ \theta = 1$ proti alternativni domnevi $H_1\colon\ \theta = 2$.

   a) Poiščite preizkus, ki ima stopnjo tveganja $\alpha = 0$. Kolikšna je njegova moč?

   b) Za $0 < t < 1$ si oglejte preizkus, ki ničelno domnevo zavrne pri $X \leq t$. Kolikšni sta njegova stopnja tveganja in moč?

   c) Naj bo spet $0 < t < 1$. Kolikšni sta stopnja tveganja in moč preizkusa, ki ničelno domnevo zavrne pri $X \geq 1 - t$?

   d) Poiščite še kakšen preizkus, ki ima enako stopnjo tveganja in moč kot tisti iz prejšnje točke.

   e) Določite območje zavrnitve pri preizkusu na podlagi razmerja verjetij v odvisnosti od predpisane maksimalne stopnje tveganja. Kdaj je ta preizkus eksakten?

   f) Kaj se zgodi s preizkusom na podlagi razmerja verjetij, če ničelno in alternativno domnevo zamenjamo, torej preizkusimo $H_0\colon\ \theta = 2$ proti $H_1\colon\ \theta = 1$?

g) Za situacijo iz prejšnje točke predlagajte še kakšen drug, eksakten preizkus in primerjajte moči obeh preizkusov.

5. Naj bosta X in Y slučajni spremenljivki z:

$$E(X) = \mu_x\,, \quad E(Y) = \mu_y\,,$$
$$\mathrm{var}(X) = \sigma_x^2\,, \quad \mathrm{var}(Y) = \sigma_y^2\,,$$
$$\mathrm{cov}(X, Y) = \sigma_{x,y}\,.$$

Denimo, da opazimo X in želimo napovedati Y.

a) Poiščite napoved oblike $\hat{Y} = \alpha + \beta X$, kjer α in β izberemo tako, da je srednja kvadratična napaka $E\left[(Y - \hat{Y})^2\right]$ minimalna. Matematični upanji, varianci in kovarianco poznamo.
Namig: velja $E\left[(Y - \hat{Y})^2\right] = \left[E(Y) - E(\hat{Y})\right]^2 + \mathrm{var}(Y - \hat{Y})$.

b) Pokažite, da se pri tako izbranih koeficientih *determinacijski koeficient* (kvadrat korelacijskega koeficienta) izraža v obliki:

$$r_{x,y}^2 = 1 - \frac{\mathrm{var}(Y - \hat{Y})}{\mathrm{var}(Y)}\,.$$