Klementina Pirc

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 $\alpha = 0.01$ kot tudi pri $\alpha = 0.05$.

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:

     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, t. j. 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. Recimo, da želimo oceniti skupno vrednost inventarja, ki sestoji iz N enot. Vsaka enota ima knjigovodsko in dejansko vrednost. Knjigovodsko vrednost vsake enote, ki jo označimo z X, dobro poznamo, težje pa je določiti dejansko vrednost, ki jo označimo z Y. Kar želimo oceniti, je vsota dejanskih vrednosti vseh enot.

   Bolj formalno, za poljubni spremenljivki U in V, definirani na enotah inventarja (ki bosta vselej funkciji spremenljivk X in Y) uvedimo naslednje oznake:

   - τ_U naj označuje vsoto vrednosti spremenljivke U na vseh enotah inventarja;
   - μ_U naj označuje povprečno vrednost spremenljivke U na celotnem inventarju, t. j. $\mu_U = \tau_U/N$;
   - σ_U naj označuje standardni odklon spremenljivke U na celotnem inventarju;
   - $\rho_{U,V}$ naj označuje korelacijski koeficient med U in V, spet na celotnem inventarju.

   Želimo torej oceniti τ_Y. Za ta namen vzamemo enostavni slučajni vzorec velikosti n in kot običajno za dano spremenljivko U označimo z $\bar{U}$ njeno vzorčno povprečje. Oglejmo si tri možne cenilke:

   - $\hat{\tau}_Y^{(0)} := N\bar{Y}$;
   - $\hat{\tau}_Y^{(1)} := \tau_X + N(\bar{Y} - \bar{X})$;
   - $\hat{\tau}_Y^{(2)} := \dfrac{\bar{Y}}{\bar{X}}\,\tau_X$ (v tem primeru privzamemo, da so knjigovodske vrednosti na vseh enotah strogo pozitivne).

   a) Dokažite, da je cenilka $\hat{\tau}_Y^{(1)}$ nepristranska, in izrazite njeno varianco s σ_X, σ_Y in $\rho_{X,Y}$.

   b) Privzemimo, da sta X in $Y - X$ na populaciji nekorelirani (če populacijo opremimo z verjetnostno mero, na kateri so vse enote enako verjetne). Kdaj ima cenilka $\hat{\tau}_Y^{(1)}$ nižjo varianco kot cenilka $\hat{\tau}_Y^{(0)}$?

   c) Privzemimo, da sta X in Y/X na populaciji neodvisni (spet če populacijo opremimo z verjetnostno mero, na kateri so vse enote enako verjetne). Kdaj ima cenilka $\hat{\tau}_Y^{(2)}$ asimptotično (v določeni limiti, ko gre n proti neskončno) nižjo varianco kot cenilka $\hat{\tau}_Y^{(0)}$? Natančneje, kdaj je:

   $$\limsup \frac{\hat{\tau}_Y^{(2)}}{\hat{\tau}_Y^{(0)}} < 1\,?$$

   Odgovor izrazite z μ_X, σ_X, μ_Q in σ_Q, kjer je $Q = Y/X$. Za asimptotično varianco cenilke $\hat{\tau}_Y^{(2)}$ glejte razdelek 7.4 v knjigi.

   Opomba. Varianco je smiselno gledati, ker je cenilka $\tau_Y^{(2)}$ asimptotično nepristranska (tega ni treba dokazovati).

3. Ta naloga se nanaša na študijo, kjer so ocenjevali velikost populacije grenlandskih kitov (*Balaena mysticetus*, Raftery in Zeh, 1993). Statistični postopki pri oceni skupnega števila skupaj s standardno napako so bili dokaj zapleteni. Ta naloga

zajema le njihov majhen del, in sicer porazdelitev hitrosti plavanja teh kitov. Zbrali so pare opažanj istega kita na dveh različnih lokacijah in na njihovi podlagi ocenili hitrost tega kita. To so storili za 210 kitov. Hitrosti so pretvorili v čase $t_1, t_2, \ldots, t_{210}$: vsak od njih pomeni čas, v katerem je ustrezen kit preplaval en kilometer. Podatki so zbrani v datoteki `Kiti`.

a) Naredite histogram opaženih časov. V nadaljevanju bomo privzeli statistični model, po katerem ima čas, v katerem kit preplava en kilometer, porazdelitev *gama*. Se vam ta model glede na dobljeni histogram zdi plavzibilen?

b) Ocenite parametra porazdelitve gama po metodi momentov.

c) Ocenite parametra porazdelitve gama po metodi največjega verjetja in primerjajte z oceno iz prejšnje točke.
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 Newtonova metoda.

d) Porazdelitvi, ocenjeni v prejšnjih dveh točkah, dorišite na histogram. Se vam prileganje zdi razumno?

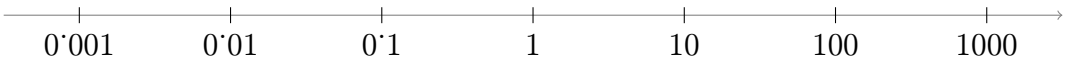
e) Histogram z dorisanima porazdelitvama narišite še na logaritemski lestvici. Lestvico transformirajte le na abscisni osi, vendar pa ustrezno transformirajte tudi dorisani gostoti. Kako je zdaj videti prileganje?

Pri histogramih združite čase oz. njihove desetiške 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. 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 so telesne temperature in pulzi porazdeljeni normalno.

a) Ocenite povprečja in standardne odklone za oba telesna parametra – tako za moške kot za ženske.

b) Za povprečja iz prejšnje točke določite 95% intervale zaupanja.

c) Za povprečno telesno temperaturo se navadno reče, da je $98{\cdot}6°\mathrm{F} = 37°\mathrm{C}$. Je to v skladu z izmerjenimi temperaturami?

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).

5. Naj bosta X in Y slučajni spremenljivki z:

$$\begin{aligned} E(X) = \mu_x\,, \quad & E(Y) = \mu_y\,, \\ \mathrm{var}(X) = \sigma_x^2\,, \quad & \mathrm{var}(Y) = \sigma_y^2\,, \end{aligned}$$

$$\mathrm{cov}(X, Y) = \sigma_{x,y}\,.$$

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 $E\big[(Y - \hat{Y})^2\big]$ minimalna. Matematični upanji, varianci in kovarianco poznamo.
Namig: velja $E\big[(Y - \hat{Y})^2\big] = \big[E(Y) - E(\hat{Y})\big]^2 + \mathrm{var}(Y - \hat{Y})$.

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

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