Projektna naloga za Statistiko

Mateja Možina
Fakulteta za matematiko in fiziko

11. september 2023

1 Prva naloga

V mestu Kibergrad živi 43.886 družin, razporejenih v 4 četrti. Zanima nas povprečno število otrok. Nalogo rešujemo v programu *Matlab* z datoteko **prva_naloga.m**.

1.a Povprečno število otrok z enostavnim slučajnim vzorčenjem

Izmed $N = 43.886$ naključno izberemo $n = 400$ družin in z

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_{400}}{400},$$

kjer je

$$X_i = \text{število otrok v } i\text{-ti družini},$$

izračunamo, da je povprečno število otrok enako 0.955. Standardno napako dobimo z

$$\widehat{SE_+} = \sqrt{\frac{N-n}{n-1}\frac{1}{N}\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

in tako je $\widehat{SE_+} \doteq 0.0557$.
Za interval zaupanja pri $\alpha = 0.05$ pa vemo, da je

$$\overline{X} - F^{-1}_{Student(n-1)}\left(1 - \frac{\alpha}{2}\right)\widehat{SE_+} < \mu < \overline{X} + F^{-1}_{Student(n-1)}\left(1 - \frac{\alpha}{2}\right)\widehat{SE_+}.$$

Z našimi podatki bo torej

$$0.8455 < \mu < 1.0644.$$

1.b Povprečno število otrok s stratificiranim vzorčenjem glede na četrt

Pri stratificiranem vzorčenju razdelimo populacijo velikosti N na k stratumov velikosti $N_1, \ldots, N_k$. V našem primeru bomo imeli 4 stratume - četrti v katerih stanuje družina. Velja, da je $N_1 + N_2 + N_3 + N_4 = N$, z $W_i = \frac{N_i}{N}$ pa označimo še velikostne deleže. Pri stratificiranem vzorčenju s proporcionalno alokacijo iz vsakega stratuma izberemo naključni vzorec n_i, tako da velja

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \frac{n_4}{N_4}.$$

Naš vzorec je enak $n = 400$, torej rabi še veljati

$$n_1 + n_2 + n_3 + n_4 = 400.$$

Iz teh pogojev dobimo, da je

$$n_1 \left(1 + \frac{10390}{10149} + \frac{13457}{10149} + \frac{9890}{10149}\right) = 400$$

oziroma, da je

$$n_1 = \frac{400}{\left(1 + \frac{10390}{10149} + \frac{13457}{10149} + \frac{9890}{10149}\right)}.$$

Tako dobimo, da so

$$n_1 = 92,\ n_2 = 95,\ n_3 = 123 \ \text{ in } \ n_4 = 90,$$

kjer so vsa števila zaokrožena, razen n_1 je zaokrožen navzdol.
Sedaj iz vsakega stratuma izberemo naključni vzorec velikosti n_i in s formulo

$$\overline{X_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

za vsak stratum izračunamo povprečno število otrok. Tako so

$$\overline{X_1} = \frac{1}{92} \sum_{j=1}^{92} X_{1j} = 0.837,$$

$$\overline{X_2} = \frac{1}{95} \sum_{j=1}^{95} X_{2j} = 1.095,$$

$$\overline{X_3} = \frac{1}{123} \sum_{j=1}^{123} X_{3j} = 1.033,$$

$$\overline{X_4} = \frac{1}{90} \sum_{j=1}^{90} X_{4j} = 1.144.$$

Tako bo povprečno število otrok

$$\overline{X} = \sum_{i=1}^{4} W_i \overline{X_i} = 1.027.$$

Da bi ocenili standardno napako, rabimo najprej izračunati variance posameznih stratumov. Za to uporabimo

$$\widehat{\sigma_i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \overline{X_i})^2,$$

kar nam da

$$\widehat{\sigma_1}^2 = 1.171,\ \widehat{\sigma_2}^2 = 1.619,\ \widehat{\sigma_3}^2 = 1.425,\ \widehat{\sigma_4}^2 = 1.788.$$

Standardno napako izračunamo s formulo

$$\widehat{SE} = \sqrt{\sum_{i=1}^{4} W_i^2 \frac{1}{n_i} \widehat{\sigma_i}^2},$$

kar nam da

$$\widehat{SE} = 0.0611.$$

Izračunajmo še interval zaupanja pri $\alpha = 0.05$. Vemo, da je

$$T = \frac{\overline{X} - \mu}{\widehat{SE}} \dot{\sim} Student(\nu),$$

kjer je

$$\nu = \frac{SE^4}{\sum_{i=1}^{4} \frac{W_i^4 \sigma_i^4}{n_i^2 (n_i - 1)}}.$$

Opomba: Ta enakost je dobljena iz spletne učilnice Statistika - Gradiva - Stratificirani modeli, čisto na koncu.

Sedaj namesto standardne napake in varianc, v enačbo damo njihove cenilke. Tako bo

$$\widehat{\nu} = 387.742$$

in aproksimirani interval zaupanja pri stopnji tveganja $\alpha = 0.05$ je

$$\overline{X} - F^{-1}_{Student(\widehat{\nu})}\left(1 - \frac{\alpha}{2}\right)\widehat{SE} < \mu < \overline{X} + F^{-1}_{Student(\widehat{\nu})}\left(1 - \frac{\alpha}{2}\right)\widehat{SE},$$

kar bo v našem primeru

$$0.907 < \mu < 1.147.$$

Vidimo, da nam stratificirano vzorčenje s proporcionalno alokacijo da večje povprečno število otrok, prav tako pa tudi večjo standardno napako. Interval zaupanja je pri enostavnem slučajnem vzorčenju manjši (pri enostavnem slučajnem vzorčenju je dolžine 0.21, pri stratificiranem s proporcionalno alokacijo pa približno 0.24), vendar ne za veliko.

2 Druga naloga

Podane imamo podatke o skokih, ki jih ptiči naredijo med dvema letoma, kjer je število skokov vedno vsaj 1, saj štejemo tudi zadnji skok, ko poleti. Nalogo rešujemo s programom **druga_naloga.m**.

2.a Geometrijska porazdelitev

Označimo X_i = število skokov pri i-tem opažanju, iščemo pa geometrijsko porazdelitev, ki se najbolje prilega našim podatkom, zato predpostavimo, da so $X_i \sim Geom(p)$ in med sabo neodvisne. Torej rabimo najti cenilko za p za

$$P(X = k) = pq^{k-1},$$

kjer je $q = 1 - p$. Vseh opažanj je $n = 130$, število vseh skokov pa 363. Besedna zveza »se najbolje prilega« nam pove, da iščemo cenilko po metodi največjega verjetja. Verjetje bo torej

$$\begin{aligned} L(p|X_1, \ldots, X_n) &= \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p = (1-p)^{\sum_{i=1}^{n} x_i - n} p^n \\ &= (1-p)^{S-n} p^n, \end{aligned}$$

kjer je $S = \sum_{i=1}^{n} x_i$ število vseh skokov. Če logaritmiramo, dobimo

$$l(p|X_1, \ldots, X_n) = \ln((1-p)^{S-n} p^n) = (S-n)\ln(1-p) + n\ln(p),$$

kar sedaj odvajamo po p

$$\frac{\partial l(p|X_1, \ldots, X_n)}{\partial p} = \frac{S-n}{p-1} + \frac{n}{p}.$$

Enačimo to z 0 in dobimo

$$\widehat{p} = \frac{n}{S},$$

kar je naša cenilka za p. Pri naših podatkih bo znašala $\widehat{p} = 0.3581$.

2.b Grafični prikaz

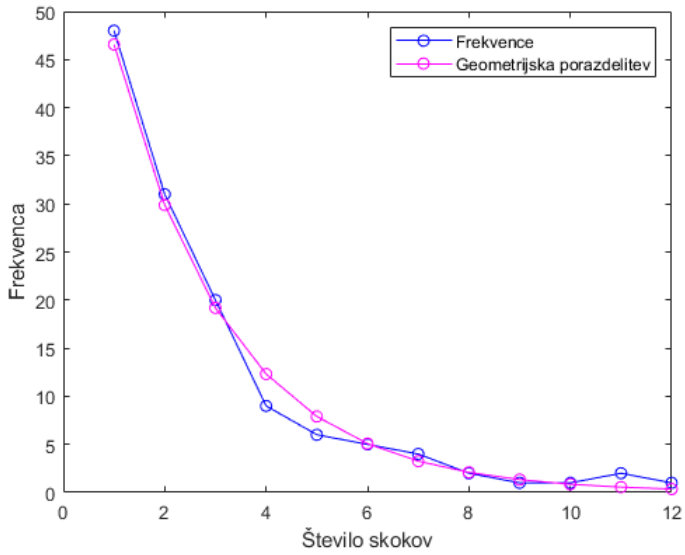
Poglejmo kako dobro se ta geometrijska porazdelitev ujema z našimi podatki. Vemo, da je verjetnost, da ptič poskoči k-krat enaka

$$P(X = k) = \widehat{p}q^{k-1}$$

za $\widehat{p} = 0.3581$, torej bomo v n-tih opažanjih videli

$$n \cdot P(X = k) = n \cdot \widehat{p}(1-\widehat{p})^{k-1}$$

opažanj oziroma frekvenc, da so ptiči poskočili k-krat.



Slika 1: Graf porazdelitev.

Vidimo, da se naši podatki in geometrijska porazdelitev dobro ujemata, največje odstopanje pa je približno 3.3.

2.c En presledek

Recimo, da smo opazili samo en presledek med letoma. Naša cenilka v tem primeru ne bo nepristranska (za $n = 1$), saj velja

$$E(\widehat{p}) = E\left(\frac{1}{X_1}\right) = \sum_{k=1}^{\infty} \frac{1}{k} p(1-p)^{k-1} = p + \underbrace{\sum_{k=2}^{\infty} \frac{1}{k} p(1-p)^{k-1}}_{\text{pozitivne vrednosti}} > p.$$

Konstruirajmo sedaj nepristransko cenilko. Iščemo torej funkcijo f, da bo veljalo

$$E(f(X)) = \sum_{k=1}^{\infty} f(k) p q^{k-1} = p$$

oziroma

$$\sum_{k=1}^{\infty} f(k) q^{k-1} = 1.$$

To nam reši funkcija

$$f(k) = \begin{cases} 1 & k = 1, \\ 0 & k > 1, \end{cases}$$

kar je naša iskana nepristranska cenilka za p (je tudi edina možna).

2.d Več presledekov

Tudi, če opazimo več presledkov $X, \ldots, X_n$, je cenilka $\widehat{p}$ še vedno pristranska. Vemo, da je vsota n geometrijskih porazdelitev $X_i \sim Geom(p)$ porazdeljena negativno binomsko, torej $Z = \sum_{i=1}^{n} X_i \sim NegBin(n, p)$. Tako je

$$\begin{aligned} E(\widehat{p}) &= nE\left(\frac{1}{Z}\right) = \sum_{k=n}^{\infty} \frac{n}{k} \binom{k-1}{n-1} p^n q^{k-n} \\ &= p \sum_{k=n}^{\infty} \frac{n}{k} \frac{k-1}{n-1} \binom{(k-1)-1}{(n-1)-1} p^{n-1} q^{(k-1)-(n-1)} \\ &> p \cdot \underbrace{\sum_{k=n}^{\infty} \binom{(k-1)-1}{(n-1)-1} p^{n-1} q^{(k-1)-(n-1)}}_{=1, \text{ ker je to vsota vseh verjetnosti v NegBin(n-1,p)}} = p \end{aligned}$$

in cenilka za p je pristranska.

Poglejmo sedaj cenilko

$$\check{p} = \frac{n-1}{S-1},$$

kjer je S skupno število opaženih skokov. Ko izračunamo pričakovano vrednost

$$\begin{aligned} E(\check{p}) &= \sum_{k=n}^{\infty} \frac{n-1}{k-1} \binom{k-1}{n-1} p^n q^{k-n} \\ &= \sum_{k=n}^{\infty} \binom{k-2}{n-2} p^n q^{k-n} \\ &= p \cdot \underbrace{\sum_{k=n}^{\infty} \binom{(k-1)-1}{(n-1)-1} p^{n-1} q^{(k-1)-(n-1)}}_{=1, \text{ ker je to vsota vseh verjetnosti v NegBin(n-1,p)}} \\ &= p, \end{aligned}$$

vidimo, da je ta cenilka nepristranska. Za naš primer bomo dobili, da je $\check{p} = 0.3564$, kar se od pristranske $\widehat{p}$ razlikuje šele na tretji decimalki.

3 Tretja naloga

Podane imamo podatke o težah piščancev, ki so bili hranjeni z različnimi dietami v roku treh tednov. Nalogo rešujemo s programom **tretja_naloga.m**.

3.a Pričakovana teža v enem dnevu

Z enostavno linearno regresijo ocenimo koliko teže pridobi piščanec v enem dnevu. Na voljo imamo podatke o petdesetih piščancih, hranjenih v treh tednih s štirimi različnimi dietami, vseh podatkov pa je $n = 578$. Naj bo

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{579} \end{bmatrix}$$

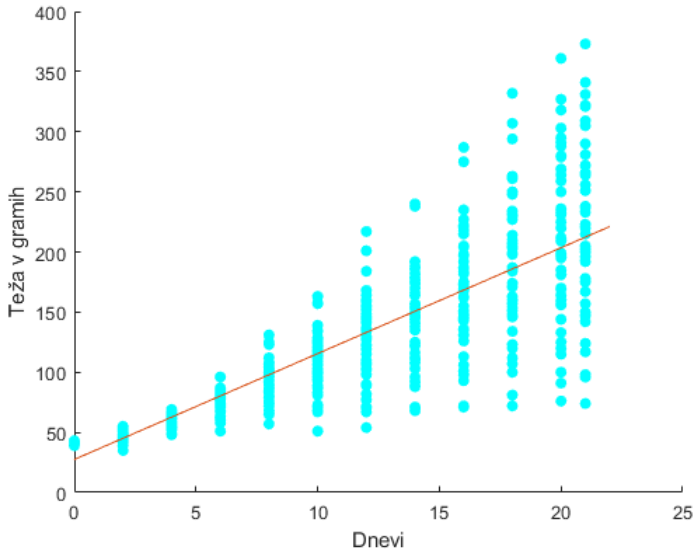
vektor vseh podanih tež in

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{579} \end{bmatrix}$$

matrika, kjer so x_i dnevi. Tako bo naša cenilka za $\widehat{\beta}$

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T Y = \begin{bmatrix} 27.467 \\ 8.803 \end{bmatrix}.$$

Torej piščanec v enem dnevu pridobi približno 8.803 gramov. Poglejmo si to še na grafu.



Slika 2: Premica naraščanja teže.

Vidimo, da so na začetku piščanci imeli zelo podobne teže - malo pod 50 g, po enaindvajstih dnevih pa so razlike velike - najmanjša izmerjena teža je bila okoli 60 g, največja pa okoli 370 g. Glede na sliko se zdi, da se premica dobro prilega našim podatkom.

Opomba: Tukaj manjkajo še napake ϵ_i. Iz podatkov, ki jih imamo na grafu, naj bi ti epsiloni bili oblike $\epsilon_i = \sigma^2 x_i^2 + \tau^2$, kjer sta sigma in tau neznani. Splošen model - $E(\epsilon_i) = 0$ in $Var(\epsilon) = \sigma^2 I$ - tako ne velja, velja samo prva enakost.

3.b Vpliv diete na težo

Poglejmo, ali dieta vpliva na težo piščanca. Naša ničelna hipoteza H_0 je, da dieta ne vpliva na težo, alternativna domneva H_1 pa je, da dieta vpliva na težo. Naj bo β_0 maša piščanca na začetku, $\beta_i, i = 1, 2, 3, 4$ pa koliko piščanec pridobi teže v enem dnevu glede na i-to dieto. Sestavimo matriko X

$$X = \begin{bmatrix} 1 & x_{1,1} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,l_1} & 0 & 0 & 0 \\ 1 & 0 & x_{2,1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{2,l_2} & 0 & 0 \\ 1 & 0 & 0 & x_{3,1} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & x_{3,l_3} & 0 \\ 1 & 0 & 0 & 0 & x_{4,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & x_{4,l_4} \end{bmatrix},$$

ki ima v prvem stolpcu same enice, v drugem podatke o dnevih za piščance hranjene s prvo dieto, v tretjem stolpcu o dnevih za piščance hranjene z drugo dieto, v četrtem o dnevih za piščance hranjene s tretjo dieto in v petem s četrto dieto. Vektor β je tako

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix},$$

kjer je β_0 začetna teža, β_i pa pričakovana pridobitev teže v enem dnevu z dieto i. Torej, naša ničelna hipoteza je, da

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$$

oziroma, da je

$$\beta = \beta_0 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \beta_1 \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Dimenzija podprostora W, ki ga razpenjata ta dva vektorja je torej $q = \dim W = 2$, dimenzija prostora $V = \mathrm{Im(X)}$ pa $p = \dim V = 5$. Naj bo H ortogonalni projektor na V in K ortogonalni projektor na W. H izračunamo z

$$H = X(X^T X)^{-1} X^T,$$

K pa podobno, samo da namesto X-a vstavimo

$$A = X \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Preizkusna statistika bo

$$F = \frac{\frac{\|(H-K)Y\|^2}{p-q}}{\frac{\|(I_n-H)Y\|^2}{n-p}},$$

domnevo H_0 pa bomo zavrnili pri stopnjah tveganja $\alpha = 0.05$ in $\alpha = 0.01$, če je

$$F \geqslant F^{-1}_{Fisher(p-q,n-p)}\,(1-\alpha)\,.$$

Opomba: Ta enakost je dobljena iz spletne učilnice Statistika - Gradiva - Statistično sklepanje pri linearni regresiji.

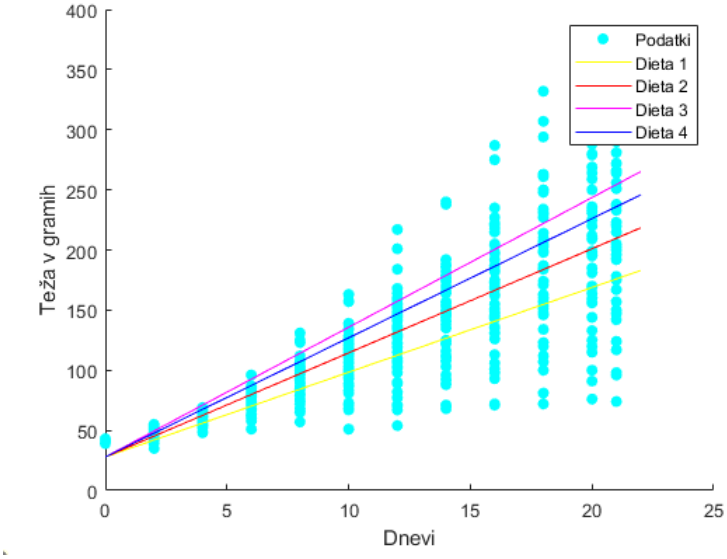
Za naše podatke dobimo, da je

$$\begin{aligned} &F = 59.32, \\ &F^{-1}_{Fisher(p-q,n-p)}\,(1-0.05) = 2.62, \\ &F^{-1}_{Fisher(p-q,n-p)}\,(1-0.01) = 3.82, \end{aligned}$$

torej v obeh primerih hipotezo H_0 zavrnemo, torej izbira diete vpliva na težo piščanca. Podobno kot v primeru a lahko izračunamo β

$$\beta = \begin{bmatrix} 27.86 \\ 7.05 \\ 8.66 \\ 10.79 \\ 9.91 \end{bmatrix}$$

in vidimo da je bila dieta 3 najbolj učinkovita. Še grafično



Slika 3: Različne diete.