1. NALOGA

a)

Povprečni dohodek

Populacija je velikosti 43886. Vzamemo enostavni slučajni vzorec 400 enot.
Sledi: $N = 43886, n = 400$.
Naj bo X_k skupni dohodek v k-ti družini.
Torej je povprečni dohodek v Kindergradu:

$$\overline{X} = \frac{X_1 + ... + X_n}{n}.$$

Standardna napaka

Vemo: $se(\overline{X}) = \sqrt{var(\overline{X})}$. Ker imamo enostavni slučajni vzorec, vemo tudi, da je

$$var(\overline{X}) = \frac{1}{n}\frac{N-n}{N-1}\sigma^2$$

kjer je σ^2 populacijska varjanca. Nepristranska cenilka za σ^2 je

$$\hat{\sigma}^2 = \frac{N-1}{N(n-1)}\sum_{k=1}^{n}(X_k - \overline{X})^2.$$

Sledi:

$$\widehat{se(\overline{X})} = \sqrt{\frac{1}{n}\frac{N-n}{N(n-1)}\sum_{k=1}^{n}(X_k - \overline{X})^2}.$$

Interval zaupanja

Iz navodil sledi, da je interval zaupanja enak $\overline{X} \pm 1,96 \cdot se(\overline{X})$.

Končne vrednosti

b)

Če stratificiramo, mora veljati

$$\frac{n_k}{n} = \frac{N_k}{N}, \sum_{k=1}^{k} n_k = n.$$

V našem primeru startificiramo po četrtih, torej $k = 4$. Vemo:
$N_1 = 10149$ (*severna četrt*),
$N_2 = 10390$ (*vzgodna četrt*),

$N_3 = 13457$ (*južna četrt*),
$N_4 = 9890$ (*zahodna četrt*). Če malo obrnemo zgornjo enakost, dobimo

$$n_k = \frac{N_k}{N} n.$$

Izračunamo za $k = 1, 2, 3, 4$ in upoštevamo vrednosti N_1, N_2, N_3, N_4 ter $N = 43886$.
Dobimo:

$$n_1 = \frac{10149}{43886} \cdot 400 = 92,5033 \rightarrow n_1 = 92,$$

$$n_2 = \frac{10390}{43886} \cdot 400 = 94,699 \rightarrow n_2 = 95,$$

$$n_3 = \frac{13457}{43886} \cdot 400 = 122,654 \rightarrow n_3 = 123,$$

$$n_4 = \frac{9890}{43886} \cdot 400 = 90,142 \rightarrow n_4 = 90.$$

Preverimo:

$$\sum_{k=1}^{4} n_k = 92 + 96 + 123 + 90 = 400.$$

Naj bo sedaj X_{kj} povprečni dohodek j-te družine v k-tem stratumu.
Povprečni dohodek družine se sedaj izraža kot:

$$\overline{X} = \frac{1}{n} \sum_{k=1}^{\#stratumov} \sum_{j=1}^{n_k} X_{kj}.$$

Standarna napaka $se(\overline{X}) = \sqrt{var(\overline{X})}$:

$$var(\overline{X}) = \sum_k w_k^2 var(\overline{X_k})) = \sum_k w_k^2 \cdot \frac{\hat{\sigma}_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1},$$

kjer je $w_k = \frac{N_k}{N}$ delež, σ_k^2 pa populacijska varjanca v k-tem stratumu. Torej:

$$\hat{\sigma}_k^2 = \frac{N_k - 1}{N_k(n_k - 1)} \sum_{j=1}^{n_k} (X_{kj} - \overline{X_k})^2,$$

kjer je X_k povprečje k-tega stratuma.
Interval zaupanja: $\overline{X} \pm 1,96 \cdot se(\overline{X})$.
Vstavimo podatke in dobimo:

c)

Variance znotraj četrti:
Variance med četrtmi dobimo tako, da izračunamo povprečje dohodka v vsaki četrti, jo primerno obtežimo in izračunamo varjanco. Torej dobimo:

2. NALOGA

3.NALOGA

a)

$$X \sim f(x,\alpha) = \begin{cases} \frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)} x^{\alpha-1}(1-x)^{2\alpha-1}, & 0 < x < 1. \\ 0, & sicer. \end{cases}$$

Vemo:

$$E(X) = \frac{1}{3}, var(X) = \frac{2}{9(3\alpha+1)}.$$

Ker je:

$$var(X) = E(X^2) - E(X)^2,$$

sledi:

$$E(X^2) = var(X) + E(X)^2 = \frac{2}{9(3\alpha+1)} + \frac{1}{9} = \frac{\alpha+1}{3(3\alpha+1)}$$

$E(X^2)$ je drugi moment slučajne spremenljivke X, torej:

$$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Zgornji enačbi izenačimo in poračunamo α:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{\alpha+1}{3(3\alpha+1)}$$

$$\frac{9\alpha+3}{n}\sum_{i=1}^{n} X_i^2 = \alpha+1$$

$$\frac{9\alpha}{n}\sum_{i=1}^{n} X_i^2 - \alpha = 1 - \frac{3}{n}\sum_{i=1}^{n} X_i^2$$

$$\alpha = \frac{1 - \frac{3}{n}\sum_{i=1}^{n} X_i^2}{\frac{9}{n}\sum_{i=1}^{n} X_i^2 - 1}$$

b)

$$f_X(x,\alpha) = \begin{cases} \frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)} x^{\alpha-1}(1-x)^{2\alpha-1}, & 0 < x < 1. \\ 0, & sicer. \end{cases}$$

$$L_1(\alpha|x) = \frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)} x^{\alpha-1}(1-x)^{2\alpha-1}$$

$$L(\alpha|x_1,...,x_n) = L_1(\alpha|x_1)\cdot ... \cdot L_1(\alpha|x_n) =$$
$$\left(\frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)}\right)^n x_1^{\alpha-1}\cdot ... \cdot x_n^{\alpha-1}(1-x_1)^{2\alpha-1}\cdot ... \cdot (1-x_n)^{2\alpha-1}$$

$$l(\alpha|x_1,...,x_n) = ln(L(\alpha|x_1,...,x_n)) = l_1(\alpha|x_1) + ... + l_1(\alpha|x_n)$$

$$l_1(\alpha|x) = ln(L_1(\alpha|x)) =$$
$$ln(\Gamma(3\alpha)) - ln(\Gamma(\alpha)) - ln(\Gamma(2\alpha)) + (\alpha-1)ln(x) + (2\alpha-1)ln(1-x)$$

$$l(\alpha|x) =$$
$$\sum_{i=1}^{n}(ln(\Gamma(3\alpha)) - ln(\Gamma(\alpha)) - ln(\Gamma(2\alpha)) + (\alpha-1)ln(x_i) + (2\alpha-1)ln(1-x_i))$$

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n}\frac{1}{\Gamma(3\alpha)}\Gamma^{'}(3\alpha)3 - \frac{1}{\Gamma(\alpha)}\Gamma^{'}(\alpha) - \frac{1}{\Gamma(2\alpha)}\Gamma^{'}(2\alpha)2 + ln(x_i) + 2ln(1-x_i)$$

$$\sum_{i=1}^{n}\frac{1}{\Gamma(3\alpha)}\Gamma^{'}(3\alpha)3 - \frac{1}{\Gamma(\alpha)}\Gamma^{'}(\alpha) - \frac{1}{\Gamma(2\alpha)}\Gamma^{'}(2\alpha)2 = ln(\frac{1}{x_i(1-x_i)})$$

Cenilka obstaja, ko ima zgornja enačba rešitev.
Uporabimo funkcijo digamma in rešimo do konca? no clue.

c)

$var(\hat{\alpha}) = \frac{1}{nI_1(\hat{\alpha})}$.

$$I_1(\hat{\alpha}) = -E\left[\frac{\partial^2 l_1(\alpha|x)}{\partial\alpha^2}\right]$$

$$\frac{\partial^2 l_1(\alpha|x)}{\partial\alpha^2} =$$
$$3\frac{\Gamma^{''}(3\alpha)\Gamma(3\alpha) - \Gamma^{'}(3\alpha)^2}{\Gamma(3\alpha)^2} - \frac{\Gamma^{''}(\alpha)\Gamma(\alpha) - \Gamma^{'}(\alpha)^2}{\Gamma(\alpha)^2} - 2\frac{\Gamma^{''}(2\alpha)\Gamma(2\alpha) - \Gamma^{'}(2\alpha)^2}{\Gamma(2\alpha)^2}$$

$$var(\hat{\alpha}) = \frac{1}{n}\frac{1}{3\frac{\Gamma^{''}(3\alpha)\Gamma(3\alpha)-\Gamma^{'}(3\alpha)^2}{\Gamma(3\alpha)^2} - \frac{\Gamma^{''}(\alpha)\Gamma(\alpha)-\Gamma^{'}(\alpha)^2}{\Gamma(\alpha)^2} - 2\frac{\Gamma^{''}(2\alpha)\Gamma(2\alpha)-\Gamma^{'}(2\alpha)^2}{\Gamma(2\alpha)^2}}$$

4.NALOGA

Za reševanje te naloge, sem si pomagala s knjigo *Mathematical statistics and data analysis*. Iz poglavja 9.5, na strani 341, izvemo, da če imamo model, kjer so verjetnosti vektorsko porazdeljene in imamo opis za hipotezo H_0, kjer je $p = p(\theta)$, kjer je θ neznam, je števec za verjetnostno funkcijo enak:

$$\max_{p\in\omega_0}\left(\frac{n!}{x_1!...x_m!}\right)p_1(\theta)^{x_1}\cdots p_m(\theta)^{x_m},$$

kjer so x_i opazovane vrednosti v m celicah. Maksimum bo dosežen pri:

$$\hat{p_i} = \frac{x_i}{n}.$$

Razmerje verjetij je tako enako:

$$\Lambda = \frac{\frac{n!}{x_1!...x_m!}p_1(\hat{\theta})^{x_1}\cdots p_m(\hat{\theta})^{x_m}}{\frac{n!}{x_1!...x_m!}\hat{p_1}^{x_1}\cdots \hat{p_m}^{x_m}}.$$

Poglejmo sedaj naš primer. Podatke imamo podane v razpredelnici za $m = 0, 1, ..., 12.$, $n = \#$ podatkov $= 6115$. Oprazovane vrednosti $x_1, ..., n_{12}$ prebermo iz drugega stolpca v tabeli. Računamo po postopku opisanem zgoraj:

$$\max\left(\frac{6115!}{7!45!181!\cdots 24!3!}\right)p_1^7p_2^{45}\cdots p_{12}^3$$

Maksimum je torej dosežen pri:

$$\hat{p}_1 = \frac{7}{6115}, \hat{p}_2 = \frac{45}{6115}, \cdots, \hat{p}_{12} = \frac{3}{6115},$$

Poračunati moramo še imenovalec ulomka, da bomo dobili razmerje verjetij. Testirati želimo hipotezo, da je število moških potomcov, ki se rodijo v družini z 12 otroci, porazdeljeno binomsko $Bin(12, p)$. Torej je verjetnostna funkcija L enaka:

$$L = \left(\frac{6115!}{7!45!181!\cdots 24!3!}\right)\left((1-p)^{12}\right)^7\left(2p(1-p)^{11}\right)^{45}\cdots\left(p^{12}\right)^3.$$

Izraz logaritmiramo in odvajamo po p, izračunamo maksimum.

$$\frac{\partial l}{\partial p} = 0.$$

Po krajšem izračunu dobimo:

$$\frac{35280}{p} = \frac{38100}{1-p}$$

Tako dobimo:

$$\hat{p} = \frac{35280}{73380} = 0.4807849550286182.$$

Torej po Wilkinsonovem izreku zapišemo:

$$\lambda = 2ln(\Lambda) = 2l(\hat{p}_0, ..., \hat{p}_{12}) - 2l(\hat{p})$$

Vstavimo zgoraj poračunane vrednosti:

$$2\left(7ln(\hat{p}_1) + 45ln(\hat{p}_2) \cdots 3ln(\hat{p}_{12})\right) - 2\left(7 * 12ln(1-\hat{p}) + 45(ln(12\hat{p}) + 11ln(1-\hat{p})) + \cdots + 3 * 12ln(\hat{p})\right)$$

Ko vstavimo podatke dobimo:

$$\lambda = 97.0065.$$

Wilksov izrek nam pove, da bomo hipotezo zavrnili, če je $\lambda > \chi^2_{1-\alpha}(dim\Omega - dim\Omega_0)$. Dimenziji prostorov sta očitni: $dim\Omega - dim\Omega_0 = 12 - 1 = 11$. Testitamo za $\alpha = 0.01$ in $\alpha = 0.05$. Iz tabele preberemo vrednosti in dobimo:

$$\chi^2_{0.99}(11) = 19.68, \chi^2_{0.95}(11) = 24.72.$$

Vidimo, da sta vrednosti v obeh primerih manjši od λ, zato hipotezo v obeh primerih zavrnemo.
Model bi lahko bil napačen iz več razlogov. Eden bi bil naprimer, da nismo upoštevali vsa rojstva, vendar le preživele otroke. Lahko, da smo zbrali podatke v različnih zgodovinskih obdobjih, katera vplivajo na rojstva otrok (vojna in podobno).

5.NALOGA

Ocen po metodi največjega verjetja

Da bomo vedeli kako so porazdeljeni Y_i-ji, poračunamo njihovo pričakovano vrednost in varjanco.

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_1$$

$$var(Y_i) = \sigma^2$$

Torej so Y_i porazdeljeni $N(\beta_0 + \beta_1 x_1, \sigma^2)$. Dobimo:

$$L_1 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_o - \beta_1 x_i)^2}{2\sigma^2}}$$

Verjetnostna funkcija je sestavljena iz produktov porazdelitvenih funkcij. Torej dobimo:

$$log(L) = -nlog(\sigma) - \frac{n}{2}log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_o - \beta_1 x_i)^2$$

. Najprej poračunamo za prvi parameter β_0. Enačbo odvajamo po β_0 in enačimo z 0.

$$\frac{\partial}{\partial\beta_0}log(L) = \frac{\partial}{\partial\beta_0}(-nlog(\sigma) - \frac{n}{2}log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_o - \beta_1 x_i)^2) =$$
$$= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_o - \beta_1 x_i)$$

Dobimo:

$$\sum_{i=1}^{n}(y_i - \beta_o - \beta_1 x_i) = 0$$

Sledi:

$$\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

Torej je končni rezultat enak:

$$\beta_o = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i.$$

Postopek ponovimo za drugi parameter β_1:

$$\frac{\partial}{\partial\beta_1}log(L) = \frac{\partial}{\partial\beta_1}(-nlog(\sigma) - \frac{n}{2}log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_o - \beta_1 x_i)^2) =$$
$$= -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)$$

Dobimo:

$$\sum_{i=1}^{n} x_i(y_i - \beta_o - \beta_1 x_i) = 0.$$

Sledi:

$$\sum_{i=1}^{n} x_i y_i - n\beta_o - \beta_1 \sum_{i=1}^{n} x_i = 0.$$

$$\sum_{i=1}^{n} x_i y_i - n\beta_o - \beta_1 \sum_{i=1}^{n} x_i = 0.$$

Vstavimo vrednost za β_0:

$$\sum_{i=1}^{n} x_i y_i - n\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i\right) - \beta_1 \sum_{i=1}^{n} x_i = 0.$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i - \beta_1\left(n\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right) = 0.$$

Torej je končni rezultat enak:

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i - \sum_{i=1}^n x_i}.$$

Vstavimo še za β_0:

$$\beta_o = \frac{1}{n}\sum_{i=1}^n y_i - \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i - \sum_{i=1}^n x_i}\sum_{i=1}^n x_i.$$

$$\beta_o = \frac{1}{n}\sum_{i=1}^n y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i - \sum_{i=1}^n x_i}.$$

$$\beta_o = \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i + \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i - \sum_{i=1}^n x_i}.$$

Ocena po metodi najmanjših kvadratov

Ker so šumi ϵ_i neodvisni za vsak $i = 1, 2, ..., n$ in $\epsilon_i \sim N(0, \sigma^2)$ lahko uporabimo izrek Gauss-Markova za ocene parametrov β_0 in β_1. Imamo:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Po izreku Gauss-Markova vemo da je najboljša cenilka za $\beta = (X^T X)^{-1} X^T Y$. Računamo:

$$\left(\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} =$$

$$= \left(\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \right)^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} =$$

$$= \frac{1}{n\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} =$$

$$= \frac{1}{n\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i} \begin{bmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ -\sum_{i=1}^n x_i \sum_{i=1}^n y_i + n\sum_{i=1}^n x_i y_i \end{bmatrix}$$

Z malo računanja in spretnosti vidimo da sta cenilki enaki.