SEMINARSKA NALOGA IZ STATISTIKE

Izak Jenko

22. avgust 2021

Seminarska naloga iz statistike pri predmetu statistika na Fakulteti za matematiko in fiziko v Ljubljani v študijskem letu 2020/21. Celoten repozitorij seminarske naloge, ki vključuje skripte in izračune, najdemo na spletnem naslovu:

`https://github.com/kazi99/STAT-seminarska`

1. naloga

Naj x_i označuje dohodek i-te družine za $1 \leq i \leq N := 43886$. Vzamemo enostavni slučajni vzorec $X_i := x_{K_i}$, kjer je slučajni vektor $(K_1, \ldots, K_n)$ enakomerno porazdeljen po množici n-teric s samimi različnimi komponentami, ki prihajajo iz množice $\{1, \ldots, N\}$. Vzorec bo velikosti $n = 400$ in bo *proporcionalno alociran*, da bomo lahko na koncu primerjali točki a) in b).

a) Populacijsko povprečje $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ točkovno ocenimo z vzorčnim povprečjem

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

ki na našem vzorcu znaša 43410,49. Označimo s $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ populacijsko varianco. Varianca cenilke $\overline{X}$ bo

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

kjer faktor $1 - \frac{n-1}{N-1}$ nastopi zaradi koreliranosti posameznih enot enostavnega slučajnega vzorca. Standardna napaka cenilke $\overline{X}$, bo torej se $= \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$. Popoulacijske variance σ^2 običajno ne poznamo, je pa

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

njena nepristranska cenilka, torej standardno napako (se) lahko ocenimo z

$$\hat{\text{se}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}.$$

Izračunana na našem vzorcu je ocena standardne napake enaka 1758,62.

Sedaj imamo zbrane vse podatke, da postavimo 95% interval zaupanja. Stopnja tveganja bo $\alpha = 0{,}05$, kar pomeni, da bo s približno takšno verjetnostjo interval zaupanja zgrešil dejansko populacijsko statistiko, ki jo ocenjujemo. Ker imamo opravka z veliko (400 – velikost vzorca) enako porazdeljenimi (porazdelitev je empirična) slučajnimi spremenljivkami, lahko zaradi centralnega limitnega izreka predpostavljamo, da je vzorčno povprečje $\overline{X}$ porazdeljeno približno normalno. Opomnimo še, da te slučajne spremenljivke niso nujno neodvisne, je pa njihova koreliranost razmeroma majhna pri tako veliki populaciji kot je naša, zato je uporaba izreka nekoliko bolj utemeljena.

Držimo se standardnih oznak in označimo s Φ kumulativno porazdelitveno funkcijo standardne normalne porazdelitve

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} dz.$$

Če bi poznali populacijsko varianco σ^2 in s tem standardno napako, bi dobili inteval zaupanja

$$\left(\overline{X} - \text{se}\,\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \overline{X} + \text{se}\,\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right). \tag{1}$$

Kar pri našem vzorcu pride $(39963{,}65; 46857{,}32)$ in je dolžine 6250,57.

Glede na to pa, da imamo tudi oceno standardne napake in recimo prave vrednosti standardne napake ne bi poznali, lahko postavimo tudi aproksimativni interval zaupanja oblike

$$\left(\overline{X} - \hat{\text{se}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \overline{X} + \hat{\text{se}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right). \tag{2}$$

ki je asimptotično eksakten. Na našem vzorcu znaša $(35937{,}86; 41475{,}80)$ in je dolžine 5537,94. V tem primeru, ko poznamo samo oceno standardne napake, si lahko pomagamo tudi s Studentovo t-porazdelitvijo in tako dobimo Studentov interval zaupanja

$$\left(\overline{X} - \hat{\text{se}}F^{-1}_{\text{Student}(n-1)}\left(1 - \frac{\alpha}{2}\right), \overline{X} + \hat{\text{se}}F^{-1}_{\text{Student}(n-1)}\left(1 - \frac{\alpha}{2}\right)\right), \tag{3}$$

kjer $F_{\text{Student}(n-1)}$ označuje kumulativno porazdelitveno funkcijo Studentove t-porazdelitve z $n-1$ prostostnimi stopnjami. Na našem vzorcu pride interval zaupanja $(35929{,}43; 41484{,}23)$, ki je dolžine 5554,79.

b) V tej točki si bomo pogledali kako dobro lahko ocenimo statistike od prej, če si pomagamo s stratificiranjem prebivalstva po četrtih. Naš vzorec velkosti 400 je stratifician s proporcionalno alokacijo, kar pomeni, da vzamemo $n_1 = 92$ enot iz severne četrti, $n_2 = 95$ iz vzhodne, $n_3 = 123$ iz južne in $n_4 = 90$ iz zahodne četrti kot kaže tabela.

	Sever	Vzhod	Jug	Zahod
n_j	92	95	123	90

Do teh števil pridemo, tako da nekoliko prilagodimo količine $n_j \approx n \cdot w_j$ (tu je w_j delež populacije, ki živi v j-ti četri). Polovico od količin n_j zaokrožimo navzgor do najbližjega celega števila, preostale pa navzdol do celega števila.

Najprej stratumsko ocenimo populacijsko povprečje μ kot

$$\overline{X}_s = \sum_{j=1}^{4} w_j \overline{X}_j, \quad \text{kjer je} \quad \overline{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$$

Stratificirana standardna napaka se izraža kot

$$\mathrm{se}_s^2 = \sum_{j=1}^{4} \frac{w_j^2 \sigma_j^2}{n_j} \left(1 - \frac{n_j - 1}{N_j - 1}\right),$$

kjer je stratificirana populacijska varianca $\sigma_j^2 = \sum_{i=1}^{n_j} \frac{(X_{ij} - \mu_j)^2}{N_j}$.

Ocena standardne napake bo tedaj

$$\hat{se}_s^2 = \sum_{j=1}^{4} \frac{w_j^2 \hat{\sigma}_j^2}{n_j} \left(1 - \frac{n_j - 1}{N_j - 1}\right),$$

kjer je $\hat{\sigma}_j^2 = \sum_{i=1}^{n_j} \frac{(X_{ij} - \overline{X}_j)^2}{n_j - 1}$ nepristranska cenilka populacijske variance v j-tem stratumu σ_j^2.

Sedaj lahko podobno kot prej postavimo 95% intervale zaupanja. Najprej če predpostavjlamo, da poznamo populacijsko varianco, dobimo

$$\left(\overline{X}_s - \mathrm{se}_s \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \overline{X} + \mathrm{se}_s \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right).$$

Na našem vzorcu je to (36604,38; 42106,27), ki je dolžine 6224,06, kar pomeni, da je krajši kot interval (1) in s tem tudi boljši za ocenejevanje povprečnega dohodka.

Naslednji je aproksimativni interval zaupanja, kjer uporabimo oceno za standardno napako $\hat{\mathrm{se}}_s$. Interval bo tedaj

$$\left(\overline{X}_s - \hat{\mathrm{se}}_s \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \overline{X} + \hat{\mathrm{se}}_s \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right),$$

ki bo pri naših podatkih (35931,97; 41480,80) dolžine 5548,84, kar je pravtako krajše kot interval (2) in tako boljše kot če ne bi stratificirali.

Nazadnje lahko postavimo še Studentov interval zaupanja za ocenejevanje povprečja, pred tem pa moramo izračunati prostostne stopnje

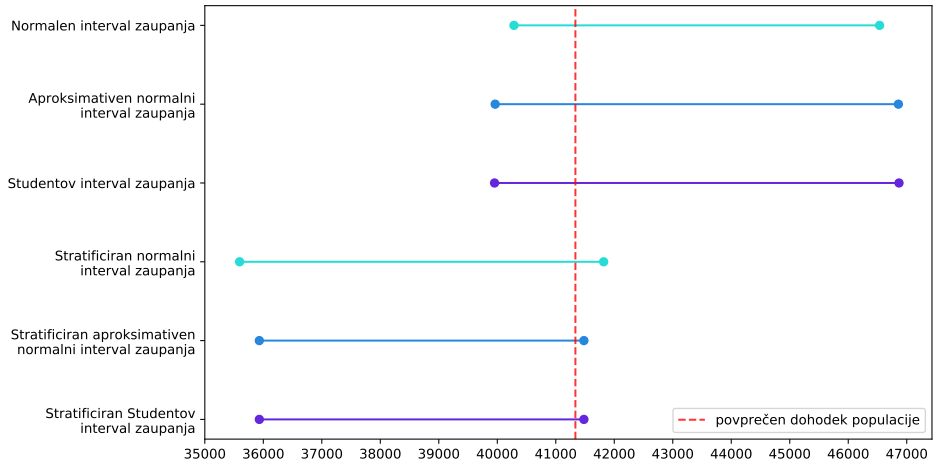
$$\hat{\nu} = \frac{\hat{\text{se}}_s^2}{\sum_{j=1}^{4} \frac{w_j^4 \hat{\sigma}_j^2}{n_j^2(n_j-1)}}.$$

Interval zaupanja bo tako oblike

$$\left(\overline{X} - \hat{\text{se}} \cdot F^{-1}_{\text{Student}(\hat{\nu})}\left(1-\frac{\alpha}{2}\right), \overline{X} + \hat{\text{se}} \cdot F^{-1}_{\text{Student}(\hat{\nu})}\left(1-\frac{\alpha}{2}\right)\right),$$

na naših podatkih bo to interval (35931,97; 41480,80) dolžine 5548,84, ki je spet boljši kot (3), ko nismo stratificirali.

Intervale zaupanja prikažemo tudi grafično.



Slika 1: 95% intervali zaupanja.

Zaključimo lahko, da je v tem primeru bilo bolj ugodno stratificirati po četrtih, saj smo tako dobili ožje intevale zaupanja, kar nam bo omogočilo postaviti bolj natančne ocene povprečnega dohodka v Kibergradu.

Do računskih razultatov pri tej nalogi smo prišli s pomočjo priloženih skript `naloga_1a.ipynb` za točko a) in `naloga_1b.ipynb` za točko b), ki analizirata isti enostaven slučajni vzorec 400 enot (s proporcionalno alokacijo) iz datoteke `vzorec.csv`.

2. naloga

Imamo vzorec 280 palic, kjer smo vsaki od njih preizkusili lomljivost na petih mestih. Sumimo, da je število mest na katerih se je palica zlomila

porazdeljeno Bin(5,p), kar bo tudi naša osnovna ničelna domneva H_0.

a) Denimo, da osnovna ničelna domneva velja, po metodi največjega verjetja pa ocenimo parameter p. Vzemimo

$$L_1(p;x) = \mathsf{P}_p(X = x) = \binom{5}{x} p^x (1-p)^{5-x},$$

kjer je slučajna spremenljivka X porazdeljena binomsko Bin(5,p). Verjetje bo tedaj funkcija

$$L(p;x_1,\ldots,x_n) = \prod_{i=1}^{n} L_1(p;x) = \prod_{i=1}^{n} \binom{5}{x_i} p^{x_i}(1-p)^{5-x_i}.$$

Cenilka za p bo torej tisti parameter, pri katerem je verjetje maksimalno. Ker je verjertje gladko odvisno od parametra $p \in (0,1)$ bomo ta ekstrem poiskali z odvodom, da bo računanje lažje pa si bomo pomagali z logaritmom verjetja

$$\begin{aligned}\ell(p;x_1,\ldots,x_n) &:= \log L(p;x_1,\ldots,x_n)\\ &= \sum_{i=1}^{n} \log\binom{5}{x_i} + \sum_{i=1}^{n} x_i \log(p) + \sum_{i=1}^{n} (5-x_i)\log(5-p)\end{aligned}$$

Z odvajanjem ℓ po p dobimo

$$\ell'(p;x_1,\ldots,x_n) = \sum_{i=1}^{n}\left(\frac{x_i}{p} - \frac{5-x_i}{1-p}\right) = \sum_{i=1}^{n} \frac{x_i - 5p}{p-p^2}.$$

Ta odvod bo enak 0 natanko tedaj, ko je $p = \frac{1}{5n}\sum_{i=1}^{n} x_i$. Za vse $p \in (0,1)$ je $L(p;x_1,\ldots,x_n) > 0$ poleg tega pa velja

$$\lim_{p\downarrow 0} L(p;x_1,\ldots,x_n) = 0 \quad \text{in} \quad \lim_{p\uparrow 1} L(p;x_1,\ldots,x_n) = 0,$$

za poljuben nabor $(x_1,\ldots,x_n) \in \{0,1,\ldots,5\}^n$ razen, kadar so vsi x_i enaki bodisi 0 bodisi 5, ko je maksimum verjetja L dosežen na robu intervala (0,1). To pomeni, da je $\frac{1}{5n}\sum_{i=1}^{n} x_i$ res edini maksimum verjetja, torej lahko za cenilko p vzamemo

$$\hat{p} = \frac{\overline{X}}{5}, \quad \text{kjer je } \overline{X} \text{ vzorčno povprečje.}$$

b) S posplošenim Pearsonovim preizkusom hi kvadrat [1, §9.5] bomo v tem delu preizkusili ničelno domnevo, da ima število lomov porazdelitev Bin(5,$\hat{p}$), kjer je $\hat{p}$ cenilka za p iz prejšnje točke. Opažanja poskusa so prikazana v tabeli.

št. lomov	št. palic
0	157
1	69
2	35
3	17
4	1
5	1

Ker Pearsonov preizkus zahteva, da je minimum po celicah najmanj 5, bomo zadnji dve celici pri 4 in 5 združili s celico 3. Naj O_i tako označuje število opažanj pri i-ti celici po zgornjem popravku (v tem primeru bo recimo $O_1 = 69$ in $O_3 = 19$).

Za vsako od celic potrebujemo še pričakovane vrednosti, ki jih dobimo na podlagi porazdlitve pod ničelno hipotezo. V našem primeru označimo $p_i = \mathsf{P}(X = i)$ za $0 \leq i \leq 2$ ter $p_3 = \mathsf{P}(X = 3) + \mathsf{P}(X = 4) + \mathsf{P}(X = 5)$, kjer je $X \sim \mathrm{Bin}(5, \hat{p})$ (da bo vsota vseh verjetnosti vselej 1, smo kot posledico združitve celic 3, 4 in 5 ustrezno prilagodili tudi verjetnost p_3). Pričakovana vrednost v vsaki od celic bo tedaj $E_i = n \cdot p_i$.

Testna statistika pri Pearsonovem preizkusu je

$$\chi^2 = \sum_{i=0}^{3} \frac{(E_i - O_i)^2}{E_i}.$$

Porazdeljena je s porazdelitvijo hi kvadrat z dvema prostostnima stopnjama, saj imamo 4 celice in en parameter $\hat{p}$, ki smo ga ocenili na podlagi podatkov, torej $\chi^2 \sim \chi^2(2)$. Testna statistika izračunana na naših podatkih ima vrednost $\chi^2 = 44{,}149$, kar je večje od kvantila $F^{-1}_{\chi^2(2)}(1-\alpha)$, tako pri stopnji tveganja $\alpha = 0{,}05$, kot tudi pri $\alpha = 0{,}01$, kjer je $F_{\chi^2(2)}$ kumulativna porazdelitvena funkcija porazdelitve $\chi^2(2)$. To pomeni, da našo ničelno hipotezo, ki pravi, da so lomi palic porazdeljeni binomsko $\mathrm{Bin}(5, \hat{p})$, lahko na podlagi tega testa zavržemo.

c) Denimo, da imamo podanih n neodvisnih opažanj $X_i \sim \mathrm{Bin}(m_i, p_i)$ z znanimi parametri m_i in neznanimi p_i. Ogledali si bomo kako s pomočjo razmerja verjetji preizkusimo ničelno domnevo, da so vsi parametri p_i enaki proti alternativni domnevi, da temu ni tako.

Naj bo *funkcija verjetja* ali samo *verjetje*

$$L(p_1, \ldots, p_n; x_1, \ldots, x_n) = \prod_{i=1}^{n} \binom{m_i}{x_i} p_i^{x_i} (1 - p_i)^{m_i - x_i}.$$

Naj bosta $\Theta = (0{,}1)^n$ in $\Theta_0 = \{(p_1, \ldots, p_n) \in \Theta \mid p_1 = \cdots = p_n\}$. Zanima nas razmerje verjertji

$$\Lambda = \frac{\sup_{\boldsymbol{p} \in \Theta_0} L(\boldsymbol{p}; \boldsymbol{x})}{\sup_{\boldsymbol{p} \in \Theta} L(\boldsymbol{p}; \boldsymbol{x})},$$

za poljuben $\boldsymbol{x} \in \{0,1,\ldots,5\}^n$, kjer imamo krajše zapisano $\boldsymbol{p} = (p_1,\ldots,p_n)$ in $\boldsymbol{x} = (x_1,\ldots,x_n)$. Označimo še

$$\begin{aligned}\ell(\boldsymbol{p};\boldsymbol{x}) &:= \log L(\boldsymbol{p};\boldsymbol{x}) \\ &= \sum_{i=1}^{n} \log \binom{m_i}{x_i} + \sum_{i=1}^{n} x_i \log(p_i) + \sum_{i=1}^{n} (m_i - x_i)\log(m_i - p_i)\end{aligned}$$

logaritem verjetja.

Lotimo se najprej računanja imenovalca razmerja verjetji $\sup_{\boldsymbol{p}\in\Theta} L(\boldsymbol{p};\boldsymbol{x})$. Ker je verjetje gladko odvisna funkcija parametra $\boldsymbol{p} \in \Theta \subseteq \mathbb{R}^n$, bo maksimum zavzela v stacionarni točki, kar pomeni, da bomo iskali ničle vseh pracialnih odvodov logaritma verjetja ℓ, saj bo tako računanje lažje. Vidimo, da velja

$$\frac{\partial \ell}{\partial p_j}(\boldsymbol{p};\boldsymbol{x}) = \frac{x_j}{p_j} - \frac{m_i - x_i}{m_i - p_i} = \frac{x_i - m_i p_i}{p_i(1-p_i)} = 0$$

natanko tedaj, ko je $p_i = \frac{x_i}{m_i}$. S podobnim argumentiranjem kot pri točki a) lahko zaključimo, da $\ell(\boldsymbol{p};\boldsymbol{x})$ oziroma ekvivalentno $L(\boldsymbol{p};\boldsymbol{x})$ doseže maksimum v točki $\boldsymbol{p} = (\frac{x_1}{m_1},\ldots,\frac{x_n}{m_n}) \in \Theta$. Tedaj je torej

$$\sup_{\boldsymbol{p}\in\Theta} L(\boldsymbol{p};\boldsymbol{x}) = \prod_{i=1}^{n} \binom{m_i}{x_i} \left(\frac{x_i}{m_i}\right)^{x_i} \left(\frac{m_i - x_i}{m_i}\right)^{m_i - x_i}.$$

Sedaj se posvetimo še računanju števca $\sup_{\boldsymbol{p}\in\Theta_0} L(\boldsymbol{p};\boldsymbol{x})$. Ožjo množico Θ_0 lahko parametriziramo z $p \mapsto (p,\ldots,p)$ za $p \in (0,1)$, torej bo $\ell(p,\ldots,p;\boldsymbol{x})$ gladka funkcija zgolj enega parametra $p \in (0,1)$. Z odvajanjem po tem parametru dobimo

$$\begin{aligned}\frac{d\ell}{dp}(p,\ldots,p;\boldsymbol{x}) &= \sum_{i=1}^{n} \left(\frac{x_i}{p} - \frac{m_i - x_i}{1-p}\right) \\ &= \sum_{i=1}^{n} \frac{x_i - m_i p}{p(1-p)} = \frac{\sum_{i=1}^n x_i - p\sum_{i=1}^n m_i}{p(1-p)},\end{aligned}$$

kar ima ničlo natanko pri $p = \sum_{i=1}^n x_i / \sum_{i=1}^n m_i$. Prepričamo se lahko, da je v tej točki tudi zares maksimum ℓ na Θ_0, torej imamo

$$\sup_{\boldsymbol{p}\in\Theta_0} L(\boldsymbol{p};\boldsymbol{x}) = \prod_{j=1}^{n} \binom{m_j}{x_j} \left(\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n m_i}\right)^{x_j} \left(\frac{\sum_{i=1}^n (m_i - x_i)}{\sum_{i=1}^n m_i}\right)^{m_j - x_j}.$$

Nazadnje je razmerje verjetji

$$\begin{aligned}\Lambda &= \frac{\sup_{\boldsymbol{p}\in\Theta_0} L(\boldsymbol{p};\boldsymbol{x})}{\sup_{\boldsymbol{p}\in\Theta} L(\boldsymbol{p};\boldsymbol{x})} \\ &= \prod_{j=1}^{n} \frac{\binom{m_j}{x_j}\left(\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n m_i}\right)^{x_j}\left(\frac{\sum_{i=1}^n (m_i-x_i)}{\sum_{i=1}^n m_i}\right)^{m_j-x_j}}{\binom{m_j}{x_j}\left(\frac{x_j}{m_j}\right)^{x_j}\left(\frac{m_j-x_j}{m_j}\right)^{m_j-x_j}} \\ &= \prod_{j=1}^{n}\left(\frac{m_j}{x_j}\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n m_i}\right)^{x_j}\left(\frac{m_j}{m_j-x_j}\frac{\sum_{i=1}^n (m_i-x_i)}{\sum_{i=1}^n m_i}\right)^{m_j-x_j}.\end{aligned}$$

Porazdelitev te testne statistike je povsem neočitna in izkaže se, da si ne moremo niti pomagati z Wilksovim izrekom, zato izberemo drugačno pot. Simuliramo veliko število testne statistike, recimo $N = 10.000$ in potem na podlagi tega definiramo porazdelitev s pomočjo histograma. Realno os razdelimo na intervale širine

$$\delta = \hat{\sigma}\sqrt[3]{\frac{24\sqrt{\pi}}{N}} \quad \text{po Scottovem pravilu,}$$

nato pa preštejemo koliko simuliranih testnih statistik pade v dani interval ($\hat{\sigma}$ predstavlja empirični standardni odklon). Frekvenca – tj. število simuliranih testnih statistik v danem intevalu deljeno z N – bo verjetnost, da slučajna spremenljivka zavzame vrednost ravno v tem intervalu. Natančneje to pomeni, da definiramo

$$\mathsf{P}(\Lambda \in [\delta k, \delta(k+1))) := \frac{\#\text{ simuliranih podatkov, ki pade v } [\delta k, \delta(k+1))}{N}$$

za poljuben $k \in \mathbb{Z}$. Na podlagi tako definirane porazdelitve slučajne spremenljivke Λ, lahko izračunamo p-vrednost testne statistike λ_{stat} našega vzorca kot

$$p = \mathsf{P}(|\Lambda| \geq \lambda_{stat}) = \mathsf{P}(\Lambda \leq \lambda_{stat}) + \mathsf{P}(\Lambda \geq \lambda_{stat}).$$

d) V zadnji točki te naloge uporabimo metodo *bootstrap* kjer simuliramo 10.000 vrednosti testne statistike Λ pod ničelno domnevo, da naš vzorec sestavlja 280 binomsko $\mathrm{Bin}(5, \hat{p})$ porazdeljenih neodvisnih opažanj $X_i \sim \mathrm{Bin}(5, \hat{p})$. Ob tem je $\hat{p} = \overline{X}/5$ cenilka za p po metodi največjega verjetja iz točke a). Glede na takšen vzorec se naša testna statistika poenostavi v

$$\Lambda = \prod_{i=1}^{n}\left(\frac{\overline{X}}{X_i}\right)^{X_i}\left(\frac{5-\overline{X}}{5-X_i}\right)^{5-X_i},$$

kjer je $\overline{X} = \sum_{i=1}^n X_i/n$ vzorčno povprečje. Na našem vzorcu testna statistika zavzame vrednost $\lambda_{stat} = 3{,}02 \cdot 10^{-97}$.

Test pokaže, da so prav vse vrednosti testne statistike na 10.000 simuliranih vzorcih vedno večje od λ_{stat} izračunane na naših konkretnih podatkih, torej bomo ničelno hipotezo, da imajo vse palice zlome porazdeljene binomsko z enakimi parametri, s p-vrednostjo 0 zavrgli.

Skripto, ki spada k tej nalogi, najdemo v datoteki `naloga_2.ipynb`.

3. naloga

V tej nalogi se bomo ukvarjali z analizo pulza pri študentih. V raziskavi so nekateri od študentov bili deležni fizične obremenitve (tek na mestu), pulz pa so jim izmerili pred in po obremeniti. Tudi študentom, ki niso tekli so za kontrolo dvakrat izmerili pulz. V naslednjih treh točkah bomo obravnavali kakšni dejavniki bi lahko vplivali na spremembo pulza med obema meritvama.

a) Za začetek bomo testirali ali je fizična obremenitev vplivala na spremembo pulza. Za spremembo pulza bomo vzeli razliko med zadnje izmerjenim pulzom in prvotno izmerjenim pulzom. Naj $\overline{X}_1$ označuje povprečno spremembo pulza v skupini študentov deležnih obremenitve, $\overline{X}_2$ pa povprečno spremembo pulza preostalih študentov.

Za ničelno hipotezo si bomo zadali, da fizična obremenitev *ne* vpliva na spremembo pulza. Pod to predpostavko bi bilo smiselno sklepati, da sta povprečji spremembe pulza v dveh skupinah študentov (z oziroma brez obremenitve) v splošnem enaki. Bolj formalno bomo torej testirali ničelno hipotezo

$$H_0 : \mu_1 = \mu_2,$$

kjer μ_1 predstavlja povprečno spremembo pulza splošnega študenta, ki je med meritvama opravljal fizično obremenitev, μ_2 pa predstavlja spremembo pulza splošnega študenta, ki fiziče obremenitve ni izvajal (teh dveh podatkov ne poznamo). Z drugimi besedami bomo preizkušali ali nam medsebojno odstopanje vzorčnih povprečji $\overline{X}_1$ in $\overline{X}_2$ pojasni posebna izbira našega vzorca ali pa je za tem razlog, da obremenitev res vpliva na spremembo pulza.

Predpostavljamo, da je pulz pri študentih porazdeljen normalno (kot je značilno za številne druge človeške karakteristike), torej bo tudi sprememba pulza kot razlika dveh normalnih podatkov normalno porazdeljena. Ob tem pa smo se odločili, da hipotezo preizkušamo z *Welchevim t-testom* – katerega opis še sledi, namesto s Studentovim t-testom. Razlog za tem pojasni veliko odstopanje varianc po skupinah. Če namreč označimo

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2$$

nepristransko cenilko variance povprečne spremembe pulza v skupini študentov z obremenitvijo in podobno definiramo s_2^2, potem vidimo, da je na

naših podatkih $s_1 > 2s_2$. Pri tem je $n_1 = 46$ študentov bilo deležnih fizične obremenitve, $n_2 = 63$ pa je bilo študentov brez obremenitve.

Pri Welchevem t-testu je testna statistika podana kot

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2_{\overline{X}_1} + s^2_{\overline{X}_2}}}, \quad \text{kjer je} \quad s^2_{\overline{X}_i} = \frac{s_i^2}{n_i}.$$

Ta statistika je aproksimirana s Studentovo t-porazdelitvijo s približno

$$\hat{v} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

prostostnimi stopnjami. To je Welch-Satterthwaitova enačba. Podobna je parametru, ki ga uporabljamo pri ocenjevanju vzorčnega povprečja pri stratificiranem vzorčenju, kjer ne poznamo stratumskih varianc (če sta stratuma samo dva). Podrobnosti o Welchevem t-testu najdemo v [2] ali [1, §11].

Ker zdaj (vsaj približno) poznamo porazdelitev testne statistike, jo lahko izvrednotimo na naših podatkih in tako izračunamo p-vrednost[1] našega opažanja, ki nam pove kako ekstremno je. Naš test bo obojestranski, torej izračunamo p-vrednost kot

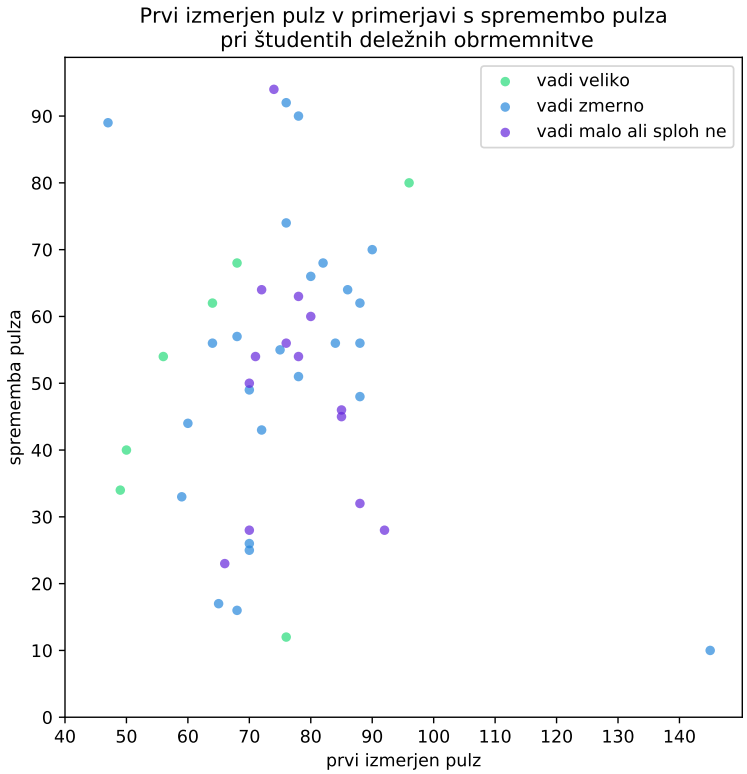
$$\begin{aligned} p = \mathsf{P}(|T| \geq t_{stat}) &= \mathsf{P}(T \leq -t_{stat}) + \mathsf{P}(T \geq t_{stat}) \\ &= 2(1 - \mathsf{P}(T < t_{stat})) = 2(1 - F_{\text{Student}(\hat{v})}(t_{stat})), \end{aligned}$$

kjer je t_{stat} vrednost statistične spremenljivke na naših podatkih, $T \sim \text{Student}(\hat{v})$, $F_{\text{Student}(\hat{v})}$ pa je kumulativna porazdelitvena funkcija Studentove porazdelitve z $\hat{v}$ prostostnimi stopnjami. V tretjem enačaju smo upoštevali simetričnost Studentove porazdelitve okoli izhodišča 0. Tako izračunana p-vrednost naših opažanj dosega velikostne rede 10^{-21}, kar je *zelo* malo in nam jasno omogoča zavreči ničelno hipotezo (pri stopnji tveganja $\alpha = 0{,}01$) ter sprejeti (dosti bolj smiselno) alternativno domnevo, da obremenitev vpliva na spremembo pulza študentov.

b) Empirično bi lahko iz grafikona 2 sklepali, da študentje izbrani za obrmeneitev niso goljufali s tem, da niso tekli. Vsi, ki so imeli majhno spremembo pulza (recimo manj kot 30) bodisi vadijo veliko in imajo natrenirano srce, ali pa so že prvo meritev opravili z neobičajno visokim pulzom in se jim zaradi tega ni mogel med obremenitvijo drastično dvigniti. Mogoče so hiteli na testiranje, ker so zamujali, in so zaradi tega že imeli povišan pulz ob prvem merjenju.

c) Najprej testirajmo, če so vzorčne variance vseh treh skupin enake – homoskedastičnost (homogenost variance). V primeru da so, bomo uporabili

[1] p-vrednost opažanja je najmanjša možna meja pri kateri ničelno hipotezo zavrnemo z vsaj tolikšnjo stopnjo tveganja.



Slika 2: Sprememba pulza v odvisnosti od prvo izmerjenega pulza.

analizo variance – krajše ANOVA – za preizkušanje domneve ali so povprečja po skupinah enaka.

Homogenost variance bomo testirali z Levenovim testom, ki je v resnici zgolj modifikacija testa ANOVA, zato si najprej poglejmo kako deluje ta. Glavni cilj analize variance je ugotoviti ali se povprečja po skupinah med sabo statistično razlikujejo ali ne. Recimo, da imamo danih n vzorecev Y_{ij}, za $1 \leq j \leq n_i$, $1 \leq i \leq k$ razdeljenih na k skupin, kjer vsaka skupina vsebuje n_i vzorcev. Z $\overline{Y}_{i\cdot}$ označimo povprečje znotraj i-te skupine, $\overline{Y}$ pa naj bo vzorčno povprečje vseh podatkov. Osnovna ideja tega testa bo pogledati razmerje variabilnosti vzorca med skupinami in variabilnosti vzorca v skupinah samih. Natančneje definirajmo

$$\sigma_B^2 = \sum_{i=1}^{k} \frac{n_i(\overline{Y}_{i\cdot} - \overline{Y})^2}{k-1},$$

imenovano tudi *pojasnjena varianca* – varianca med skupinami vzorca, ter

$$\sigma_W^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(Y_{ij} - \overline{Y}_{i\cdot})^2}{n-k},$$

ki se imenuje *nepojasnjena varianca* – varianca znotraj skupin vzorca. Testno statistiko tako vpeljemo kot

$$F = \frac{\sigma_B^2}{\sigma_W^2}.$$

Izkaže se, da sta pojasnjena in nepojasnjena varianca hi kvadrat porazdeljeni in sicer je $(k-1)\sigma_B^2 \sim \chi^2(k-1)$ ter $(n-k)\sigma_W^2 \sim \chi^2(n-k)$. Tako spoznamo novo porazdelitev imenovano *F-porazdelitev s parametroma d_1 in d_2*, ki je definirana kot porazdelitev spremenljivke

$$\frac{U_1/d_1}{U_2/d_2} \sim F(d_1, d_2),$$

kjer sta $U_1 \sim \chi^2(d_1)$ in $U_2 \sim \chi^2(d_2)$ neodvisni. Ničelno hipotezo tedaj zavrnemo, če je p-vrednost, ki jo izračunamo kot

$$p = \mathsf{P}(F \geq f_{stat}) = 1 - F_{F(k-1,n-k)}(f_{stat})$$

kjer je $F \sim F(k-1, n-k)$, $F_{F(k-1,n-k)}$ kumulativna porazdelitvena funkcija te spremenljivke, f_{stat} pa vrednost statistične spremenljivke izračunane na našem vzorcu, manjša od zadane stopnje tveganja α.

Omenili smo že, da je Levenov test zgolj prilagojena različica tega testa in sicer spremenljivke $\overline{Y}_{ij}$ zamenjamo z $Z_{ij} = |Y_{ij} - \overline{Y}_{i\cdot}|$. Tako dobimo testno statistiko

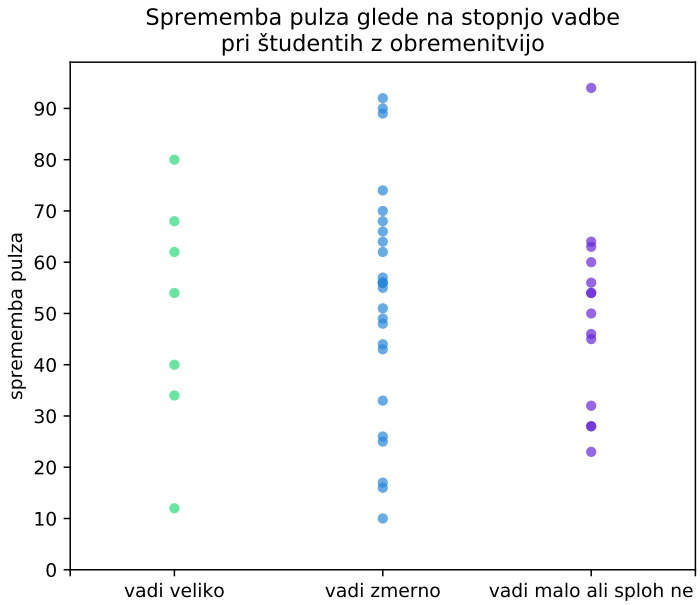
$$W = \frac{n-k}{k-1} \cdot \frac{\sum_{i=1}^{k} n_i(\overline{Z}_{i\cdot} - \overline{Z})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \overline{Z}_{i\cdot})^2},$$

ki je $F(k-1, n-k)$-porazdeljena. V enačbi $\overline{Z}_{i\cdot}$ predstavlja povprečje spremenljivk Z_{ij} v i-ti skupini, $\overline{Z}$ pa je povprečje vseh Z_{ij}.

S p-vrednostjo 0,660, lahko po Levenovem testu sklepamo, da imajo tri skupine študentov z obremenitvijo glede na to koliko vadijo približno enake variance spremembe pulza. Z drugimi besedami je sprememba pulza v vseh treh skupinah približno enako razpršena okrog povprečja, kot lahko empirično do neke mere potrdimo iz grafikona 3.

Sedaj bomo za študente, ki so bili deležni obrmemenitve, testirali ničelno hipotezo

$$H_0 : \text{vadba nima vpliva na spremembo pulza.}$$



Slika 3: Razpršenost spremembe pulza po skupinah glede na vadbo.

Z drugimi besedami, bi bilo smiselno za ničelno hipotezo vzeti $H_0 : \mu_1 = \mu_2 = \mu_3$, kjer so μ_1, μ_2, μ_3 povprečja splošnih študentov, ki vadijo veliko, zmerno oziroma malo ali skoraj nič. To hipotezo bomo testirali z analizo variance. Ta test prepostavlja naslednje:

- Ekspirimentalna napaka pri merjenju podatkov je približno normalno porazdeljena (to bomo privzeli).
- Homogenost variance (to nam je potrdil Levenov test zgoraj)
- Posamezne enote vzorca so neodvisne drug od druge (tudi to bomo privzeli)

Vidimo, da je p-vrednost 0,907 precej velika, kar pomeni, da bi vsaj tako ekstremen rezulat, kot je naš, dobili dokaj pogosto – naš vzorec torej ni preveč ekstremen. Na podlagi tega testa lahko tako statistično sklepamo, da je ničelna hipoteza resnična, kar pomeni, da vadba nima vpliva na spremembo pulza pri študentih z obrmenitvijo.

Skripto za izračune in analizo zadnje naloge najdemo v datoteki `naloga_3.ipynb`.

Literatura

[1] J. Rice, *Mathematical statistics & data analysis*, third edition, Duxbury, 2007

[2] *Welch's t-test*, v: Wikipedia, The Free Encyclopedia, [ogled 21. 8. 2021], dostopno na `https://en.wikipedia.org/wiki/Welch%27s_t-test`.