Seminarska naloga iz statistike

Žiga Gladek

September 23, 2021

Povzetek

V tem dokumentu so povzete obravnave nalog, ki sem jih preučeval v sklopu predmeta statistika na programu Matematika UN. Pri reševanju sem si pomagal s programskim jezikom Python, z orodjem Jupyter Notebook in s knjižnico pandas za branje datotek tipa csv. Za prikaze podatkov sem uporabljal knjižnico matplotlib. Za večino uporabljenih metod se sklicujem na zapiske iz predavanj, za preostanek pa na priporočeno literaturo [1].

1. naloga: Kibergrad

Točka a)

Ustrezne datoteke za to in ostale točke pri tej nalogi se nahajajo v datoteki **Kibergrad.ipynb**. Najprej moramo vzeti enostavni slučajni vzorec 200 družin in oceniti delež družin v Kibergradu, v katerih vodja gospodinjstva nima mature. To pomeni, da je v tabeli takega gospodinjstva stopnja izobrazbe ≤ 38. Pravi delež bomo označili s p. Nepristranska ocena za p na podlagi dobljenega vzorca je:

$$\hat{p} = \frac{\text{št. gospodinjstev v vzorcu brez mature}}{200}.$$

Lahko si mislimo, da smo v enostavnem slučajnem vzorcu gospodinjstvom, v katerih ima vodja vsaj maturo, priredili število 0, tistim, v katerih ima nižjo izobrazbo pa 1. V tem primeru je ta ocena le povprečje vrednosti v dobljenem vzorcu. Dobimo, da je $\hat{p} = 0,18500$. Seveda pri različnih vzorcih lahko dobimo drugačno oceno.

Točka b)

Ocenili bomo standardno napako, ki je po definiciji enaka $se = \sqrt{var(\hat{p})}$. Naj bo N velikost populacije, n velikost vzorca in σ^2 populacijska varianca. Potem vemo, da pri enostavnem slučajnem vzorcu velja formula

$$se = \sqrt{\frac{N-n}{N-1}\frac{\sigma^2}{n}}.$$

V resnici pa σ^2 še ne poznamo, zato moramo najprej oceniti še to. V našem primeru, kjer smo vse družine v vzorcu razdelili v dve skupini, velja $\sigma^2 = p(1-p)$. To znamo oceniti s pomočjo $\hat{p}$ kot $\hat{\sigma}^2 = \hat{p}(1-\hat{p})$, vendar pa ta ocena ni nepristranska. Iz predavanj vemo, da jo lahko do nepristranske popravimo na naslednji način:

$$\hat{\sigma}^2_+ = \frac{N-1}{N}\frac{n}{n-1}\hat{p}(1-\hat{p}).$$

S tem se lahko dokopljemo še do nepristranske ocene za standardno napako, tako da v formuli za standardno napako σ^2 zamenjamo s $\hat{\sigma}^2_+$. Dobimo:

$$\hat{se}_+ = \sqrt{\frac{N-n}{N}\frac{1}{n-1}\hat{p}(1-\hat{p})}.$$

V našem primeru je $N = 43886$ in $n = 200$. Pri danem vzorcu je $\hat{se}_+$ enaka približno $0,02746$. Na podlagi te ocene sedaj lahko izpeljemo še 95% interval zaupanja za p. Da ga dobimo, bomo upoštevali, da je približno $\frac{\hat{p}-p}{\hat{se}_+} \sim Student(n-1)$. Pri stopnji tveganja $\alpha = 0,05$ dobimo aproksimativni interval zaupanja za p:

$$p \in \left[\hat{p} - F^{-1}_{Student(199)}(0,975)\hat{se}_+, \quad \hat{p} + F^{-1}_{Student(199)}(0,975)\hat{se}_+\right].$$

Ta je pri danih podatkih enak približno $[0,13062, 0,23938]$. Pri tem smo na podlagi tabele kvantilov Studentove porazdelitve upoštevali, da je $F^{-1}_{Student(199)}(0,975) \doteq 1,98$. Dejansko smo uporabili kvantil, ki ustreza 120 prostostnim stopnjam, vendar je to za naše namene dovolj dober približek.

Točka c)

Poglejmo populacijski delež gospodinjstev, v katerih vodja gospodinjstva nima srednješolske izobrazbe. Ta delež je enak:

$$p = \frac{\text{št. gospodinjstev brez mature}}{N} \doteq 0,21150.$$

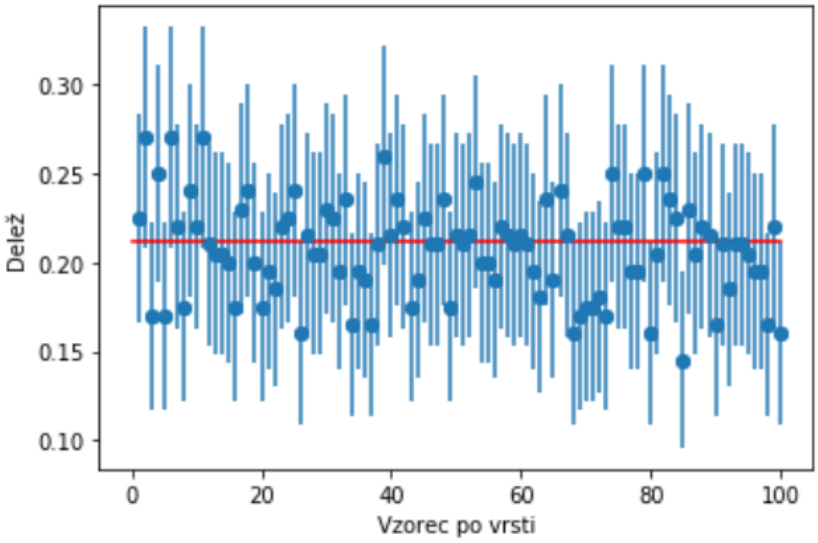
Naša točkovna ocena se od točne vrednosti torej razlikuje za slabe 3%, opazimo pa tudi, da prej dobljeni aproksimativni interval zaupanja pokrije populacijski delež. S tem da poznamo p, lahko izračunamo tudi pravo standardno napako pri vzorcih velikosti 200. Upoštevajmo, da je $\sigma^2 = p(1-p)$. Tedaj je:

$$se = \sqrt{\frac{N-n}{N-1}\frac{\sigma^2}{n}} = \sqrt{\frac{N-n}{N-1}\frac{p(1-p)}{n}} \doteq 0,02881.$$

Približek $\hat{se}_+$ se torej od točne vrednosti razlikuje šele na tretji decimalki.

Točka d)

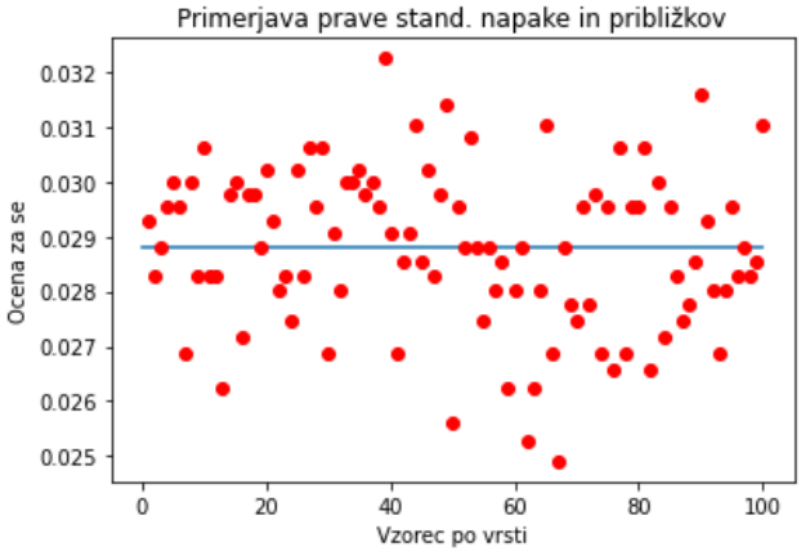
Vzeli bomo še 99 enostavnih slučajnih vzorcev in za vsakega od teh določili 95% interval zaupanja za p. S pomočjo programa dobimo, da je delež intervalov, ki pokrije populacijski delež enak $\frac{95}{100}$. Poglejmo si to še s sliko:



Rdeča črta na sliki prikazuje populacijski delež, intervali zaupanja pa so narisani vertikalno, pri čemer je vrednost na sredini intervala dodatno označena z modro piko. Število intervalov, ki je pokrilo populacijski delež ni presenetljivo in kvečjemu potrdi, da smo pravilno določilli interval zaupanja, saj 95% interval zaupanja pomeni ravno to, da bo populacijski delež v povprečju pokril v 95 od 100 primerih.

Točka e)

Izračunajmo standardne odklone vzorčnih deležev iz prejšnjih 100 vzorcev in jih primerjajmo s pravo standardno napako pri vzorcih velikosti 200, ki jo poznamo že od prej. Standardni odklon vzorčnega deleža $\hat{p}$ je enak $\sqrt{var(\hat{p})}$, kar pa je ravno standardna napaka. To vrednost lahko ocenimo na podlagi vsakega intervala posebej. Primerjavo teh s pravo vrednostjo nato lahko ponazorimo z grafikonom:



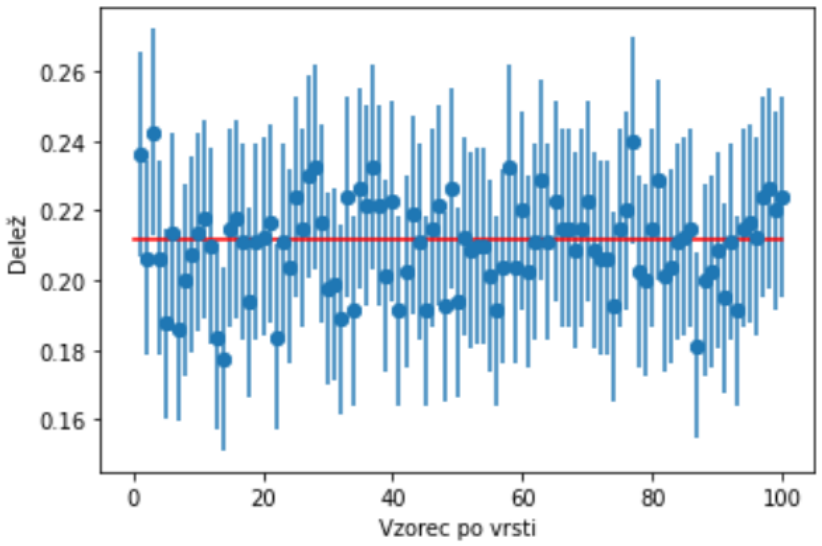
Pri tem modra vodoravna črta označuje pravo standardno napako pri vzorcih te velikosti.

Točka f)

Ponovimo prejšnji dve točki še za primer, ko so vzorci velikosti 800, torej $n = 800$. Interval zaupanja za p je tedaj oblike

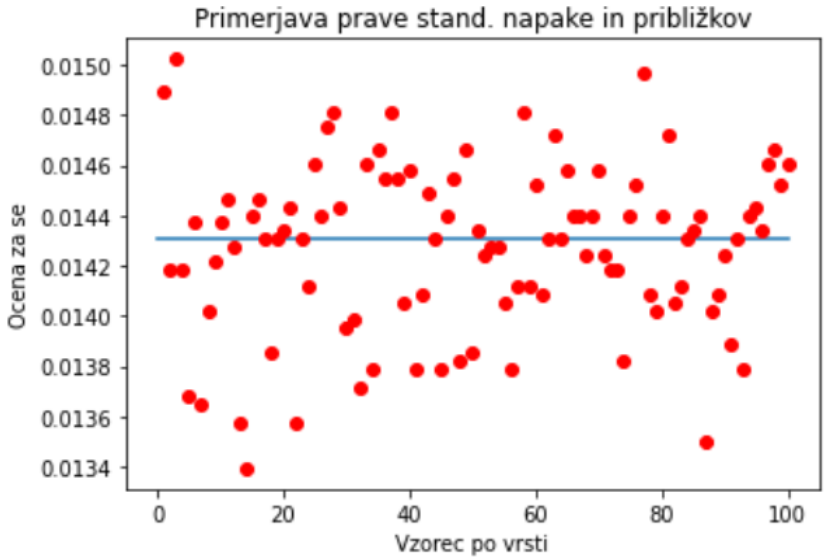
$$p \in \left[\hat{p} - F^{-1}_{Student(799)}(0,975)\hat{se}_+, \quad \hat{p} + F^{-1}_{Student(799)}(0,975)\hat{se}_+\right],$$

pri čemer bomo vzeli $F^{-1}_{Student(799)}(0,975) = 1,97$. Seveda to ni točna vrednost, ampak bo za naše namene dovolj dober približek. Verjetno bi lahko tu vzeli tudi $1,98$ ali pa vrednost, ki je nekje vmes med tema dvema. Delež intervalov, ki pokrije populacijski delež je znova enak $\frac{95}{100}$. Oglejmo si jih še s sliko:



Na prvi pogled ni vidnih razlik od prejšnjega primera, vendar pa so ti novi intervali dejansko ožji od tistih prej. Povprečna širina teh novih intervalov je enaka približno $0,05633$ v primerjavi s povprečjem v prejšnjem primeru, ki je bilo približno enako $0,11371$. Dejansko so intervali skoraj dvakrat ožji. Izkazalo se bo, da se to zgodi na račun približno dvakrat manjše standardne napake.

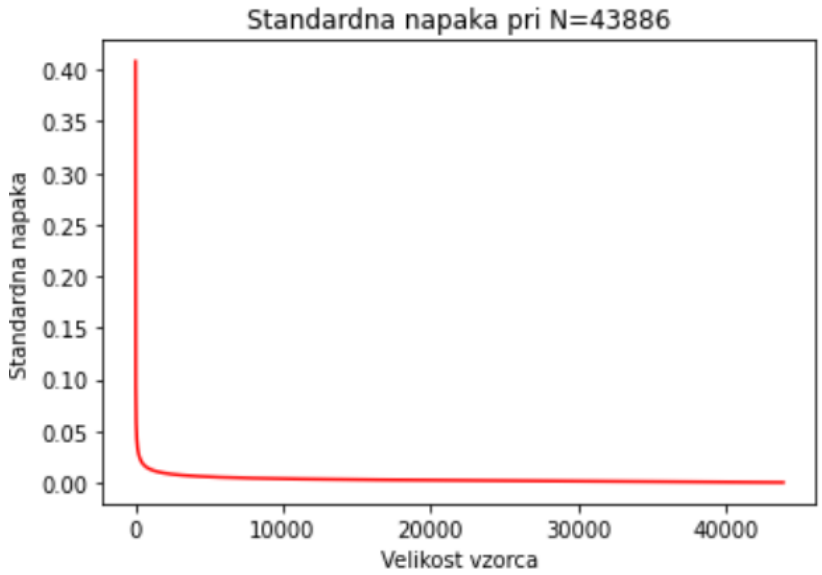
Populacijski delež seveda ni odvisen od vzorca in je enak kot prej. Za razliko od tega je prava standardna napaka pri vzorcu velikosti 800 drugačna kot pri vzorcu velikosti 200. Pri $n = 800$ dobimo, da je $se \doteq 0,01431$, kar je občutno manj kot v prejšnjem primeru. Poglejmo še, kako izgledajo približki za se, ki jih dobimo na podlagi dobljenih vzorcev.



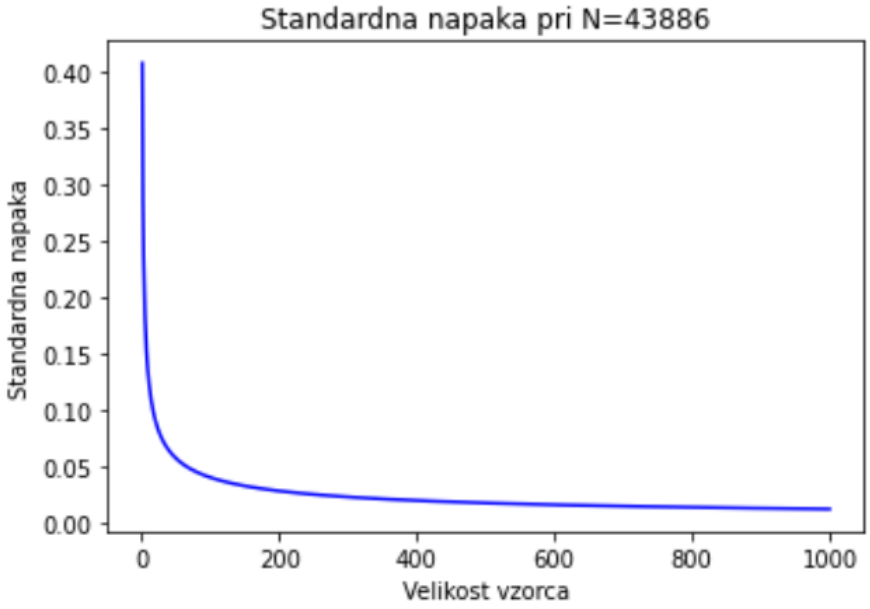
Sliki sicer izgledata podobno, vendar pa so približki v tem drugem primeru dejansko manj razpršeni okrog prave vrednosti, kot v prvem primeru, kar se vidi, če pogledamo vrednosti na y-osi. Da je standardna napaka v tem primeru manjša, ni presenetljivo, saj je le-ta merilo za to, kako natančno vzorčni delež aproksimira pravi delež. Že intuitivno je jasno, da če bo vzorec večji, bo natančnost boljša, se pa to vidi tudi iz eksplicitnega zapisa. Spomnimo se, da velja

$$se = \sqrt{\frac{N-n}{N-1}\frac{\sigma^2}{n}}.$$

Očitno je, da gre ta izraz proti 0, ko gre n proti N. Oglejmo si odvisnost SE od velikosti vzorca za dani N še z grafom.



Podrobeneje si lahko pogledamo, kako hitro pada pri vrednostih, s katerimi imamo opravka.



Seveda je standardna napaka odvisna tudi od populacijske variance σ^2. Manjša kot je, manjše vzorce rabimo, da dobimo dobre približke. Opazimo pa še nekaj. Standardna napaka pri vzorcu velikosti 800 je približno dvakrat

manjša kot pri vzorcu velikosti 200 in to ni naključje. Če zgornjo formulo zapišemo nekoliko drugače, dobimo

$$se = \sqrt{1 - \frac{n-1}{N-1}} \frac{\sigma}{\sqrt{n}}.$$

Če je n majhen v primerjavi z N, je izraz pod korenom približno enak 1, kar pomeni da je

$$se \approx \frac{\sigma}{\sqrt{n}}.$$

Iz tega izraza se vidi, da za dvakrat manjšo standardno napako, potrebujemo štirikrat večji vzorec, kar pa je ravno tisto, kar smo v našem primeru naredili. To razloži tudi razliko v širinah intervalov. Dolžina posameznega intervala je enaka $2 \cdot 1,98 \cdot \hat{se}_+$. Ker je standardna napaka približno dvakrat manjša, se to odraža tudi v približkih $\hat{se}_+$, zato so intervali približno dvakrat ožji, kot v prvem primeru. Na kratko komentirajmo še kako je z rapršenostjo približkov za standardno napako okrog prave vrednosti. Omenili smo že, da je v drugem primeru ta razpršenost razvidno manjša. Da bi to kvantificirali, bi morali izračunati varianco približka $\hat{se}_+$. Ta je zaradi nepristranskosti, ravno enaka srednji kvadratični napaki te cenilke, kar pa vemo, da je v nekem smislu merilo za to, kako dobra je cenilka.

$$MSE(\hat{se}_+) = var(\hat{se}_+) + Bias(\hat{se}_+)^2 = var(\hat{se}_+).$$

Tega ne bomo popolnoma do konca izračunali, vendar se odvisnost od n delno vidi že brez tega. Velja namreč:

$$var(\hat{se}_+) = var\left(\frac{N-n}{N(n-1)}\hat{p}(1-\hat{p})\right) = \left(\frac{N-n}{N(n-1)}\right)^2 var(\hat{p}(1-\hat{p})).$$

Seveda je $\hat{p}$ tudi odvisen od velikosti vzorca, vendar pa lahko še vedno dobimo približek, če izraz v oklepaju zapišemo drugače:

$$\frac{N-n}{N}\frac{1}{n-1} = \left(1 - \frac{n}{N}\right)\frac{1}{n-1} \approx \frac{1}{n-1}.$$

V zadnjem koraku smo upoštevali, da je v našem primeru n majhen v primerjavi z N. Tako je

$$var(\hat{se}_+) \approx \frac{1}{(n-1)^2} var(\hat{p}(1-\hat{p}))$$

Tu se že nekoliko vidi, kako bo razpršenost padala, ko vzorec povečujemo. Pričakujemo namreč, da bodo z večjimi vzorci boljši tudi približki za populacijsko varianco in da bo tudi razpršenost teh približkov padala.

2. naloga: TempPulz

Točka a)

Ustrezne datoteke za to in ostale točke pri tej nalogi se nahajajo v datoteki **TempPulz.ipynb**. Dane imamo odčitke telesnih temperatur in pulzov 65 moških in 65 žensk. Pri tem pa dodatno predpostavljamo, da sta telesna temperatura in pulz pri moških in pri ženskah porazdeljena normalno. Seveda ni nujno, da so te normalne porazdelitve enake. Za namene naloge se bomo omejili na obravnavo telesnih temperatur. Naj bo telesna temperatura moških porazdeljena $\sim N(\mu_1, \sigma_1^2)$, telesna temperatura žensk pa $\sim N(\mu_2, \sigma_2^2)$. Poleg tega označimo z X_i odčitek telesne temperature za i-tega moškega in z Y_j odčitek telesne temperature za j-to žensko. Skozi celoten potek bomo predpostavljali še, da so vse meritve med seboj neodvisne. Za začetek ocenimo povprečje in standardni odklon za telesno temperaturo žensk in moških posebej. Za oceni povprečij vzamemo kar empirični povprečji odčitkov:

$$\hat{\mu}_1 = \frac{1}{65}\sum_{i=1}^{65} X_i, \quad \hat{\mu}_2 = \frac{1}{65}\sum_{j=1}^{65} Y_j.$$

V programu to lahko izračunamo z uporabo funkcije *mean* na ustreznih seznamih. Dobimo $\hat{\mu}_1 \doteq 98,10462°F$ in $\hat{\mu}_2 \doteq 98,39385°F$. Med Celzijevimi in Fahrenheitovimi stopinjami lahko prehajamo s pomočjo formule $y = 5(x-32)/9$, kjer $x°F = y°C$. V Celzijevih stopinjah sta oceni torej enaki $\hat{\mu}_1 \doteq 36,72479°C$ in $\hat{\mu}_2 \doteq 36,88547°C$.

Ocenimo še standardna odklona σ_1 in σ_2. Če se osredotočimo na odčitke temperatur pri moških, znamo to znova oceniti z empiričnim standardnim odklonom, vendar pa ta ocena ni nepristranska. Vzemimo:

$$\hat{\sigma}_1 = \sqrt{\frac{1}{65-1}\sum_{i=1}^{65}(X_i - \hat{\mu}_1)^2}, \quad \hat{\sigma}_2 = \sqrt{\frac{1}{65-1}\sum_{j=1}^{65}(Y_j - \hat{\mu}_2)^2}.$$

Dobimo $\hat{\sigma}_1 \doteq 0,69876°F = 0,38819°C$ in $\hat{\sigma}_2 \doteq 0,74349°F = 0,41305°C$.

Točka b)

Sedaj bomo določili 95% interval zaupanja za μ_1 in μ_2. Postavimo stopnjo tveganja $\alpha = 0,05$ in postavimo pogoj zaupanja

$$P(|\hat{\mu}_1 - \mu_1| < c) = 1 - \alpha.$$

Ker so X_i vsi enako porazdeljeni in neodvisni, velja $\hat{\mu}_1 \sim N(\mu_1, (\frac{\sigma_1}{65})^2)$. Označimo $se_1 = \frac{\sigma_1}{65}$. Potem velja $\frac{\hat{\mu}_1 - \mu_1}{se_1} \sim N(0,1)$. Od tu lahko nadaljujemo na vsaj dva načina. Problem je, ker σ_1 ne poznamo, vendar pa na njegovi podlagi vseeno lahko določimo interval zaupanja in šele zatem uporabimo približek. Najprej izračunamo

$$P(|\hat{\mu}_1 - \mu_1| < c) = P\left(\frac{|\hat{\mu}_1 - \mu_1|}{se_1} < \frac{c}{se_1}\right) = 2\Phi\left(\frac{c}{se_1}\right) - 1.$$

Iz pogoja zaupanja dobimo, da je $c = se_1 \cdot \Phi^{-1}(1 - \frac{\alpha}{2})$. Ker pa se_1 ne poznamo, ga na tej točki zamenjamo z nepristransko oceno $\hat{se}_1 = \frac{\hat{\sigma}_1}{65}$. Tako smo dobili aproksimativni 95% interval zaupanja

$$\left[\hat{\mu}_1 - \hat{se}_1\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \hat{\mu}_1 + \hat{se}_1\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right] = \left[\hat{\mu}_1 - \hat{se}_1 1,96, \hat{\mu}_1 + \hat{se}_1 1,96\right],$$

ki pa je asimptotično eksakten. Popolnoma analogno dobimo tak interval zaupanja za μ_2. Program nam pove, da je za μ_1 ta interval enak približno $[97,93474°F, 98,27449°F] = [36,63041°C, 36,81916°C]$, za μ_2 pa približno $[98,21309°F, 98,57459°F] = [36,78505°C, 36,98588°C]$.

Druga možnost je, da upoštevamo $\frac{\hat{\mu}_1 - \mu}{\hat{se}_1} \sim Student(65-1)$.

$$P(|\hat{\mu}_1 - \mu_1| < c) = P\left(\frac{|\hat{\mu}_1 - \mu_1|}{\hat{se}_1} < \frac{c}{\hat{se}_1}\right) = 2F_{Student(64)}\left(\frac{c}{\hat{se}_1}\right) - 1.$$

Iz pogoja zaupanja dobimo $c = \hat{se}_1 \cdot F^{-1}_{Student(64)}(1 - \frac{\alpha}{2}) = 2 \cdot \hat{se}_1$. V zadnjem enačaju smo upoštevali, da je $F^{-1}_{Student(64)}(1 - \frac{\alpha}{2}) \doteq 2,00$. V resnici smo uporabili kvantil pri 60 prostostnih stopnjah, ki je za nas dovolj dober približek. Za razliko od prejšnjega intervala, je ta eksakten. Za μ_1 je približno enak $[97,93128°F, 98,27796°F] = [36,62849°C, 36,82109°C]$, za μ_2 pa približno $[98,20941°F, 98,57828°F] = [36,78301°C, 36,98793°C]$.

Točka c)

Pri danih podatkih bomo sedaj preizkusili domnevo, da imajo moški in ženske v povprečju enako telesno temperaturo. Preizkusili jo bomo pri stopnjah tveganja $0,05$ in $0,01$. Ničelna domneva se torej glasi $H_0 : \mu_1 = \mu_2$. Na podlagi $\mu_1 - \mu_2$ bomo zasnovali dva testa. Za prvega bomo nekoliko idealizirali pogoje in predpostavili še $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Tedaj vemo, da je

$$\mu_1 - \mu_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{64} + \frac{1}{64}\right)\right) = N\left(\mu_1 - \mu_2, \frac{1}{32}\sigma^2\right).$$

V našem primer σ^2 ne poznamo, zato rabimo zanj približek, ki pa bo upošteval vzorčni varianci moških in žensk. Za splošna vzorca iz dveh takih porazdelitev, velikosti n in m lahko vzamemo

$$\sigma_p^2 = \frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{n+m-2}.$$

To je res nepristranska ocena za σ^2, saj je

$$\begin{aligned}\mathbb{E}\left(\frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{n+m-2}\right) &= \frac{n-1}{n+m-2}\mathbb{E}(\hat{\sigma}_1^2) + \frac{m-1}{n+m-2}\mathbb{E}(\hat{\sigma}_2^2) \\ &= \frac{n-1}{n+m-2}\sigma^2 + \frac{m-1}{n+m-2}\sigma^2 \\ &= \sigma^2.\end{aligned}$$

Pri tem smo upoštevali, da sta $\hat{\sigma}_1^2$ in $\hat{\sigma}_2^2$ nepristranski cenilki za σ^2.

Pod našimi predpostavkami velja izrek, ki nam zagotovi, da velja:

$$T = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sigma_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim Student(n+m-2).$$

V našem primeru, ko je $n = m = 64$, dobimo $T \sim Student(126)$. To dejstvo lahko uporabimo za testiranje domnev. Test je pri alternativni domnevi $H_1 : \mu_1 \neq \mu_2$ naslednji. Če upoštevamo H_0, potem je

$$T = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sigma_p\sqrt{\frac{1}{64} + \frac{1}{64}}} = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sigma_p\sqrt{\frac{1}{32}}} \sim Student(126).$$

H_0 pri stopnji tveganja α zavrnemo, če je $|T| \geq F^{-1}_{Student(126)}(1 - \frac{\alpha}{2})$. Pri testu bomo rajši gledali kvantile pri 120 prostostnih stopnjah, ki bodo dobri približki za resnične. Pri danih podatkih je $|T| \doteq 2,26779$, pri stopnji tveganja $\alpha = 0,05$ je desna zgornje neeankosti enaka približno $1,9800$, pri $\alpha = 0,01$ pa približno 2.6200. Zato v prvem primeru ničelno domnevo zavrnemo, v drugem pa jo sprejmemo.

Zares pa lahko alternativno hipotezo nekoliko spremenimo. Iz opazovanja podatkov pri prvih dveh točkah se nam namreč dozdeva, da bo μ_2 kvečjemu večji od μ_1. To upoštevamo tako, da postavimo alternativno domnevo $H_1 : \mu_2 > \mu_1$, H_0. Tokrat pri stopnji tveganja α zavrnemo H_0, če $T \leq -F^{-1}_{Student(126)}(1 - \alpha)$. V tem primeru je $T \doteq -2,26779$, desna stran

neenačbe pa je pri $\alpha = 0,05$ približno enaka $-1,6600$, pri $\alpha = 0,01$ pa približno $-2,3600$, kar pomeni, da ponovno zavrnemo H_0 pri $\alpha = 0,05$ in sprejmemo H_0 pri $\alpha = 0,01$.

Drugi test bo upošteval tudi možnost $\sigma_1 \neq \sigma_2$. Ker predpostavljamo, da so meritve X_i in Y_j vse neodvisne, velja

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2 + \sigma_2^2}{64}\right),$$

kar pomeni, da je

$$\frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{64}}} \sim N(0,1).$$

Problem je znova, da imenovalca ne poznamo. Označimo $se^2 = \frac{\sigma_1^2+\sigma_2^2}{64}$. To znamo nepristransko oceniti s $\hat{se}_+^2 = \frac{\hat{\sigma}_1^2+\hat{\sigma}_2^2}{64}$. Iz predavanj potem vemo, da je

$$T = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\hat{se}_+} \sim Student(64-1).$$

Prostostne stopnje so take kot so, ker je $\hat{\mu}_1 - \hat{\mu}_2$ povprečje 64 vrednosti $X_i - Y_i$ za $i \in \{1, 2, \ldots, 64\}$. Torej je $\mu_1 - \mu_2$ ravno povprečje takega sumanda, $\hat{\sigma}_1^2 + \hat{\sigma}_2^2$ pa je nepristranska ocena za njegovo varianco. Če upoštevamo ničelno domnevo H_0, je $T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{se}_+}$. Pri alternativni domnevi $H_1 : \mu_1 \neq \mu_2$ in pri stopnji tveganja α bomo zavrnilli H_0, če bo $|T| \geq F^{-1}_{Student(63)}(1 - \frac{\alpha}{2})$. Znova bomo vzeli le približek za pravi kvantil in vzeli kvantile, ki ustrezajo 60 prostostnim stopnjam. Pri danih podatkih je $|T| \doteq 2,26779$, pri $\alpha = 0,05$ je desna stran približno enaka $2,00000$, pri $\alpha = 0,01$ pa približno $2,66000$. Torej je situacija enaka kot prej, v prvem primeru domnevo zavrnemo, v drugem pa jo sprejmemo. Naredimo še enostranski test, torej $H_1 : \mu_2 > \mu_1$. H_0 bomo pri stopnji tveganaj α zavrnili, če bo $T \leq -F^{-1}_{Student(63)}(1-\alpha)$. Desna stran je pri $\alpha = 0,05$ enaka približno $-1,67000$, pri $\alpha = 0,01$ pa približno $-2,39000$, zato spet v prvem primeru domnevo zavrnemo, v drugem pa jo sprejmemo.

3. naloga: Pulz

Točka a)

Ustrezne datoteke za to in ostale točke pri tej nalogi se nahajajo v datoteki **Pulz.ipynb**. Dana je tabela meritev pulzov posameznih študentov.

Vsakemu študentu so pulz izmerili dvakrat, nekateri od njih pa so bili med obema meritvama fizično obremenjeni. Zanima nas, ali je obremenitev pri študentih, ki so jo imeli, vplivala na spremembo pulza. Mislimo si, da vse prve meritve pulzov prihajajo iz neke porazdelitve s povprečjem μ_1 in varianco σ_1^2, vse druge pa iz neke porazdelitve s povprečjem μ_2 in z varianco σ_2^2. V luči tega lahko ničelno domnevo postavimo kot $H_0 : \mu_1 = \mu_2$. Označimo z X_i prvo meritev pulza i-tega študenta z obremenitvijo, z Y_i pripadajočo drugo meritev in z n število študentov, ki so bili deležni obremenitve. V našem primeru je $n = 46$. Predpostavili bomo še, da so X_i paroma neodvisne, da so Y_i paroma neodvisne, in da sta X_i in Y_j za $i \neq j$ neodvisni. Smiselna testna statistika je nato $\hat{\mu}_1 - \hat{\mu}_2$, kjer sta

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Problem je, ker porazdelitev, iz katerih izvirajo meritve, ne poznamo. Najbolj idealno bi bilo, če bi bili te porazdelitvi kar normalni. Potem bi veljalo

$$X_i - Y_i \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2cov(X_i, Y_i)).$$

V prejšnji nalogi, smo člen $-2cov(X_i, Y_i)$ opuščali, saj smo predpostavljali neodvisnost vseh meritev, verjetno pa ni najbolj smiselno predpostaviti, da sta dve meritvi pri istem človeku neodvisni, zato moramo tu upoštevati še ta člen. Sumandi $X_i - Y_i$ so za različne i paroma neodvisni, $cov(X_i, Y_i)$ pa je enaka ne glede na i, zato sledi

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2 + \sigma_2^2 - 2cov(X_i, Y_i)}{n}\right).$$

Označimo z se standardni odklon te normalne porazdelitve. Potem je

$$\frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{se} \sim N(0, 1).$$

Problem je, ker se ne poznamo, vendar pa lahko se^2 nepristransko ocenimo kot

$$\hat{se}^2 = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\widehat{cov}}{n},$$

kjer so

$$\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_1)^2, \quad \hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_2)^2,$$

in

$$\widehat{cov} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_1)(Y_i - \hat{\mu}_2).$$

Tedaj vemo, da je

$$T = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\hat{se}} \sim Student(n-1) = Student(45).$$

Če upoštevamo ničelno domnevo, je $T = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\hat{se}}$, zato lahko test izvedemo kot v prejšnji nalogi. Alternativna hipoteza naj bo $H_1 : \mu_1 \neq \mu_2$. Potem pri $\alpha = 0,05$ zavrnemo H_0, če je $|T| \geq 2,02$, pri $\alpha = 0,01$ pa, če je $|T| \geq 2,69$. Ker je $|T| \doteq 5,31503$, domnevo zavrnemo pri obeh stopnjah tveganja. Kot pri prejšnji nalogi bi bilo smiselno to preizkusiti še z enostranskim testom, saj je v tem kontekstu bolj smiselna alternativna hipoteza $H_1 : \mu_2 > \mu_1$. V tem primeru H_0 zavrnemo pri $\alpha = 0,05$, če je $T \leq -1,68$, pri $\alpha = 0,01$ pa, če je $T \leq -2,42$. Ker je $T \doteq -5,31503$, tudi v tem primeru H_0 zavrnemo pri obeh stopnjah tveganja, kar pa je tudi smiselno, saj pričakujemo, da vadba zares vpliva na pulz.

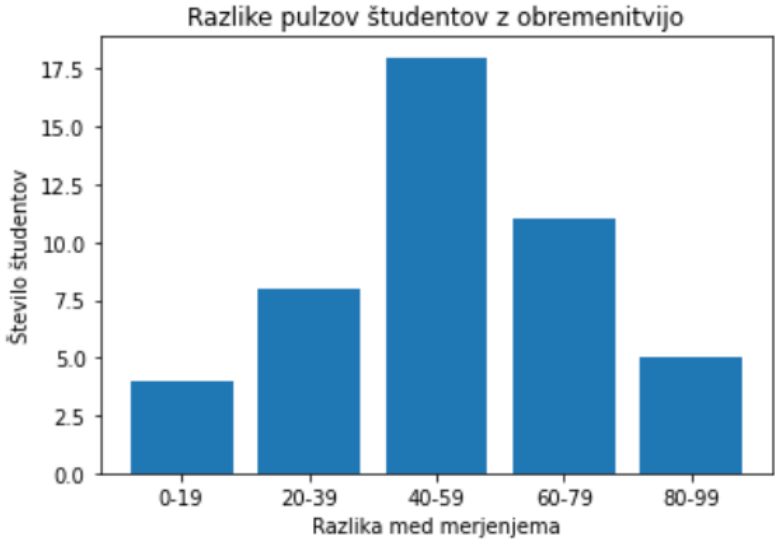
Za ta test smo sicer predpostavljali, da meritve X_i in Y_j prihajajo iz normalnih porazdelitev, vendar pa je zgornji potek v resnici utemeljen že brez te predpostavke, če uporabimo centralni limitni izrek. Spremenljivke $X_i - Y_i$ so namreč enako porazdeljene za vsak i in so poleg tega paroma neodvisne. Zato v resnici res približno velja

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2 + \sigma_2^2 - 2cov(X_i, Y_i)}{n}\right).$$

in je zato zgornji postopek v tem približku korekten.

Točka b)

Pri tej točki bomo raziskali, če se nam na podlagi podatkov zdi, da je kakšen od študentov, ki so bili deležni obremenitve, goljufal in med obema meritvama sploh ni tekel na mestu. V ta namen si bomo najprej ogledali stolpični grafikon, ki prikazuje število študentov, v odvisnosti od razlike pulzov. Največja razlika pulzov pri posameznem študentu je 94, za širino posameznega stolpca pa bomo vzeli 20.



Podrobneje si bomo sedaj ogledali tiste študente, pri katerih je bila razlika pulzov strogo manjša od 20. Te lahko razvrstimo v naslednjo tabelo skupaj z vsemi relevantnimi podatki.

Kadi	Alkohol	Vadba	obremenitev	Pulz1	Pulz2	Leto
Ne	Pije	Vadi veliko	Da	76	88	93
Ne	Ne Pije	Vadi zmerno	Da	68	84	96
Ne	Pije	Vadi zmerno	Da	65	82	96
Ne	Pije	Vadi zmerno	Da	145	155	97

Oglejmo si vsakega od teh študentov posebej. Prvi v tabeli ima zares majhno razliko v pulzih, vendar pa tudi veliko vadi, kar pomeni, da bi morda potreboval večjo obremenitev, da bi njegov pulz narastel bolj kot je. Vpliv vadbe bomo bolj natančno raziskali v naslednji točki. Pri naslednjih dveh študentih je situacija morda že bolj vprašljiva, saj ni nobenega tako prepričljivega razloga, da bi njun pulz narastel le toliko, zato bi morda lahko sklepali, da sta ta dva zares goljufala, saj razlika njunih pulzov še vedno zelo odstopa od povprečja. Če vadba nima bistvenega vpliva, potem bi enakovredno lahko sklepali, da je goljufal prvi študent. V nasprotju s tem zadnji študent skoraj sigurno ni goljufal. Podatki nam dajo misliti, da je pred merjenjem vadil že sam in je zato merjenje izvedel s povišanim pulzom. Pri višjih pulzih nasploh rabimo večjo obremenitev, da se pozna razlika, zato je smiselno, da mu od le ene minute teka na mestu pulz ni pretežno narastel.

Točka c)

V prejšnji točki smo že videli, da smo za študente beležili tudi koliko vadijo. V tej točki pa nas bo zanimalo, ali vadba pri študentih, ki so bili deležni obremenitve, vpliva na spremembo pulza. Glede na nivo vadbe smo študente tako razdelili v tri skupine. Prva skupina vadi veliko, druga zmerno in tretja malo ali pa sploh ne. Ker imamo tri skupine, bomo morali narediti tri primerjave. Za začetek vpeljimo nekaj oznak. Z $X_i^{(j)}$ bomo označilli prvo meritev, z $Y_i^{(j)}$ pa drugo meritev i-tega študenta iz j-te skupine. Poleg tega bomo velikosti posameznih skupin označili z n_1, n_2, n_3. Sedaj lahko ocenimo povprečja razlik pulzov v posamezni skupini:

$$\hat{\mu}^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(X_i^{(1)} - Y_i^{(1)} \right),$$

$$\hat{\mu}^{(2)} = \frac{1}{n_2} \sum_{i=1}^{n_2} \left(X_i^{(2)} - Y_i^{(2)} \right),$$

$$\hat{\mu}^{(3)} = \frac{1}{n_3} \sum_{i=1}^{n_3} \left(X_i^{(3)} - Y_i^{(3)} \right).$$

Ta obravnava se razlikuje od tiste v prvi točki, saj v resnici tle ne predpostavljamo, da so $Y_i^{(j)}$ vsi enako porazdeljeni. Da ima ta naloga sploh smisel, moramo namreč dovoliti, da za različni skupini druge meritve lahko prihajajo iz različnih porazdelitev. Za $X_i^{(j)}$ bomo še vedno predpostavljali, da vse prihajajo iz iste porazdelitve, saj predvidevamo, da bo vadba k razliki prispevala le po teku na mestu. Za začetek se omejimo na primerjavo skupin 1 in 2. Ničelno domnevo postavimo kot $H_0 : \mu^{(1)} = \mu^{(2)}$, kjer je $\mu^{(1)}$ pravo povprečje porazdelitve, iz katere izvirajo $X_i^{(1)} - Y_i^{(1)}$, $\mu^{(2)}$ pa pravo povprečje porazdelitve, iz katere izvirajo $X_i^{(2)} - Y_i^{(2)}$. Smiselna testna statistika bi potem bila na primer $\hat{\mu}^{(1)} - \hat{\mu}^{(2)}$, od katere pa še ne poznamo porazdelitve. Tu ne moremo preprosto normalno aproksimirati, saj niso vsi sumandi enako porazdeljeni, lahko pa aproksimiramo posebej $\hat{\mu}^{(1)}$ in posebej $\hat{\mu}^{(2)}$, nato pa izračunamo porazdelitev vsote teh aproksimacij. Tako je približno

$$\hat{\mu}^{(1)} \sim N\left(\mu^{(1)}, \frac{\sigma^2_{(1)}}{n_1} \right), \quad \hat{\mu}^{(2)} \sim N\left(\mu^{(2)}, \frac{\sigma^2_{(2)}}{n_2} \right),$$

kjer sta $\sigma^2_{(1)}$ in $\sigma^2_{(2)}$ varianci, ki pripadata $X_i^{(1)} - Y_i^{(1)}$ in $X_i^{(2)} - Y_i^{(2)}$. Tedaj je približno

$$\frac{(\hat{\mu}^{(1)} - \hat{\mu}^{(2)}) - (\mu^{(1)} - \mu^{(2)})}{\sqrt{\frac{\sigma^2_{(1)}}{n_1} + \frac{\sigma^2_{(2)}}{n_2}}} \sim N(0,1).$$

Sedaj bi lahko nepristransko ocenili imenovalec in znova izvedli Studentov test, vendar pa ta metoda, ki jo trenutno opazujemo, ni najboljša, ker so $n_1 = 7$, $n_2 = 25$, $n_3 = 14$. Vzorci s katerimi delami so torej zelo majhni in zato taka normalna aproksimacija verjetno ne bo najbolša. Druga možnost je, da bi uporabili močnejšo različico centralnega limitnega izreka, ki pravi, da normalna aproksimacija deluje tudi, če niso vse spremenljivke enako porazdeljene, ampak je tudi natančnost te vprašljiva. Skratka, na ta način nastane veliko težav, zato se bomo celotnega problema lotili drugače.

Poznamo še en način, kako preverjati ali so povprečja večih porazdelitev enaka, ki mu pravimo *analiza variance*. Osnovna metoda je sicer zastavljena za skupine istih velikosti, vendar pa se to zlahka priredi tudi za skupine različnih velikosti. Za uporabo te metode moramo določene stvari še predpostaviti. Naše meritve bomo za namen te metode modelirali kot

$$Z_{ji} = X_i^{(j)} - Y_i^{(j)} = \mu + \alpha_j + \epsilon_{ji},$$

pri čemer je μ skupno povprečje, α_j je *efekt skupine* j, $\epsilon_{ij} \sim N(0, \sigma^2)$ pa je šum. Poleg tega predpostavljamo, da so šumi paroma neodvisni. Predstavljamo si lahko, da šum nastane pri merjenju pulza z napravo. Poleg tega dodamo še pogoj $\alpha_1 + \alpha_2 + \alpha_3 = 0$. Z J označimo število skupin, z I_j pa število meritev v j-ti skupini. Analiza variance temelji na naslednji formuli:

$$\sum_{j=1}^{J}\sum_{i=1}^{I_j}(Z_{ji} - \bar{Z}_{..})^2 = \sum_{j=1}^{J}\sum_{i=1}^{I_j}(Z_{ji} - \bar{Z}_{.j})^2 + \sum_{j=1}^{J} I_j(\bar{Z}_{.j} - \bar{Z}_{..})^2,$$

pri čemer sta

$$\bar{Z}_{.j} = \frac{1}{I_j}\sum_{i=1}^{I_j} Z_{ij}, \quad \bar{Z}_{..} = \frac{1}{\sum_{j=1}^{J} I_k}\sum_{j=1}^{J}\sum_{i=1}^{I_j} Z_i j.$$

Označimo

$$SS_W = \sum_{j=1}^{J}\sum_{i=1}^{I_j}(Z_{ji} - \bar{Z}_{.j})^2, \quad SS_B = \sum_{j=1}^{J} I_j(\bar{Z}_{.j} - \bar{Z}_{..})^2.$$

SS_W je *nepojasnjena varianca*, SS_B pa *pojasnjena varianca*. Na podlagi tega lahko sedaj definiramo testno statistiko

$$F = \frac{SS_B/(J-1)}{SS_W/\sum_{j=1}^{J}(I_j - 1)}.$$

Izkaže se, da sta SS_B in SS_W hi-kvadrat porazdeljeni. Prva z $J-1$ prostostnimi stopjami, druga pa z $\sum_{j=1}^{J}(I_j - 1)$ prostostnimi stopnjami. Zato je

$$F \sim Fischer(J-1, \sum_{j=1}^{J}(I_j - 1)).$$

To sedaj lahko izkoristimo za testiranje domneve $H_0 : \mu^{(1)} = \mu^{(2)} = \mu^{(3)}$. Tega se bomo lotili tako, da bomo izračunali p-vrednost opažanja. Domnevo bomo zavrnili natanko tedaj, ko bo p-vrednost $\leq \alpha$, kjer je α dana stopnja tveganja. Izračun nam na danih podatkih pove, da je p-vrednost enaka približno $0,90652$, kar pomeni, da tako pri $\alpha = 0,05$ kot pri $\alpha = 0,01$ domnevo sprejmemo. Torej sklepamo, da vadba nima nekega bistvenega vpliva na spremembo pulza.

Literatura

[1] J. A. Rice, *Mathematical Statistics and Data Analysis, Third Edition*, Thomson Brooks/Cole, Duxbury, 2007