Seminarska naloga iz statistike

Žiga Gladek

September 21, 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)

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. Enostavni slučajni vzorec lahko v programu dobimo s pomočjo ukaza *sample*. Nepristranska ocena za p 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. Če to implementiramo, nam program pove, 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 $n = 200$. Program nam sedaj pove, da je pri danem vzorcu $\hat{se}_+$ 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:

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

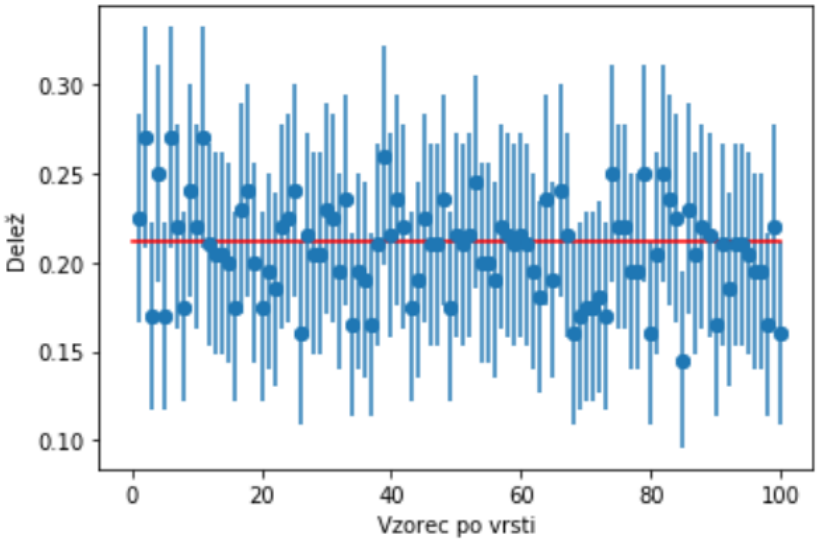
Naša točkovna ocena se od točne vrednosti torej razlikuje za slabe 3%, opazimo pa tudi, da prej dobljeni aproksimativni interval zaupanja pokrije populacijski delež. S tem da poznamo p, lahko izračunamo tudi pravo standardno napako pri vzorcih velikosti 200. Upoštevajmo, da je $\sigma^2 = p(1-p)$. Tedaj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

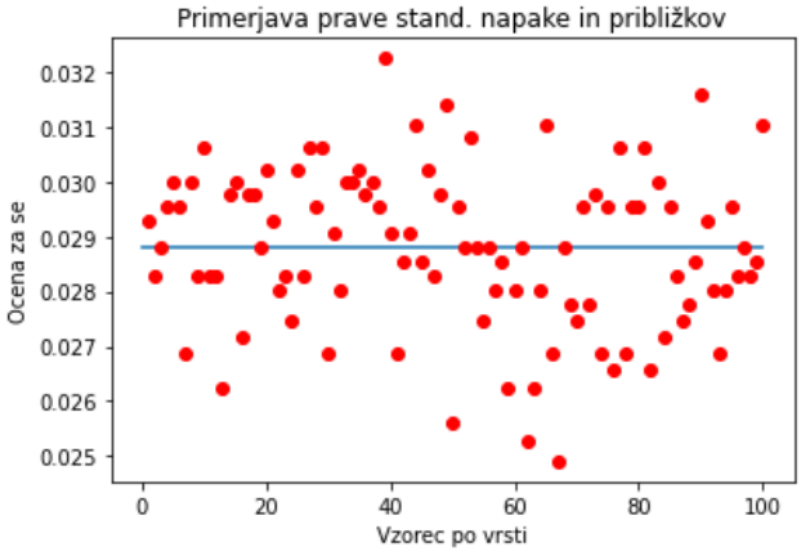
Vzeli bomo še 99 enostavnih slučajnih vzorcev in za vsakega od teh določili 95% interval zaupanja za p. S pomočjo programa dobimo, da je delež intervalov, ki pokrije populacijski delež enak $\frac{95}{100}$. Poglejmo si to še s sliko:



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:



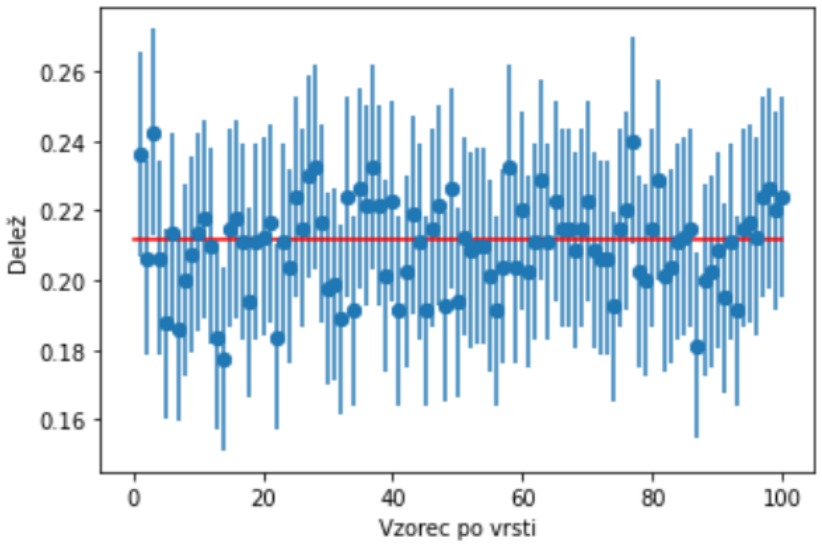
Pri tem modra vodoravna črta označuje pravo standardno napako pri vzorcih te velikosti.

Točka f)

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

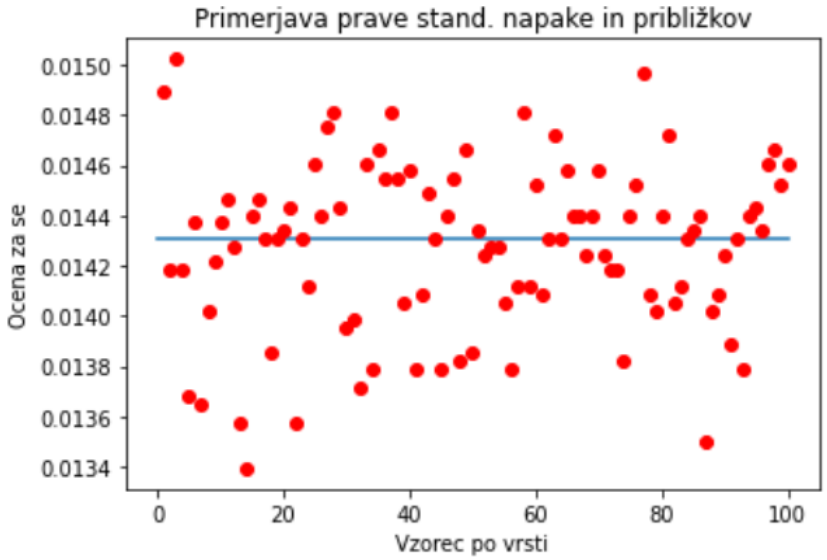
$$p \in \left[\hat{p} - F^{-1}_{Student(799)}(0,975)\hat{se}_+, \quad \hat{p} + F^{-1}_{Student(799)}(0,975)\hat{se}_+\right],$$

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.

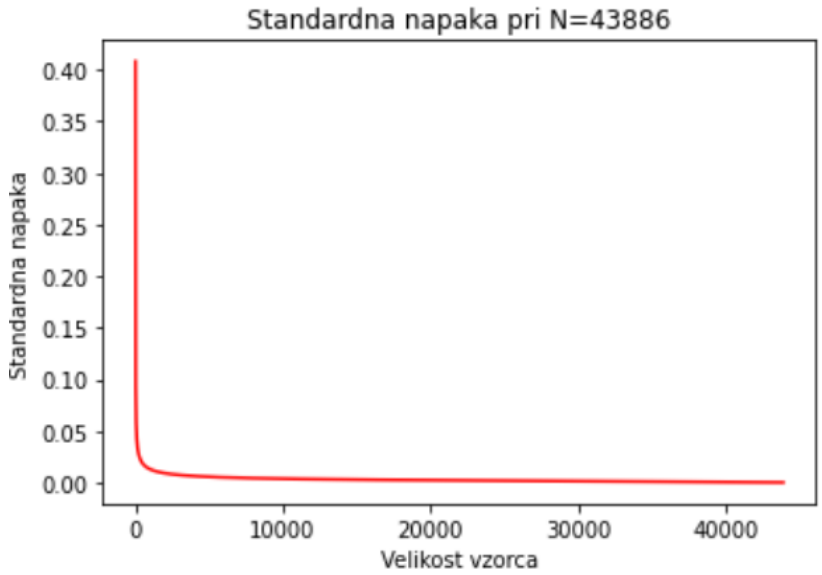
Populacijski delež seveda ni odvisen od vzorca in je enak kot prej. Za razliko od tega je prava standardna napaka pri vzorcu velikosti 800 drugačna kot pri vzorcu velikosti 200. Pri $n = 800$ dobimo, da je $se \doteq 0,01431$, kar je občutno manj kot v prejšnjem primeru. Poglejmo še, kako izgledajo približki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



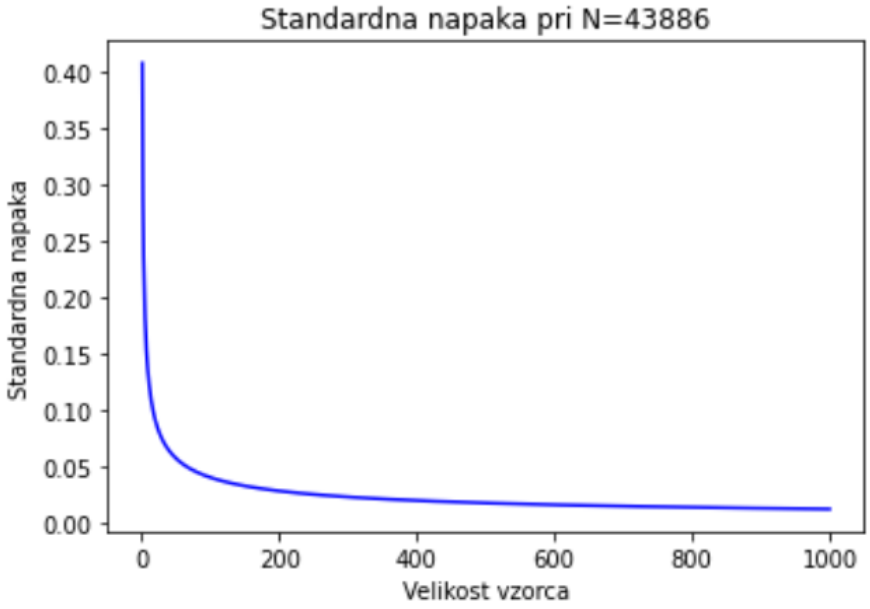
Sliki sicer izgledata podobno, vendar pa so približki v tem drugem primeru dejansko manj razpršeni okrog prave vrednosti, kot v prvem primeru, kar se vidi, če pogledamo vrednosti na y-osi. Da je standardna napaka v tem primeru manjša, ni presenetljivo, saj je le-ta merilo za to, kako natančno vzorčni 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)

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. Iz predavanj vemo, da sta nepristranski oceni enaki

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

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

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

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

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

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

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

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 $\frac{\hat{\mu}_1 - \mu}{\hat{se}_1} \sim Student(65 - 1)$.

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

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)

Na podlagi danih podatkov 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 $H_0 : \mu_1 = \mu_2$. Smiselna testna statistika je na primer $\hat{\mu}_1 - \mu_2$, lahko pa bi vzeli tudi $\hat{\mu}_2 - \mu_1$.

Porazdelitve $\hat{\mu}_1$ ne poznamo, saj ne poznamo σ_1, lahko pa znova upoštevamo, da je $\frac{\hat{\mu}_1-\mu_1}{\hat{se}_1} \sim Student(64)$. Če torej upoštevamo ničelno domnevo, je

$$T = \frac{\hat{\mu}_1 - \mu_2}{\hat{se}_1} \sim Student(64).$$

T vzamemo za novo testno statistiko. Na podlagi te sedaj lahko izvedemo obojestranski T-preizkus, kar pomeni, da bomo domnevo H_0 preizkusili proti alternativni domnevi $H_1 : \mu_1 \neq \mu_2$. Naj bo $\alpha = 0,05$. Ničelno domnevo bomo zavrnili, če bo na danih podatkih $|T| \geq F^{-1}_{Student(64)}(1 - \frac{\alpha}{2}) \doteq 2,00$. Pri danih podatkih je $|T| \doteq 3.33715$, zato domnevo zavrnemo. Pri $\alpha = 0,01$ je $F^{-1}_{Student(64)}(1 - \frac{\alpha}{2}) \doteq 2.66$, zato tudi v tem primeru zavrnemo. Dodatno, če za T vzamemo $\frac{\hat{\mu}_2-\mu_1}{\hat{se}_2}$, potem dobimo $|T| \doteq 2.91998$, torej bi tudi s to testno statistiko domnevo zavrnili pri obeh stopnjah tveganja.

Lahko poskusimo še nekaj. Na podlagi ocen iz prvih dveh točk namreč pričakujemo, da je povprečna telesna temperatura žensk kvečjemu večja od povprečne telesne temperature moških, zato bi se morda bolj splačalo narediti enostranski test. Zato lahko postavimo novo alternativno domnevo $H_1^+ : \mu_2 > \mu_1$. Vzemimo spet testno statistiko $T = \frac{\hat{\mu}_1-\mu_2}{\hat{se}_1}$. Pri stopnji tveganja α bomo H_0 tokrat zavrnili, če bo $T \geq F^{-1}_{Student(64)}(1 - \alpha)$. Pri $\alpha = 0,05$ je $F^{-1}_{Student(64)}(1 - \alpha) \doteq 1,67$, pri $\alpha = 0,01$ pa $F^{-1}_{Student(64)}(1 - \alpha) \doteq 2,39$. Za dane podatke je $T = -3,33715$, torej domnevo sprejmemo pri obeh stopnjah tveganja. Podobno lahko sestavimo še enostranski test za testno statistiko $\frac{\hat{\mu}_2-\mu_1}{\hat{se}_2}$ in ugotovimo popolnoma enako. Razlika je namreč le to, da pri isti alternativni domnevi kot prej tu zavrnemo H_0 takrat, ko je $\frac{\hat{\mu}_2-\mu_1}{\hat{se}_2} \leq -F^{-1}_{Student(64)}(1 - \alpha)$, $\frac{\hat{\mu}_2-\mu_1}{\hat{se}_2}$ pa je na danih podatkih približno $3,33715$.

Literatura

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