Projektna naloga

Martin Žust

18. 8. 2023

Povzetek

Ta dokument je poročilo moje projektne naloge pri predmetu Statistika.

1 Kibergrad

1.1 Izbira slučajnega vzorca iz vsake izmed četrti

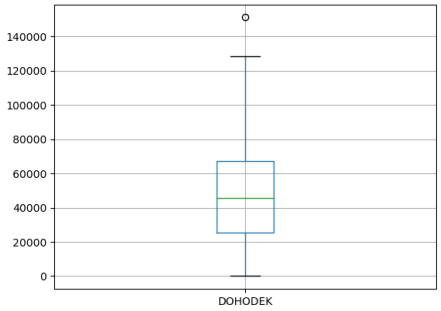
Za delo s podatki sem uporabil Pythonovo knjižnico Pandas. Najprej sem uvozil podatke iz csv datoteke in jih uredil, da so bili nared za obdelavo. Ustvaril sem štiri nove tabele: prva_cetrt, druga_cetrt, tretja_cetrt in cetrta_cetrt. Na vsakem od teh sem nato uporabil funkcijo sample z argumentom sto (torej v slučajni vzorec sem vzel 100 vrstic) in vsakega shranil v novo tabelo: sample_prvi, sample_drugi, sample_tretji in sample_cetrti.

1.2 Risanje škatel z brki

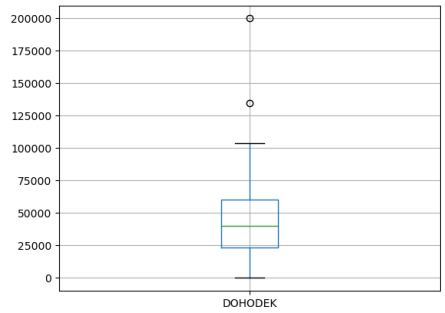
Najprej bi nekaj besed posvetil Pandasovi funkciji *boxplot*, ki nariše škatlo z brki. Funkcija sprejme argument oblike DataFrame. Če ima ta DataFrame en sam stolpec z numeričnimi vrednostmi, ga uporabi za risanje tega grafa. Škatla se razteza od prvega kvartila, ki je spodnja meja, do tretjega kvartila, ki je zgornja meja. Pod in nad škatlo so t.i. brki, ki so v bistvu sestavljeni iz dveh ravnih črt. Črti sta lahko dolgi največ 1,5 interkvartilnega razmika, to je razlike med tretjim in prvim kvartilom. Ostale vrednosti, ki ne sežejo niti v škatlo niti v brke, so t.i. osamelci. Vsakega od osamelcev narišemo posebej, označimo ga z majhnim krožcem.

Najprej sem narisal škatlo z brki za vsakega izmed štirih slučajnih vzorcev. Te grafe si lahko pogledate na Slikah 1 - 4.

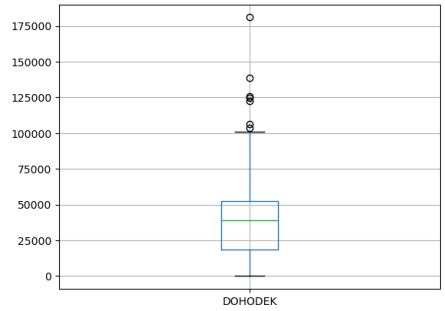
Slika 1: Dohodek v prvi četrti



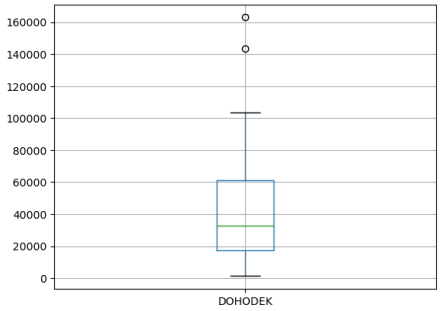
Slika 2: Dohodek v drugi četrti



Slika 3: Dohodek v tretji četrti



Slika 4: Dohodek v četrti četrti

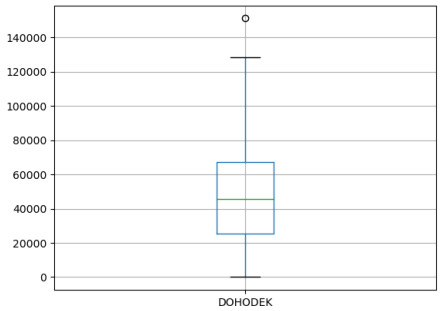


Opazimo, da so vse štiri škatle z brki precej podobne. Poleg tega vidimo, da sta škatli v prvi in četrti četrti večji in škatli v drugi in tretji četrti manjši. Četrti lahko ločimo tudi po številu osamelcev. V tem primeru imamo največ osamelcev v tretji četrti, kjer jih imamo kar šest. V ostalih pa imamo le enega ali dva.

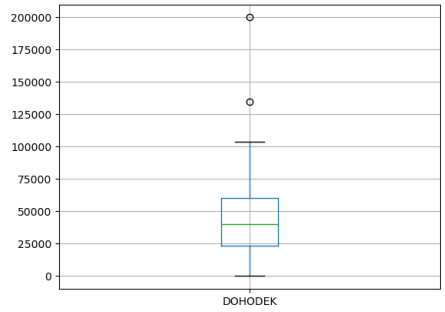
Zanimivo je, da so vsi grafi nesimetrični v smer večjega dohodka, to pomeni, da so brki na strani nad mediano daljši, poleg tega pa imamo osamelce vedno le nad tretjim kvartilom in nikoli pod prvim kvartilom. To opažanje nikakor ni nenavadno, saj je znano, da je premoženje med prebivalstvom običajno razporejeno po Paretovi porazdelitvi. Majhen delež prebivalstva poseduje velik delež premoženja.

Nato sem bolj specifično raziskal prvo četrt in vzel še štiri slučajne vzorce za prvo četrt in tudi za vsakega od njih narisal škatlo z brki. Ti grafi se nahajajo na Slikah 5-8.

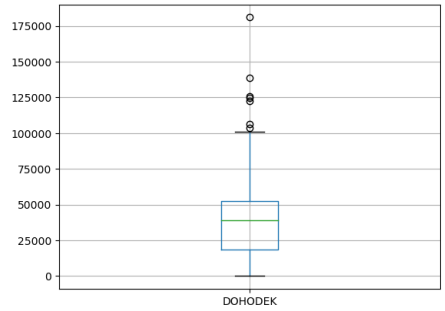
Slika 5: Dohodek v prvi četrti - drugi vzorec



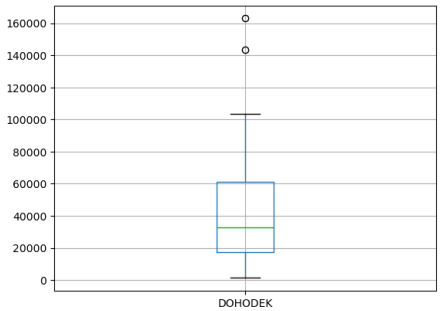
Slika 6: Dohodek v prvi četrti - tretji vzorec



Slika 7: Dohodek v prvi četrti - četrti vzorec



Slika 8: Dohodek v prvi četrti - peti vzorec



Sedaj, ko imamo več slučajnih vzorcev se zdi, da se škatle z brki teh petih vzorcev iz iste četrti ne razlikujejo nič manj, kot so se razlikovale škatle z brki po različnih četrtih. Sklepamo, da je porazdelitev premoženja po celem mestu približno podobna in ni revnih in bogatih četrti.

Vendar hitro vidimo, da se motimo, če izračunamo povprečje vsake četrti posebej. Prva četrt ima poprečni dohodek na gospodinjstvo 45758, druga 41234, tretja 37473 in četrta 42157. Vidimo predvsem, da v negativno smer izstopa tretja četrt in v pozitivno smer prva četrt. Še bolj pa o razlikah med četrtmi pričata varianca dohodka, pojasnjena s četrtmi, in preostala (rezidualna) varianca. Reziualna varianca je vsota varianc znotraj razredov. Varianca dohodka pojasnjenega s četrtmi pa ima nekoliko bolj kompliciranо formulo. Označimo z n_j število gospodinjstev v j-ti četrti. Nadalje označimo z X_j povprečni dohodek v j-ti četrti in z X povprečen dohodek v celotnem mestu. Potem je varianca, pojasnjena s četrtmi enaka:

$$varianca = \frac{\sum_{j=1}^{4} n_j (X - X_j)^2}{4}$$

.

V našem primeru, dobimo za slednjo varianco število 101518440354 in za residualno varianco 4112294846. Kvocien teh dveh števil je približno 25, kar pomeni, da razredi, kar se tiče razdelitve dohodka, niti približno niso enaki.

Namreč, večje kot je to razmerje, z večjo gotovostjo lahko sklepamo, da razlika med razredi obstaja.

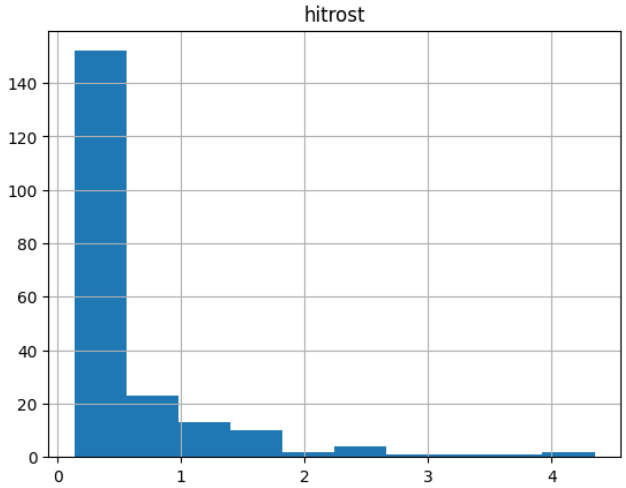
Razloge za zmotno sklepanje pri škatlah z brki lahko iščemo pri majhni velikosti vzorca. Četrti namreč vsebujejo nekaj velikostnih redov več gospodinjstev kot 100. Vsaka četrt namreč vsebuje približno 10000 gospodinjstev. Ker vidimo, da je tudi znotraj četrti velika razpršenost bogastva, nam vzorec 100 gospodinjstev ne daje dobre informacije o celotni četrti. Dokaz za to je pestrost škatel z brki za prvo četrt.

2 Grenlandski kiti

2.1 Histogram hitrosti kitov

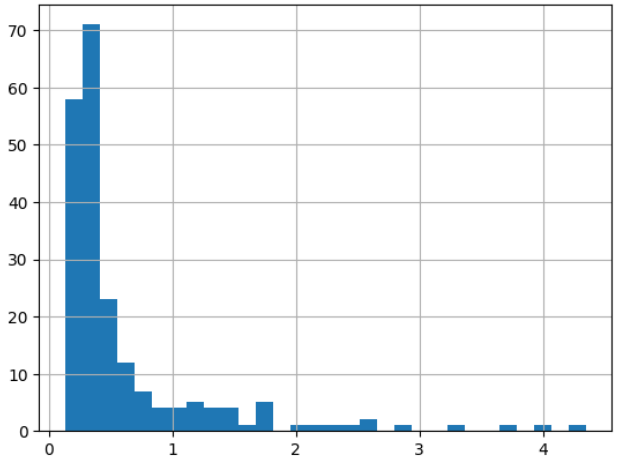
Histogram hitrosti kitov z desetimi stolpci se nahaja na Sliki 9.

Slika 9: Histogram 1



Iz tega histograma se zdi smiselna porazdelitev, ki ji sledijo hitrosti kitov eksponentna, saj je gostota eksponentne porazdelitve precej podobna. Vendar sumimo, da je lahko histogram zavajujoč, saj je večina podatkov zbranih v prvem pravokotniku od leve, torej je večina vrednosti relativno majhnih glede na celoten razpon časov. Ne moremo pa neposredno sklepati, kakšna je porazdelitev časov znotrej tega stolpca. Zato sem narisal še en histogram, ki pa vsebuje 30 razredov. Širino razredov pri tem histogramu sem določil s pomočjo Freedman-Diaconicovega prabivila, ki pravi, da je optimalna izbira širine razreda enaka $\frac{2,6 \cdot IQR}{\sqrt[3]{n}}$. Pri tem je IQR interkvartilni razmik, torej razlika med tretjim in prvim kvartilom. V našem primeru širina razreda pride 0.1486. Če dolžino celotnega razpona 4.5, delimo s to širino, dobimo optimalno število razredov, to je v tem primeru 30. Ta histogram lahko vidite na Sliki 10.

Slika 10: Histogram 2

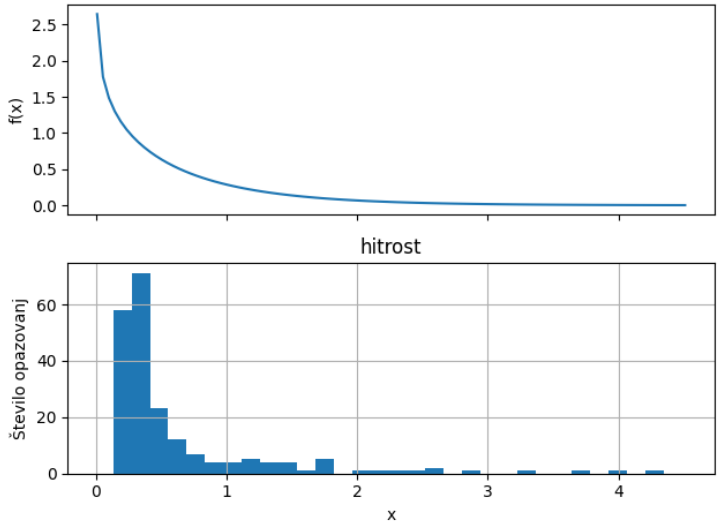


Iz tega histograma vidimo, da bi bila smiselna porazdelitev tudi kakšna druga gama porazdelitev, poleg eksponentne, saj se zdi mogoče, da gostota ne bi bila zgolj padjoča na celotnem intervalu pozitivnih števil.

2.2 Ocena parametrov gama porazdeltive po metodi momentov

Recimo, da je slučajna spremenljivka X porazdeljena $\Gamma(a, \lambda)$. Potem vemo, da je $E(X) = \frac{a}{\lambda}$ in $var(X) = \frac{a}{\lambda^2}$. Iz tega izpeljemo, da je ocena za λ po metodi momemntov $\hat{\lambda} = \frac{n\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ in $\hat{a} = \hat{\lambda}\bar{X}$. Pandas nam po izračunu na konkretnih podatkih vrne $\hat{\lambda} = 1.32$ in $\hat{a} = 0.80$. Kako se ujemata obliki funkcije frekvence in gostote te gama porazdelitve, lahko opazujemo na Sliki 11.

Slika 11: Ujemanje gostote po metodi momentov



2.3 Ocena parametrov gama porazdelitve po metodi največjega verjetja

Najprej bomo naredili izpeljavo cenilk za parametre gama porazdelitve po metodi največjega verjetja. Funkcija gama porazdelitve za pozitivne argumente je:

$$f_X(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}$$

Iz tega sledi, da je funkcija verjetja:

$$L(\lambda, a | X_1, X_2, \ldots, X_n) = \frac{\lambda^{na}}{(\Gamma(a))^n} (X_1 X_2 \cdots X_n)^{a-1} e^{-\lambda(X_1 + X_2 + \ldots X_n)}$$

iz česar sledi, da je logaritem verjetja enak:

$$l(\lambda, a | X_1, X_2, \ldots, X_n) = na \ln(\lambda) - n \ln(\Gamma(a)) + (a-1) \sum_{i=1}^{n} \ln(X_i) - \lambda \sum_{i=1}^{n} X_i$$

Iščemo maksimum zadnjega izraza v odvisnosti od a in λ. V drugem členu se pojavi logaritem funkcije gama. Ko odvajamo ta logaritem po parametru a dobimo t.i. funkcijo digama ($F(a)$). Izračunajmo odvoda funkcije po λ in a. Za stacionarne točke dobimo naslednji sistem enačb:

$$l_\lambda(\lambda, a | X_1, X_2, \ldots, X_n) = \frac{na}{\lambda} - \sum_{i=1}^{n} X_i = 0$$

$$l_a(\lambda, a|X_1, X_2, \ldots, X_n) = n\ln(\lambda) - n\digamma(a) + \sum_{i=1}^{n} \ln(X_i) = 0.$$

Iz prve enačbe sledi:

$$\hat{\lambda} = \frac{\hat{a}}{\bar{X}}.$$

Vendar, ko to vstavimo v drugo enačbo, dobimo:

$$n\ln(\hat{a}) - n\ln(\bar{X}) - n\digamma(\hat{a}) + \sum_{i=1}^{n} \ln(X_i) = 0.$$

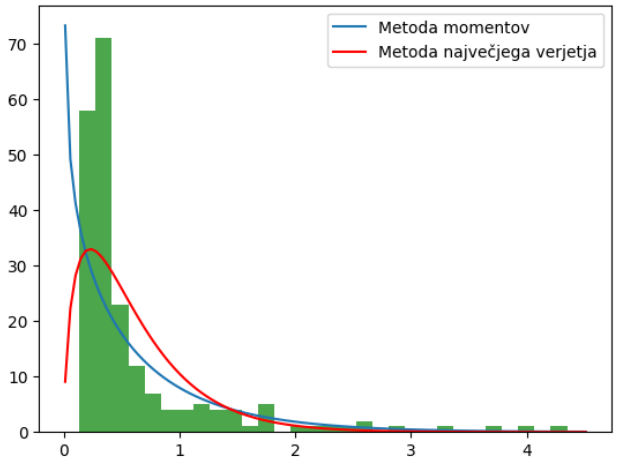
Te enačbe se ne da rešiti analitično, zato se poslužimo numerične metode. Najprej vstavimo konkretne številke iz našega primera, ko imamo $n = 209$, $\bar{X} = 0.6078$ in $\sum_{i=1}^{209} \ln(X_i) = -176.15$, in dobimo:

$$209\ln(\hat{a}) - 209\ln(0,6078) - n\digamma(\hat{a}) - 176,15 = 0$$

$$209\ln(\hat{a}) - 209\digamma(\hat{a}) - 72,0869 = 0.$$

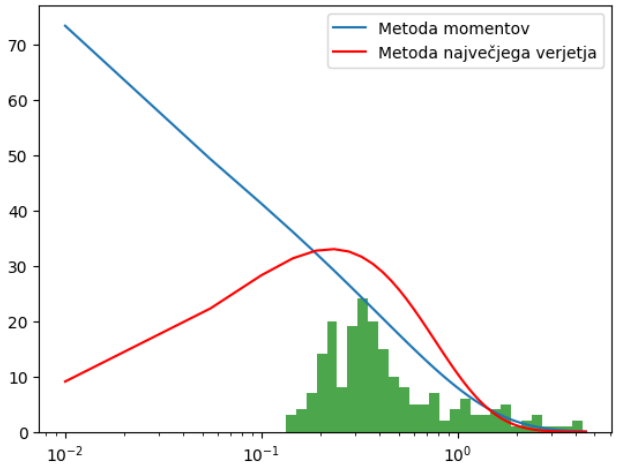
To enačbo sem rešil v Mathematici (priloga) z uporabo funkcije *FindRoot* in dobil rešitev $\hat{a} = 1,596$. Iz tega sledi, da je $\hat{\lambda} = \frac{1,596}{0,6078} = 2,626$. Nato sem v datoteki *Grenlandski_kiti.ipynb* narisal histogram hitrosti kitov s 30 razredi in ustrezno skalirane gostote. Graf si lahko pogledate na Sliki 12. Ujemanje ni preveč natančno v nobenem izmed primerov.

Slika 12: Ujemanje gostot in histograma



Nato sem isti graf narisal še v logaritemski lestvici. Tokrat sta prileganji na desni videti boljši, vendar je še bolj poudarjeno neprileganje za majhne vrednsoti argumenta. Graf se nahaja na Sliki 13.

Slika 13: Ujemanje gostot in histograma - logaritemska lestvica



3 Zobje

V tabeli *Zobje* imamo 59 vrstic podatkov o dolžini odentoblastov in količini in načinu dodajanja vitamina C. Želimo ugotoviti ali dodajanje vitamina C vpliva na dolžino odentoblastov. Postavimo ničelno hipotezo:

H_0**: Dodajanje vitamina C ne vpliva na dolžino odentoblastov.**

in alternativno hipotezo:

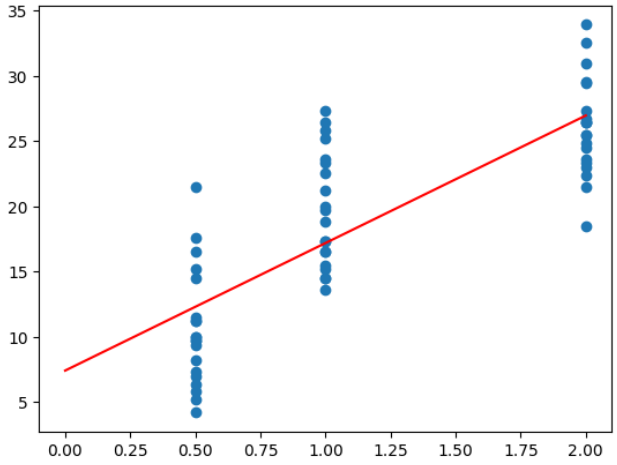
H_1**: Dodajanje vitamina C vpliva na dolžino odentoblastov.**

Najprej si poglejmo analizo variance. Pri tej metodi podatke razporedimo v tri kategorije glede na količino vitamina C, ki smo ga dodali. Za vsako kategorijo izračunamo varianco. Seštevek teh varianc nam da residualno varianco. Izračunamo tudi varianco, razloženo s kategorijami. Izkaže se, da je razmerje med slednjo in residualno varianco spet preveliko, da ne bi zavrnili ničelne hipoteze, količnik je namreč 11. Ta metoda nam je dala zgolj slutiti, da bomo ničelno hipotezo zavrnili. Eksakten preizkus pa izvedemo s pomočjo linearne regresije.

3.1 Linearna regresija

Najprej sem narisal graf dolžine odontoblastov v odvisnosti od količine vitamina C, ki smo ga dodali. Opazimo lepo premosorazmernost. Zato se poslužimo linearne regresije. Izračunamo cenilko za naklon premice $\hat{b} = 9,764$ in cenilko za začetno vrednost $\hat{a} = 7,42$. Tako dobimo graf, ki je prikazan na Sliki 14.

Slika 14: Linearna regresija



Vidimo, da je premica dobljana z linearno regresijo precej strma. Če bi bila njena strmina 0 oziroma blizu 0, potem bi lahko razmišljali o potrditvi ničelne hipoteze.

Preizkus naredimo za stopnjo tveganja $\alpha = 0,05$ in $\alpha = 0,01$. Poiskali bomo interval zaupanja za b. Če bo 0 vsebovana v tem intervalu, potem ničelne hipoteze ne zavrnemo, sicer jo. Iz predavanj vemo (ob predpostavki, da je šum Gaussov), da je interval zaupanja za b enak

$$[\hat{b} - F^{-1}_{Student(n-2)}\left(1-\frac{\alpha}{2}\right)\cdot S\hat{E}_+, \hat{b} - F^{-1}_{Student(n-2)}\left(1-\frac{\alpha}{2}\right)\cdot S\hat{E}_+],$$

kjer je $\hat{SE} = \sigma_+ \frac{n}{\sqrt{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2}}$ in $\sigma_+ = \frac{\sum_{i=1}^n \left(Y_i - \hat{b}X_i - \hat{a}\right)^2}{n-2}$

V datoteki *zobje.ipynb* sem s pomočjo Pandasa izračunal, da je v našem primeru $\hat{SE} = 0,9073$. Posledično je interval zaupanja za $\alpha = 0,05$ enak $[5,9454; 11,5746]$. Ker ta interval ne zajame 0, ničelno hipotezo zavrnemo. Ker bi pri stopnji tveganja $0,01$ dobili še ožji interval, ničelno hipotezo zavrnemo tudi v tem primeru.

Sedaj pa se lahko vprašamo, kater način dodajanja vitamina C sproži večje učinke. V ta namen sem podatke razporedil v dve novi podtabeli. V prvi so shranjeni vse vstice z pomarančnim sokom kot načinom dodajanja vitamina C in v drugi vrstice s podatki, pri katerih je bila uporabljena askorbinska kislina, torej neposredno dodajanje vitamina C. Ko izračunamo strmini linearnih regresij teh dveh skupin, ugotovimo, da ima druga večji naklon ($\hat{b} = 11.72$) kot prva ($\hat{b} = 7.81$). Tako za učinkovitejšo proglasimo naposredno dodajanje vitamina C.