Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Matematički odsjek

ŠESTI PRAKTIČNI ZADATAK IZ PREDMETA

Odabrane primjene vjerojatnosti i statistike

Radio: Ante Ćubela (1191239103)

svibanj 2020.

Sadržaj

1 Uvod

U ovome ću radu napraviti jedan praktični zadatak iz predmeta Odabrane primjene vjerojatnosti i statistike.
U prvome dijelu zadatka napravit ću više simulacija Reed i Frostovog modela za populaciju od 500 jedinki. Napravit ću simulacije za više različitih vjerojatnosti prijenosa zaraze i više različitih udjela inicijalno imunih jedinki u populaciji, koji će varirati od 0 do 75%. Za sve te vrijednosti ću napraviti nekoliko stotina simulacija kako bih vidio kako se epidemija približno ponaša u tim slučajevima, koje ću zatim grafički prikazati. Navedene ishode ću prokomentirati i probati usporediti s teoretski očekivanim ishodima.
U drugome dijelu zadatka ću za neke od tih vjerojatnosti prijenosa zaraze koje daju superkritični režim epidemije ($Np>1$, gdje je N veličina populacije osjetljive na zarazu, a p vjerojatnost prijenosa zaraze u kontaktu osjetljive jedinke s već zaraženom jedinkom) odrediti intervalnu procjenu za očekivani ukupni broj zaraženih jedinki u populaciji nakon 100 vremenskih jedinica (generacija).

2 O Reed-Frost modelu

Reed-Frost model je jedan od najjednostavnijih stohastičkih epidemioloških modela. Spada u skupinu SIR modela koje smo već obradili na predmetu Matematičko modeliranje u biologiji. Smislili su ga Lowell Reed i Wade Frost 1928. godine u neobjavljenome znanstvenom radu. Model opisuje razvoj epidemije kroz vrijeme u određenoj populaciji.
Svaka jedinka u generaciji t ($t\in \mathbb{N}_0$) nezavisno može zaraziti svaku drugu osjetljivu jedinku u populaciji s vjerojatnošću p. Jedinke koje su zarazne

u generaciji t se u generaciji $t+1$ uklone iz aktivno promjenjivog dijela modela, dok sve jedinke koje su se u kontaktu s njima zarazile postanu zarazne u generaciji $t+1$ i tako iterativno sve dok ne dođemo do generacije u kojoj se nitko nije zarazio i tu proces prestaje. Reed-Frostov model je jako pojednostavljen model, gdje ćemo mi u našem zadatku između ostaloga pretpostavljati sljedeće bitne stvari:

- Zaraza se širi izravno sa zaraženih jedinki na druge jedinke određenom vrstom kontakta te nikako drugačije.
- Ukoliko se neimuna jedinka u kontaktu sa zaraženom zarazi, ona postaje zarazna nakon jednog vremenskog perioda (u sljedećoj generaciji).
- U svakome vremenskom periodu, svaka zaražena jedinka dolazi u dodir sa svakom osjetljivom jedinkom i prilikom svakog takvog susreta (koji su nezavisni jedni od drugih) ima vjerojatnost p da prenese zarazu na osjetljivu jedinku.
- Jedinke su potpuno odvojene od svega izvan naše populacije.
- Sve prethodne pretpostavke ostaju iste tijekom cijele epidemije.

Vrijednosti koje unosimo i mjenjamo u modelu su:

- veličina populacije, oznaka: N
- broj inicijalno imunih jedinki, oznaka: Im
- broj inicijalno zaraženih jedinki, oznaka: I_0 (u mojim simulacijama uvijek 1)
- vjerojatnost prijenosa zaraze u slučaju kontakta, oznaka: p

Daljnje oznake koje ćemo mi koristiti u našem modelu su: I_t - broj inficiranih jedinki u vremenu t, S_0 - broj inicijalno osjetljivih jedinki (mi ga računamo kao $N - Im - I_0$), S_t - broj osjetljivih jedinki u vremenu t, $q = 1 - p$ - vjerojatnost da se osjetljiva jedinka neće zaraziti u kontaktu sa zaraženom jedinkom.

Model:

$$S_{t+1} = \sum_{s=1}^{S_t} \mathbb{1}_P \sim \ Bin(S_t, q^{I_t}),$$

gdje P označava skup događaja u kojima se jedinka s iz S_t nije zarazila u kontaktu sa svim jedinkama iz I_t.

$$I_{t+1} = S_t - S_{t+1} \sim \ Bin(S_t, 1 - q^{I_t})$$

Nadalje, na predavanjima smo pokazali da je (I_t, S_t) Markovljev lanac na $\{0,1,...,N\}^2$.

3 R kod

Za simulaciju Reed-Frost modela koristio sam sljedeći R kod:

```
n<-250
N<-500
Im<-0
I0<-1
S0<-N-I0-Im
S<-St<-rep(S0,n)
I<-It<-rep(I0,n)
V<-Vt<-rep(I0,n)
p<-0.005
q=1-p
time<-0
while(sum(It)>0){
  It<-rbinom(n,St,1-(q^It))
  St<-St-It
  Vt<-Vt+It
  I<-rbind(I,It)
  S<-rbind(S,St)
  V<-rbind(V,Vt)
  time<-time+1
}
```

```
I<-as.matrix(I)
S<-as.matrix(S)
V<-as.matrix(V)

matplot(I,type="l",col="grey")
matplot(S,type="l",col="grey")
matplot(V,type="l",col="grey")
```

Oznake varijabli su jednake onima koje je profesor stavio u svoj kod za ovaj model. Varijable koje sam ja dodao su: Im (označava broj inicijalno imunih u populaciji), N (ukupna populacija), V/Vt (matrica koja će sadržavati podatke o ukupnome broju zaraženih jedinki kroz vremenske jedinice, vektor koji sadrži broj ukupno zaraženih u određenoj vremenskoj jedinici t unutar petlje). Kroz više simulacija ću mjenjati vrijednosti varijabli p (između 0 i 1) i Im (između 0 i 375), dok će ostale varijable ostati iste kroz sve izvedbe koda. Također, argumenti funkcije matplot će se mijenjati kroz pokretanja ovisno o potrebi.

Za crtanje kvantila, medijana i prosjeka svojih simulacija koristio sam neke od sljedećih linija R koda:

```
lines(apply(V, 1, min), col = "black", lwd = 2)
lines(apply(V, 1, quantile)[2,], col = "orangered", lwd = 2)
lines(apply(V, 1, median), col = "mediumblue", lwd = 2)
lines(apply(V, 1, mean), col = "black", lwd = 2)
lines(apply(V, 1, quantile)[4,], col = "orangered", lwd = 2)
lines(apply(V, 1, max), col = "black" , lwd = 2)

i <- 1
gkvantil <- c(1:time)

while(i<=time+1){
  gkvantil[i] <- quantile(V[i, ],c(0.9))
  i <- i + 1
}
lines(gkvantil, col = "orangered", lwd = 2)
```

Naravno, ovdje može biti i matrica I ili S, ovisno na kojem grafu želimo crtati kvantile.

U svrhu crtanja naših grafova (kasnije i u svrhu drugog dijela zadatka) smo koristili i sljedeći kod koji će nam našu matricu V i naš vektor koji sadrži kvantil produžiti do određenog broja redaka/brojeva. Uočimo to da ukoliko se u našoj simulaciji npr. 250 epidemija svaka epidemija ugasila prije 30. generacije, da možemo matricu V proširiti i do 100. generacije, s time da će u 100. generaciji za svaku simulaciju vrijednost V_{100} biti ista kao i vrijednost V_{30}, zbog toga što od tridesete generacije uopće nismo imali novih zaraženih jedinki.

```
for (i in time:99){
  V<-rbind(V, V[time, ])
}

matplot(V,type="l",col="springgreen", xlim=c(0,30), ylim=c(0,S0))

i <- 1
gkvantil <- c(1:100)

while(i<=100){
  gkvantil[i] <- quantile(V[i, ],c(0.9))
  i <- i + 1
}

lines(gkvantil, col = "orangered", lwd = 2)
```

Dalje slijedi zadnji dio R koda kojeg ću koristiti isključivo za drugi dio zadatka.

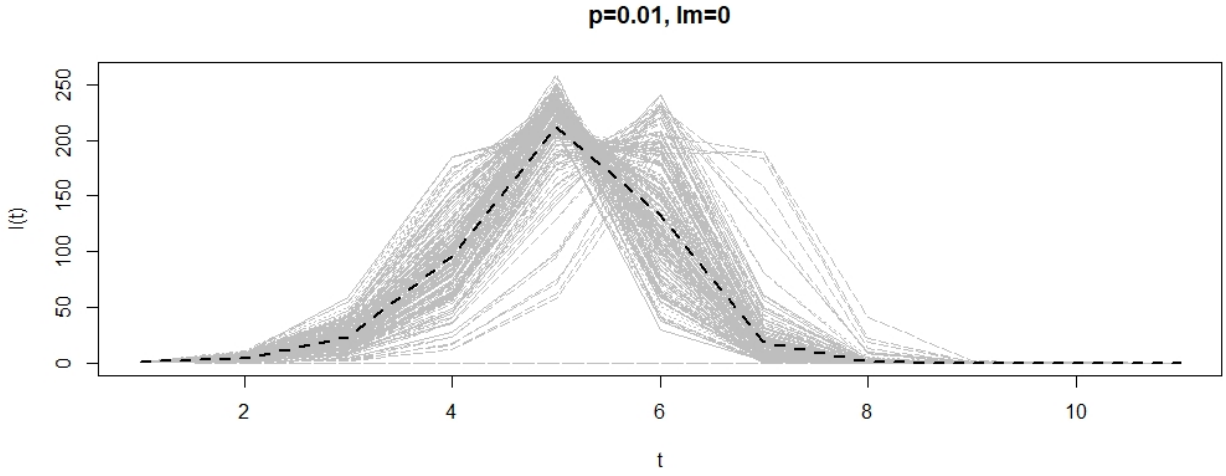
```
fi <- ##vrijednost iz tablice
ikap <- mean(V[100,])
sigma <- sqrt(var(V[100,]))
dg <- ikap-(fi*sigma)/(sqrt(n))
gg <- ikap+(fi*sigma)/(sqrt(n))
```

Ovaj dio koda će biti poprilično jasan kada se prisjetimo teoretske pozadine podzadatka (ii).

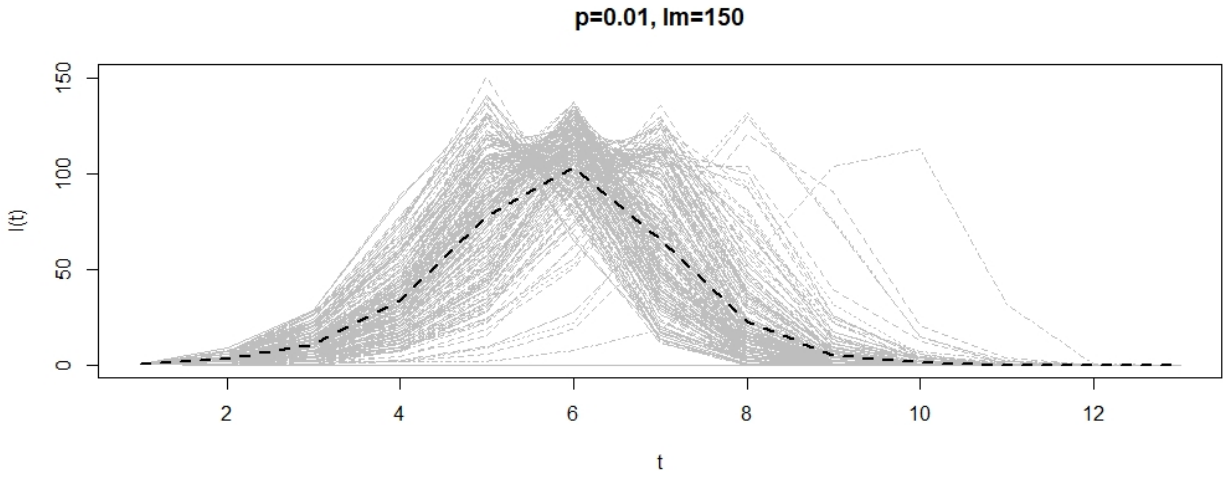
NAPOMENA: U našim grafovima ću koristiti iste boje koje su navedene u gornjem primjeru koda, uz moguće neke druge manje izmjene izgleda grafova. Na grafovima na kojima prikazujemo podatke iz matrice I, najčešće imamo nacrtanu krivulju prosjeka redaka te matrice (generacija), dok na grafovima na kojima prikazujemo podatke iz matrice V, koristimo i medijan i 90%tni kvantil.

4 Podzadatak (i)

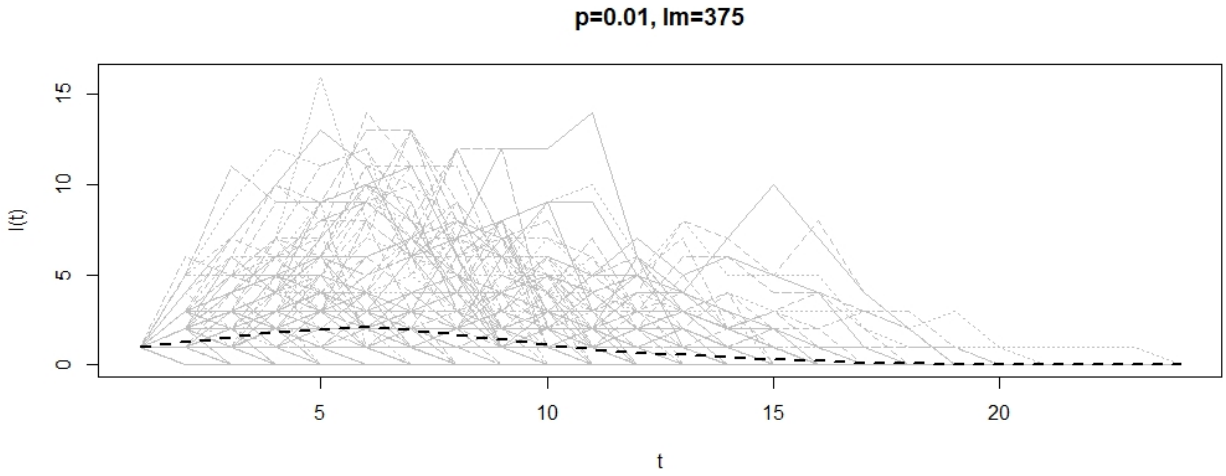
U prvome dijelu zadatka ćemo simulirati Reed-Frostov model za različite vrijednosti p i Im.
Naredne tri slike prikazuju grafove za vrijednost p=0.01 i udjele imunih jedinki u populaciji redom 0%, 30% i 75%.



Slika 1. Broj zaraženih u trenutku t za $p = 0.01$, $S_0 = 499$



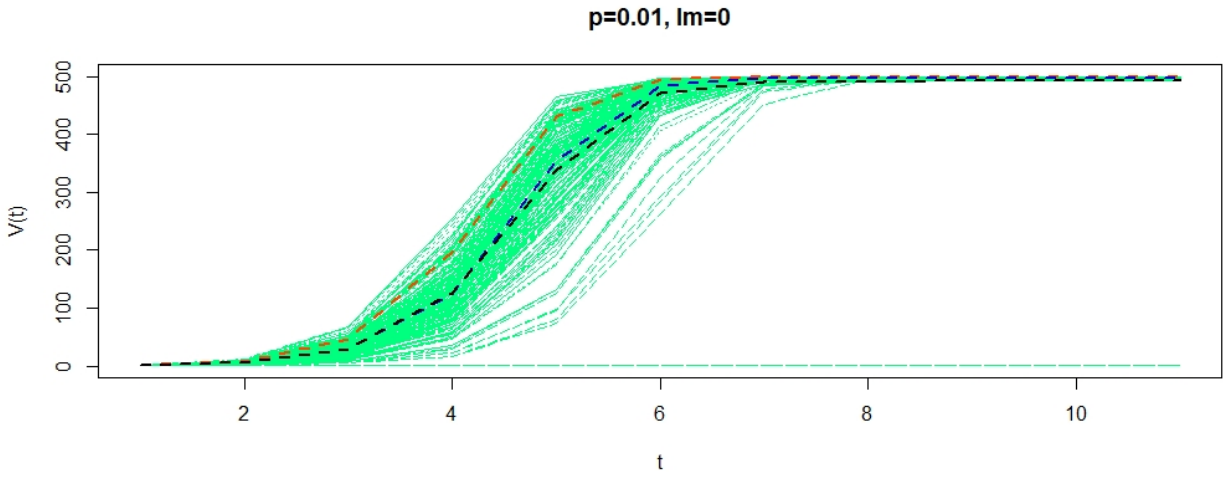
Slika 2. Broj zaraženih u trenutku t za $p = 0.01$, $S_0 = 349$



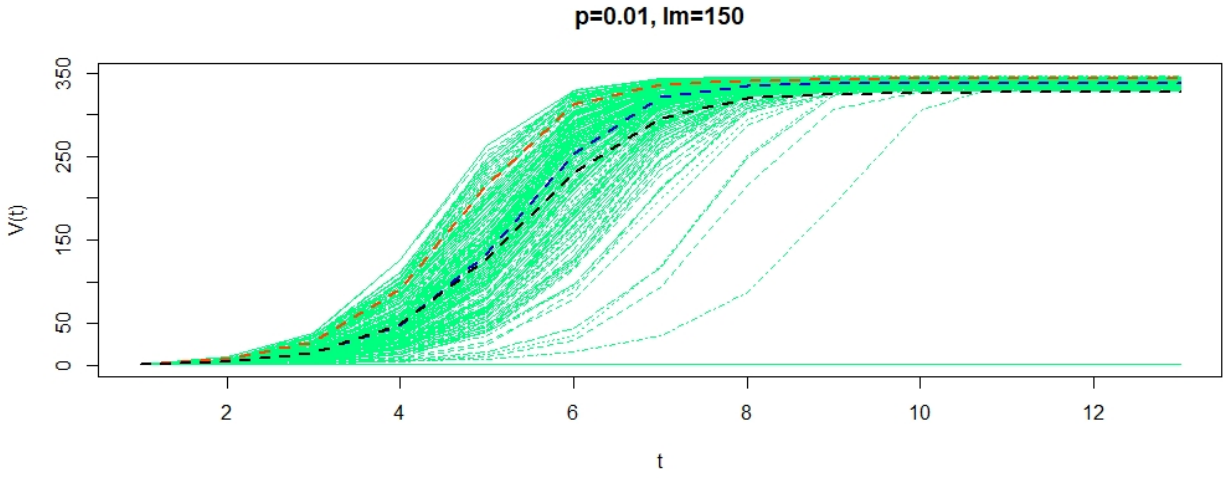
Slika 3. Broj zaraženih u trenutku t za $p = 0.01$, $S_0 = 124$

Grafove sa slika 1., 2., 3. ćemo respektivno nazivat brojevima 1, 2 i 3. Vidimo da su za te grafove koeficijenti λ=S_0p redom $\lambda_1 \approx 5$, $\lambda_2 \approx 3.5$, $\lambda_1 \approx 1.25$. Za sve te grafove je λ_i>1, što znači da imamo superkritični režim. I rezultati i izgled naših simulacija daju očekivane rezultate. U grafu 3 krivulja prosjeka se znatno izravnava u odnosu na grafove 1 i 2, što je opet očekivano jer je λ_3 "blizu" 1.

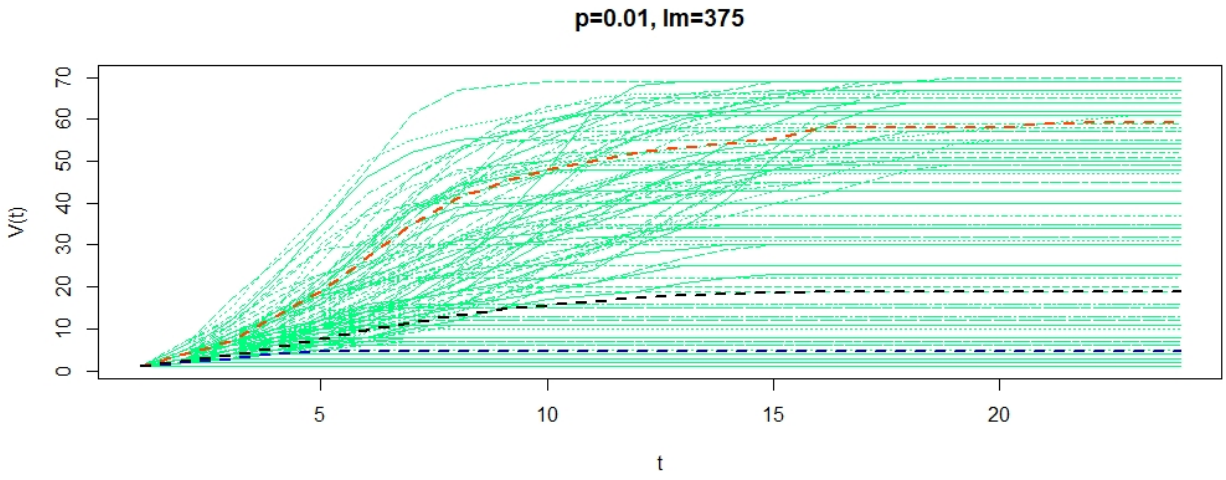
Na sljedećim slikama opet potvrđujemo očekivano ponašanje superkritičnoga režima, a to je da će se tokom epidemije zaraziti "velika većina" jedinki osjetljivih na zarazu, dok je za koeficijent λ koji je bliži broju 1, rezultat donekle različit i niti u jednoj simulaciji broj ukupno zaraženih jedinki ne pređe 75 jedinki, tj 60% populacije osjetljive na zarazu.



Slika 4. Ukupni broj zaraženih za $p = 0.01$, $S_0 = 499$



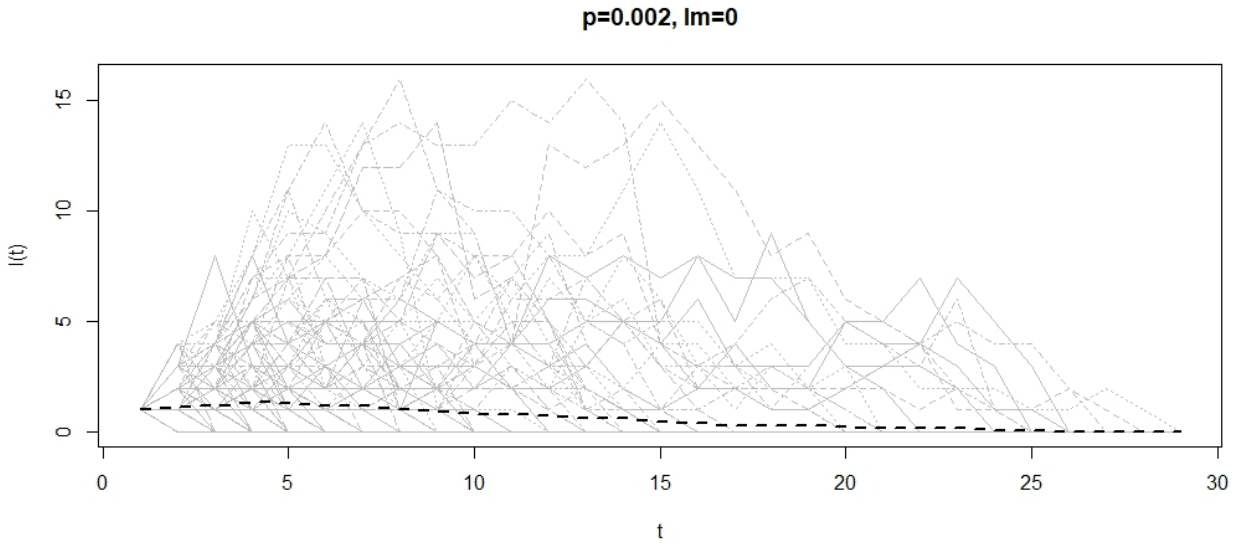
Slika 5. Ukupni broj zaraženih za $p = 0.01$, $S_0 = 349$



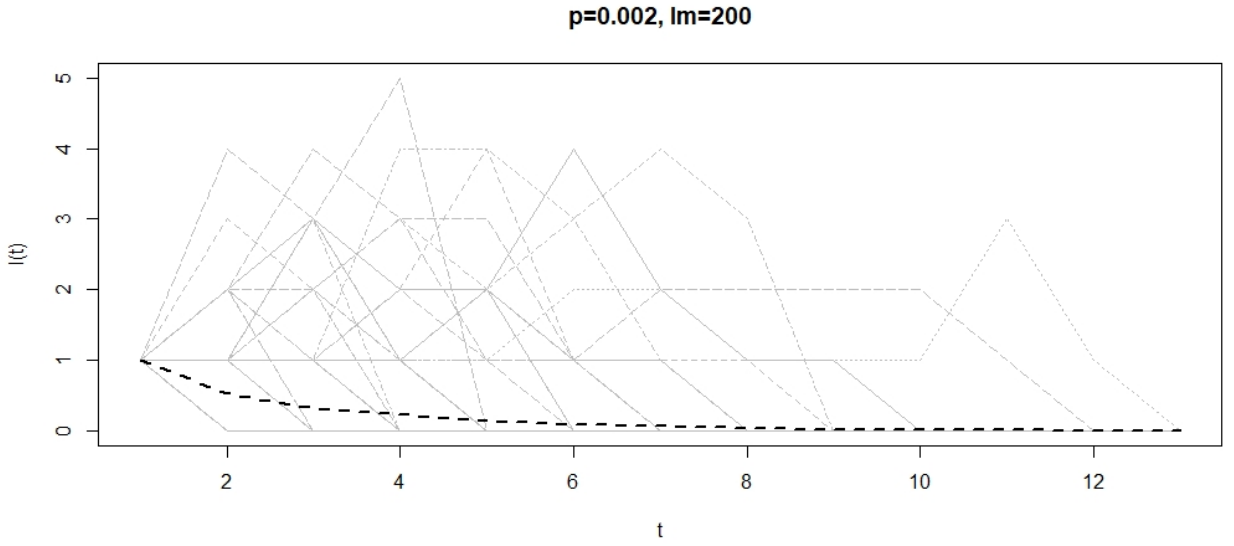
Slika 6. Ukupni broj zaraženih za $p = 0.01$, $S_0 = 124$

Također primijetimo da će generalno u uvjetima na grafu 1 i 2 epidemija trajati znatno kraće nego u uvjetima na grafu 3. To naravno nije lijepo kao što zvuči jer je jedini uzrok toga činjenica da su epidemije s grafa 1 i 2 znatno "agresivnije" i "brže" nego ona s manjim kojeficijentom λ na grafu 3.

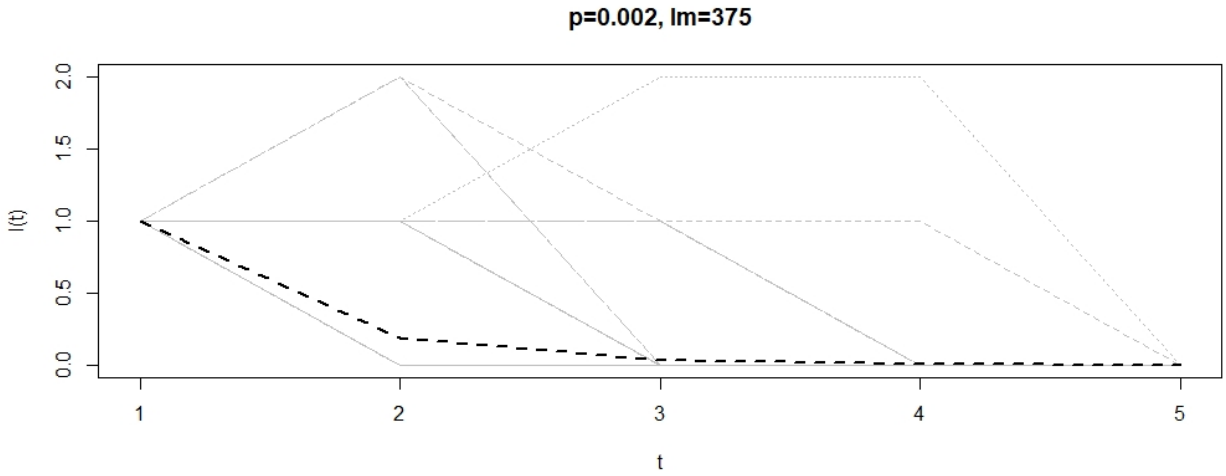
Sljedeće tri slike prikazuju grafove za vrijednosti p=0.002 i udjele imunih jedinki u populaciji redom 0%, 40% i 75%.



Slika 7. Broj zaraženih u trenutku t za $p = 0.002$, $S_0 = 499$



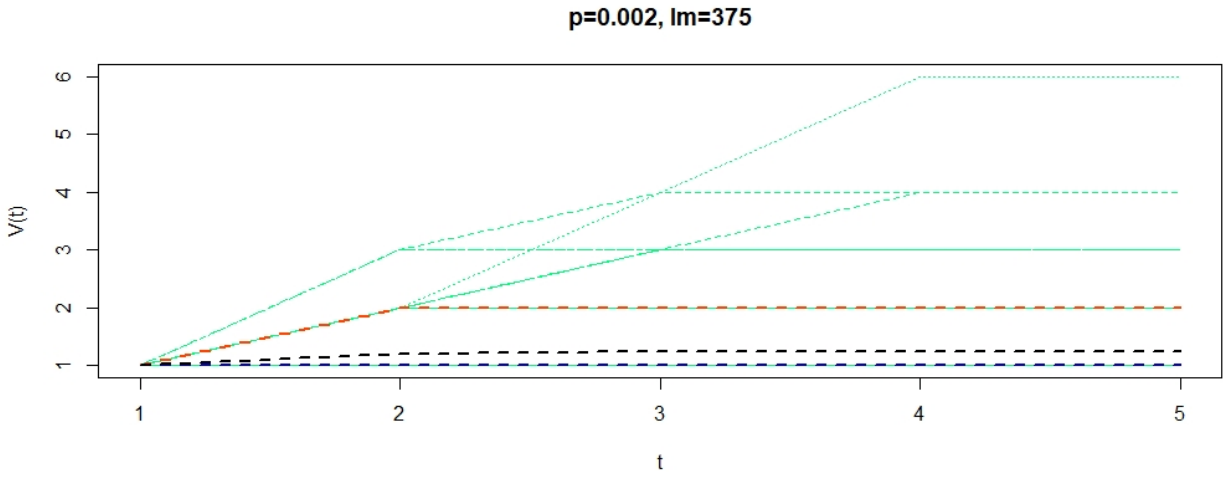
Slika 8. Broj zaraženih u trenutku t za $p = 0.002$, $S_0 = 299$



Slika 9. Broj zaraženih u trenutku t za $p = 0.002$, $S_0 = 124$

Nazovimo grafove na slikama 7., 8., 9., respektivno brojevima 4, 5, 6. Za te simulacije Reed-Frost modela, gore navedeni koeficijent λ je redom, $\lambda_4 \approx 1$, $\lambda_5 \approx 0.6$, $\lambda_6 \approx 0.25$.

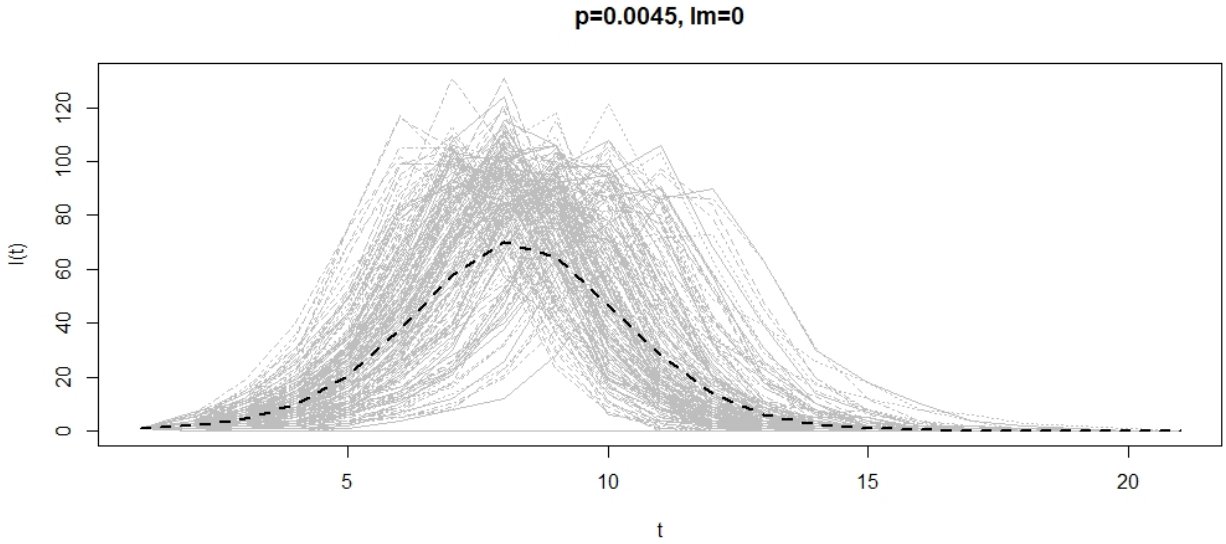
Za grafove 5 i 6 imamo sada subkritični režim, stoga naše krivulje (možemo gledati krivulju prosjeka) izgledaju kako smo i očekivali vidjevši ih na predavanjima. Za graf 6, od 200 simulacija modela niti jedna od tih 200 epidemija nije u jednome trenutku t dosegla više od dvije zarazne jedinke. Za takve subkritične režime očekujemo da će broj ukupno zaraženih jedinki tokom epidemije biti relativno malen s obzirom na veličinu osjetljive populacije. Za graf 6 ćemo vidjeti to i na sljedećem grafu:



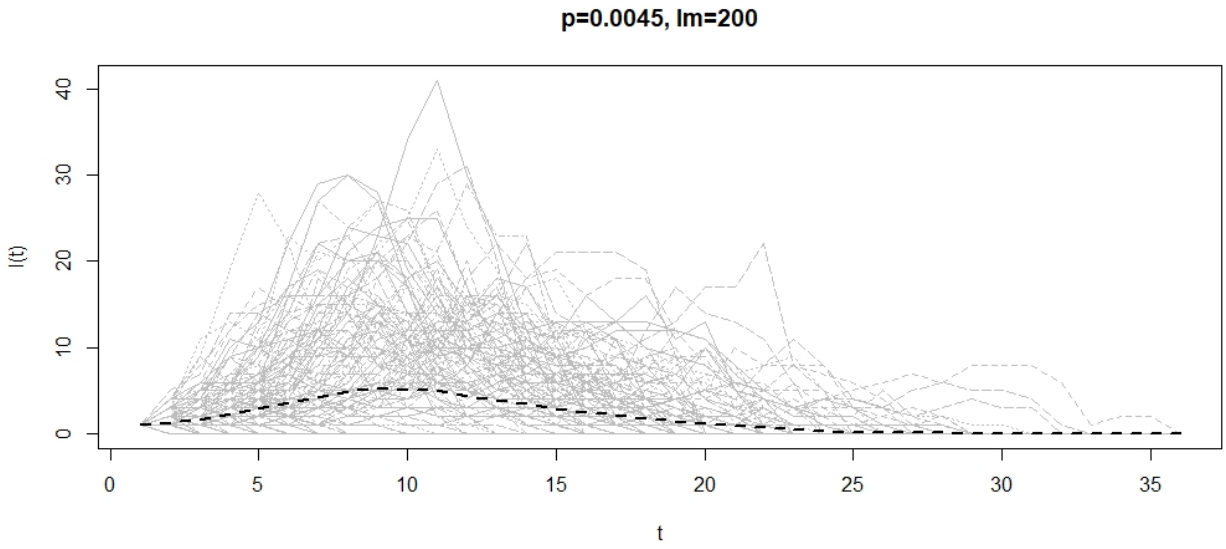
Slika 10. Ukupan broj zaraženih za $p = 0.002$, $S_0 = 124$

Vidimo da nijedna od tih 200 epidemija nije na kraju zarazila više od 6 jedinki od 124 jedinke osjetljive na zarazu.

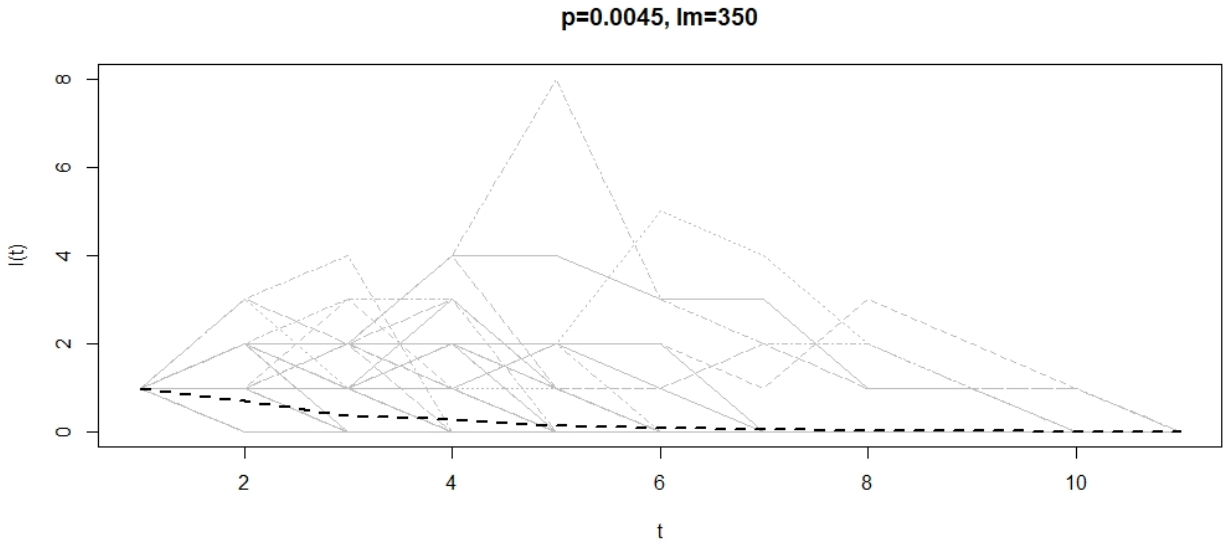
Sljedeće tri slike prikazuju grafove za vrijednosti p=0.0045 i udjele imunih jedinki u populaciji redom 0%, 40% i 60%.



Slika 11. Broj zaraženih u trenutku t *za* $p = 0.0045$, $S_0 = 499$



Slika 12. Broj zaraženih u trenutku t za $p = 0.0045$, $S_0 = 299$



Slika 13. Broj zaraženih u trenutku t za $p = 0.0045$, $S_0 = 199$

Grafove sa slika 11, 12 i 13, ćemo redom nazivati graf 7, 8, 9. Na te tri slike vidimo tri veoma različita grafa za isti koeficijent prijenosa zaraze p. No to nas ne bi trebalo previše iznenaditi, ukoliko opet obratimo

pozornost na prethodno spomenuti koeficijent λ. Za te grafove redom dobijamo vrijednosti $\lambda_7 \approx 2.25$, $\lambda_8 \approx 1.35$ i $\lambda_9 \approx 0.9$. Uz gore prethodno objašnjeno ponašanje grafova ovo i jest rezultat koji očekujemo. Ovi grafovi su dobar primjer kako bi razumjeli to da u Reed-Frostovom modelu ishod simulacija epidemije ne ovisi isključivo o vjerojatnosti prijenosa zaraze između jedinki, nego baš o tome parametru λ, tj. umnošku parametra p i broja jedinki u populaciji osjetljivih na zarazu S_0.

Uz tu napomenu možemo i opravdati korištenje I_0=1 u svim našim simulacijama, jer epidemija ne ovisi I_0. Povećavanje I_0 bi jedino "ubrzalo" epidemiju, ali naše krivulje bi se i dalje generalno ponašale isto, samo bi zapravo počinjale iz neke druge pozicije na grafu.

5 Podzadatak (ii)

U drugome dijelu zadatka ću Monte Carlo simulacijama odrediti intervalnu procjenu za očekivani ukupni broj zaraženih u superkritičnom režimu ($Np > 1$) do trenutka $T = 100$, za razne parametre p.
Na predavanjima smo argumentirali kako kada imamo (X_i) nezavisne jednako distribuirane varijable, možemo njihovo očekivanje aproksimirati s

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Po jakom zakonu velikih brojeva tada slijedi da

$$\hat{I_n} \xrightarrow[n\to\infty]{g.s.} I.$$

To dalje po centralnom graničnom teoremu povlači da

$$\sqrt{n}(\hat{I_n} - I) \xrightarrow[n\to\infty]{distr.} Norm(0, \sigma_g^2).$$

To sada znači da je

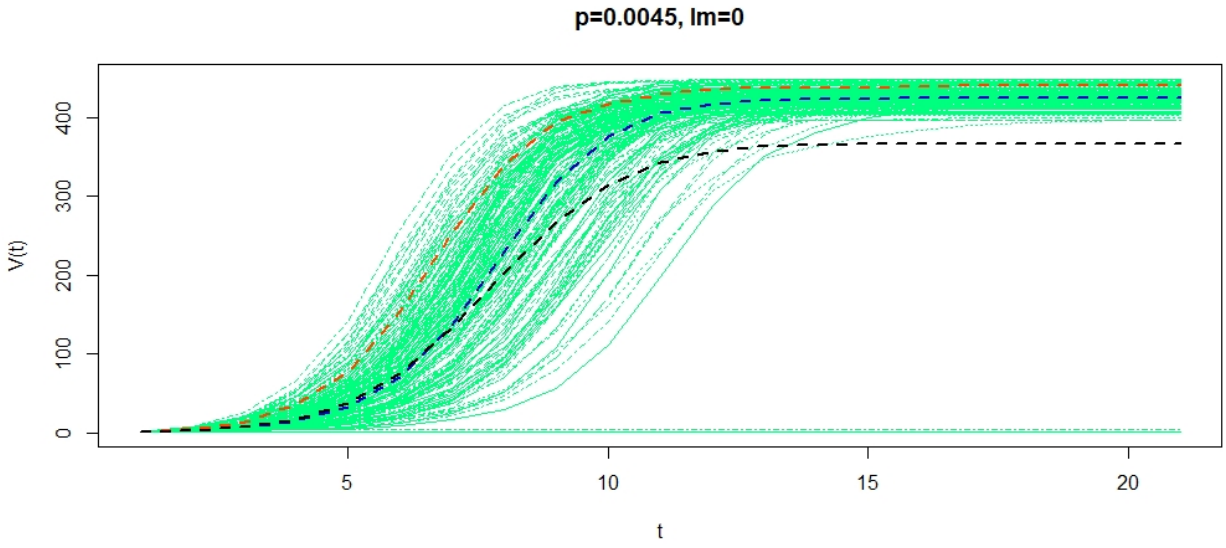
$$\mathbb{P}(-\sigma_g z_{\frac{\alpha}{2}} < \sqrt{n}(\hat{I}_n - I) < \sigma_g z_{\frac{\alpha}{2}}) \approx (1-\alpha),$$

tj. vrijedi:

$$\mathbb{P}(I \in (\hat{I}_n - \frac{\sigma_g}{\sqrt{n}} z_{\frac{\alpha}{2}}, \hat{I}_n + \frac{\sigma_g}{\sqrt{n}} z_{\frac{\alpha}{2}})) \approx (1-\alpha).$$

Ovaj interval sada zovemo $(1-\alpha)100\%$ pouzdani interval za I. Što n više raste u beskonačnost, "širina" intervala se smanjuje. Sada ću za neke od naših primjera, u kojima imamo superkritični režim ($S_0p>1$ aproksimirati taj interval i obrazložiti dobivene rezultate. Napomena da ćemo ovdje koristiti iste grafove kao u podzadatku 1, iz toga razloga što ti grafovi prikazuju simulacije do vremena t gdje zadnja zaraza izumre. Stoga se nakon toga vremena t do $t = 100$ naša vrijednost V_{100} očito neće mjenjati. Stoga je dovoljno gledati V_t, gdje je t infimum skupa t, za koje je $I_t = 0$.

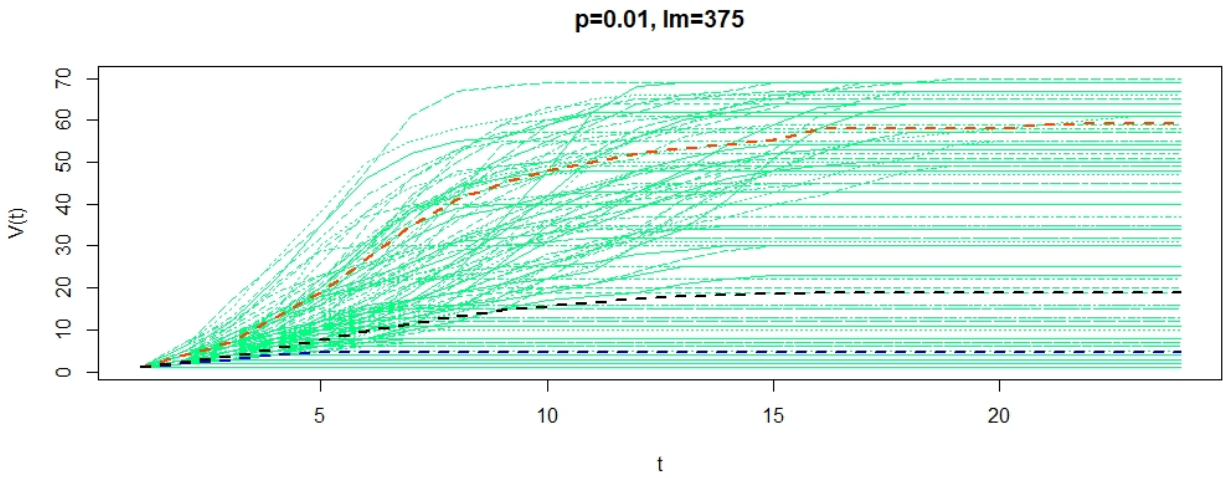
Sada ću Monte Carlo simulacijama odrediti intervale pouzdanosti za grafove i simulacije s prethodnih slika. Primijetimo da je koeficijent $\lambda = S_0p > 1$ u svim našim primjerima, tj. sve su to primjeri superkritičnog režima.



Slika 14. Ukupni broj zaraženih za $p = 0.0045$, $S_0 = 499$ ($\lambda \approx 2.25$)

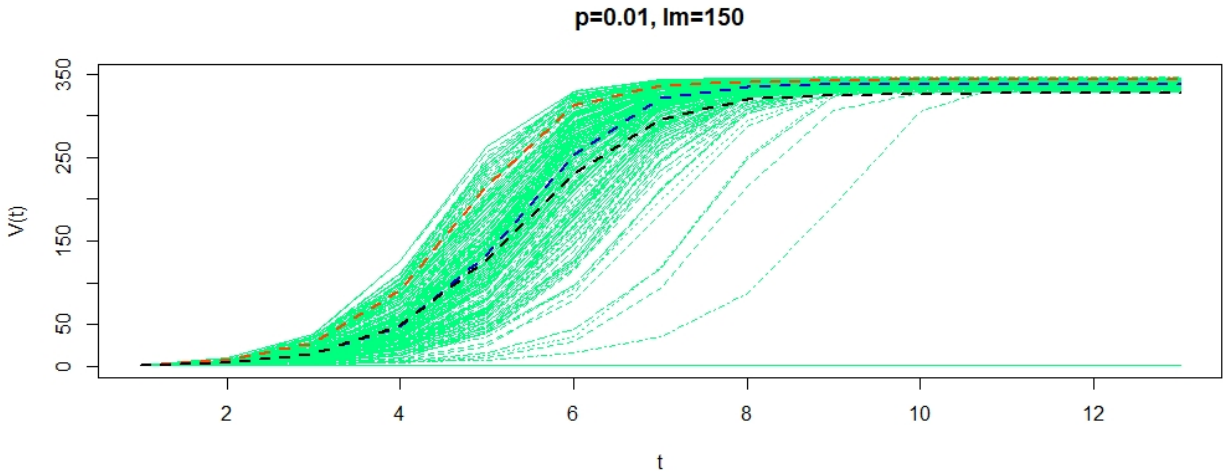
Za graf na prethodnoj slici iz našeg koda navedenog gore dobijemo da su granice za 90%pouzdani interval $[348, 384]$, dok su za 80%pouzdani interval $[353, 380]$. Ukoliko vidimo vrijednosi V_{100} raznih simulacija dobijamo sljedeće brojeve: 404, 420, 426, 415, 436, 422, 427, 434, 399, 1, 426, 426, 418, 1, 430, 427, 417, 1, 422, 445, 423, 440, 420, 433, 416, 427, 1, 429, 435, 437, 431, 438, 413, 428, 434, 443, 1, 419, 413, 1, 2, 433, 436, 426, 438, 1, 431, 402, 1, 425... (uzorak prvih 50 od 200 vrijednosti V_{100} u simulacijama). Čini se da naš 90% pouzdani interval nije ni približno toliko pouzdan. Objašnjenje donekle možemo naći na grafu i ovim vrijednostima. Jasno je da većina vrijednosti V_{100} (obrazloženo zašto je to na ovom grafu V_{30}), leži otprilike između vrijednosti 400 i 450, no povremeno se pojavi simulacija u kojoj epidemija se nije uspjela proširiti s početne jedinke il početnih par generacija, pa za koje ćemo dobit $V_{100}<4$. Zbog tih relativno jako velikih varijacija ćemo dobiti dosta lošu varijancu, što je na grafu donekle vidljivo iz razlike između medijana i prosjeka. Zbog tih jednostavno rečeno "neuspjelih epidemija", naše Monte Carlo simulacije

neće dati veoma dobru aproksimaciju intervala pouzdanosti za vrijednost V_{100}.



Slika 15. Ukupni broj zaraženih za $p = 0.01$, $S_0 = 124$ $(\lambda \approx 1.25)$

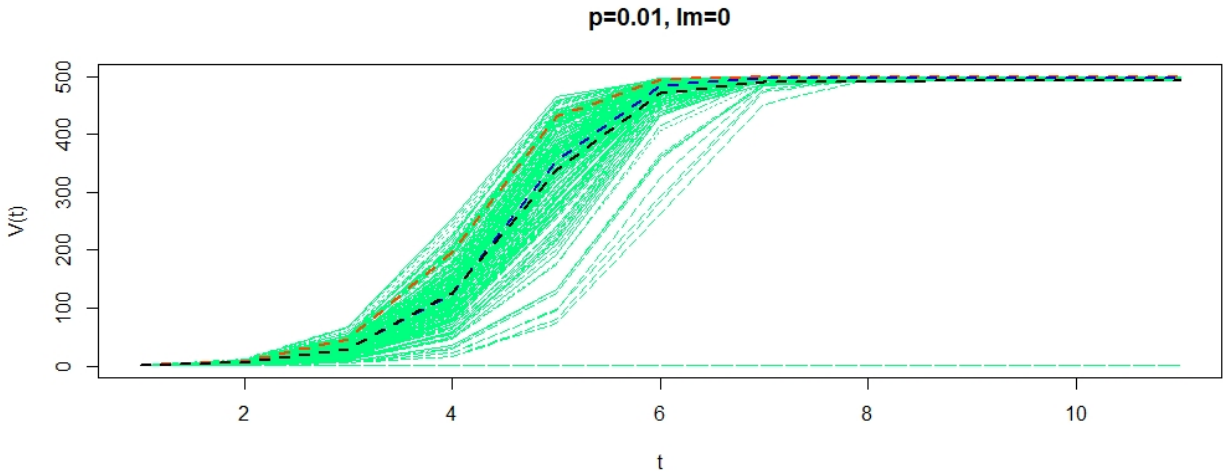
Za graf na prethodnoj slici iz našeg koda dobijemo da je 90%pouzdan interval procjene za V_{100} $[16, 22]$, dok je 80%pouzdan interval $[18, 21]$. Ukoliko vidimo vrijednosi V_{100} raznih simulacija dobijamo sljedeće brojeve: 48, 1, 47, 22, 57, 64, 1, 1, 61, 1, 62, 1, 48, 58, 6, 4, 1, 8, 7, 1, 69, 69, 66, 1, 3, 43, 1, 10, 1, 3, 4, 6, 2, 45, 15, 1, 3, 3, 2, 43, 1, 15, 4, 1, 51 ,1 ,20 ,1 ,1 ,1... (uzorak prvih 50 od 200 vrijednosti V_{100} u simulacijama). Prema viđenome, naši intervali pouzdanosti nisu baš pouzdani. Ovdje je još gora situacija nego prije jer je otprilike polovina epidemija zapravo izumrla na samom početku, pa je time broj ukupno zaraženih jako mali. Tolika nestabilnost epidemije je zapravo očekivana, jer je vrijednost λ blizu 1.



Slika 16. Ukupni broj zaraženih za $p = 0.01$, $S_0 = 349$ $(\lambda \approx 3.5)$

Za graf na prethodnoj slici iz našeg koda dobijemo da je 90%pouzdan interval procjene za V_{100} $[319, 334]$, dok je 80%pouzdan interval $[321, 332]$. Ukoliko vidimo vrijednosi V_{100} raznih simulacija dobijamo sljedeće brojeve: 342, 339, 340, 338, 337, 345, 338, 333, 337, 341, 337, 2, 342, 337, 343, 340, 338, 339, 335, 339, 332, 341, 329, 344, 338, 336, 330, 342, 340, 342, 339, 334, 337, 335, 331, 335, 335, 343, 340, 341, 334, 333, 338, 339, 341, 334, 336, 341, 1, , 339... (uzorak prvih 50 od 200 vrijednosti V_{100} u simulacijama). Vidimo da je procjena naših intervala pouzdanosti bolja nego u prethodna dva primjera, ali i dalje očito ne zadovoljavajuća.

Već uz ova tri primjera možemo i zaključiti jednu stvar koju smo donekle mogli i očekivati, a to je da s rastom koeficijenta λ će se naša procjena intervala pouzdanosti poboljšati. To možemo povezati sa smanjenjem broja epidemija koje se "ugase" pri samom početku, što vodi do manje razlike između medijana i prosjeka ovih simulacija, što vidimo i sa prethodnih slika.



Slika 17. Ukupni broj zaraženih za $p = 0.01$, $S_0 = 499$ $(\lambda \approx 5)$

Za graf na prethodnoj slici iz našeg koda dobijemo da je 90%pouzdan interval procjene za V_{100} $[486, 498]$, dok je 80%pouzdan interval $[487, 496]$. Ukoliko vidimo vrijednosi V_{100} raznih simulacija dobijamo sljedeće brojeve: 492, 498, 496, 498, 495, 498, 498, 499, 495, 496, 498, 497, 494, 498, 494, 497, 499, 497, 494, 497, 498, 499, 497, 499, 494, 492, 496, 495, 498, 497, 497, 497, 495, 494, 498, 496, 495, 1, 497, 496, 497, 492, 498, 497, 496, 498, 498, 497, 497, 498... (uzorak prvih 50 od 200 vrijednosti V_{100} u simulacijama). Sada smo po prvi puta zapravo "zadovoljni" našim intervalima pouzdanosti i to opet prepisujemo porastu koeficijenta λ koji sada iznosi 5. To znači da se mnogo manje epidemija nego u prethodnim primjerima gasi na početku i samim time će prosjek simulacija kojeg smo koristili za računanje očekivanja biti mnogo bolji prikaz stvarnog stanja na grafu nego do sada.

Radeći u programskom paketu R, vjerojatno bi dobili puno bolje rezultate računajući očekivanje pomoću funkcije median umjesto funkcije mean.

6 Zaključak

Za kraj, možemo iz samo par primjera zaključiti da Reed-Frostov model kao ovakav, ne može zadovoljavajuće prikazati epidemiju. No, postoje mnogi drugi epidemiološki modeli, nastali iz ovoga modela, koji se danas (pogotovo posljednjih pola godine) koriste za predviđanje ponašanja epidemije. Preveliki zahtjevi modela kojeg smo obrađivali jednostavno ne opisuju nikakvu realnu situaciju za koju bismo ga htjeli možda nekada iskoristiti. Ipak, čak i ovakav jednostavan model izumljen davne 1928. godine i dan danas nalazi razne primjene u statistici što je pomalo ironično ako se sjetimo da sami autori, Reed i Frost, nisu bili svjesni važnosti modela kojega su izumili.