

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103006/B/2021/36122163606865668

**DISKRÉTNÉ ZLOŽENÉ ROZDELENIA AKO NÁSTROJ
ZÍSKANIA VYBRANÝCH INFORMÁCIÍ O POČTOCH
NÁKAZY KORONAVÍRUSOM V ZVOLENEJ LOKALITE**

Bakalárska práca

2021

Juraj Škopek

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

**DISKRÉTNE ZLOŽENÉ ROZDELENIA AKO NÁSTROJ
ZÍSKANIA VYBRANÝCH INFORMÁCIÍ O POČTOCH
NÁKAZY KORONAVÍRUSOM V ZVOLENEJ LOKALITE**

Bakalárska práca

Študijný program:	Manažérske rozhodovanie
Študijný odbor:	Ekonómia a manažment
Školiace pracovisko:	Katedra matematiky a aktuárstva
Vedúci záverečnej práce:	doc. RNDr. Galina Horáková, CSc.

Bratislava 2021

Juraj Škopek

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Juraj Škopek
Študijný program: manažérske rozhodovanie (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: ekonómia a manažment
Typ záverečnej práce: Bakalárska záverečná práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Diskrétno zložené rozdelenia ako nástroj získania vybraných informácií o počtoch nákazy koronavírusom v zvolenej lokalite

Anotácia: Téma bakalárskej práce umožňuje v teoretickej rovine tvorivo spracovať podstatu zložených diskretných rozdelení, ktoré sú vhodným modelom pre spresnenie stanovenia pravdepodobností nastatia počtu sledovaných javov. Na základe zvolených dostupných informácií, resp. vlastným výskumom o nákuze jedinca koronavírusom v konkrétnej lokalite a o jeho následnom ďalšom šírení, je možné využiť toto rozdelenie na opísanie počtu nakazených. Jednou z možných eventualít riešenia môže byť sledovanie nakazených v súvislosti s cestovateľskou anamnézou kontra počty nakazených domácich prípadov. Na základe spracovania údajov je možné odvodiť konkrétny zákon rozdelenia vrátane charakteristík, potrebný na priblíženie sledovaného procesu.

Vedúci: doc. RNDr. Galina Horáková, CSc.
Oponent: doc. RNDr. Lea Škrovánková, PhD.
Katedra: KMA FHI - Katedra matematiky a aktuárstva FHI
Vedúci katedry: doc. Ing. Michal Páleš, PhD.
Dátum zadania: 24.03.2020

Dátum schválenia: 25.03.2020

doc. Ing. Michal Páleš, PhD.
vedúci katedry

PodĎakovanie

Týmto spôsobom by som sa chcel hlavne poďakovať doc. RNDr. Galine Horárovej, CSc., za jej ochotu, súčinnosť a konštruktívne poznámky smerované k tejto práci. Ďalej sa chcem poďakovať mojim rodičom za vytvorenie kreatívneho prostredia pri písaní tejto práce, mojej priateľke za podporu a v neposlednom rade môjmu bratovi za cenné rady.

Abstrakt

ŠKOPEK, Juraj: *Diskrétné zložené rozdelenia ako nástroj získania vybraných informácií o počtoch nákazy koronavírusom v zvolenej lokalite.* – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra matematiky a aktuárstva. – Vedúci záverečnej práce: doc. RNDr. Galina Horáková, CSc. – Bratislava: FHI EU, 2021, 60 s.

Cieľom bakalárskej práce je spracovať podstatu diskretných zložených rozdelení, ako vhodných modelov na spresnenie stanovenia pravdepodobnosti počtu nakazených osôb infekčnou nákazou akou je koronavírus a získať tak užitočné informácie o priebehu tejto nákazy. Predpokladom pre opis šírenia nákazy je informácia o počte nakazených jedným jedincom vyjadrená diskretným rozdelením. Následným n -násobným zložením tohto rozdelenia môžeme získať rozdelenie počtu nakazených n -tej generácie opísaných vetviacim procesom. Potom môžeme porovnať výsledky s prípadom, keď zohľadníme počet prípadov s cestovateľskou anamnézou opísaných diskretnou náhodnou premennou. Tiež vieme porovnať účinnosť rôznych kontrolných mechanizmov, ktoré zväčšujú pravdepodobnosť zániku nákazy ľubovoľnej generácie. V aplikačnej časti našej práce uvádzame model vetviaceho procesu, ktorý opisuje stochastické šírenie nákazy v jazyku $\mathbb{R}$, v ktorom pomocou metódy Monte Carlo simulácie počítame diskretné zložené rozdelenia.

Kľúčové slová: diskretné zložené rozdelenia, koronavírus, rozdelenie počtu potomkov, vetviaci proces, pravdepodobnosť zániku, Monte Carlo simulácia

Abstract

ŠKOPEK, Juraj: *Discrete compound distributions as a tool for obtaining selected information on the number of coronavirus infections in a given location.* – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Mathematics and Actuarial Science – Supervisor: doc. RNDr. Galina Horáková, CSc. – Bratislava: FHI EU, 2021, 60 p.

The aim of this bachelor thesis is to process the fundamentals of discrete compound distributions, as suitable models for more accurate determination of the probability of the number of persons infected with an infectious disease, such as the coronavirus and to obtain useful information about the course of this disease. The assumption for describing disease spread is the information about the number of infected by one individual, expressed by a discrete distribution that we call the offspring distribution. With the subsequent n -fold compounding of this distribution, we can obtain the distribution of the number of infected of the n -th generation described by a branching process. We can then compare results with the case when we account for the number of cases with travel history, described by a discrete random variable. Also, we can compare the effectiveness of different control mechanisms that magnify the probability of extinction of any given generation. In the application part of our thesis, we present a branching process model that describes stochastic disease spread in $\mathbb{R}$ language, in which we compute discrete compound distributions using the Monte Carlo simulation method.

Key words: discrete compound distributions, coronavirus, offspring distribution, branching process, probability of extinction, Monte Carlo simulation

Obsah

ÚVOD.....	10
1 SÚČASNÝ STAV RIEŠENEJ PROBLEMATIKY DOMA A V ZAHRANIČÍ.....	12
1.1 POROVNANIE DETERMINISTICKÉHO A STOCHASTICKÉHO PRÍSTUPU PRI MODELOVANÍ ŠÍRENIA NÁKAZY.....	12
1.2 DISKRÉTNÁ NÁHODNÁ PREMENNÁ A JEJ ZÁKONY ROZDELENIA	14
1.2.1 Momenty diskkrétnej náhodnej premennej.....	15
1.3 ZLOŽENÉ ROZDELENIA.....	17
1.3.1 Definícia zloženej náhodnej premennej a jej zákonov rozdelenia.....	17
1.3.2 Momenty a vytvárajúce funkcie zloženej náhodnej premennej.....	19
1.3.3 Metódy určenia zložených rozdelení	21
1.4 PREDPOKLADY VETVIACEHO PROCESU	22
2 CIELE PRÁCE	23
3 METODIKA PRÁCE A METÓDY SKÚMANIA.....	24
3.1 MODELÝ POČTU POTOMKOV	25
3.1.1 Useknuté a modifikované rozdelenia počtu potomkov.....	27
3.1.2 Odhad parametrov.....	29
3.1.3 Test dobrej zhody.....	30
3.2 MODEL ŠÍRENIA NÁKAZY AKO VETVIACI PROCES.....	31
3.2.1 Rozdelenie náhodnej premennej Z_n	31
3.2.2 Momenty náhodnej premennej Z_n	32
3.2.3 Pravdepodobnosť zániku nákazy	34
3.2.4 Model s prípadmi cestovateľskej anamnézy	35
3.2.5 Metóda simulácie Monte Carlo na určenie rozdelenia počtu nakazených n-tej generácie	36
3.2.6 Kontrolované šírenie vo vetviacom procese	38
4 VÝSLEDKY PRÁCE A DISKUSIA	40
4.1 ZOSTAVENIE VETVIACEHO PROCESU ŠÍRENIA NÁKAZY V JAZYKU $\mathbb{R}$	41

4.2	VÝBER MODELU POČTU POTOMKOV KORONAVÍRUSU NA ZÁKLADE ÚDAJOV ..	42
4.2.1	Výpočet pravdepodobnosti eventuálneho zániku ohniska.....	44
4.3	ROZDELENIE POČTU IMPORTOVANÝCH PRÍPADOV KORONAVÍRUSU Z ÚDAJOV ..	45
4.4	ANALÝZA ŠÍRENIA KORONAVÍRUSU V HYPOTETICKEJ LOKALITE.....	46
4.5	ANALÝZA EFEKTIVITY KONTROLNÝCH MECHANIZMOV.....	50
ZÁVER.....		57
BIBLIOGRAFICKÉ ODKAZY		59
PRÍLOHY		60

Zoznam obrázkov a tabuliek

Obrázok 1.1 Jednoduchý deterministický model šírenia nákazy.

Obrázok 1.2 Jednoduchý stochastický model šírenia nákazy.

Obrázok 1.3 Graf distribučnej funkcie zloženého rozdelenia v prípade diskkrétnej sekundárnej náhodnej premennej (Zdroj: Vlastné spracovanie).

Obrázok 1.4 Metódy určovania zloženého rozdelenia.

Obrázok 3.1 Vzájomné vzťahy základných diskrétnych rozdelení (Zdroj: ^[9])

Obrázok 3.2 Kód v jazyku R na realizáciu MC simulácie podľa prezentovaných krokov.

Obrázok 4.1 Výpočet hodnôt p_i a E_i podľa NB_i rozdelenia.

Obrázok 4.2 Chí-kvadrát test dobrej zhody pre NB_i rozdelenie.

Obrázok 4.3 Počet importovaných prípadov koronavírusu a počet dní.

Obrázok 4.4 Výstup simulácie počtu nakazených 4-tej generácie, s prípadom $Z_0 = 1$.

Obrázok 4.5 Histogram počtu nakazených Z_4 , pre prípad $Z_0 = 1$.

Obrázok 4.6 Distribučná funkcia počtu nakazených Z_4 , pre prípad $Z_0 = 1$.

Obrázok 4.7 Výstup simulácie počtu nakazených 4-tej generácie, s prípadom Z_0 je NP.

Obrázok 4.8 Histogram počtu nakazených Z_4 , pre prípad Z_0 je NP.

Obrázok 4.9 Distribučná funkcia počtu nakazených Z_4 , pre prípad Z_0 je NP.

Obrázok 4.10 Histogram 16-tej generácie nakazených koronavírusom v prípade kontroly celej populácie s $c = 0,2488534$.

Obrázok 4.11 Histogram 14-tej generácie nakazených koronavírusom v prípade ind.-špec. kontroly s $c = 0,2258155$.

Tabuľka 3.1 Zákony rozdelenia a vlastnosti základných diskrétnych rozdelení ako modelov počtu potomkov. (Zdroj: Vlastné spracovanie podľa použitej literatúry).

Tabuľka 3.2 Panjerové konštanty pre triedu rozdelení $(a, b, 0)$.

Tabuľka 4.1 Údaje na odhad rozdelenia počtu potomkov koronavírusu.

Tabuľka 4.2 Výsledok analýzy rozdelenia počtu potomkov pre danú lokalitu.

Tabuľka 4.3 Rozdelenie náhodnej premennej Z_0 .

Tabuľka 4.4 Hodnoty pravd. funkcie počtu potomkov, pred a po ind.-špec. kontrole.

Úvod

Od vypuknutia epidémie spôsobenej koronavírusom SARS-CoV-2 v čínskom meste Wuhan na konci roku 2019, ktorá neskôr v marci 2020 vyústila do pandémie COVID-19, sa poradné orgány vlád rôznych krajín snažia na základe dostupných údajov predpovedať možný vývoj tejto nákazy. Pozorovaním nebezpečného charakteru tejto nákazy, vyšli nariadenia vládnych orgánov viacerých krajín zatvoriť prevádzky a inštitúcie, čo spôsobilo obrovský pokles v dopyte po tovaroch a službách a v konečnom dôsledku aj globálnu ekonomickú krízu. Vo fázach lepšej epidemickej situácie vzniká spoločenský nátlak na otvorenie ekonomík adresovaný vládny špičkám, avšak nie vždy vedia poradné orgány presne odhadnúť vývoj ani aktuálny stav pandémie a kvôli existencii potenciálneho rizika zvýšenia počtu nakazených, hospitalizácií a úmrtí nemôžu v tomto smere robiť úľavy.

Odpovede na niektoré otázky v súvislosti s epidemiou, alebo pandemiou môžeme získať pomocou matematických modelov. Aby takéto modely poskytli kvalitný zdroj informácií, potrebujú kvalitné vstupné údaje, ktoré sú častým nedostatkom modelov v tejto oblasti. Základný model, ktorým sa modeluje priebeh vlny epidémie je takzvaný *SIR model*. V ňom rozdelíme populáciu do troch základných skupín, *S* (z angl. susceptibles) – jedinci, ktorí sa môžu potenciálne nakaziť (vnímaví), *I* (z angl. infectives) – infikovaní a *R* (z angl. removed class) – tzv. vyradení jedinci, ktorí sa buď vyliečili, alebo zomreli ^[12]. Podstatou modelu je pomocou systému rovníc zachytiť dynamiku šírenia, kde jedinci postupne prechádzajú z jednotlivých skupín v čase, až kým sa nebude mať kto nakaziť, teda nastane koniec epidémie. Takýto model potom dokáže na základe vstupných parametrov predpovedať kedy nákaza zanikne. Model je však **deterministický** a predpokladá, že infikovaný jedinec ďalej infikuje vnímavých konštantnou mierou, čo v praxi nebýva pravda. Z takéhoto modelu sa dá odvodiť významný ukazovateľ pri šírení nákazy, čo je **základné reprodukčné číslo** (R_0), ktoré vyjadruje priemerný počet nakazených jedným infekčným jedincom.

V našej práci sa zameriame na iný prístup, **stochastický**, kedy počet nakazených budeme skúmať rozdeleniami pravdepodobnosti, pričom nástroje, ktorými budeme tieto pravdepodobnosti analyzovať sú **diskrétné zložené rozdelenia** a ich charakteristiky.

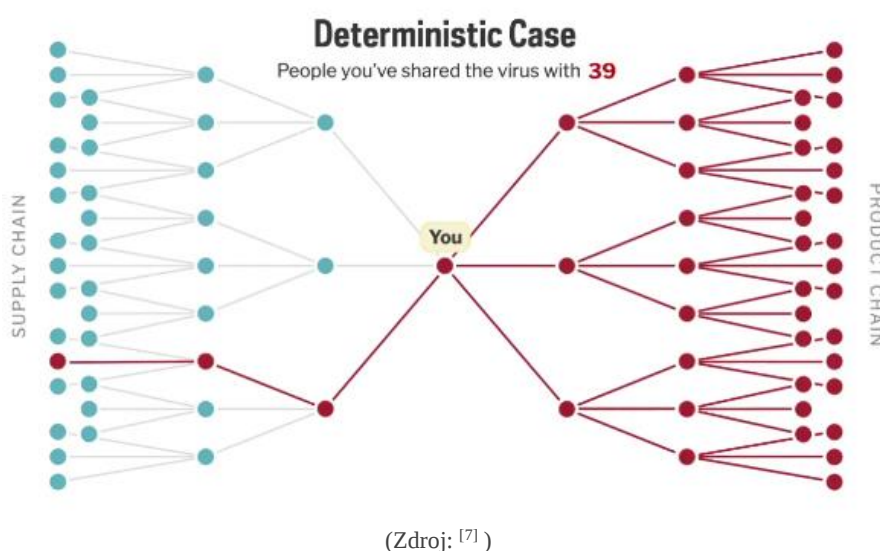
Práca je rozdelená do štyroch kapitol. Prvá kapitola sa zaoberá rozlíšením deterministického a stochastického prístupu pri modelovaní šírenia nákazy. Zároveň opisuje diskretnú náhodnú premennú, diskkrétne zložené rozdelenia a základné predpoklady vetviaceho procesu, ktorý použijeme na opis generácií počtu nakazených. Druhá kapitola obsahuje ciele našej práce, kde priblížime aj aplikačnú časť. Tretia kapitola sa zaoberá metodikou, kde najprv definujeme jednotlivé modely počtu potomkov a metódy ich odhadu z údajov. Ďalej sa venujeme viacnásobne zloženým rozdeleniam, ktoré v rámci našej práce chápeme ako vetviaci proces, pričom tieto rozdelenia budeme určovať metódou Monte Carlo simulácie. Štvrtá, záverečná kapitola, je venovaná výsledkom našej práce a diskusii. Z vetviaceho procesu, ktorý realizujeme v jazyku $\mathbb{R}$ získame informácie o pravdepodobnosti zániku koronavírusu v danej lokalite, čo nám poslúži ako predpoklad na porovnanie efektivity kontrolných mechanizmov na zamedzenie šírenia koronavírusu v hypotetickej lokalite.

1 Súčasný stav riešenej problematiky doma a v zahraničí

1.1 Porovnanie deterministického a stochastického prístupu pri modelovaní šírenia nákazy

Spôsob, akým budeme skúmať počet nakazených a šírenie nákazy, resp. veľkosť epidémie, je cez generácie, kde jednotlivé generácie vznikajú ako výsledok šírenia nákazy. Počiatočnú intuíciu môžeme získať z obrázku 1.1 kde prezentujeme jednoduchý model šírenia nákazy deterministickým prístupom.

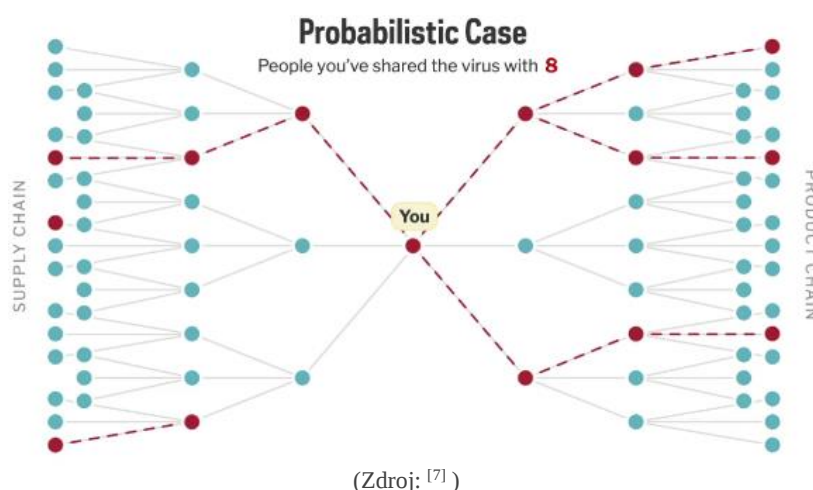
Obrázok 1.1 Jednoduchý deterministický model šírenia nákazy.



Na obrázku 1.1 ide o jednoduchý deterministický model šírenia nákazy so základným reprodukčným číslom $R_0 = 3$ (ak berieme do úvahy iba pravú stranu obrázku). Predstavme si, že máme 3 kontakty a každý náš kontakt má ďalšie 3 kontakty atď. Ak v danom modeli vystupujeme ako prvý uzol siete, resp. nultá generácia, tak vidíme, že tretia generácia od nás obsahuje už 3^3 teda 27 nakazených a nepriamo sme zdieľali vírus s 39 ľuďmi.

Iný prístup pri modelovaní nákazy, ktorému sa v našej práci budeme venovať, je tzv. stochastické modelovanie, kedy jednotlivé parametre šírenia a iné epidemiologické ukazovatele skúmame náhodnými premennými, teda nevieme dopredu stanoviť koľko vnímavých osôb nakazí infekčná osoba.

Obrázok 1.2 Jednoduchý stochastický model šírenia nákazy.



Reálnejšia situácia je na obrázku 1.2, kde vidíme, že niektorí ľudia nenakazia nikoho, niektorí nakazia len jednu, alebo dve osoby. Daný model predpokladá najviac 3 kontakty, pričom v realite je tento počet individuálny, čo sa v modeloch veľmi ťažko zohľadňuje. Z obrázku 1.2 nevieme určiť R_0 od pohľadu, pretože nemáme dostatočné množstvo informácií. V prípade, že by sme poznali podmienené pravdepodobnostné rozdelenie prenosu nákazy za predpokladu stretu s danou osobou a pravdepodobnostné rozdelenie stretu osôb, tak potom vieme R_0 vypočítať pomocou definície podmienenej pravdepodobnosti a strednej hodnoty. Tieto údaje sú však náročné na získanie kvôli ťažko merateľnému charakteru týchto informácií.

Spôľahlivé nástroje, ktoré nám slúžia na analýzu počtu nakazených sú diskkrétne pravdepodobnostné rozdelenia, ktoré vieme získať spracovaním údajov o počte nakazených. Potom základné reprodukčné číslo R_0 vieme vyjadriť strednou hodnotou z rozdelenia počtu potomkov, z angl. *offspring distribution* ^[2], alebo inak povedané rozdelenie prvej generácie počtu nakazených, teda sekundárne prípady nákazy spôsobené jedným primárnym (indexovým) prípadom ^[1]. Autori ^[2], vytvorili teoretický rámec, ako v závislosti od určitých informácií o danej nákeze možno vhodne vybrať model počtu potomkov. V nadväznosti na vhodný výber modelu počtu potomkov potom môžeme analyzovať dané generácie nakazených pomocou špeciálnej skupiny diskrétnych rozdelení. Tieto rozdelenia alebo modely sa nazývajú *diskrétne zložené rozdelenia* a predstavujú špeciálny prístup pri opisovaní náhodných udalostí.

1.2 Diskrétna náhodná premenná a jej zákony rozdelenia

Diskrétna náhodná premenná (na ktorú sa budeme odvolávať aj skratkou DNP), nadobúda určité číselné hodnoty, pričom týchto hodnôt môže byť konečné alebo spočítateľné množstvo. Náhodnú premennú jasne opisujú zákony rozdelenia. V prípade diskkrétnej náhodnej premennej sú týmito zákonmi najmä pravdepodobnostná a distribučná funkcia, ale aj vytvárajúce funkcie.

Definícia 1.1 Náhodná premenná X sa nazýva diskrétna náhodná premenná, ak existuje nanajvýš spočítateľná množina hodnôt H taká, že $\sum_{x \in H} P(X = x) = 1$. Množina H sa nazýva obor diskkrétnej náhodnej premennej, funkcia $p: R \rightarrow \langle 0; 1 \rangle$, rozdelením pravdepodobnosti v podobe pravdepodobnostnej funkcie, pričom

$$p_X(x) = \begin{cases} P(X = x), & \text{pre } x \in H \\ 0, & \text{pre } x \notin H \end{cases} \quad (1.1)$$

Pravdepodobnostnou funkciou vyjadrujeme pravdepodobnosť, že náhodná premenná X nadobudne práve hodnotu x .

Definícia 1.2 Distribučnou funkciou diskkrétnej náhodnej premennej X rozumieme funkciu $F: R \rightarrow R$ definovanú vzťahom:

$$F_X(x) = P(X \leq x), x \in R \quad (1.2)$$

Veta 1.1 Pre diskrétnu náhodnú premennú X s pravdepodobnostnou funkciou podľa (1.1) platí

$$F_X(x) = \sum_{t \leq x} p_X(t), \quad \forall x \in R \quad (1.3)$$

Veta 1.2 Vlastnosti distribučnej funkcie:

$$1. \forall x \in R: 0 \leq F_X(x) \leq 1 \quad (1.4)$$

$$2. \lim_{x \rightarrow -\infty} F_X(x) = 0 \quad (1.5)$$

$$3. \lim_{x \rightarrow \infty} F_X(x) = 1 \quad (1.6)$$

$$4. \forall x_1, x_2 \in R: x_1 < x_2 \Rightarrow P(x_1 < X \leq x_2) = F(x_2) - F(x_1) \quad (1.7)$$

$$5. \text{Funkcia } F \text{ je neklesajúca funkcia.} \quad (1.8)$$

$$6. \text{Funkcia } F \text{ je spojitá sprava.} \quad (1.9)$$

Definícia 1.3 Pravdepodobnostná vytvárajúca funkcia $P_X(t)$ náhodnej premennej X je stredná hodnota premennej t^X

$$P_X(t) = E(t^X) = \sum_{x \in H} t^x \cdot p_X(x) \quad (1.10)$$

Definícia 1.4 Momentová vytvárajúca funkcia $m_X(t)$ náhodnej premennej X je stredná hodnota premennej e^{tX}

$$m_X(t) = E(e^{tX}) = \sum_{x \in H} e^{tx} \cdot p_X(x) \quad (1.11)$$

za predpokladu, že:

$$\sum_{x \in H} e^{tx} \cdot p_X(x) < \infty \quad (1.12)$$

1.2.1 Momenty diskkrétnej náhodnej premennej

Definícia 1.5 Nech X je diskrétna náhodná premenná nadobúdajúca hodnoty x s pravdepodobnosťami $p_X(x)$, $x \in H$. Potom výraz:

$$\sum_{x \in H} x^k p_X(x), \quad k \in N \quad (1.13)$$

označujeme $v_k(X)$ a nazývame začiatkový moment k -teho rádu, ak platí:

$$\sum_{x \in H} |x|^k p_X(x) < \infty \quad (1.14)$$

Začiatkový moment pre ktorý platí $k = 1$ (ak platí (1.14)) sa nazýva **Stredná hodnota** (očakávaná hodnota), ktorú označujeme $E(X)$.

Definícia 1.6 Nech X je diskretná náhodná premenná nadobúdajúca hodnoty x s pravdepodobnosťami $p_X(x) = P(X = x)$, $x \in H$. Potom výraz:

$$\sum_{x \in H} (x - v_1(X))^k \cdot p_X(x), \quad k \in \mathbb{N} \quad (1.15)$$

Označujeme $\mu_k(X)$ a nazývame ho *centrálny moment k -teho rádu*, ak platí:

$$\sum_{x \in H} |(x - v_1(X))^k| \cdot p_X(x) < \infty \quad (1.16)$$

Disperzia alebo **rozptyl** je špeciálny prípad centrálného momentu pre $k = 2$. Pre disperziu náhodnej premennej X platí

$$D(X) = v_2(X) - v_1^2(X) = E(X^2) - E^2(X) \quad (1.17)$$

1.3 Zložené rozdelenia

Motiváciou k použitiu týchto rozdelení (modelov) zvyčajne býva informácia o výskyte počtu zhlukov, ktoré spôsobujú, resp. generujú udalosť, ktorú sledujeme a agregovaním takýchto zhlukov potom dostaneme udalosť celkovú.

Zložené rozdelenia, ale aj všeobecne pravdepodobnostné rozdelenia patria medzi obľúbené nástroje na analýzu javov vo viacerých oblastiach, kde výsledok predmetných javov je dopredu neistý. Takýto výsledok môže predstavovať určité riziko, čo nemusí v každej situácii predstavovať niečo negatívne. Tradične sa tieto modely aplikujú vo finančnej oblasti, napr. v poistnej praxi môže ísť o riziko vyplývajúce z rozdielu medzi očakávanými a skutočnými poistnými plneniami a súvisiacimi nákladmi ^[9]. Zložené rozdelenia v takom prípade umožňujú poisťovní pomocou škodovej frekvencie a výšky škody predvídať náklady súvisiace s daným portfóliom poistných zmlúv za určitý čas a zabezpečiť si konkurenčnú výhodu. O niečo menej tradičné, ale stále vhodné oblasti na aplikáciu týchto modelov sú napr. telekomunikácie (celkový počet stiahnutých megabajtov daného portfólia zákazníkov ^[15]), zdravotníctvo (celková kapacita pacientov v nemocnici), marketing (celková dĺžka hovoru s klientmi v call centre) ^[14] a v neposlednom rade aj v epidemiológii, ako nástroj na analýzu počtu nakazených, ako výsledok vetviaceho procesu.

1.3.1 Definícia zloženej náhodnej premennej a jej zákonov rozdelenia

Nasledujúcu definíciu uvádzame podľa ^[9] s mierne odlišnou interpretáciou:

Definícia 1.7 Nech $\{X_i\}_{i=1}^{\infty}$ je postupnosť nezávislých, nezáporných a identicky rozdelených náhodných premenných. Nech náhodná premenná N , ktorá má diskrétné rozdelenie generuje náhodnú premennú S a zároveň je nezávislá od X . Potom náhodnou premennou S vyjadríme súčet

$$S = X_1 + X_2 + \dots + X_N \quad (1.18)$$

a hovoríme, že má *zložené rozdelenie*. Náhodná premenná N má *primárne rozdelenie* a náhodná premenná X *sekundárne rozdelenie*. Bez špecifikácie jednotlivých rozdelení použijeme štandardný zápis $S \sim Co(p_N(n); F_X(x))$.

Poznámka 1.1 Ak náhodná premenná X je v takomto modeli diskretná (môže byť aj spojitá), potom je diskretnou náhodnou premennou aj náhodná premenná S a má *diskrétno zložené rozdelenie*. My budeme ďalej len pracovať len s týmto typom zloženého rozdelenia.

Keďže náhodná premenná S je diskretná, potom pravdepodobnostná funkcia $p_S(x)$ náhodnej premennej S má tvar:

$$p_S(x) = \begin{cases} \sum_{n=0}^{\infty} p_N(n) \cdot p_X^{*n}(x), & x = 0, 1, 2, \dots \\ 0, & x \neq 0, 1, 2, \dots \end{cases} \quad (1.19)$$

kde p_X^{*n} je n -násobnou konvolúciou pravdepodobnostnej funkcie $p_X(x)$. Pre nultú konvolúciu p_X^{*0} platí

$$p_X^{*0}(x) = \begin{cases} 1, & x = 0 \\ 0, & x \neq 0 \end{cases} \quad (1.20)$$

Pri zložených (diskrétnych) rozdeleniach môže náhodná premenná X všeobecne nadobudnúť hodnoty nezáporné, ale môže nadobudnúť aj hodnoty kladné, pričom je dôležité tieto dve podmienky rozlišovať. Všeobecný prípad opíšeme podmienkou (1.21) pretože $S = 0 \Rightarrow N = n$ a $\sum_{i=1}^n X_i = 0$ potom

$$p_S(0) = p_N(0) + \sum_{n=1}^{\infty} p_N(n) \cdot p_X^{*n}(0) \quad (1.21)$$

Prípad, že náhodná premenná nadobúda hodnoty kladné opíšeme podmienkou $P(S = 0) = P(N = 0)$, resp. $p_S(0) = p_N(0)$, pretože $S = 0 \Rightarrow N = 0$.

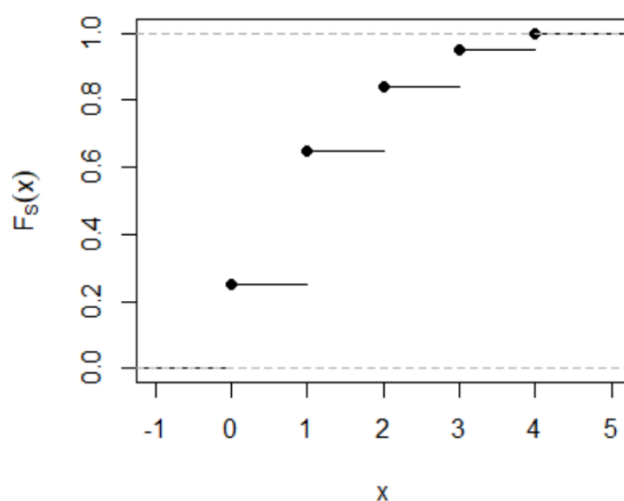
V diskretnom prípade je distribučná funkcia $F_S(x)$ ako v prípade (1.3) „skokovou“ funkciou pre ktorú platia rovnaké vzťahy, teda

$$F_S(x) = P(S \leq x), \quad x \in \langle k; k+1 \rangle, \quad k \in Z_0^+ \quad (1.22)$$

resp.

$$F_S(x) = \sum_{n=0}^{\infty} p_N(n) \cdot F_X^{*n}(x), \quad x \geq 0 \quad (1.23)$$

Obrázok 1.3 Graf distribučnej funkcie zloženého rozdelenia v prípade diskkrétnej sekundárnej náhodnej premennej s piatimi bodmi nespojitosti (Zdroj: Vlastné spracovanie).



1.3.2 Momenty a vytvárajúce funkcie zloženej náhodnej premennej

Nasledujúce dôkazy platných vzťahov preberáme od ^[9].

Stredná hodnota $E(S)$. ^[10] Veta o úplnej strednej hodnote dáva do platnosti vzťah

$$E(S) = E[E(S|N)] \quad (1.24)$$

Využitím vlastnosti strednej hodnoty, pre súčet $X_1 + X_2 + \dots + X_n$ identicky rozdelených, nezávislých NP uvádzame vzťah

$$E(S|N = n) = E\left(\sum_{i=1}^n X_i\right) = n \cdot E(X) \quad (1.25)$$

ktorý pre $n = 0, 1, 2, \dots$ vyjadríme ako

$$E(S|N) = N \cdot E(X) \quad (1.26)$$

a výsledný vzťah zapíšeme v podobe

$$E(S) = E[E(S|N)] = E[N \cdot E(X)] = E(N) \cdot E(X) \quad (1.27)$$

Disperzia $D(S)$. ^[10] Veta o úplnej disperzii dáva do platnosti vzťah

$$D(S) = E[D(S|N)] + D[E(S|N)] \quad (1.28)$$

Podľa vzťahu (1.17) pre súčet $X_1 + X_2 + \dots + X_n$ identicky rozdelených, nezávislých NP

$$D(S|N = n) = D\left(\sum_{i=1}^n X_i\right) = n \cdot [E(X^2) - E^2(X)] \quad (1.28)$$

Pre $n = 0, 1, 2, \dots$ platí vzťah

$$D(S|N) = N \cdot [E(X^2) - E^2(X)] \quad (1.29)$$

a po dosadení vzťahov (1.25) a (1.28) do (1.27) dostaneme

$$D(S) = E(N) \cdot D(X) + E^2(X) \cdot D(N) \quad (1.30)$$

Pravdepodobnostná vytvárajúca funkcia $P_S(t)$. Z vety o úplnej strednej hodnote vyplýva vzťah

$$P_S(t) = E(t^S) = E[E(t^S|N)] \quad (1.31)$$

Pre súčet $X_1 + X_2 + \dots + X_n$ identicky rozdelených, nezávislých NP platí

$$E(t^S|N = n) = \prod_{i=1}^n E(t^{X_i}) = (P_X(t))^n \quad (1.32)$$

Vzhľadom na $E(t^S) = P_S(t)$, dosadením (1.32) do (1.31) pre $n = 0, 1, 2, \dots$ dostaneme

$$P_S(t) = E\left[(P_X(t))^N\right] = P_N(P_X(t)) \quad (1.33)$$

Analogickým vyjadrením by sme prišli k výslednému vzťahu pre **momentovú vytvárajúcu funkciu $m_S(t)$** náhodnej premennej S , teda platí vzťah

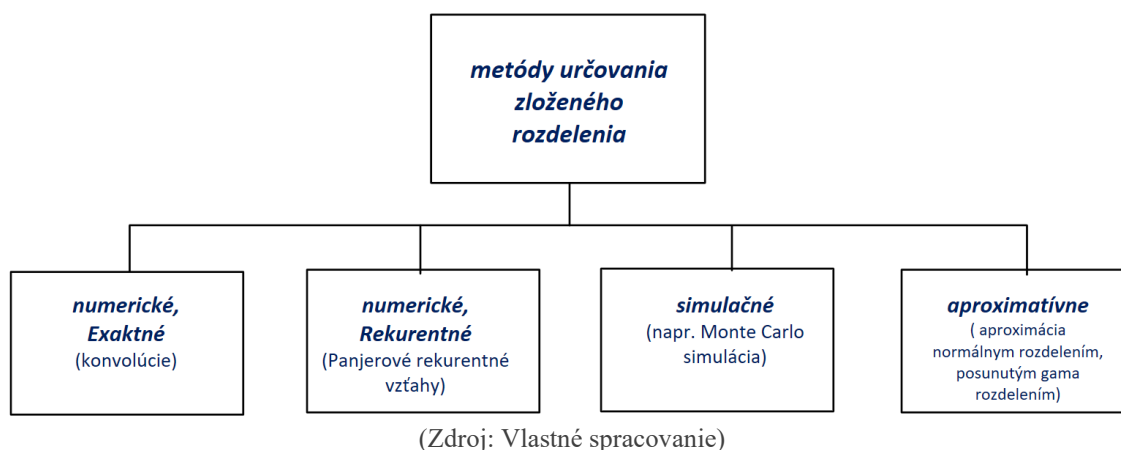
$$m_S(t) = E\left[e^{N \cdot \ln(m_X(t))}\right] = m_N(\ln(m_X(t))) \quad (1.34)$$

1.3.3 Metódy určenia zložených rozdelení

Zložené rozdelenia si vyžadujú komplexnejší prístup pri výpočtoch hodnôt zákonov rozdelenia, nakoľko berieme do úvahy dve, resp. viacero náhodných premenných. Určenie zloženého rozdelenia môžeme exaktne získať *numerickou metódou* cez *konvolúcie* použitím vzťahov (1.19), resp. (1.23). Táto metóda je na jednu stranu presná, avšak s rozsahom hodnôt začína byť numericky náročná a v mnohých prípadoch neuskutočniteľná [9,14]. Alternatívnou *numerickou metódou* sú napr. *Panjerové rekurentné vzťahy*, ktorými vieme zabezpečiť výpočet hodnôt blízkym k exaktným. Značným obmedzením tejto metódy je, že rozdelenie primárnej náhodnej premennej musí patriť do určitej skupiny rozdelení a rozdelenie sekundárnej náhodnej premennej musí byť diskrétné.

S výkonom dnešnej výpočtovej techniky možno uplatniť metódy *simulačné*, z ktorých jednou je napríklad *Monte Carlo simulácia*. Znalosťou určitých výpočtových nástrojov je ľahko aplikovateľná aj v prípade určovania zložených rozdelení, ktoré vznikajú skladaním viacerých, všeobecne n rozdelení. Ďalej existujú aj metódy *aproximatívne*, napr. aproximácia normálnym rozdelením, resp. posunutým gama rozdelením, ktoré sú za určitých podmienok vhodné a ľahko aplikovateľné. Prehľad týchto metód prezentujeme na obrázku 1.4.

Obrázok 1.4 Metódy určovania zloženého rozdelenia.



1.4 Predpoklady vetviaceho procesu

Predstavme si objekty, ktoré dokážu generovať ďalšie objekty rovnakého druhu. Pod takými objektami môžeme mať na mysli mužov, rozmnožujúce sa baktérie alebo neutróny v reťazovej reakcii ^[8]. Vetviaci proces predstavuje spôsob, ktorým možno pomerne jednoducho opísať takúto situáciu. Tento stochastický model, tiež nazývaný *Galton-Watsonov proces* (ďalej len G-W proces), vychádza zo štatistického skúmania vyhynutia priezvisk aristokratických rodov v Anglicku. Práve touto otázkou *F. Galton* položil základy tohto modelu v roku 1874 a neskôr sa mu venoval aj *Watson*. Definíciu základného jedno typového G-W procesu uvádzame podľa ^[8] v bodoch:

- Proces začína s jednotlivcom, ktorý je *nultou generáciou* populácie.
- Tento jednotlivec stvorí náhodný počet potomkov $X \in \mathbb{Z}_0^+$, ktorí predstavujú *prvú generáciu* populácie.
- Každý jednotlivec z prvej generácie stvorí ďalších potomkov nezávisle od seba s identicky rozdelenými náhodnými premennými X , výsledkom čoho vznikne *druhá generácia* populácie. Táto generácia potom podľa rovnakých pravidiel stvorí *tretiu generáciu* atď.
- Každý jednotlivec žije presne jednu časovú periódu, teda jednu „generáciu“.

V rámci našej práce priblížime aj špeciálny prípad, keď predpokladáme, že takýto proces nezačne jednotlivec, ale niekoľko jedincov, ktorí prichádzajú zvonka do systému náhodne. V epidemiológii predstavujú takýto jedinci prípady tzv. cestovateľskej anamnézy, ktorí sa z dostupných údajov dajú opísať náhodnou premennou.

2 Ciele práce

Z prirodzene stochastického charakteru šírenia vysoko infekčných nákaz, ktorými sú rôzne kmene koronavírusov, je potrebné predpovedať, do akej miery môže potenciálne ohnisko nákazy narásť a kedy očakávame koniec epidémie, resp. pandémie spôsobenej týmto kmeňom. Ako uvádzame v prvej kapitole, šírenie nákazy infikovaného jednotlivca, ktorý môže nakaziť náhodný počet jedincov, resp. nemusí nakaziť nikoho, je vhodné opísať náhodnou premennou. Túto náhodnosť budeme opisovať spomínaným rozdelením počtu potomkov, pre vybranú lokalitu. Pomocou údajov o nákaze jednotlivca za danú lokalitu môžeme využitím štatistických metód odhadnúť rozdelenie počtu potomkov. So znalosťou tohto rozdelenia vieme prostredníctvom vetviaceho procesu opísať ako sa táto nákaza šíri jednotlivými generáciami počtu nakazených. Nástroj, ktorým vieme ľubovoľné generácie opísať a získať tak pre danú generáciu niektoré informácie sú práve *diskrétno zložené rozdelenia* a ich číselné charakteristiky.

Hlavným cieľom našej práce je prostredníctvom diskrétnych zložených rozdelení analyzovať pravdepodobnosť zániku nákazy, pričom vychádzame z predpokladu, že nákazu v danej lokalite začal jeden infikovaný jedinec (indexový prípad). Tento predpoklad však v realite splnený nie je. Ak sa skúmaná lokalita úplne neizoluje, napr. uzavretím hraníc krajiny, tak nové prípady nákazy budú stále do skúmanej lokality prichádzať „zvonka“ a každý nový infikovaný jedinec tak predstavuje potenciálne riziko vytvorenia ďalšieho lokálneho ohniska. Tieto prípady sa odborne nazývajú *cestovateľská anamnéza* a taktiež ich môžeme skúmať náhodnou premennou. V takomto prípade vetviaci proces nezačne jedným indexovým prípadom, ale náhodným počtom „importovaných prípadov“, ktoré zásadne menia dynamiku šírenia a zvyšujú tak riziko epidémie. Keďže pôjde o istú komplexnosť výpočtov pri viacnásobnom skladaní rozdelenia počtu potomkov, náhodné premenné v rámci nami vybudovaného vetviaceho procesu budeme simulovať prostredníctvom Monte Carlo simulácie, ktorá je aj jednou zo spomínaných metód určenia zloženého rozdelenia. Tieto komplexné výpočty, teda simulovanie vetviaceho procesu budeme vykonávať v konkrétnom prostredí jazyka $\mathbb{R}$ a výsledky tak budeme prezentovať v rôznych formách výstupov, ktoré dané prostredie ponúka.

3 Metodika práce a metódy skúmania

Z uvedeného cieľa je samozrejmé, že prvotne je treba poznať rozdelenie (modely) počtu potomkov infekčnej nákazy pre danú lokalitu. Model, ktorým budeme opisovať počet potomkov predstavuje diskretná náhodná premenná. V našej práci budeme označovať počet potomkov náhodnou premennou X , a jej hodnoty x . Spracovaním údajov vieme určiť konkrétne rozdelenie, ktorým sa počet potomkov riadi. Parametre týchto rozdelení budeme odhadovať z údajov *metódou momentov* a vhodnosť navrhovaného teoretického rozdelenia budeme testovať *Pearsonovým chí-kvadrát testom dobrej zhody*.

V rámci našej analýzy budeme tiež skúmať, ako môžu kontrolné mechanizmy meniť chod šírenia nákazy, čo vplýva na pravdepodobnosť zániku nákazy. Tieto implementácie vieme vykonať prostredníctvom úpravy pôvodného rozdelenia počtu potomkov v parametroch. Kvôli týmto úpravám si predstavíme *useknuté a modifikované* diskretné rozdelenia.

So známym rozdelením počtu potomkov môžeme skonštruovať vetviaci proces. Výsledkom vetviaceho procesu je rozdelenie počtu nakazených osôb n -tej generácie, ktoré vzniklo n -krát zložením rozdelenia počtu potomkov. Túto náhodnú premennú budeme označovať Z_n a jej hodnoty z . Takisto vieme na základe vzťahov definovaných v prvej časti priblížiť toto rozdelenie základnými číselnými charakteristikami a pravdepodobnostnou vytvárajúcou funkciou, pomocou ktorej možno analyticky vypočítať pravdepodobnosť eventuálneho zániku nákazy.

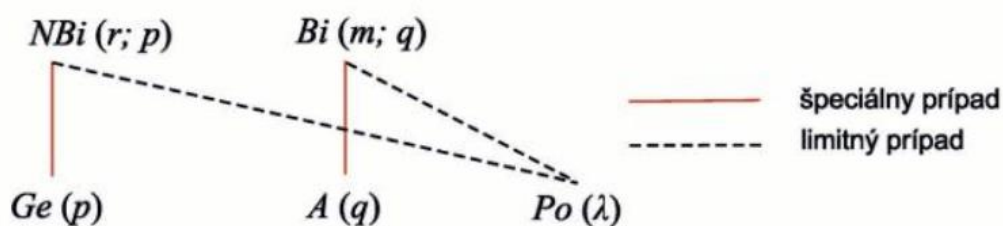
3.1 Modely počtu potomkov

Rozdelenie počtu potomkov je pravdepodobnostné rozdelenie počtu sekundárnych prípadov náhodnej premennej X spôsobených každým infekčným hostiteľom, ktoré sa všeobecne riadi Poissonovým rozdelením, avšak líši sa rôznorodosťou spôsobenou či už hostiteľom, alebo inými faktormi. Na adresovanie spomínanej rôznorodosti, hodnota pre infekčný profil daného jednotlivca je potom očakávaná hodnota počtu sekundárnych prípadov, ktoré spôsobí, resp. jeho individuálne reprodukčné číslo R_0 . Kvôli vplyvu okolností na prenos choroby, R_0 nie je nevyhnutne pevným atribútom každého jednotlivého hostiteľa, ale skôr vlastnosť konkrétnej infekčnej anamnézy daného hostiteľa (t.j. okolností počas celého infekčného obdobia tohto hostiteľa) ^[3].

Podľa ^[3] sú vhodnými kandidátnymi modelmi diskkrétne rozdelenia *Poissonovo*, *geometrické*, resp. *negatívne binomické*. Autori predkladajú teoretický rámec, kde rozoberajú, v akých situáciách možno aplikovať jednotlivé rozdelenia. Avšak pre jednoduchosť našej práce a faktu, že preferujeme prístup cez analýzu údajov o počte potomkov, budeme sa riadiť kritériom a overovaním vhodnosti vybraného modelu stanovenými metódami. Z uvedeného vyplýva, že nevylučujeme ani rozdelenie *binomické*. Na strane 25, v tabuľke 3.1 je prehľad základných diskkrétnych rozdelení.

Tieto základné diskkrétne rozdelenia spolu veľmi úzko súvisia. Limitné vlastnosti sa prejavujú vtedy, keď rozsah súboru, alebo parametre nadobúdajú určitú veľkú ($\theta \rightarrow \infty$), resp. malú hodnotu ($\theta \rightarrow 0$). Za takých predpokladov, môžeme spomenuté rozdelenia aproximovať navzájom. Okrem limitných prípadov, vieme niektoré rozdelenia vyjadriť ako špeciálne prípady druhých, napr. geometrické rozdelenie je špeciálnym prípadom negatívne binomického, ak $r = 1$.

Obrázok 3.1 Vzájomné vzťahy základných diskkrétnych rozdelení (Zdroj: ^[9])



Tabuľka 3.1 Zákony rozdelenia a vlastnosti základných diskretných rozdelení ako modelov počtu potomkov. (Zdroj: Vlastné spracovanie podľa použitej literatúry).

Rozdelenie	Parametre	Pravdepodobnostná funkcia	Stredná hodnota	Disperzia	Pravdepodobnostná vytvárajúca funkcia	Limitný vzťah
$X \sim Po(\lambda)$	$\lambda > 0$	$p_X(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda}, & x = 0, 1, 2, \dots \\ 0, & x \neq 0, 1, 2, \dots \end{cases}$	$E(X) = \lambda$	$D(X) = \lambda$	$P_X(t) = e^{\lambda(t-1)}$	-
$X \sim Bi(m; q)$	$m \in N$ $q \in (0; 1)$	$p_X(x) = \begin{cases} \binom{m}{x} \cdot q^x \cdot p^{m-x}, & x = 0, 1, 2, \dots, m \\ 0, & x \neq 0, 1, 2, \dots, m \end{cases}$; $p = 1 - q$	$E(X) = m \cdot q$	$D(X) = m \cdot q \cdot p$	$P_X(t) = (p + q \cdot t)^m$	$m \rightarrow \infty; q \rightarrow 0$ $Bi \approx Po$
$X \sim NBi(r; p)$	$r > 0$ $p \in (0; 1)$	$p_X(x) = \begin{cases} \binom{x+r-1}{r-1} \cdot p^r \cdot q^x, & x = 0, 1, 2, \dots \\ 0, & x \neq 0, 1, 2, \dots \end{cases}$; $q = 1 - p$	$E(X) = r \cdot \frac{q}{p}$	$D(X) = r \cdot \frac{q}{p^2}$	$P_X(t) = \left(\frac{p}{1 - q \cdot t} \right)^r$	$r \rightarrow \infty; p \rightarrow 0$ $NBi \approx Po$
$X \sim Ge(p)$	$p \in (0; 1)$	$p_X(x) = \begin{cases} q^x \cdot p, & x = 0, 1, 2, \dots \\ 0, & x \neq 0, 1, 2, \dots \end{cases}$; $q = 1 - p$	$E(X) = \frac{q}{p}$	$D(X) = \frac{q}{p^2}$	$P_X(t) = \frac{p}{1 - q \cdot t}$	-

3.1.1 Useknuté a modifikované rozdelenia počtu potomkov

Rekurenciou v matematike možno vyjadriť postupnosť tak, že podľa určitého pravidla, alebo kľúča vypočítame nasledujúci člen postupnosti jeho predchádzajúcimi členmi. Rozdelenia z časti 3.1 môžeme vyjadriť rekurzívnou metódou *Harry. H. Panjera*. Tento kanadský aktuár prišiel so vzťahmi pre výpočet diskretných rozdelení, ktoré nazval trieda rozdelení $(a,b,0)$. Túto triedu rozšíril o ďalšie rozdelenia, ktoré zoskupil do triedy $(a,b,1)$.

Modifikované rozdelenia patriace do triedy $(a,b,1)$ budú pre nás modelmi počtu potomkov, v prípade zavedenia špeciálneho typu kontrolného mechanizmu. V tejto časti si ukážeme najmä techniky, pomocou ktorých môžeme získať modifikované rozdelenia, pretože v rámci individuálnej-špecifickej kontroly budeme pôvodné modely počtu potomkov modifikovať v nule. Kontrolné mechanizmy predstavíme v časti 3.2.6.

Trieda rozdelení $(a,b,0)$

Podľa ^[9], pravdepodobnostné funkcie rozdelení triedy $(a,b,0)$ vieme vyjadriť cez *Panjerov rekurzívny vzťah*

$$p(x) = \left(a + \frac{b}{x}\right)p(x-1), \quad x = 1, 2, 3, \dots \quad (3.1)$$

kde a, b sú konštanty, pričom začiatočná hodnota pre rekurzívny výpočet hodnôt $p(0) > 0$. Jediné diskretné rozdelenia, nadobúdajúce nezáporné celočíselné hodnoty, ktoré vyhovujú vzťahu (3.1) sú Poissonovo, binomické a negatívne binomické rozdelenie ^[9].

Tabuľka 3.2 Panjerové konštanty pre triedu rozdelení $(a,b,0)$.

Rozdelenie	a	b	$p_x(0)$
$X \sim Po(\lambda)$	0	λ	$e^{-\lambda}$
$X \sim Bi(m; q)$	$-\frac{q}{1-q}$	$\frac{q}{1-q}(m+1)$	$(1-q)^m$
$X \sim NBi(r; p)$	$1-p$	$(1-p)(r-1)$	p^r
$X \sim Ge(p)$	$1-p$	0	p

(Zdroj: Vlastné spracovanie podľa ^[9])

Trieda rozdelení $(a,b,1)$

Pre túto triedu rozdelení platí $p(0) = 0$ a nazývame ich rozdelenia v nule useknuté (z angl. *zero-truncated distributions*). Okrem useknutých, patria do tejto triedy aj rozdelenia v nule modifikované (z angl. *zero-modified distributions*).

Pre túto triedu rozdelení hodnoty pravdepodobnostnej funkcie môžeme vypočítať rekurzívne podľa vzťahu

$$p(x) = \left(a + \frac{b}{x}\right) p(x-1) \text{ pre } x = 2, 3, 4, \dots \quad (3.2)$$

kde a a b sú konštanty. Zo vzťahu vyplýva, že štartovacia hodnota pre rekurzívny výpočet je $p(1)$. Rozlišujeme dva prípady, resp. podtriedy týchto rozdelení a metódy ich získania, ktoré rozlišujú podľa toho, či $p(0) = 0$, alebo $p(0) > 0$.

Prvá metóda sa nazýva *nulové skrátenie*, resp. *useknutie* (z angl. *zero-truncation*). Ak $p_X(x)$ sú funkčné hodnoty pravdepodobnostnej funkcie rozdelení z triedy $(a,b,0)$, tak pre modifikované hodnoty platí vzťah ^[9]

$$p_{X^T}(x) = \frac{p_X(x)}{1 - p_X(0)} \text{ pre } x = 1, 2, 3, \dots \quad (3.3)$$

Druhá metóda sa nazýva *nulová modifikácia* (z angl. *zero-modification*). Ak $p_X(x)$ sú hodnoty pravdepodobnostnej funkcie rozdelení z triedy $(a,b,0)$, je nulová modifikácia daná vzťahom ^[9]

$$p_{X^M}(x) = \frac{1 - p_{X^M}(0)}{1 - p_X(0)} \cdot p_X(x) \text{ pre } x = 1, 2, 3, \dots \quad (3.4)$$

kde $p_X(0)$ je nahradená $p_{X^M}(0)$ a zvyšné hodnoty $p_X(x)$, $x = 1, 2, 3, \dots$ sú „preškálované“ tak, aby $\sum_{x \in H} p_{X^M}(x) = 1$.

3.1.2 Odhad parametrov

Parametre rozdelení možno odhadnúť *metódou momentov*, ktorá na odhad využíva znalosť charakteristík strednej hodnoty a disperzie (za predpokladu, že existujú), ktoré vypočítame z výberového súboru ^[11].

Ak $X \sim Po(\lambda)$, pre odhad parametra λ platí:

$$\hat{\lambda} = \bar{x} = \frac{\sum_{k=0}^{\infty} k \cdot n_k}{\sum_{k=0}^{\infty} n_k} \quad (3.5)$$

kde $\bar{x} \approx R_0$ je výberový aritmetický priemer, n_k je počet pozorovaní (počet potomkov) s frekvenciou k .

Ak $X \sim Ge(p)$, pre odhad parametra p platí:

$$\hat{p} = \frac{1}{1 + \bar{x}} \quad (3.6)$$

Ak $X \sim NBi(p)$, pre odhad parametrov r a p platí:

$$\hat{p} = \frac{\bar{x}}{s^2} ; \quad \hat{r} = \frac{\bar{x}^2}{s^2 - \bar{x}} \quad (3.7)$$

kde s^2 je výberový rozptyl vypočítaný vzťahom:

$$s^2 = \frac{\sum_{k=0}^{\infty} k^2 \cdot n_k}{\sum_{k=0}^{\infty} n_k} - \left(\frac{\sum_{k=0}^{\infty} k \cdot n_k}{\sum_{k=0}^{\infty} n_k} \right)^2 = \overline{x^2} - \bar{x}^2 \quad (3.8)$$

Poznámka 3.1 ^[9] Prvotným orientačným kritériom pri výbere vhodného rozdelenia počtu potomkov je vzťah medzi strednou hodnotou $E(X)$ a $D(X)$. V prípade,

- ak $E(X) = D(X)$, potom $X \sim Po(\lambda)$,
- ak $E(X) < D(X)$, potom $X \sim NBi(r; p)$, resp. $X \sim Ge(p)$,
- ak $E(X) > D(X)$, potom $X \sim Bi(m; q)$.

3.1.3 Test dobrej zhody

Ak máme špecifikované, resp. odhadnuté parametre daného modelu, je dôležité overiť vhodnosť tohto modelu porovnaním s empirickou distribúciou daných údajov. Rozdelenia môžeme overovať tzv. testami dobrej zhody. Pomocou nich testujeme nulovú hypotézu:

H_0 : Údaje sa riadia modelovým (teoretickým) rozdelením.

H_1 : Údaje sa neriadia modelovým (teoretickým) rozdelením.

Pearsonov chí-kvadrát test dobrej zhody:

Množinu empirických hodnôt rozdelíme do intervalov $\langle k_{i-1}; k_i \rangle, i = 1, 2, \dots, s$, kde pre každý z týchto intervalov uvažujeme dva druhy početností^[11]:

O_i (z angl. *Observed* – pozorované, absolútne početnosti n_i) hodnôt výberového súboru,

E_i (*expected*) – očakávané, tzv. teoretické početnosti, ktoré vypočítame vzťahom

$$E_i = n \cdot p_i \quad (3.9)$$

kde p_i sú relatívne teoretické početnosti (pravdepodobnosti) i -teho intervalu hodnôt a n je rozsah výberového súboru.

Podľa testovacej charakteristiky

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (3.10)$$

zamietame nulovú hypotézu vtedy ak $\chi^2 > \chi^2_{1-\alpha}(v)$, v je počet stupňov voľnosti $s - 1 - p$, kde s je počet intervalov a p je počet parametrov modelu. Tento test je vhodné použiť v prípade väčšieho rozsahu výberového súboru, pričom hranice intervalov (triedy) sa navrhujú tak, aby aspoň 80 % z nich obsahovalo minimálne 5 hodnôt. Ak je počet intervalov aspoň 8, test poskytuje spoľahlivé výsledky.

3.2 Model šírenia nákazy ako vetviaci proces

Ako sme načrtli v prvej kapitole, vetviacim procesom budeme skúmať ohnisko nákazy. V epidemiológii je ohniskom náhle zvýšenie výskytu choroby v konkrétnom čase a na danom mieste ^[6]. Ohnisko pretrváva tak dlho, kým v ňom nevyhynie posledný infikovaný jedinec. Práve informácia o tom kedy „skončí“ resp. zanikne nákaza, je ako spomíname v úvode a cieľoch práce, jedna z najžiadanejších. Vetviacim procesom a diskretným zloženým rozdelením možno potom odpovedať na otázku kedy a či vôbec táto udalosť eventuálne nastane, teda počnúc koľkou generáciou a s akou pravdepodobnosťou a taktiež môžeme nájsť všeobecné podmienky, ktoré túto udalosť podmieňujú.

V tejto časti rozvineme model, ktorého predpoklady sme predstavili v podkapitole 1.4. Okrem základného aparátu predstavíme aj metódu *Monte Carlo simulácie*, podľa ktorej môžeme takýto model realizovať za pomoci výpočtovej techniky.

3.2.1 Rozdelenie náhodnej premennej Z_n

Definícia 3.1 Nech $\{X_i\}_{i=1}^{\infty}$ je postupnosť nezávislých a identicky rozdelených náhodných premenných počtu potomkov. Náhodnou premennou Z_{n-1} ($n \in \mathbb{N}$) označíme počet nakazených generácie $n - 1$ nezávislej od X . Potom náhodnú premennú Z_n opisujúcu počet nakazených n -tej generácie zapíšeme rekurzívnou reláciou

$$Z_n = \sum_{i=1}^{Z_{n-1}} X_i \quad (3.11)$$

kde $Z_0 = 1$ a $X \geq 0$. Postupnosť $\{Z_n\}_{n \in \mathbb{Z}_0^+}$ sa nazýva *Galton-Watsonov proces* s rozdelením potomkov X a jednotlivé členy postupnosti sú zložené diskkrétne náhodné premenné, teda $Z_1 \sim X$ a platí

$$p_{Z_n}(0) = p_{Z_{n-1}}(0) + \sum_{z=1}^{\infty} p_{Z_{n-1}}(z) \cdot p_X^{*z}(0) \quad , \quad n \in \mathbb{N} \quad (3.12)$$

Veta 3.1 Nech $P_X(t) = E(t^X)$ je pravdepodobnostná vytvárajúca funkcia NP X . Podľa vzťahu (1.33) ak $Z_n = X_1 + X_2 + \dots + X_{Z_{n-1}}$, pravd. vytvárajúca funkcia NP Z_n je

$$P_{Z_n}(t) = P_{Z_{n-1}}(P_X(t)) \quad (3.13)$$

Veta 3.2 ^[4] Nech $P_X(t) = E(t^X)$ je pravdepodobnostná vytvárajúca funkcia NP X . Nech $Z_0 = 1$ je počiatočná hodnota a nech Z_n je počet nakazených n -tej generácie, $n = 0, 1, 2, \dots$. Nech $P_{Z_n}(t)$ je pravdepodobnostná vytvárajúca funkcia NP Z_n . Potom

$$P_{Z_n}(t) = P_X \left(P_X \left(P_X \left(\dots P_X(t) \dots \right) \right) \right) \quad (3.14)$$

Poznámka 3.2 Z pravdepodobnostnej vytvárajúcej funkcie $P_{Z_n}(t)$ možno získať pravdepodobnostné rozdelenie náhodnej premennej Z_n , avšak v prípade veľkého n je veľmi nepravdepodobné, že toto rozdelenie dokážeme vyjadriť v jednoduchšej explicitnej forme. Napríklad, uvažujme $X \sim Po(\lambda)$, potom pre generáciu $n = 3$, dostaneme pravdepodobnostnú vytvárajúcu funkciu v nasledujúcom tvare:

$$P_{Z_n}(t) = e^{\lambda(e^{\lambda(e^{\lambda(t-1)}-1)}-1)}$$

3.2.2 Momenty náhodnej premennej Z_n

Pre lepšie pochopenie takéhoto procesu je potrebné vedieť vypočítať strednú hodnotu, resp. rozptyl pre ľubovoľnú generáciu, pričom si pomôžeme rozdelením počtu potomkov. Dôkazy nasledujúcich charakteristík uvádzame podľa ^[5].

Označme $E(X) = \mu$, $D(X) = \sigma^2$ strednú hodnotu a rozptyl počtu potomkov.

Veta 3.3 Nech $\{Z_n\}_{n \in \mathbb{Z}_0^+}$ je vetviaci proces s $Z_0 = 1$. Nech X je rozdelenie počtu potomkov a nech $E(X) = \mu$. Pretože $Z_n = X_1 + X_2 + \dots + X_{Z_{n-1}}$, potom stredná hodnota n -tej generácie nakazených je

$$\begin{aligned} E(Z_n) &= E(X) \cdot E(Z_{n-1}) \\ &= \mu \cdot E(Z_{n-1}) \\ &= \mu^2 \cdot E(Z_{n-2}) \\ &\vdots \\ &= \mu^{n-1} \cdot E(Z_1) \\ &= \mu^{n-1} \cdot \mu \\ &= \mu^n \end{aligned} \quad (3.15)$$

Veta 3.4 Nech $\{Z_n\}_{n \in \mathbb{Z}_0^+}$ je vetviaci proces s $Z_0 = 1$. Nech X je rozdelenie počtu potomkov a nech $D(X) = \sigma^2$ a $E(X) = \mu$. Pretože $Z_n = X_1 + X_2 + \dots + X_{Z_{n-1}}$, potom rozptyl n -tej generácie nakazených vyjadríme v tvare

$$D(Z_n) = \begin{cases} n \cdot \sigma^2, & \mu = 1 \\ \sigma^2 \mu^{n-1} \cdot \left(\frac{1 - \mu^n}{1 - \mu} \right), & \mu \neq 1 \end{cases} \quad (3.16)$$

Dôkaz:

Nech $D_n = D(Z_n)$. Podľa (1.30) vypočítame $D(Z_n)$ ako

$$D(Z_n) = E(X) \cdot D(Z_{n-1}) + E^2(Z_{n-1}) \cdot D(X)$$

$$\Rightarrow D_n = \mu^2 D_{n-1} + \sigma^2 E(Z_{n-1})$$

$$\Rightarrow D_n = \mu^2 D_{n-1} + \sigma^2 \mu^{n-1},$$

keďže

$$D_1 = D(Z_1) = D(X) = \sigma^2$$

potom

$$D_1 = \sigma^2$$

$$D_2 = \mu^2 D_1 + \sigma^2 \mu = \mu^2 \sigma^2 + \mu \sigma^2 = \mu \sigma^2 (1 + \mu)$$

$$D_3 = \mu^2 D_2 + \sigma^2 \mu^2 = \mu^2 \sigma^2 (1 + \mu + \mu^2)$$

$\vdots$

$$D_n = \mu^{n-1} \sigma^2 (1 + \mu + \mu^2 + \dots + \mu^{n-1})$$

$$= \mu^{n-1} \sigma^2 \left(\frac{1 - \mu^n}{1 - \mu} \right) \quad \text{ak } \mu \neq 1.$$

$$D_n = 1^{n-1} \sigma^2 (1^0 + 1^1 + \dots + 1^{n-1}) = \sigma^2 \cdot n \quad \text{ak } \mu = 1.$$

3.2.3 Pravdepodobnosť zániku nákazy

Ďalej priblížime originálny problém, ktorý sa snažil vyriešiť Galton, teda aká je pravdepodobnosť zániku rodu. V súvislosti s našou prácou môžeme vypočítať pravdepodobnosť zániku ohniska nákazy skôr než začne byť epidémiou. Ak $Z_n = 0$, hovoríme, že ohnisko zaniklo n -tou generáciou a platí $Z_m = 0$ pre $m \geq n$ ^[5].

Na začiatku vetviaceho procesu nás zaujíma aká je *pravdepodobnosť eventuálneho zániku*, teda pravdepodobnosť zániku nákazy n -tou generáciou, pre ľubovoľnú hodnotu n ^[5].

Definícia 3.2 Nech γ je pravdepodobnosť eventuálneho zániku a μ ľubovoľné nezáporné reálne číslo, potom

$$\lim_{n \rightarrow \infty} P_{Z_n}(t) = \gamma, \quad 0 \leq t < 1 \quad (3.17)$$

Veta 3.5 Nech γ je pravdepodobnosť eventuálneho zániku. Potom γ je najmenším nezáporným riešením rovnice $P_X(t) = t$.

Zo zistení Galtona, Watsona a neskôr Harrisa ^[8], vieme na základe hodnoty $\mu = E(X) \in (0; +\infty)$ rozdeliť G-W proces na 3 situácie:

- 1) Ak $\mu < 1$, zánik je istý ($\gamma = 1$). Proces sa nazýva *subkritický*.
- 2) Ak $\mu = 1$, zánik je istý ($\gamma = 1$), resp. ak $P(X = 1) = 1$, potom $\gamma = 0$. Proces sa nazýva *kritický* a platí, že $E(Z_n) = \mu^n = 1, \forall n$.
- 3) Ak $\mu > 1$, zánik nie je istý ($\gamma < 1$) a existuje pravdepodobnosť prežitia nákazy ($1 - \gamma$). Proces sa nazýva *superkritický* a platí, že $E(Z_n) = \mu^n \rightarrow \infty$ ak $n \rightarrow \infty$.

Príklad 3.1 ^[5] Nech $\{Z_0 = 1, Z_1, Z_2, \dots\}$ je vetviaci proces s počtom potomkov $X \sim Ge(0,25)$. Úlohou je nájsť pravdepodobnosť eventuálneho zániku γ .

Riešenie:

$$X \sim Ge(p) \Rightarrow P_X(t) = \frac{p}{1 - qt}$$

$$\text{Dosadením } p = \frac{1}{4} \Rightarrow P_X(t) = \frac{\frac{1}{4}}{1 - (\frac{3}{4})t} = \frac{1}{4 - 3t}$$

hľadáme najmenšie riešenie rovnice $P_X(t) = t$:

$$\frac{1}{4-3t} = t$$

$$4t - 3t^2 = 1$$

$$3t^2 - 4t + 1 = 0$$

$$(t-1) \cdot (3t-1) = 0$$

Koreňmi rovnice sú: $t_1 = 1$, $t_2 = \frac{1}{3} \Rightarrow \gamma = \frac{1}{3}$.

3.2.4 Model s prípadmi cestovateľskej anamnézy

V praxi sa z viacerých dôvodov, najmä v zmysle opatrení, modeluje šírenie nákazy v rámci hraníc krajín a iných územných celkov. Základný *G-W proces* predpokladá, že nultá generácia obsahuje práve jedného nakazeného, teda $P(Z_0 = 1) = 1$. V zmysle novovzniknutej epidémie, môžeme takéhoto jedinca nazvať *indexový prípad*. O tomto jedincovi sa potom predpokladá, že sa infikoval niekde mimo skúmaného územia a odborne sa takýto prípad nazýva *cestovateľská anamnéza*, resp. nákazu importoval zo zahraničia. Viacerí autori zaoberajúci sa vetviacimi procesmi (napr. pozri ^[8]) nazývajú takýto prípad *imigráciou*, v rámci vetviacich procesov.

Nemôžeme však predpokladať, že ide výlučne len o jednu takúto osobu, lebo na začiatku epidémie (ale aj v neskoršom priebehu) môže byť takýchto importovaných prípadov viacej, kde každý môže spustiť vlastný proces šírenia. Dá sa povedať, že každý importovaný prípad alebo „imigrant“ predstavuje potenciálne riziko lokálneho ohniska nákazy, preto má zmysel analyzovať, koľko takýchto prípadov sa môže vyskytnúť za časovú periódu. Ak zohľadníme tento fakt, potom neplatí $Z_1 \sim X$, pretože Z_0 je náhodnou premennou.

Definícia 3.3 Nech $\{X_i\}_{i=1}^{\infty}$ je postupnosť nezávislých, nezáporných, identicky rozdelených diskretných náhodných premenných opisujúcich počet potomkov danej nákazy, Z_0 je náhodná premenná počtu imigrantov nezávislá od X . Potom náhodná premenná Z_1 opisujúca počet nakazených prvej generácie je generovaná náhodnou premennou Z_0 a vyjadríme ju vzťahom

$$Z_1 = X_1 + X_2 + \dots + X_{Z_0} \quad (3.18)$$

náhodná premenná Z_1 má *diskrétné zložené rozdelenie*, Z_0 má primárne a náhodná premenná X sekundárne rozdelenie.

3.2.5 Metóda simulácie Monte Carlo na určenie rozdelenia počtu nakazených n -tej generácie

Základná myšlienka simulácií vychádza z priameho napodobňovania systému, ktorá nám umožní riešiť úlohy pravdepodobnostného charakteru bez využitia analytických riešení^[9]. Napodobňovanie systému uskutočňujeme generátorom pseudonáhodných čísel, čiže generovanie čísel podľa deterministického algoritmu, ktorý nie je klasickými štatistickými testami rozlíšiteľný od náhodného^[9]. Ak tento systém opakujeme niekoľko veľakrát, spracovaním takto nasimulovaných hodnôt možno získať rozdelenie modelovaného javu, ale aj iné informácie, pričom takto získané výsledky sú potom veľmi presné, resp. porovnateľné s exaktnými metódami.

Niektoré metódy generovania pseudonáhodných čísel generujú hodnoty z rovnomerného normovaného rozdelenia, teda výskyt ľubovoľného čísla z intervalu $(0; 1)$ je rovnako pravdepodobný a mimo tohto intervalu je pravdepodobnosť výskytu nulová. Čísla s rovnomerným normovaným rozdelením možno transformovať na čísla s požadovaným rozdelením^[14].

Na určenie rozdelenia náhodnej premennej Z_n predkladáme nasledujúci algoritmus, resp. postup jednotlivých krokov *Monte Carlo simulácie* v tomto poradí, ktoré sme upravili podľa^[13]:

1. Vygenerujeme počet nakazených generácie $n-1$, z_1 z rozdelenia náhodnej premennej Z_{n-1} využitím generátora pseudonáhodných čísel.

Poznámka 3.3 V rámci jazyka $\mathbb{R}$ existujú predprogramované pseudonáhodné generátory, teda so znalosťou príslušných parametrov môžeme funkciou `r***` vygenerovať počet potomkov konkrétného rozdelenia `***` n -krát. V prípade, že nedisponujeme funkciou konkrétného rozdelenia priamo, ale poznáme počet a príslušnú pravdepodobnosť, je možné použiť kombináciu funkcií `simul` a `subset` za predpokladu správne rozvrhutej syntaxe^[14].

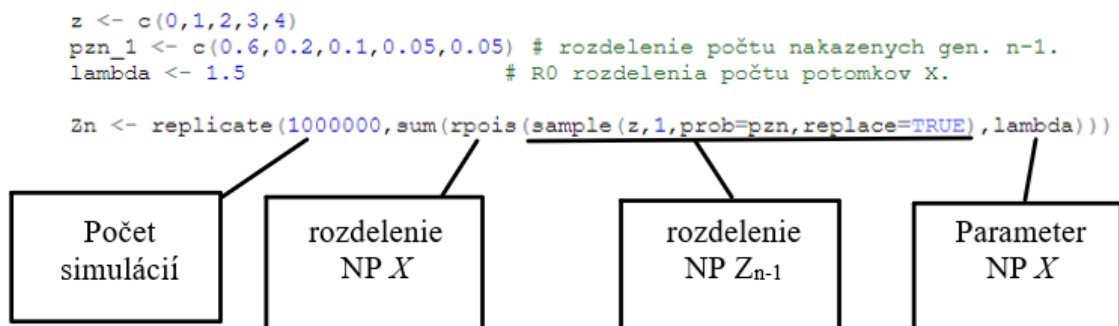
2. Zo známeho rozdelenia počtu potomkov X vygenerujeme len z_1 hodnôt počtu potomkov infikovaného jednotlivca $x_1, x_2, \dots, x_{z_1}$.
3. Súčet $z_{n_1} = x_1 + x_2 + \dots + x_{z_1}$ určuje prvé náhodné číslo z_{n_1} počtu nakazených n -tej generácie Z_n .
4. Kroky 1 až 3 opakujeme m -krát čím získame pseudonáhodné generované čísla $z_{n_1}, z_{n_2}, \dots, z_{n_m}$ z neznámeho rozdelenia náhodnej premennej Z_n .

Prezentované kroky *Monte Carlo Simulácie* možno realizovať v jazyku $\mathbb{R}$ využitím funkcií `replicate` a `r*** (simul, subset)`. Obrázok 3.2, ktorý sme vytvorili podľa ^[14], znázorňuje syntax funkcie `replicate` pre simulovanie hodnôt náhodnej premennej Z_n počtu nakazených n -tej generácie, kde počet potomkov generujeme cez `r***` a počet nakazených $n-1$ -tej generácie cez `sample`. Teda ak napríklad $X \sim Po(\lambda = 1,5)$ a rozdelenie náhodnej premennej Z_{n-1} je dané tabuľkou:

z	0	1	2	3	4
$p_{Z_{n-1}}(z)$	0,6	0,2	0,1	0,05	0,05

Potom $Z_n \sim Co(p_{Z_{n-1}}(z); X \sim Po(1,5))$.

Obrázok 3.2 Kód v jazyku R na realizáciu MC simulácie podľa prezentovaných krokov.



(Zdroj: Vlastné spracovanie podľa ^[14])

3.2.6 Kontrolované šírenie vo vetviacom procese

Doteraz sme uvažovali o situácii, keď sa nákaza šíri nekontrolovane, teda bez zásahu vonkajších vplyvov. Zrejmy cieľ pri zasahovaní do náhodného procesu šírenia je minimalizovať, resp. úplne zastaviť šírenie, čo je v tejto hraničnej situácii veľmi ťažko dosiahnuteľný cieľ. Tieto zásahy sú v praxi realizované konkrétnymi vládnyimi orgánmi, ktorých úlohou je monitorovať stav, ale aj navrhovať riešenia a opatrenia, na zamedzenie šírenia novovzniknutej infekčnej nákazy vo forme intervencií. Takéto intervencie majú zvyčajne charakter celoplošných, resp. regionálnych zákazov a obmedzení, alebo priamych kontrol vybranej časti populácie. ^[2] Prišli s metodikou, ako kvantifikovať dopady týchto dvoch spôsobov intervencií na šírenie nákazy v rámci vetviaceho procesu, ktorému sme sa venovali v podkapitole 3.2, modifikáciou rozdelenia počtu potomkov.

Predpokladajme epidémiu, ktorá má prirodzené (nekontrolované) rozdelenie počtu potomkov $X \sim NBi(r; p)$.

Poznámka 3.4 V tomto prípade uvádzame alternatívnu parametrizáciu negatívne binomického rozdelenia, s disperzným parametrom r a základným reprodukčným číslom $R_0 = E(X)$. Pravdepodobnostnú vytvárajúcu funkciu vyjadríme potom v tvare

$$P_X(t) = \left(1 + \frac{R_0}{r}(1-t)\right)^{-r} \quad (3.19)$$

^[2] Prvý spôsob intervencie nazvali **kontrola celej populácie** (z angl. *population-wide control*), čo predstavuje redukciiu infekčnosti každého jedinca o faktor $c \in (0; 1)$ tak aby ich očakávaný počet potomkov, teda ich základné reprodukčné číslo bolo zredukované na $R_c^{pop} = (1-c)R_0$. Pri tejto kontrole sa teda zmení iba parameter R_0 a disperzný parameter r zostane nezmenený. Potom počet potomkov $X_c^{pop} \sim NBi((1-c)R_0; r)$. Pomer medzi rozptylom a strednou hodnotou, resp. disperzný koeficient (označíme ako D) pri X_c^{pop} vieme vyjadriť vzťahom

$$D_{pop} = 1 + (1-c) \cdot \frac{R_0}{r} \quad (3.20)$$

a klesá monotónne pri náraste faktora c .

^[2] Ďalším spôsobom intervencie je **individuálna-špecifická kontrola** (z angl. *individual-specific control*), kde každý infikovaný jedinec je kontrolovaný dokonale (tak, že

nespôsobí prenos nákazy) s pravdepodobnosťou c . Implementácia individuálnej kontroly ovplyvňuje prenos iba pre zlomok $1 - p_0$ všetkých jedincov, ktorých prirodzený počet potomkov je väčší ako nula. Z týchto jedincov, pre zlomok „odchytených“ c bude platiť $X_c^{ind} = 0$ a pre zlomok ostatných $1 - c$ sa nič nezmení, teda $X_c^{ind} = X$. Takýto model sa dá vyjadriť buď vo forme zmesi dvoch rozdelení ako uvádzajú autori [2], alebo ako vieme z časti 3.1.1, môžeme takýto model získať pomocou *nulovej modifikácie* pôvodného rozdelenia. Pravdepodobnosť, že jedinec ďalej nikoho neinfikuje vyjadríme ako

$$p_0^{ind} = p_0 + c(1 - p_0) \quad (3.21)$$

čiže počet potomkov $X_c^{ind} \sim ZMNBi(R_0; r)$ (z angl. *Zero-modified negative binomial distribution*) s modifikovanou hodnotou pravdepodobnostnej funkcie v nule $p_{X_c^{ind}}(0) = p_0^{ind}$. V tomto prípade disperzný koeficient vyjadríme ako

$$D_{ind} = 1 + \frac{R_0}{r} + c \cdot R_0 \quad (3.22)$$

a rastie monotónne pri náraste faktora c .

Teda ak $X \sim NBi(R_0; r)$, po prevedení individuálnej-špecifickej kontroly môžeme vyjadriť pravdepodobnostnú vytvárajúcu funkciu v tvare

$$P_X(t) = c + (1 - c) \cdot \left(1 + \frac{R_0}{r}(1 - t)\right)^{-r} \quad (3.23)$$

Veta 3.6 Pre všetky $c \in (0; 1 - \frac{1}{R_0})$, pravdepodobnosť zániku n -tej generácie $p_{Z_n}(0)$ je vždy väčšia v prípade *individuálnej-špecifickej kontroly* ako v prípade *kontroly celej populácie* a platí $\gamma^{ind} > \gamma^{pop}$ pre identické rozdelenie počtu potomkov X . Ak $c > 1 - \frac{1}{R_0}$, potom $R_c^{pop} = R_c^{ind} < 1$, čiže $\gamma^{ind} = \gamma^{pop} = 1$.

Z hľadiska jednoduchosti našej práce sa nebudeme venovať dôkazu tejto vety, avšak čitateľ ju môže vyhľadať v [2].

4 Výsledky práce a diskusia

V praktickej časti našej práce vytvoríme vlastný program v jazyku $\mathbb{R}$, pomocou ktorého získame rozdelenie počtu nakazených koronavírusom n -tej generácie, teda budeme simulovať vetviaci proces. Zloženým rozdelením budeme skladať rozdelenie $n-1$ -tej generácie s rozdelením počtu potomkov aby sme získali rozdelenie počtu nakazených n -tej generácie. Prostriedok, ktorým budeme realizovať zložené rozdelenia v jednotlivých iteráciách je *Monte Carlo simulácia*. Postup praktickej časti vieme opísať nasledujúcimi krokmi:

- Vytvoríme algoritmus, ktorý realizuje vetviaci proces nekontrolovaného šírenia nákazy v hypotetickej lokalite s prípadom, keď $Z_0 = 1$, resp. keď Z_0 je náhodnou premennou opisujúcou denný počet importovaných prípadov.
- Z údajov o počte prenosov infekcií jednotlivca v hypotetickej lokalite získame konkrétny zákon rozdelenia počtu potomkov a vieme tak vypočítať pravdepodobnosť eventuálneho zániku ohniska. Z údajov o denných prípadoch s cestovateľskou anamnézou získame zákon rozdelenia počtu importovaných prípadov.
- So znalosťou parametrov šírenia (parametre rozdelenia počtu potomkov) budeme simulovať ohnisko nákazy prostredníctvom vytvoreného algoritmu.
- Výsledok, ktorý získame pre 4-tú generáciu počtu nakazených na tento druh koronavírusu pre prípad $Z_0 = 1$ porovnáme s prípadom, keď Z_0 opisuje počet importovaných prípadov za jeden deň.
- Vetviaci proces s prípadom $Z_0 = 1$ použijeme na analýzu efektivity kontrolných mechanizmov (pozri časť 3.2.6) s vopred stanovenou minimálnou pravdepodobnosťou zániku.

Poznámka 4.1 Údaje na odhad počtu potomkov, resp. na odhad denného počtu importovaných prípadov sme adekvátne nasimulovali tak, aby približne zodpovedali charakteru údajov, ktoré sme odpozorovali vo viacerých publikáciách s rovnakou problematikou.

4.1 Zostavenie vetviaceho procesu šírenia nákazy v jazyku $\mathbb{R}$

Algoritmus vetviaceho procesu najprv predstavíme pre prípad $Z_0 = 1$. Tento algoritmus skonštruujeme zovšeobecnením algoritmu *Monte Carlo simulácie* prezentovanej v časti 3.2.5, resp. vnorením funkcie `replicate` do cyklu. Každá iterácia predstavuje generovanie pseudonáhodného počtu potomkov $n-1$ -tej generácie na získanie novej, n -tej generácie počtu nakazených. Keďže prvá generácia sa bude počítať mimo cyklu, cyklus bude obsahovať $n-1$ iterácií na získanie pseudonáhodných hodnôt n -tej generácie. Vstupné parametre nášho programu budú:

n – počet simulácií,

R_0 – základné reprodukčné číslo R_0 , stredná hodnota počtu potomkov danej nákazy (pre prípad nekontrolovaného šírenia).

a ďalšie parametre v závislosti od zvoleného rozdelenia počtu potomkov. V prípade negatívne binomického, resp. geometrického rozdelenia špecifikujeme aj disperzný parameter r . V prípade, že budeme pracovať s binomickým rozdelením počtu potomkov, parameter R_0 nešpecifikujeme, nakoľko v prostredí $\mathbb{R}$ nevieme vyjadriť pseudonáhodný generátor čísel pre binomické rozdelenie cez jeho strednú hodnotu. V prípade rozdelenia Poissonovho, geometrického a negatívne binomického nám dané prostredie ponúka túto špeciálnu parametrizáciu cez strednú hodnotu.

Výstup z programu predstavuje vektor pseudonáhodných hodnôt počtu nakazených n -tej generácie opísaný náhodnou premennou Z_n , ktorý vieme viacerými príkazmi na manipuláciu údajov spracovať do podoby tabuľky. Táto tabuľka bude okrem hodnôt vygenerovaných, obsahovať početnosť týchto hodnôt, ich relatívnu početnosť, resp. pravdepodobnostnú funkciu (získame po predelení početností jednotlivých hodnôt počtom simulácií) a distribučnú funkciu (nakumulované hodnoty pravdepodobnostnej funkcie).

Pre prípad, keď Z_0 je náhodnou premennou, použijeme inú funkciu na získanie pseudonáhodných hodnôt náhodnej premennej Z_1 a aj všetkých nasledujúcich generácií. Konkrétne ide o funkciu `sample` (vzorkovanie). Pre tento prípad, bude počet iterácií algoritmu rovný počtu generácie, ktorú počítame, pretože Z_1 počítame už vo vnútri cyklu. Kód algoritmu, ktorý sme v tejto časti predstavili, uvádzame v prílohe č.1 s komentármi.

4.2 Výber modelu počtu potomkov koronavírusu na základe údajov

V tabuľke 4.1 je 500 pozorovaní počtu prenosov infekcií spôsobených jedným nakazeným jedincom na koronavírus v hypotetickej lokalite.

Tabuľka 4.1 Údaje na odhad rozdelenia počtu potomkov koronavírusu.

Počet prenosov infekcií	Počet nakazených
0	220
1	117
2	58
3	39
4	23
5	16
6	9
7	6
8	6
9	1
11	2
12	2
13	1

Po načítaní údajov vypočítame odhad základného reprodukčného čísla R_0 , resp. vypočítame výberový priemer cez funkciu `mean` a výberový rozptyl funkciou `var`.

$$\widehat{R}_0 = \bar{x} = 1,468; \quad s^2 = 4,353683 \quad (4.1)$$

Keďže $E(X) < D(X)$, údaje sa budú pravdepodobne riadiť negatívne binomickým rozdelením, resp. geometrickým. Ďalej odhadneme parametre negatívne binomického rozdelenia r , p metódou momentov z údajov prostredníctvom vzťahov (3.7), teda dostaneme hodnoty

$$\hat{r} = 0,7467985; \quad \hat{p} = 0,3371858 \quad (4.2)$$

Po odhadnutí parametrov môžeme ísť testovať hypotézu o tom, či údaje pochádzajú z navrhovaného rozdelenia. Chí-kvadrát test dobrej zhody môžeme vykonať v jazyku $\mathbb{R}$ vypočítaním testovacej štatistiky χ^2 a následným porovnaním s kritickou hodnotou so zvolenou hladinou významnosti α a s počtom stupňov voľnosti $v = 10$, alebo porovnaním s p -hodnotou. Vypočítame teoretické početnosti, teda $E_i = n \cdot p_i$, podľa negatívne binomického rozdelenia. Na výpočet hodnôt pravdepodobnostnej funkcie použijeme funkciu `dnbinom`, ktorej argumenty sú odhadnuté parametre $\hat{r}$ a $\hat{p}$ a vektor pozorovaných hodnôt.

Vypočítané hodnoty pravdepodobnostnej funkcie upravíme tak, aby ich súčet bol rovný jednej.

Obrázok 4.1 Výpočet hodnôt p_i a E_i podľa NBi rozdelenia.

```
pnbi <- dnbinom(k, size = est_r, prob = est_p) #vypocet hodnot pf NBi
pnbi_resc <- pnbi*1/sum(pnbi)                #uprava hodnot
Enbi <- pnbi_resc*sum(nk)                    #vypocet ocek.pocet. podľa NBi
```

Pomocou vzťahu (3.10) vypočítame testovaciu štatistiku χ^2 , ktorú dosadíme do vzťahu $1 - F(\chi^2; v = 10)$, čím získame p -hodnotu. Ak bude táto hodnota väčšia ako $\alpha = 0,01$, potom nemôžeme zamietnuť hypotézu, že počet potomkov sa riadi navrhnutým rozdelením.

Obrázok 4.2 Chí-kvadrát test dobrej zhody pre NBi rozdelenie.

```
chisq <- sum((nk - Enbi)^2 / Enbi) #vypocet statistiky
df <- length(k) - 3               # stupne volnosti
p_value <- 1 - pchisq(q = chisq, df) # p-hodnota
p_value
```

Tabuľka 4.2 Výsledok analýzy rozdelenia počtu potomkov pre danú lokalitu.

Počet prenosov infekcií	Počet nakazených	NBi
0	220	223,12
1	117	110,44
2	58	63,93
3	39	38,80
4	23	24,09
5	16	15,16
6	9	9,62
7	6	6,15
8	6	3,95
9	1	2,54
10+	5	2,20
<i>Spolu</i>	500	500
χ^2		7,121146
$\chi^2(0,99; 10)$		23,20925
p -hodnota		0,7139607
parametre		$\hat{r} = 0,7467985$; $\hat{p} = 0,3371858$

Vypočítaná p -hodnota = 0,7139607 > 0,01, teda počet potomkov opísaný náhodnou premennou $X \sim NBi(0,7467985; 0,3371858)$.

4.2.1 Výpočet pravdepodobnosti eventuálneho zániku ohniska

S Pravdepodobnosťou eventuálneho zániku ohniska γ pre odhadnuté rozdelenie počtu potomkov na daný druh koronavírusu môžeme potom analyzovať „ako ďaleko“ je pravdepodobnosť zániku danej generácie od eventuálnej. γ vypočítame vzťahom:

$$\lim_{n \rightarrow \infty} P_{Z_n}(t) = \gamma, \quad 0 \leq t < 1$$

Keďže

$$X \sim NBi(0,7467985; 0,3371858); \widehat{R}_0 = 1,468 \Rightarrow P_X(t) = \left(1 + \frac{1,468}{0,7467985}(1-t)\right)^{-0,7467985}$$

potom hľadáme najmenší koreň rovnice

$$\left(1 + \frac{1,468}{0,7467985}(1-t)\right)^{-0,7467985} - t = 0$$

Korene rovnice môžeme nájsť prostredníctvom funkcie `uniroot.all` po inštalácii balíka `rootSolve` v jazyku $\mathbb{R}$. Ďalej uvádzame kód v jazyku $\mathbb{R}$ na výpočet pravdepodobnosti eventuálneho zániku (`est_r` = 0,7139607 a `est_R0` = 1,468):

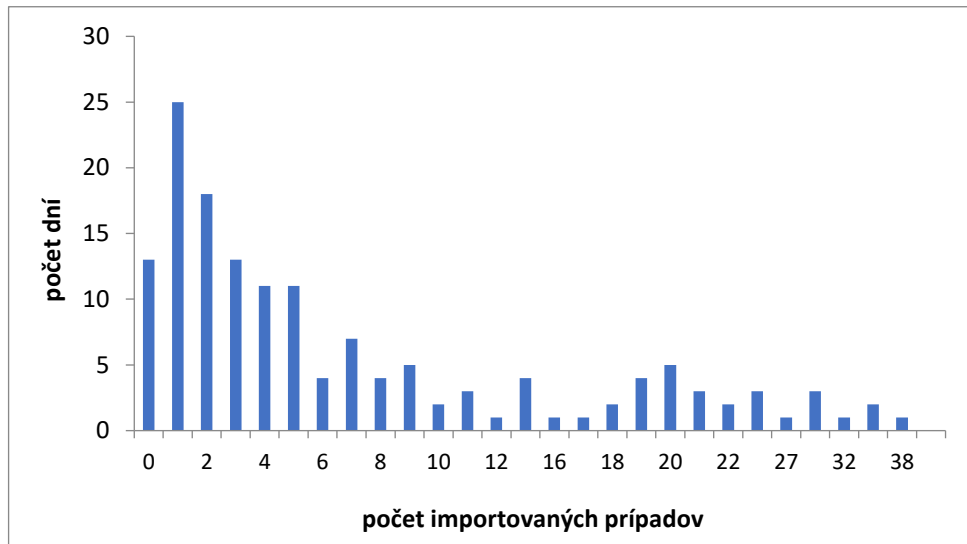
```
install.packages("rootSolve")
library(rootSolve)
eq <- function(t) {
  (1 + (est_R0/est_r)*(1-t))^( -est_r) - t
}
korene <- uniroot.all(eq, interval = c(0,2),
  lower = 0,
  upper = 2,
  tol = 0.00000000001)
```

Pravdepodobnosť eventuálneho zániku ohniska koronavírusu vzhľadom na dané rozdelenie počtu potomkov pre hypotetickú oblasť je $\gamma = 0,722331$. Pravdepodobnosť zániku nákazy konverguje k tejto hodnote pre Z_n , kde $n \rightarrow \infty$.

4.3 Rozdelenie počtu importovaných prípadov koronavírusu z údajov

Ďalej rozoberme situáciu, keď nultá generácia Z_0 je náhodnou premennou a opisuje nezávislý výskyt počtu importovaných prípadov, teda prípadov s cestovateľskou anamnézou. Na obrázku 4.3 je znázornený histogram 150 pozorovaní výskytu denného počtu importovaných prípadov koronavírusu v našej hypotetickej lokalite.

Obrázok 4.3 Počet importovaných prípadov koronavírusu a počet dní.



Po importovaní údajov do prostredia $\mathbb{R}$ spracujeme údaje pomocou nasledujúcich príkazov, čím získame hodnoty a ich pravdepodobnosť, teda dostaneme rozdelenie Z_0 , ktoré uvádzame v tabuľke 4.3.

```
import <- read.csv("import.csv", header = FALSE)
import <- as.matrix(import)
imp_tab <- as.data.frame(table(import))
days <- as.integer(as.vector(imp_tab$Freq))
zI <- imp_tab$import # získanie hodnôt
pZI <- days/sum(days) # výpočet pravd.
```

Tabuľka 4.3 Rozdelenie náhodnej premennej Z_0 .

z	0	1	2	3	4	5	6	7	8
$p_{Z_0}(z)$	0,0867	0,1667	0,1200	0,0867	0,0733	0,0733	0,0267	0,0467	0,0267
z	9	10	11	12	15	16	17	18	19
$p_{Z_0}(z)$	0,0333	0,0133	0,0200	0,0067	0,0267	0,0067	0,0067	0,0133	0,0267
z	20	21	22	23	27	28	32	33	38
$p_{Z_0}(z)$	0,0333	0,0200	0,0133	0,0200	0,0067	0,0200	0,0067	0,0133	0,0067

4.4 Analýza šírenia koronavírusu v hypotetickej lokalite

S odhadnutým rozdelením počtu potomkov na koronavírus získame simuláciou rozdelenie počtu nakazených 4-tej generácie v hypotetickej lokalite. Najskôr vytvoríme simuláciu vetviaceho procesu šírenia, ktorý začína jedným indexovým prípadom, teda $Z_0 = 1$. Pomocou generátora pseudonáhodných čísel v jazyku $\mathbb{R}$, vygenerujeme prvú generáciu nakazených z náhodnej premennej $Z_1 \equiv X$ počtu potomkov, ktorá sa riadi negatívne binomickým rozdelením s parametrami $\hat{r} = 0,7467985$ a $\hat{p} = 0,3371858$. Keďže budeme ďalej implementovať do modelu intervencie a modifikovať základné reprodukčné číslo R_0 , použijeme alternatívnu parametrizáciu negatívne binomického rozdelenia so strednou hodnotou $E(X) = R_0$ a disperzným parametrom r . Strednú hodnotu, resp. základné reprodukčné číslo koronavírusu sme odhadli z údajov, $\widehat{R}_0 = 1,468$. Po zadaní vypočítaných parametrov, počtu iterácií (v našom prípade 3 iterácie pre 4-tú generáciu) a počtu simulácii 100 000, spustíme základný algoritmus, ktorým vygenerujeme pseudonáhodné hodnoty počtu nakazených 4-tej generácie. Na obrázku 4.4 uvádzame výstup vo forme tabuľky, ktorá obsahuje absolútnu početnosť, hodnoty pravdepodobnostnej a distribučnej funkcie prvých 10 hodnôt.

Obrázok 4.4 Výstup simulácie počtu nakazených 4-tej generácie, s prípadom $Z_0 = 1$.

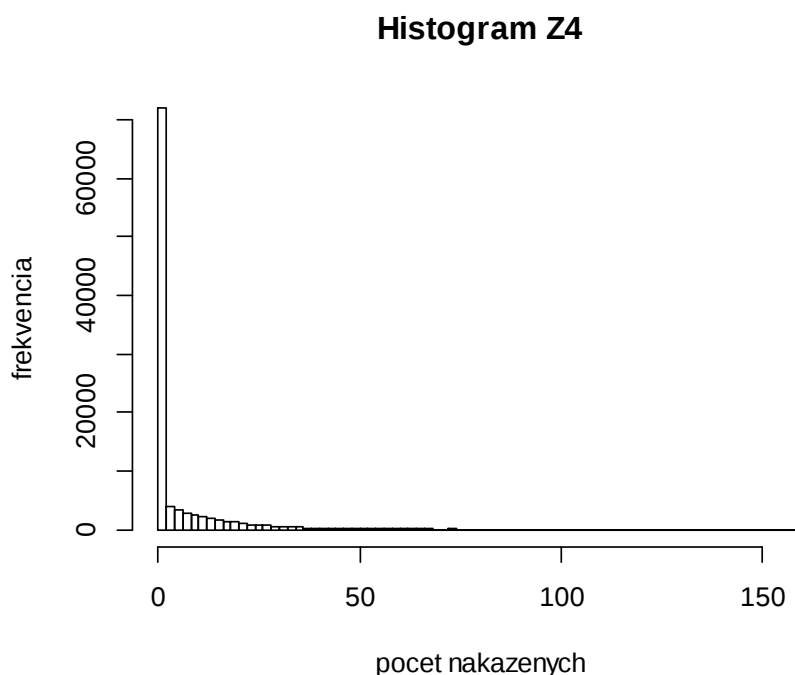
```
> Ztab
```

	z	Freq	pz	Fz
1	0	67197	0.67197	0.67197
2	1	2535	0.02535	0.69732
3	2	2252	0.02252	0.71984
4	3	2034	0.02034	0.74018
5	4	1921	0.01921	0.75939
6	5	1748	0.01748	0.77687
7	6	1536	0.01536	0.79223
8	7	1473	0.01473	0.80696
9	8	1416	0.01416	0.82112
10	9	1335	0.01335	0.83447

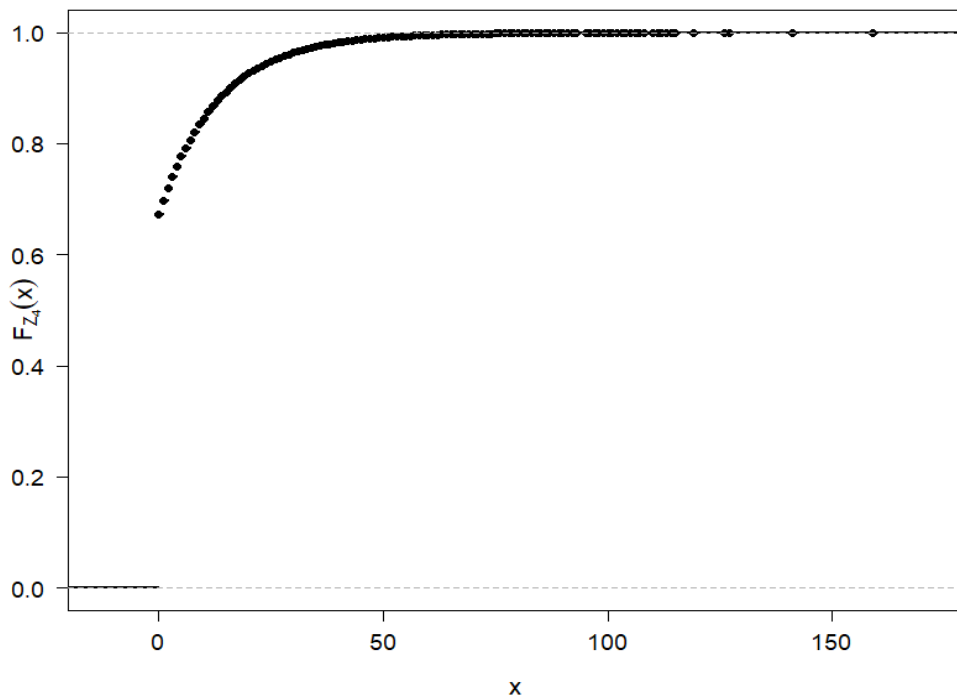
Teda získali sme rozdelenie počtu nakazených 4-tej generácie Z_4 s počtom potomkov $X \sim NBi(0,7467985; 0,3371858)$ pre prípad $Z_0 = 1$. Z výstupu si môžeme ako prvé všimnúť pravdepodobnosť zániku nákazy danou generáciou, teda $p_{Z_4}(0) = 0,67197$. Vzhľadom na vypočítanú hodnotu $\gamma = 0,722331$, môžeme sa domnievať, že pravdepodobnosť zániku danej generácie nemá ďaleko od pravdepodobnosti eventuálneho zániku. Pravdepodobnosť zániku štvrtou generáciou je o niečo väčšia, ako pravdepodobnosť, že nakazená osoba nikoho nenakazí, čo je pre tento koronavírus $p_X(0) = 0,44403$.

Očakávaný počet nakazených tejto generácie (simulovaná hodnota) je $E_{sim}(Z_4) = 4,592$, čo je počet približne rovný exaktnej hodnote $E_{ex}(Z_4) = 1,468^4 = 4,644$. Ďalšou zaujímavosťou je vysoká disperzia $D_{sim}(Z_4) = 106,4227$ (simulovaná hodnota). Túto hodnotu môžeme porovnať s exaktnou hodnotou disperzie, ktorú vypočítame pomocou vzťahu (3.16), teda dostaneme $D_{ex}(Z_4) = 107,2464$. Väčšina hodnôt sa sústreďuje okolo strednej hodnoty a maximálna hodnota je 159. Pozorujeme silné pravostranné zošikmenie, pretože počet nakazených nepresiahne hodnotu 80 s pravdepodobnosťou 0,99874. Toto rozdelenie môžeme priblížiť graficky histogramom, resp. distribučnou funkciou.

Obrázok 4.5 Histogram počtu nakazených Z_4 , pre prípad $Z_0 = 1$.



Obrázok 4.6 Distribučná funkcia počtu nakazených Z_4 , pre prípad $Z_0 = 1$.



Ďalej porovnáme dosiahnuté výsledky pre 4-tú generáciu s prípadom, keď náhodná premenná Z_0 opisuje denný počet importovaných prípadov a riadi sa rozdelením, ktoré sme odhadli v časti 4.3.

Rozdelenie náhodnej premennej Z_0 z tabuľky 4.3 pridáme do algoritmu na tak, že zložíme náhodnú premennú Z_0 s náhodnou premennou počtu potomkov X , teda dostaneme náhodnú premennú počtu nakazených prvej generácie v tvare

$$Z_1 = X_1 + X_2 + \dots + X_{Z_0}$$

Rozdelenie náhodnej premennej Z_1 určíme metódou Monte Carlo simulácie vo forme diskrétného zloženého rozdelenia, pretože ju budeme počítat' v rámci cyklu.

Znovu nás bude zaujímať rozdelenie počtu nakazených štvrtej generácie. V tomto prípade v algoritme zohľadníme importované prípady nákazy. Opäť vykonáme 100 000 simulácií pre každú generáciu s rovnakým rozdelením počtu potomkov $X \sim NBi(0,7467985; 0,3371858)$. Na obrázku opäť uvádzame prvých 10 hodnôt počtu nakazených 4-tej generácie, teda aj s absolútnymi početnosťami a hodnotami pravdepodobnostnej, resp. distribučnej funkcie.

Obrázok 4.7 Výstup simulácie počtu nakazených 4-tej generácie, s prípadom Z_0 je NP.

```
> ztab
```

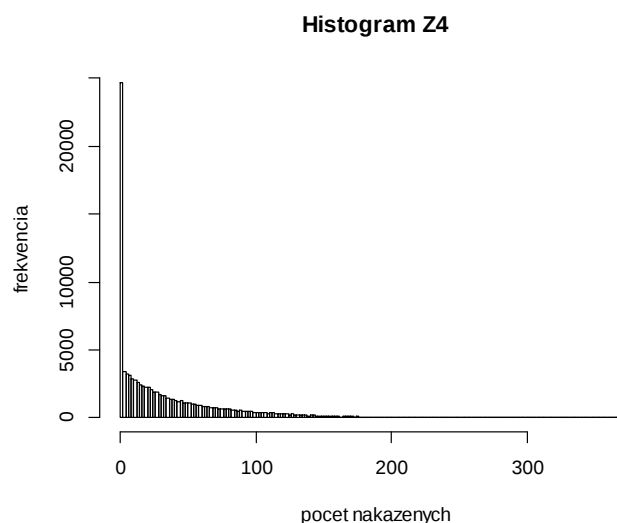
	z	Freq	pz	Fz
1	0	20791	0.20791	0.20791
2	1	1991	0.01991	0.22782
3	2	1877	0.01877	0.24659
4	3	1738	0.01738	0.26397
5	4	1638	0.01638	0.28035
6	5	1666	0.01666	0.29701
7	6	1566	0.01566	0.31267
8	7	1638	0.01638	0.32905
9	8	1525	0.01525	0.34430
10	9	1521	0.01521	0.35951

Z výstupu môžeme vidieť, aký vplyv má rozdelenie počtu importovaných prípadov v rámci nášho modelu. V tomto prípade je pravdepodobnosť zániku nákazy štvrtou generáciou $p_{Z_4}(0) = 0,20791$. V porovnaní s prípadom $Z_0 = 1$, bola táto pravdepodobnosť 0,67197. Ďalej vidíme, že simulovaná stredná hodnota počtu nakazených je $E_{sim}(Z_4) = 36,49$. Táto hodnota je približne rovná exaktnej strednej hodnote, ktorú vypočítame vzťahom

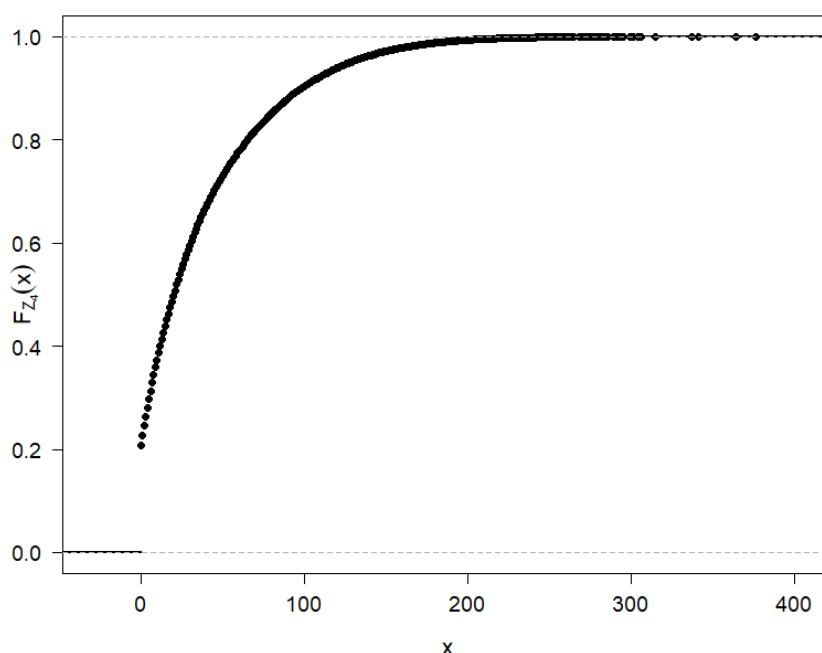
$$E_{ex}(Z_4) = E(Z_0) \cdot R_0^4 = 7,61333 \cdot 1,468^4 = 35,3573$$

Maximálny počet nakazených je 376 a počet nakazených nebude väčší ako 200 s pravdepodobnosťou 0,99348. Exaktnú hodnotu disperzie pre tento prípad nevieme vypočítať. V prílohe č.2 prezentujeme tabuľku, v ktorej porovnávame vývoj základných charakteristík a pravdepodobnosti zániku daného koronavírusu medzi jednotlivými prípadmi prvých dvadsiatich generácií počtu nakazených v hypotetickej lokalite.

Obrázok 4.8 Histogram počtu nakazených Z_4 , pre prípad Z_0 je NP.



Obrázok 4.9 Distribučná funkcia počtu nakazených Z_4 , pre prípad Z_0 je NP.



4.5 Analýza efektivity kontrolných mechanizmov

V tejto časti budeme na odhadnutom rozdelení počtu potomkov pre daný koronavírus aplikovať kontrolné mechanizmy, ktoré sme definovali v časti 3.2.6. V tejto časti budeme pracovať s prípadom $Z_0 = 1$ a pre negatívne binomické rozdelenie budeme používať parametre $R_0 = E(X)$ a r .

Analyzujeme efektivitu kontrolných mechanizmov pri dosiahnutí stanoveného cieľa, teda aby sme zaručili zaniknutie epidémie s minimálnou pravdepodobnosťou 0,9. Za efektívnejšiu kontrolu budeme považovať tú, ktorá zánik epidémie zabezpečí so stanovenou pravdepodobnostnou hranicou v nižšej generácii nakazených a s menším faktorom kontrolných mechanizmov c . Na dosiahnutie tohto cieľa si zvolíme konvergenčné kritérium, teda pravdepodobnosť eventuálneho zániku zvolíme ako $\gamma = 0,92$.

Implementáciou *kontroly celej populácie* v danej hypotetickej lokalite redukuje individuálne reprodukčné číslo každého jedinca o faktor c . Vzhľadom na nastavenú pravdepodobnosť eventuálneho zániku $\gamma = 0,92$ a odhad disperzného parametra

$\hat{r} = 0,7467985$, môžeme faktor c vypočítať pomocou pravdepodobnostnej vytvárajúcej funkcie z rovnice

$$\left(1 + \frac{(1-c) \cdot 1,468}{0,7467985} \cdot (1 - 0,92)\right)^{-0,7477985} - 0,92 = 0$$

Daná rovnica má jeden koreň, ktorý nájdeme nasledujúcim kódom v jazyku R:

```
eq2 <- function(c) {
  (1 + ((1-c)*R0/est_r)*0.08)^(-est_r) - 0.92
}
c_pop <- uniroot.all(eq2, interval = c(0,2),
  lower = 0,
  upper = 2,
  tol = 0.00000000001)
```

Po zbehnutí hore uvedeného kódu dostaneme $c = 0,2488534$. Vypočítame reprodukčné číslo po celoplošnej kontrole

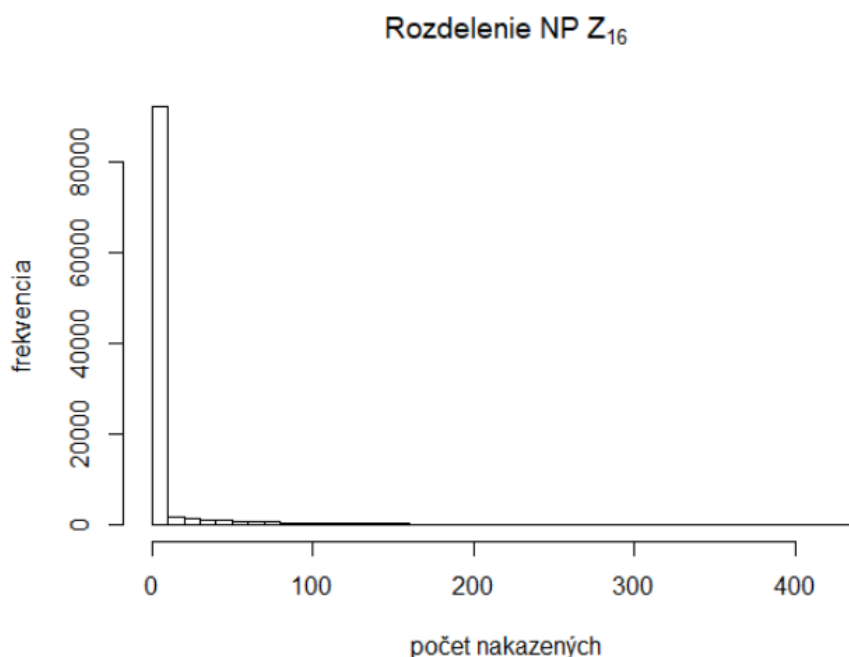
$$R_c^{pop} = (1 - 0,2488534) \cdot 1,468 = 1,10268$$

Náhodná premenná počtu potomkov sa bude riadiť nasledovným rozdelením

$$X_c^{pop} \sim NBi(R_c^{pop} = 1,10268; r = 0,7467985)$$

Simuláciu vetviaceho procesu prevedieme analogickým spôsobom ako v časti 4.4 s reprodukčným číslom R_c^{pop} po kontrole celej populácie a rovnakým disperzným parametrom (kód v jazyku R pre tento výpočet uvádzame v časti prílohy). Analyticky určiť počnúc koľkou generáciou dosiahne táto generácia stanovený cieľ je náročná úloha, preto sme sa rozhodli modifikovať algoritmus tak, aby sa cyklus zastavil keď dosiahne stanovený cieľ, teda $p_{Z_n}(0) \geq 0,9$. Výsledky naznačujú, že tento koronavírus s danými parametrami šírenia a s faktorom kontrolných mechanizmov $c = 0,2488534$ dosiahne zánik s pravdepodobnosťou 0,90153 počnúc šestnástou generáciou, ak presadzujeme kontrolu celej populácie.

Obrázok 4.10 Histogram 16-tej generácie nakazených koronavírusom v prípade kontroly celej populácie s $c = 0,2488534$.



Implementáciou *individuálnej-špecifickej kontroly* náhodne izolujeme jedincov s potenciálom šíriť nákazu, teda modifikujeme ich reprodukciu na stav $X_c^{ind} = 0$. Obdobne ako pri celoplošnej kontrole, opäť si stanovíme pravdepodobnosť eventuálneho zániku ako $\gamma = 0,92$. Koeficient kontrolného mechanizmu c vypočítame z pravdepodobnostnej vytvárajúcej funkcie v tvare (3.23). Po dosadení hodnôt $R_0 = 1,468$, $t = 0,92$ a $r = 0,7467985$, riešime nasledovnú rovnicu (pod rovnicou uvádzame aj kód v jazyku $\mathbb{R}$ na výpočet c v tomto prípade)

$$c + (1 - c) \cdot \left(1 + \frac{1,468}{0,7467985} (1 - 0,92) \right)^{-0,7467985} - 0,92 = 0$$

```
eq3 <- function(c) {
  c + (1 - c) * (1 + (R0/est_r) * 0.08) ^ (-est_r) - 0.92
}
c_ind <- uniroot.all(eq3, interval = c(0, 2),
  lower = 0,
  upper = 2,
  tol = 0.00000000001)
```

Vypočítaná hodnota $c = 0,2258155$ a reprodukčné číslo po zavedení individuálnej-špecifickej kontroly vypočítame ako

$$R_c^{ind} = (1 - 0,2258155) \cdot 1,486 = 1,136503$$

Ako sme uviedli v metodickkej časti našej práce, v prípade individuálnej-špecifickej kontroly budeme pracovať s rozdelením počtu potomkov, ktoré má modifikovanú hodnotu pravdepodobnostnej funkcie v nule. Nakoľko odhadujeme negatívne binomické rozdelenie, v tomto prípade budeme pracovať s negatívne binomickým rozdelením modifikovaným v nule. Teda $X_c^{ind} \sim NBi(r; p)$ a $p_{X_c^{ind}}(0)$ vypočítame vzťahom $p_{X_c^{ind}}(0) = p_X(0) + c(1 - p_X(0))$.

Po dosadení hodnôt dostaneme $p_{X_c^{ind}}(0) = 0,5695775$. Počet potomkov opíšeme rozdelením

$$X_c^{ind} \sim NBi(0,7467985; 0,3371858); p_{X_c^{ind}}(0) = 0,5695775;$$

$$R_c^{ind} = E_{ex}(X_c^{ind}) = 1,136503$$

V tabuľke 4.4 uvádzame na ukážku niekoľko hodnôt pravdepodobnostnej funkcie počtu potomkov pred zavedením kontroly $p_X(x)$ a po zavedení individuálnej-špecifickej kontroly $p_{X_c^{ind}}(x)$. Tieto hodnoty sme vypočítali nulovým useknutím a následne nulovou modifikáciou, teda pomocou vzťahu (3.4), pričom $p_{X^M}(0) = p_{X_c^{ind}}(0)$.

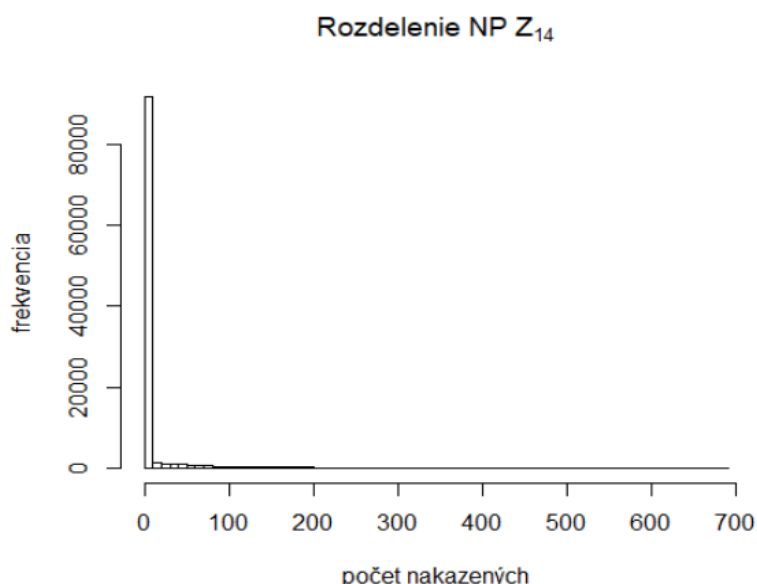
Tabuľka 4.4 Hodnoty pravd. funkcie počtu potomkov, pred a po ind.-špec. kontrole.

x	$p_X(x)$	$p_{X_c^{ind}}(x)$
0	0,44403	0,56957
1	0,21979	0,17016
2	0,12724	0,09851
3	0,07722	0,05978
4	0,04794	0,03711
5	0,03017	0,02336
6	0,01915	0,01483
7	0,01223	0,00947
$\vdots$	$\vdots$	$\vdots$

Balík `actuar` obsahuje funkciu `rzmnbi`, teda generátor pseudonáhodných hodnôt z negatívne binomického rozdelenia modifikovaného v nule, pomocou ktorého budeme generovať počet potomkov. Vo funkcii špecifikujeme nami vypočítané odhady $\hat{r} = 0,7467985$, $\hat{p} = 0,3371858$ a taktiež špecifikujeme pravdepodobnosť v nule $p_{x_c^{ind}}(0) = 0,5695775$.

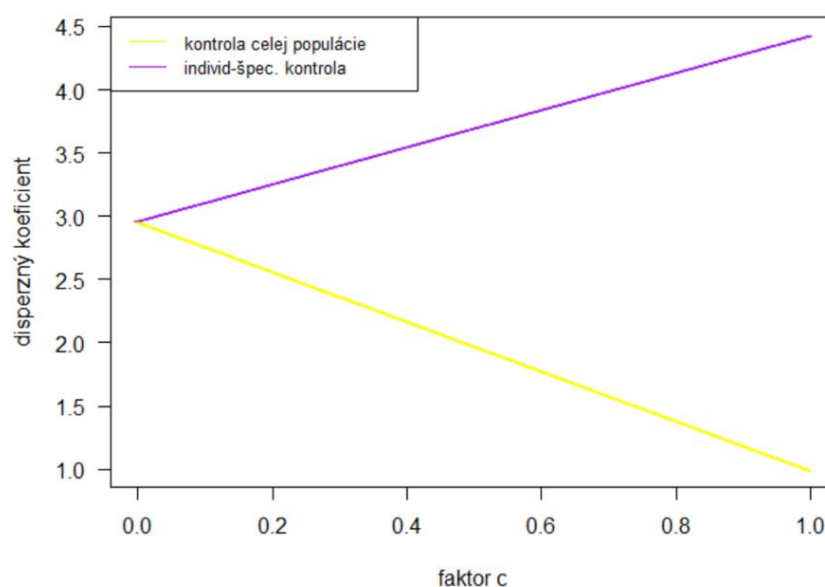
Po zbehnutí kódu, ktorý uvádzame v prílohe, dostaneme výsledok, že epidémia s danými parametrami šírenia v hypotetickej lokalite zanikne s pravdepodobnosťou 0,90324 v štrnástej generácii, ak presadzujeme individuálnu-špecifickú kontrolu.

Obrázok 4.11 Histogram 14-tej generácie nakazených koronavírusom v prípade ind.-špec. kontroly s $c = 0,2258155$.



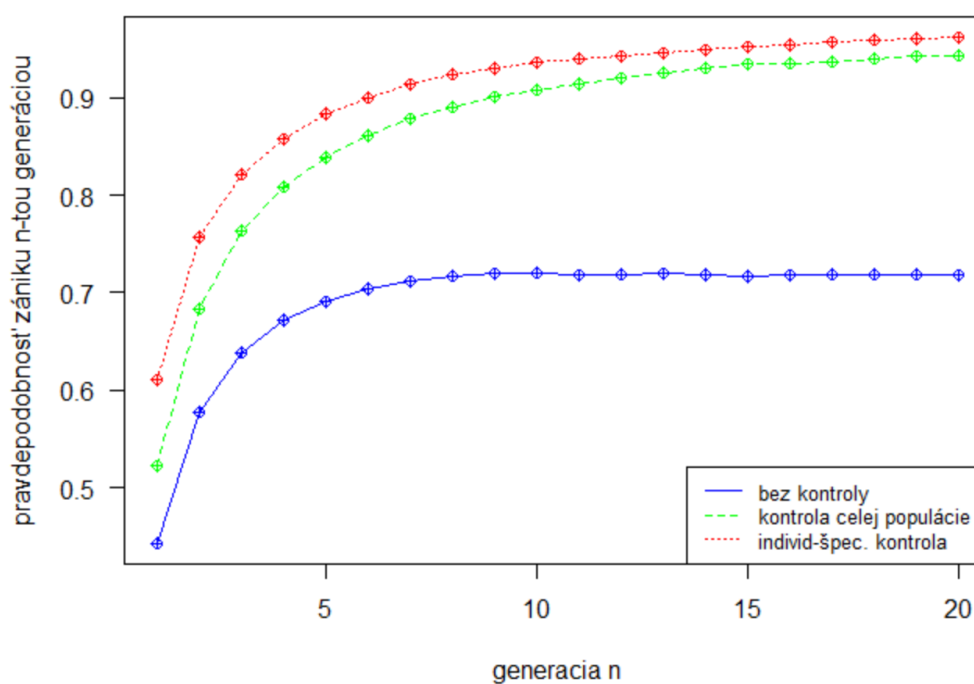
Disperzným koeficientom, teda podielom disperzie a strednej hodnoty porovnáme relatívnu variabilitu jednotlivých modelov počtu potomkov. V nasledujúcom grafe uvádzame túto mieru ako funkciu faktora c pre jednotlivé kontrolné mechanizmy.

Obrázok 4.12 Disperzný koeficient ako funkcia faktora c pri $R_0 = 1,468$, $r = 0,7468$.



Každou nasledujúcou generáciou počtu nakazených pozorujeme vyššiu pravdepodobnosť zániku. Túto pravdepodobnosť môžeme graficky porovnať pre jednotlivé modely počtu potomkov, teda pre prípad žiadnej kontroly, kontroly celej populácie a pre prípad individuálnej-špecifickej kontroly pri hodnote faktora $c = 0,3$.

Obrázok 4.13 Vývoj pravdepodobnosti zániku pre jednotlivé modely ($c = 0,3$).



Z hľadiska nami stanoveného cieľa, ktorým je zánik n -tou generáciou s minimálnou pravdepodobnosťou 0,9, nespĺňa toto kritérium voľne šíriaci sa koronavírus s odhadnutými parametrami šírenia a s pravdepodobnosťou eventuálneho zániku 0,72. Preto je potrebné zaviesť intervencie vo forme kontrolných mechanizmov, ktoré vieme kvantifikovať pomocou faktora c . Pravdepodobnosť eventuálneho zániku sme nadhodnotili o 2% ($\gamma = 0,92$) aby sme mohli faktor c vypočítať využitím pravdepodobnostnej vytvárajúcej funkcie počtu potomkov. V prípade **kontroly celej populácie** nám tento faktor vyšiel približne 25% a následnou simuláciou sme určili, že prvá generácia, ktorá v tomto prípade spĺňa dané kritérium je šestnásta. V prípade **individuálnej-špecifickej kontroly**, pri $\gamma = 0,92$ nám faktor c vyšiel približne 23% a zaručil v simulácii zánik koronavírusu s minimálnou pravdepodobnosťou 0,9 počnúc štrnástou generáciou. Naša analýza potvrdila zistenia autorov publikácie ^[2], teda individuálna-špecifická kontrola je efektívnejšia v porovnaní s kontrolou celej populácie ak považujeme za kritérium efektivity pravdepodobnosť zániku nákazy. Tieto zistenia môžu naznačovať, že menší faktor c znamená menšie úsilie na kontrolu čo sa môže potenciálne pretaviť do menších nákladov na ich realizáciu. Z praxe vieme, že kontrolné opatrenia celej populácie oproti opatreniam na individuálnejšej báze ako napr. trasovanie kontaktov alebo pravidelné testovanie vo veľkých závodoch, prinášajú väčšie ekonomické škody. Preto aj z ekonomického hľadiska, môžeme na základe nášho modelu tvrdiť, že opatrenia viacej zamerané na potenciálnych šíriteľov sú úspornejšie a efektívnejšie v porovnaní s kontrolovaním celej, v našom prípade hypotetickej lokality.

Autori uvádzajú aj tretí typ kontroly, ktorou je **cielená individuálna-špecifická kontrola** (z angl. *targeted individual-specific control*). Táto kontrola odstraňuje nedostatky individuálnej-špecifickej kontroly tak, že náhodne vybranú vzorku potenciálnych šíriteľov c rozdelí na skupiny slabších a silnejších šíriteľov. Slabší šíriteľa sú ohodnotení menšou pravdepodobnosťou a silnejší väčšou pravdepodobnosťou pri náhodnom výbere. Autori pri tejto kontrole uplatnili Paretovo pravidlo 80/20, a polovicu kontrolného úsilia (c) upriamili na 20% najinfekčnejších jedincov a ich výsledky naznačujú zvýšenú efektivitu oproti ostatným kontrolným mechanizmom. Avšak v prípade, že skúmaná nákaza sa šíri homogénne a vhodný model počtu potomkov sa javí ako Poisson (podľa limitnej vety: $NB_i \approx Po$ ak $r \rightarrow \infty$), jednotlivé kontrolné mechanizmy majú podobnú efektivitu. Čo je však v praxi veľmi ojedinelý jav.

Záver

Modelovanie ohnisk infekčných nákaz a z nich vznikajúcich epidémií nie je v dnešnej dobe žiadnou novinkou. Keďže klasické deterministické modely nedokážu zachytiť variabilitu v šírení nákazy jednotlivých osôb, ktorá vyplýva zo stochastického správania nakazených osôb a obzvlášť ťažko identifikovateľných šíriacich predispozícií, možno v takom prípade pristúpiť ku stochastickým modelom. Cieľom našej práce je poukázať na neštandardné využitie diskretných zložených rozdelení v oblasti epidemiológie, ktoré túto variabilitu v šírení vedia zachytiť a spresniť pravdepodobnosť počtu nakazených.

Pomocou negatívne binomického rozdelenia sme opísali šírenie koronavírusu jednotlivca v hypotetickej lokalite a dokázali sme cez pravdepodobnostnú vytvárajúcu funkciu tohto rozdelenia určiť pravdepodobnosť eventuálneho zániku nákazy. Následným skladaním tohto rozdelenia sme získali užitočné informácie, ako napríklad pravdepodobnosť zániku nákazy danou generáciou počtu nakazených. Dosiahnuté výsledky pre prípad jedného indexového prípadu ($Z_0 = 1$), sme porovnali s prípadom, keď Z_0 je tiež náhodnou premennou, opisujúcou počet denných importovaných prípadov a mohli sme pozorovať zvýšenú variabilitu v počte nakazených pre rovnakú generáciu. Všetky výpočty a algoritmus vetviaceho procesu, teda viacnásobne zložené diskretné rozdelenie sme získali pomocou jazyka $\mathbb{R}$, kde jednotlivé iterácie sme realizovali metódou *Monte Carlo simulácie*. Jednotlivé generácie počtu nakazených sme priblížili pravdepodobnostnou, distribučnou funkciou a základnými číselnými charakteristikami strednou hodnotou a disperziou, pričom tieto charakteristiky sme určili simuláciou, ale aj exaktne.

Kontrolné mechanizmy, ktoré sme najprv predstavili teoreticky v metodologickej časti, sme aplikovali na konkrétnych údajov. Potvrdili sme známe zistenia, že *individuálna-špecifická kontrola* je efektívnejšia v porovnaní s *kontrolou celej populácie*. Implementáciou týchto kontrol sme potvrdili, že ak je indikátorom lepšie kontrolovanej nákazy nižšia pravdepodobnosť zániku a menší počet nakazených, tak individuálna-špecifická kontrola dokáže vopred stanovenú minimálnu hranicu tejto pravdepodobnosti zabezpečiť v rámci nižšej generácie počtu nakazených a s menším úsilím, ktoré sme kvantifikovali pomocou faktora c . Domnievame sa, že opatrenia, ktoré sú cielené na celú populáciu (napr. lockdown celej krajiny), sú z hľadiska zabezpečenia skoršieho zániku nákazy, v porovnaní s opatreniami zameranými na osoby s potenciálom šíriť nákazu

(trasovanie kontaktov, testovanie vo veľkých závodoch) menej efektívne a taktiež ekonomicky nákladnejšie.

Bibliografické odkazy

- [1] HE, D. et al. Low dispersion in the infectiousness of COVID-19 cases implies difficulty in control. In *BMC Public Health* . 2020. Vol. 20, no. 1, s. 1558. .
- [2] LLOYD-SMITH, J.O. et al. Superspreading and the effect of individual variation on disease emergence. In *Nature* . 2005. Vol. 438, no. 7066, s. 355–359. .
- [3] LLOYD-SMITH, J.O. et al. Superspreading and the impact of individual variation on disease emergence. In . s. 31. .
- [4] PANJER, H.H. - WILLMOT, G.E. *Insurance risk models*. . Schaumburg, IL: Society of Actuaries, 1992. 442 s. ISBN 978-0-938959-25-0.
- [5] [Online]. . [cit. 2021-04-15]. Dostupné na internete: <<https://www.stat.auckland.ac.nz/~fewster/325/notes/325book.pdf>>.
- [6] [Online]. 2021. [cit. 2021-04-08]. Dostupné na internete: <https://en.wikipedia.org/w/index.php?title=Disease_outbreak&oldid=1010827527>.
- [7] Econometrics researcher's calculations: Why we should social distance. In [online]. [cit. 2021-04-09]. Dostupné na internete: <<https://www.preventionweb.net/go/71160>>.
- [8] Harris T.E. (1963): *The Theory of Branching Processes*, Springer Verlag, Berlin. .
- [9] HORÁKOVÁ, Galina, Michal PÁLEŠ a František SLANINKA. *Teória rizika v poistení*. Bratislava: Wolters Kluwer, 2015, 420 s. [21,07 AH]. *Ekonomía*. ISBN 978-80-8168-273-5. .
- [10] HUŤKA, Vladimír. *Teória pravdepodobnosti 2*. 2. vyd. Bratislava: Vydavateľstvo EKONÓM, 2005, 204 s. ISBN 80-225-2040-3. .
- [11] KOTLEBOVÁ, Eva, Martin LUKÁČIK, Michal PÁLEŠ a Erik ŠOLTÉS. *Aktuárska štatistika*. Bratislava: Letra Edu, 2020, [278 s.] [15,2 AH]. ISBN 978-80-89962-55-6. .
- [12] Matematika koronaviru: Matematické modely šírení epidémie. In [online]. [cit. 2021-04-08]. Dostupné na internete: <<https://www.matfyz.cz/clanky/matematika-koronaviru-matematicke-modely-sireni-epidemie>>.
- [13] PACÁKOVÁ, Viera. Simulation of collective risk model. Management information systems: International scientific journal. Subotica, s. 60-63. In . .
- [14] PÁLEŠ, Michal. *Jazyk R pre aktúarov*. Recenzenti: Gábor Szücs, František Slaninka. 1. vydanie. Bratislava: Vydavateľstvo Letra Edu, 2019. 349 s. [13,28 AH]. VEGA 1/0120/18, VEGA 1/0647/19, KEGA 021EU-4/2019. ISBN 978-80-89962-26-6. .
- [15] PETHÖ, Jakub. 2017. *Praktické využitie charakteristík zložených rozdelení: bakalárska práca*. Bratislava, 37 s. Fakulta hospodárskej informatiky EU. Vedoucí práce Galina Horáková. .

Prílohy

Kompletný zdrojový kód a údaje použité v tejto práci možno dohľadať v jednotlivých súboroch s príponou **.R** a **.csv** v tomto repozitári: <https://github.com/youuto/covid-branching-process>.

Príloha č.1 Program vetviaceho procesu v jazyku $\mathbb{R}$ s negatívne binomickým rozdelením počtu potomkov pre prípad $Z_0 = 1$. Komentáre k nasledujúcim riadkom obsahuje zdrojový kód, ktorý na vyššie uvedenom odkaze možno nájsť v súbore **model.R**.

```
#prípád Z0 = 1
set.seed(1234)
n <- 100000
R <- R0
r <- est_r
Z <- rnbinom(n, mu = R, size = r)
Ztab <- as.data.frame(table(Z))
z <- as.integer(as.vector(Ztab$Z))
pZ <- as.vector(Ztab$Freq/sum(Ztab$Freq))
for (i in 1:3) {
  cat("generating generation", i+1, "\n")
  Z <- replicate
    (n, sum(rnbinom
              (sample(z, 1, prob = pZ,
                     replace = TRUE),
                mu = R, size = r)))
  Ztab <- as.data.frame( table(Z))
  z <- as.integer( as.vector(Ztab$Z))
  pZ <- as.vector( Ztab$Freq / sum(Ztab$Freq))
}
FZ <- numeric(length(pZ))
FZ[1] <- pZ[1]
for (i in 2:length(pZ))
{
  FZ[i] <- FZ[i-1]+pZ[i]
}
Ztab <- cbind(Ztab, pZ, FZ)
```

Príloha č.2 Program vetviaceho procesu v jazyku $\mathbb{R}$ s negatívne binomickým rozdelením počtu potomkov pre prípad keď Z_0 je náhodná premenná opisujúca počet denných importovaných prípadov koronavírusu v hypotetickej lokalite. Komentáre k nasledujúcim riadkom obsahuje zdrojový kód, ktorý na vyššie uvedenom odkaze možno nájsť v súbore **model_import.R**.

```
#pripad Z0 je NP
import <- read.csv("import.csv", header = FALSE)
import <- as.matrix(import)
imp_tab <- as.data.frame(table(import))
days <- as.integer(as.vector(imp_tab$Freq))
set.seed(2345)
n <- 100000
RI <- R0
rI <- est_r
zI <- imp_tab$import
pZI <- days/sum(days)
for (i in 1:4)
{
  cat("generating generation",i,"\n")
  ZI <- replicate
    (n, sum( rnbinom
              ( sample(zI, 1, prob = pZI,
                      replace = TRUE),
                mu = RI,size = rI)))
  ZItab <- as.data.frame(table(ZI))
  zI <- as.integer(as.vector(ZItab$ZI))
  pZI <- as.vector(ZItab$Freq/sum(ZItab$Freq))
}
FZI <- numeric(length(pZI))
FZI[1] <- pZI[1]
for (i in 2:length(pZI))
{
  FZI[i] <- FZI[i-1]+pZI[i]
}
ZItab <- cbind(ZItab,pZI,FZI)
```

Zdrojový kód modelu, kde sme počítali zmeny po kontrole celej populácie nájdeme v súbore **model_popc.R**, resp. zmeny po individuálnej-špecifickej kontrole nájdeme v súbore **model_ind_spec.R** na vyššie uvedenom odkaze, aj s komentármi jednotlivých riadkov (oba modely využívajú rovnaký algoritmus, ktorý sme použili v modeli v prílohe č.1)

Príloha č.3 Vývoj simulovanej pravdepodobnosti zániku $p_{Z_n}(0)$, strednej hodnoty $E_{sim}(Z_n)$ a disperzie $D_{sim}(Z_n)$ prvých dvadsiatich generácií počtu nakazených na koronavírus v hypotetickej lokalite, pre prípad $Z_0 = 1$, resp. Z_0 je náhodná premenná počtu denných importovaných prípadov.

generácia n	$p_{Z_n}(0);$ $Z_0 = 1$	$E_{sim}(Z_n);$ $Z_0 = 1$	$D_{sim}(Z_n);$ $Z_0 = 1$	$p_{Z_n}(0);$ Z_0 je NP	$E_{sim}(Z_n);$ Z_0 je NP	$D_{sim}(Z_n);$ Z_0 je NP
1	0,4428	1,4703	4,3920	0,1960	11,2524	191,9554
2	0,5773	2,1495	16,0940	0,2545	16,6090	464,5771
3	0,6387	3,1237	42,8657	0,2887	24,3209	1074,2478
4	0,6720	4,5918	106,4227	0,3123	35,4723	2401,9405
5	0,6905	6,7777	252,6813	0,3263	52,1394	5362,8701
6	0,7039	9,8885	570,3753	0,3341	76,6384	11819,1990
7	0,7116	14,5002	1272,4010	0,3410	113,1694	26115,7130
8	0,7171	21,2002	2819,9188	0,3468	165,5974	56292,7954
9	0,7197	31,3128	6246,7187	0,3504	243,3049	123884,9495
10	0,7204	45,9196	13272,1627	0,3518	356,9937	266271,6312
11	0,7181	67,9907	28825,1819	0,3511	526,1460	578892,3358
12	0,7189	100,1118	62015,4796	0,3521	776,0469	1259851,2317
13	0,7196	146,5417	133251,6219	0,3503	1139,8853	2716187,0015
14	0,7185	217,1595	292196,7724	0,3495	1682,9440	5895463,0670
15	0,7169	321,3138	642228,2938	0,3478	2483,4218	12757405,5860
16	0,7179	476,2470	1415026,3590	0,3488	3654,8181	27542969,3305
17	0,7189	696,8386	3036405,6493	0,3474	5389,4669	59824835,1067
18	0,7191	1013,2845	6401848,7912	0,3484	7906,4333	129357359,8186
19	0,7189	1478,1799	13488463,2128	0,3493	11578,0529	276423794,7182
20	0,7182	2185,0822	29452256,0533	0,3505	17019,4048	601029988,3146