UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 2. stopnja

Jan Neveda

Negativni binomski model

Magistrsko delo

Mentor: izr. prof. dr. Janez Bernik

Ljubljana, 2018

Kazalo

Program dela v
1. Uvod 1
2. Osnove zavarovalništva 2
   2.1. Kolektivni model 4
3. Osnovne lastnosti porazdelitev 6
   3.1. Poissonova porazdelitev 6
   3.2. Negativna binomska porazdelitev 7
4. Testiranje hipotez 16
   4.1. Kolmogorov - Smirnov test za negativno binomsko porazdelitev 16
      4.1.1. Računanje p-vrednosti 18
      4.1.2. Freyev algoritem 20
      4.1.3. Primeri Kolmogorov - Smirnovega testa 21
   4.2. Testiranje vzorcev v praksi 27
      4.2.1. Metoda momentov 28
      4.2.2. Metoda največjega verjetja 30
      4.2.3. Testiranje kvalitete aproksimacije 33
5. Modeliranje števila škod 37
   5.1. Spreminanje intenzivnosti skozi čas 40
   5.2. Ocenjevanje intenzivnosti na podlagi zgodovinskih podatkov 40
   5.3. Intenzivnost μ kot slučajna spremenljivka 41
   5.4. Ocenjevanje upanja in variance intenzivnosti μ 42
   5.5. Negativni binomski model 45
   5.6. Aproksimiranje negativne binomske porazdelitve 46
   5.7. Primer iz zavarovalništva 48
Literatura 53

Program dela

Predstavite negativno binomsko porazdelitev, njene lastnosti in uporabo v zavarovalništvu.

Osnovna literatura:

E. Bolviken, *Computation and Modelling in Insurance and Finance*, (2014) strani 279–313.

Podpis mentorja:

Negativni binomski model

Povzetek

Na začetku tega dela so opisane glavne značilnosti dveh porazdelitev, ki se najpogosteje uporabljata pri napovedovanju števila škodnih zahtevkov. To sta negativna binomska porazdelitev in Poissonova porazdelitev. Uvodni predstavitvi sledi analiza Kolmogorov-Smirnovega testa, s katerim preverimo, ali je določen vzorec porazdeljen negativno binomsko. Pri tem se uporabi poseben algoritem, ki se imenuje Freyev algoritem. Celoten algoritem je stestiran na določenih podatkih, opremljen pa je tudi s psevdokodo, ki je napisana v programu R. V nadaljevanju sta nato predstavljeni metoda momentov in metoda največjega verjetja, ki ju uporabimo pri ocenjevanju parametrov neznane porazdelitve. Na koncu četrtega poglavja spoznamo še hi-kvadrat test, s katerim preverimo pravilnost ocenjenih parametrov in navsezadnje tudi kvaliteto aproksimacije. V zaključku dela pa je na zavarovalniških podatkih predstavljen postopek napovedovanja pričakovanega števila škod, s pomočjo negativnega binomskega modela.

Negative binomial model

Abstract

At the beginning main features of negative binomial and Poisson distribution are described, which are mainly used for modelling claim frequency. In order to test, if some data is distributed negative binomial, Kolmogorov-Smirnov test is introduced in chapter 4.1. For the calculation of p-value Frey algorithm is used on some data and pseudocode written in program R is presented. Two other methods are then defined in chapter 4.2, which are mainly used in practice to estimate parameters of some distribution. These are method of moment and method of maximum likelihood estimation. In the next chapter, quality of approximation is tested with chi-squared test. At the end, negative binomial model is used to predict claim frequency of some actuarial data.

Math. Subj. Class. (2010): 62E15, 62F03

Ključne besede: zavarovalnica, škodni dogodek, Kolmogorov-Smirnov test, Freyev algoritem, negativna binomska porazdelitev, hi-kvadrat test

Keywords: insurance company, claim, Kolmogorov-Smirnov test, Frey algorithm, negative binomial distribution, chi-squared test

1. Uvod

V današnjem svetu je razvitih že veliko modelov, ki jih lahko uporabimo na različnih področjih. K njihovi uporabnosti zagotovo pripomore tudi bolj razvita tehnologija, brez katere bi bilo modeliranje bistveno težje. Zavarovalnice so vsekakor ena izmed skupin, ki ti dve dejstvi vsakodnevno uporabljajo pri svojem delu. Še posebej se to odraža pri ocenjevanju tveganja, ki ga zavarovalnice sprejemajo pri sklenjanju razmerij s svojimi strankami. Z matematičnega stališča lahko negotovosti lepo predstavimo z raznimi slučajnimi spremenljivkami. Prav tu se mi je porodila ideja za mojo temo magistrskega dela. Pri prebiranju raznih člankov v zvezi z modeliranjem škodnih zahtevkov hitro opazimo, da sta pri neživljenjskem zavarovanju dve glavni porazdelitvi, ki se v večini primerov uporabljata za namen modeliranja števila škod. To sta Poissonova porazdelitev in negativna binomska porazdelitev. Jaz sem se odločil, da bom v svojem delu predstavil načine, s katerimi lahko s pomočjo negativne binomske porazdelitve poiščemo bodoče frekvence pojavljanja škod. Vprašanje, ki se nam pojavi je, kako lahko zavarovalnice pri svojem delovanju delujejo z dobičkom, kljub temu da morajo svojim strankam omogočati denarna sredstva v primeru škodnega dogodka. Odgovor na to vprašanje in še nekatere druge osnovne značilnosti zavarovalnic bom predstavil v začetnem poglavju tega dela, saj le-to predstavlja osnovo za dobro razumevanje. Po kratkem uvodu v zavarovalništvo bo sledila predstavitev Poissonove in negativne binomske porazdelitve. V tem delu bom tudi definiral osnovne lastnosti, ki jih bom potem v nadaljevanju pri raznih modelih uporabljal. Še preden se bom osredotočil na ocenjevanje števila škodnih zahtevkov, pa bom v 4. poglavju predstavil tudi nekatere statistične teste, s katerimi preverimo, ali je naključen vzorec porazdeljen negativno binomsko. Vse to moramo opraviti z namenom, da ima potem uporaba modela smisel. Namreč, če bi uprabljali negativni binomski model na določenih podatkih, ki prav nič ne ustrezajo negativni binomski porazdelitvi, bi lahko dobili povsem napačne napovedi. Tega si pa seveda ne želimo. Te statistične teste bom uporabil tudi na določenih primerih in preveril njihovo učinkovitost. Na koncu dela pa se bom v 5. poglavju bolj posvetil modeliranju in napovedovanju pričakovanega števila škodnih zahtevkov s pomočjo negativne binomske porazdelitve. V ta namen bom uporabil tudi določen primer iz zavarovalništva, ki bo še najbolj nazorno prikazal, kako se ocenjevanja lotimo v praksi.

2. Osnove zavarovalništva

Zavarovalnice se pri svojem delovanju vsakodnevno srečujejo z določenim tveganjem, ki ga prevzamejo s sklenitvijo zavarovanja z določeno stranko. Zaradi tega je obvladovanje tveganja in ocenjevanje le-tega, ena izmed njenih poglavitnih nalog. Skozi zgodovino se je oblikovalo kar nekaj modelov, ki se jih poslužujejo zavarovalnice pri tem opravilu. Za delovanje sistema zavarovalništva sta pomembni dve dejstvi. Zavarovancem sklenitev zavarovanja predstavlja varnost, saj se jim v primeru nesreče povrne škoda, na drugi strani pa zavarovalnici to pomeni dobiček. Ob sklenjanju zavarovanj namreč zavarovalnice prejemajo premije, ki jih določijo z oceno pričakovanih škod. Pri sklenitvi zavarovanja se podpiše tudi zavarovalna pogodba. S to pogodbo se zavarovalec zavezuje, da bo zavarovalnici plačal zavarovalno premijo, zavarovalnica pa se zavezuje, da bo, če se zgodi dogodek, ki pomeni zavarovalni primer, izplačala zavarovancu ali nekomu tretjemu zavarovalnino ali odškodnino ali storila kaj drugega. Opazimo lahko, da se v zavarovalni pogodbi uporablja tudi izraz zavarovalec, ki ga pogosto enačimo z izrazom zavarovanec. Gre namreč za fizično ali pravno osebo, ki sklene zavarovalno pogodbo z zavarovalnico. Zavarovalnica kot stranka v zavarovalni pogodbi, ki prevzema tveganje, pa se imenuje zavarovatelj. Zelo pogost izraz, ki ga bom uporabljal tudi v nadaljevanju, je zavarovalna polica, ki predstavlja listino o zavarovalni pogodbi. V polici morajo biti navedeni naslednji podatki:

- pogodbeni stranki,
- zavarovana stvar oziroma zavarovana oseba,
- nevarnosti (riziki), ki jih krije zavarovanje,
- trajanje zavarovanja in doba kritja,
- zavarovalna vsota, ki je lahko tudi neomejena,
- zavarovalna premija ali prispevek,
- datum izdaje police,
- podpisa pogodbenih strank.

Zavarovalništvo navadno razdelimo v dve skupini. Eno predstavljajo življenjska zavarovanja, drugo pa neživljenjska zavarovanja. Jaz se bom v svojem delu bolj osredotočil na neživljenjska zavarovanja, kamor spadajo npr. avtomobilske nesreče, požari, ipd. Ker škode ne moremo vnaprej natančno določiti, jo bomo označevali s slučajno spremenljivko X. Dogodek, da je $X = 0$ torej pomeni, da se ni zgodil noben škodni zahtevek. Pri opredeljevanju zavarovanja je pogost izraz, ki ga srečamo, tudi zavarovalna vsota. Le ta predstavlja zgornjo mejo, do katere je zavarovano določeno premoženje. Z drugimi besedami je to maksimalna količina, ki jo zavarovalnica izplača zavarovancu. Z F bomo označili franšizo, ki predstavlja dogovorjen znesek, do katerega zavarovalnica škode ne plača. Najbolj pogosto se franšiza določi v absolutnem znesku ali pa kot odstotek od zavarovalne vsote. Če škoda franšizo preseže, se pojavita naslednji dve možnosti:

- če zavarovalnica povrne celotno škodo, govorimo o navadni oziroma integralni franšizi - F_i,
- če zavarovalnica plača razliko med celotno škodo in franšizo govorimo o odbitni franšizi - F_o.

S pojmom celotna škoda imamo v mislih zavarovalnino, ki bi jo zavarovalnica plačala tudi v primeru, če ne bi bilo franšize. Bolj nazorno lahko ta dva primera prikažemo na naslednji način. Označimo z Z zavarovalno vsoto, z F_i in F_o, tako kot zgoraj, integralno in odbitno franšizo ter naj Y označuje zavarovalnino. Potem v prvem primeru, kjer govorimo o integralni franšizi velja:

$$Y = \begin{cases} 0, & \text{če } X \leq F_i \\ X, & \text{če } F_i < X < Z \\ Z, & \text{če } X \geq Z \end{cases}. \tag{2.1}$$

Za drugi primer pa velja naslednje:

$$Y = \begin{cases} 0, & \text{če } X \leq F_o \\ X - F_o, & \text{če } F_o < X < Z \\ Z - F_o, & \text{če } X \geq Z \end{cases}. \tag{2.2}$$

To sta dva najbolj osnovna tipa zavarovalnine. Kot sem že zgoraj omenil, mora zavarovanec za to, da se mu omogoči izplačilo zavarovalnine, plačati premijo. Premija na nek način predstavlja mero tveganja, pri kateri je zavarovalnica pripravljena sprejeti tveganje. Določena mora biti tako, da je večja od pričakovanega števila škod in stroškov, saj bi v nasprotnem primeru zavarovalnica na dolgi rok propadla. Prav tako pa premija ne sme presegati največje mogoče škode, kajti nihče se ne bi hotel zavarovati, če bi premija presegala največje mogoče izplačilo v primeru škode. Pričakovana vrednost škod je osnova za določanje premije. Zato definiramo

$$\pi^{pu} = E(X),$$

kar imenujemo čista premija. Če bi zavarovalnice prejemale tako premijo, bi v odsotnosti vseh ostalih stroškov na dolgi rok nekaj dobile, lahko bi pa tudi izgubile. Izkaže se še več. Na dolgi rok bi z verjetnostjo 1 celo propadle. To je posledica zakona v teoriji verjetnosti, ki govori o oscilacijah martingalov. Seveda pa interes zavarovalnic ni poslovati z izgubo, zato k tej premiji dodajo še dodatek, ki ga označimo z γ. Premija se v tem primeru zapiše kot

$$\pi = (1 + \gamma)\pi^{pu}.$$

Na dodatek $\gamma\pi^{pu}$ lahko gledamo kot strošek tveganja. Na njegovo višino vpliva trenutno stanje na trgu.

S tem smo definirali ključne pojme, ki se uporabljajo v zavarovalništvu. Bolj podrobno se v izračune premij in ostalih stvari v mojem delu ne bom spuščal. Večjo pozornost bom namenil raznim modelom, ki se uporabljajo pri ugotavljanju porazdelitev števila škodnih zahtevkov. S tem se zavarovalnica lahko pripravi na bodoča izplačila škod. Najbolj uporabljena porazdelitev, ki se pri ugotavljanju tega uporablja, je Poissonova porazdelitev. Poleg te, pa bom jaz nekoliko več časa posvetil še eni pogosto uporabljeni porazdelitvi, ki se imenuje negativna binomska porazdelitev.

Za začetek pa si poglejmo zgolj osnoven model, v katerem bomo samo definirali ključne pojme in si pogledali kako pravzaprav pričnemo z ocenjevanjem škod. Ta model se imenuje kolektivni model.

2.1. **Kolektivni model.**

Pri kolektivnem modelu modeliramo število škodnih zahtevkov v določenem obdobju. Število teh zahtevkov predstavlja slučajna spremenljivka N. Vrednost posameznega škodnega zahtevka označuje slučajna spremenljivka X. Za škodne zahtevke predpostavimo naslednje dve lastnosti:

- škodni zahtevki so med seboj neodvisni in enako porazdeljeni (imajo porazdelitveno funkcijo slučajne spremenljivke X),
- škodni zahtevki so neodvisni od števila škod N.

Sedaj lahko zapišemo kumulativno škodo S, ki je enaka:

$$S = \sum_{i=1}^{N} X_i.$$

S pomočjo defincije matematičnega upanja lahko tako izrazimo pričakovano škodo $E(S)$:

$$\begin{aligned} E(S) &= E\Big(E(S|N)\Big) = E\Big(E(\sum_{i=1}^{N} X_i|N)\Big) \\ &= \sum_{k=0}^{\infty} E(\sum_{i=1}^{N} X_i|N=k)\cdot P(N=k) \\ &= \sum_{k=0}^{\infty} E(\sum_{i=1}^{k} X_i|N=k)\cdot P(N=k). \end{aligned} \tag{2.3}$$

Pri tem smo v prvem koraku uporabili formulo za pogojno matematično upanje, drugo vrstico pa smo dobili po definiciji za popolno matematično upanje. Uporabimo še lastnost, da so dogodki X_i enako porazdeljeni in neodvisni od N, zato lahko v zgornjem izrazu za pogojno upanje izpustimo dogodek $N = k$ in dobimo:

$$\begin{aligned} E(S) &= \sum_{k=0}^{\infty} E(\sum_{i=1}^{k} X_i)\cdot P(N=k) = \sum_{k=0}^{\infty} k\cdot E(X)\cdot P(N=k) \\ &= E(X)\cdot E(N). \end{aligned} \tag{2.4}$$

Dobili smo, da je pričakovana škoda enaka produktu pričakovane vredosti števila škodnih zahtevkov in pričakovane vrednosti višine škodnih zahtevkov. To dobljeno formulo imenujemo tudi prva Waldova identiteta. Iz tega lahko sedaj dokaj hitro izrazimo tudi $Var(S)$. Izračunati moramo zgolj še $E(S^2)$, ki ga dobimo na podoben način, kot smo to storili pri $E(S)$:

(2.5)

$$\begin{aligned} E(S^2) &= E\Big(E(S^2|N)\Big) = \sum_{k=0}^{\infty} E(S^2|N=k)\cdot P(N=k) \\ &= \sum_{k=0}^{\infty} E(\sum_{i=1}^{k} X_i)^2\cdot P(N=k) = \sum_{k=0}^{\infty} E\Big(\sum_{i=1}^{k} X_i^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{k} X_i\cdot X_j\Big)\cdot P(N=k) \\ &= \sum_{k=0}^{\infty} k\cdot E(\sum_{i=1}^{k} X^2)\cdot P(N=k) + \sum_{k=0}^{\infty}\sum_{\substack{i,j=1\\ i\neq j}}^{k} E(X_i\cdot X_j)\cdot P(N=k). \end{aligned}$$

V naslednjem koraku upoštevamo, da so slučajne spremenljivke X_i neodvisne in enako porazdeljene, zato velja $E(X_i \cdot X_j) = E(X_i) \cdot E(X_j) = E(X)^2$, za $i, j = 1, ..., k$, $i \neq j$. Tako dobimo

$$\begin{aligned} E(S^2) &= E(X^2) \cdot E(N) + \sum_{k=0}^{\infty} E(X)^2 \cdot (k^2 - k) \cdot P(N = k) \\ &= E(X^2) \cdot E(N) + E(X)^2 \cdot \Big(E(N^2) - E(N) \Big). \end{aligned} \tag{2.6}$$

Varianco izračunamo po definiciji in dobimo:

$$\begin{aligned} Var(S) &= E(X^2) \cdot E(N) + E(X)^2 \cdot \Big(E(N^2) - E(N) \Big) - E(X)^2 \cdot E(N)^2 \\ &= E(N) \cdot \Big(E(X^2) - E(X)^2 \Big) + E(X)^2 \cdot \Big(E(N^2) - E(N)^2 \Big) \\ &= E(N) \cdot Var(X) + E(X)^2 \cdot Var(N), \end{aligned} \tag{2.7}$$

kar je druga Waldova identiteta. Ker se bom jaz v svojem delu bolj posvetil modeliranju števila škodnih zahtevkov, nas ne bo toliko zanimala pričakovana škoda, temveč zgolj obnašanje slučajne spremenljivke N. Še preden pa se povsem posvetimo temu delu, moramo najprej definirati osnovne lastnosti Poissonove in negativne binomske porazdelitve, s pomočjo katerih bom v svojem delu predstavil postopke, ki jih uporabimo za modeliranje števila škodnih zahtekov.

3.1. **Poissonova porazdelitev.**

Poissonova porazdelitev je zelo uporabna porazdelitev, ki se uporablja predvsem pri procesih štetja. Je namreč diskretna porazdelitev, pri kateri opazujemo verjetnost, da se pojavi določeno število dogodkov v določenem časovnem intervalu. Če označimo z λ pričakovano število dogodkov v časovnem intervalu, potem je verjetnost, da se bo zgodilo točno k dogodkov enaka

$$p_k = \frac{\lambda^k}{k!}e^{-\lambda} \text{ za } k = 0, 1, 2, ...$$

kjer p_k imenujemo funkcija verjetnosti. Tako kot za vse diskretne slučajne spremenljivke, tudi za Poissonovo velja, da mora zadoščati naslednjemu pogoju:

$$\sum_{k=0}^{\infty} p_k = 1, \tag{3.1}$$

za vsak parameter Poissonove porazdelitve. To enakost lahko hitro preverimo tako, da zapišemo eksponentno funkcijo v vsoto. Po definiciji dobimo:

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

Sedaj delimo obe strani enačbe z e^{λ} in zapišemo:

$$1 = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=0}^{\infty} p_k$$

Naj bo sedaj slučajna spremenljivka X porazdeljena po Poissonu s parametrom λ. Izračunajmo njeno pričakovano vrednost. Po osnovni formuli za matematično upanje velja

$$E(X) = \sum_{k=0}^{\infty} kp_k = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=0}^{\infty} k\frac{\lambda\lambda^{k-1}}{k(k-1)!}e^{-\lambda} = \lambda\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda}.$$

Vpeljimo sedaj novo spremenljivko $l = k - 1$ in z uporabo zgornje enakosti, da je $\sum_{l=0}^{\infty} p_l = 1$ dobimo

$$E(X) = \lambda\sum_{l=0}^{\infty} \frac{\lambda^l}{l!}e^{-\lambda} = \lambda \cdot 1 = \lambda. \tag{3.2}$$

Tako smo dobili, da je pričakovana vrednost Poissonove slučajne spremenljivke kar enaka njenemu parametru λ. Poiščimo sedaj še njeno varianco. Za to potrebujemo

najprej izračunati vrednost $E(X^2)$

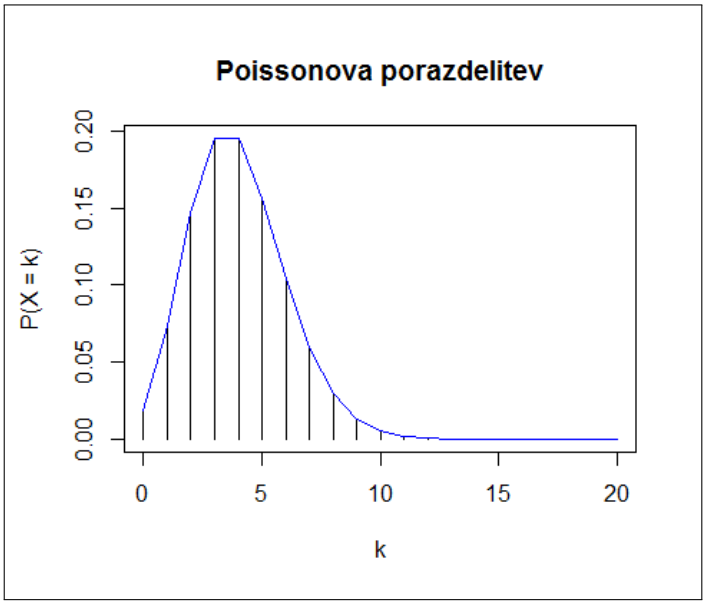
$$
\begin{aligned}
E(X^2) &= \sum_{k=0}^{\infty} k^2 p_k = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\
&= \lambda \Big(\sum_{k=1}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \Big) \\
&= \lambda \Big(\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \Big) \\
&= \lambda \Big(\lambda \sum_{l=0}^{\infty} \frac{\lambda^{l}}{l!} e^{-\lambda} + \sum_{l=0}^{\infty} \frac{\lambda^{l}}{l!} e^{-\lambda} \Big) = \lambda(\lambda + 1) \\
&= \lambda^2 + \lambda.
\end{aligned} \tag{3.3}
$$

Z uporabo osnovne formule za varianco dobimo

$$
Var(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \tag{3.4}
$$

Opazimo lahko, da je varianca enaka pričakovani vrednosti. Tukaj velja omeniti še eno pomembno lastnost. Za Poissonovo porazdelitev je namreč značilno, da je primer limitne porazdelitve iz zakona redkih dogodkov. V 5. poglavju bom tudi prikazal primer, ko pridemo do Poissonove porazdelitve s pomočjo limite binomske porazdelitve.

To so najbolj pomembne lastnosti Poissonove porazdelitve, ki jih bomo v nadaljevanju tudi uporabljali. Na spodnji sliki je prikazan primer Poissonove verjetnostne funkcije:



Slika 1. Primer Poissonove porazdelitve s parametrom $\lambda = 4$

3.2. **Negativna binomska porazdelitev.**

Prav zaradi zadnje omenjene lastnosti pri Poissonovi porazdelitvi, ima ta določene omejitve v praksi. Bolj redko namreč dobimo naključni vzorec, ki ima enako upanje

in varianco. Zato moramo vpeljati novo porazdelitev, ki dopušča tudi možnosti, da sta upanje in varianca različna. Za najboljšo rešitev se največkrat izkaže prav negativna binomska porazdelitev. Gre za diskretno porazdelitev in jo lahko definiramo na več različnih načinov, zato moramo biti še posebej pozorni, katero parametrizacijo uporabljamo. Najbolj osnovni dve parametrizaciji sta naslednji. Imejmo zaporedje neodvisnih Bernoullijevih poskusov s parametrom p. V prvem primeru označimo z X slučajno spremenljivko, ki predstavlja število poskusov, potrebnih za to, da se zgodi M-ti uspeh. Funkcija verjetnosti take slučajne spremenljivke se potem glasi:

$$p_k = P(X = k) = \binom{k-1}{M-1} p^M (1-p)^{k-M}, \quad k = M, M+1, ... \tag{3.5}$$

kjer je M neko naravno število in s p označimo verjetnost uspeha. Rečemo, da je X porazdeljena negativno binomsko (M, p). Porazdelitev definirano na tak način včasih imenujemo tudi Pascalova porazdelitev. Obstaja pa še ena oblika, ki se skoraj najbolj pogosto uporablja. V tem primeru označimo z Y slučajno spremenljivko, ki predstavlja število neuspelih poskusov pred M-tim uspehom. Ta dogodek je statistično ekvivalenten zgornji parametrizaciji, saj lahko zapišemo $Y = X - M$. Po tej definiciji ima negativna binomska porazdelitev naslednjo obliko:

$$p_j = P(Y = j) = \binom{M+j-1}{j} p^M (1-p)^j, \quad j = 0, 1, ... \tag{3.6}$$

Lahko vidimo, da je Pascalova porazdelitev pravzaprav zgolj poseben primer te druge, bolj stanadardne parametrizacije negativne binomske porazdelitve. Seveda mora v obeh primerih veljati, da je

$$\sum_{k=0}^{\infty} p_k = 1. \tag{3.7}$$

Vzemimo pri parametrizaciji (3.5) za M vrednost 1. Poglejmo si, kaj dobimo:

$$p_k = p(1-p)^{k-1}. \tag{3.8}$$

Opazimo, da pri tej izbiri M-ja dobimo znano porazdelitev, ki se imenuje geometrijska porazdelitev.

Jaz se bom v svojem delu bolj osredotočil na parametrizacijo (3.6). Značilnost te parametrizacije je tudi ta, da do nje pridemo s pomočjo mešanja Poissonove in Gamma porazdelitve. Sedaj si kar poglejmo postopek, kako to naredimo. Definirajmo slučajno spremenljivko Λ, ki ima Gamma porazdelitev s parametroma M in c, kjer velja $M > 0$ in $c > 0$:

$$\Lambda \sim Gamma(M, c). \tag{3.9}$$

Gostota Gamma porazdelitve je definirana na naslednji način:

$$f_\Lambda(x) = \frac{c^M}{\Gamma(M)} \cdot x^{M-1} \cdot e^{-cx}. \tag{3.10}$$

Slučajna spremenljivka Y pogojno na Λ, pa naj ima Poissonovo porazdelitev s parametrom Λ:

$$Y|\Lambda \sim Pois(\Lambda). \tag{3.11}$$

Porazdelitev slučajne spremenljivke Y izračunamo na naslednji način:

$$\begin{aligned} P(Y=k) &= E(\mathbf{1}_{Y=k}) = E\big(E(\mathbf{1}_{Y=k}|\Lambda)\big) = E\big(P(Y=k|\Lambda)\big) \\ &= E\big(e^{-\Lambda}\cdot\frac{\Lambda^k}{k!}\big) \\ &= \int_0^\infty \frac{x^k}{k!}e^{-x}\cdot\frac{c^M}{\Gamma(M)}x^{M-1}e^{-cx}dx \\ &= \frac{c^M}{k!\cdot\Gamma(M)}\int_0^\infty x^{k+M-1}e^{-x(c+1)}dx. \end{aligned} \tag{3.12}$$

Uvedemo novo spremenljivko $z = x(c+1)$, kar pomeni $dz = (1+c)dx$. Dobimo:

$$P(Y=k) = \frac{c^M}{k!\cdot\Gamma(M)(1+c)^{k+M}}\int_0^\infty z^{k+M-1}e^{-z}dz. \tag{3.13}$$

Dobljeni integral je po definiciji enak $\Gamma(k+M)$. Sledi

$$\begin{aligned} P(Y=k) &= \frac{\Gamma(k+M)}{k!\Gamma(M)}\cdot\frac{c^M}{(1+c)^{k+M}} = \frac{\Gamma(k+M)}{k!\Gamma(M)}\cdot\Big(\frac{c}{1+c}\Big)^M\cdot\Big(\frac{1}{1+c}\Big)^k \\ &= \frac{(k+M-1)!}{k!(M-1)!}p^M(1-p)^k \\ &= \binom{M+k-1}{k}p^M(1-p)^k, \end{aligned} \tag{3.14}$$

kjer smo označili $p = c/(1+c)$. Dobili smo, da ima Y negativno binomsko porazdelitev s porazdelitveno funkcijo, ki smo jo definirali v (3.6).

Definicija 3.1. Slučajno spremenljivko X, ki je porazdeljena negativno binomsko s parametroma M in p ter ima porazdelitveno funkcijo enako enačbi 3.14, zapišemo na naslednji način:

$$X \sim NBin(M,p). \tag{3.15}$$

S pomočjo mešanja Poissonove in Gamma porazdelitve lahko sedaj tudi dokaj hitro izračunamo momentno rodovno funkcijo negativne binomske porazdelitve. Pri tem bomo potrebovali tudi momentno rodovni funkciji Poissonove in Gamma porazdelitve. Najprej izračunajmo za primer Poissonove porazdelitve. Označimo slučajno spremenljivko z X. Po definiciji velja:

$$E(e^{Xt}) = \sum_{i=0}^\infty e^{it}\cdot\frac{\lambda^i}{i!}e^{-\lambda} = e^{-\lambda}\sum_{i=0}^\infty\frac{(\lambda e^t)^i}{i!}. \tag{3.16}$$

Dobljena vsota je znana, saj je tipična za eksponentno funkcijo $e^x = \sum_{i=0}^\infty \frac{x^i}{i!}$. Tako smo dobili:

$$E(e^{Xt}) = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}. \tag{3.17}$$

Momentno rodovno funkcijo Gamma porazdelitve dobimo z izračunom naslednjega integrala:

$$\begin{aligned} M_\Lambda(t) = E(e^{\Lambda t}) &= \int_0^\infty e^{\lambda t}\cdot\frac{c^M}{\Gamma(M)}\cdot\lambda^{M-1}\cdot e^{-c\lambda}d\lambda \\ &= \frac{c^M}{\Gamma(M)}\int_0^\infty e^{-\lambda(c-t)}\cdot\lambda^{M-1}d\lambda. \end{aligned} \tag{3.18}$$

Uvedemo novo spremenljivko $u = \lambda(c-t)$. Iz tega dobimo $du = (c-t)d\lambda$ oziroma $d\lambda = du/(c-t)$. Uporabimo to v zgornjem izrazu in dobimo:

$$\begin{aligned} M_\Lambda(t) &= \frac{c^M}{\Gamma(M)} \int_0^\infty e^{-u} \cdot (\frac{u}{c-t})^{M-1} \cdot \frac{1}{c-t} du \\ &= \frac{c^M}{\Gamma(M)(c-t)^M} \int_0^\infty e^{-u} \cdot u^{M-1} du = \frac{c^M}{\Gamma(M)(c-t)^M} \cdot \Gamma(M) \\ &= \big(\frac{c}{c-t}\big)^M. \end{aligned} \tag{3.19}$$

S pomočjo dobljenih dveh momentno rodovnih funkcij Poissonove in Gamma porazdelitve, lahko sedaj izračunamo tudi momentno rodovno funkcjio negativne binomske porazdelitve. Dobimo jo na podoben način, kot smo prišli do rezultata v (3.12). Torej uporabimo lastnost, da je $Y|\Lambda$ porazdeljena po Poissonu in dobimo:

$$\begin{aligned} M_Y(t) = E(e^{Yt}) &= E\Big(E(e^{Yt}|\Lambda)\Big) = E\Big(E(e^{Poisson(\Lambda)\cdot t})\Big) \\ &= E\Big(e^{\Lambda(e^t-1)}\Big). \end{aligned} \tag{3.20}$$

Dobili smo, da je momentno rodovna funkcija negativne binomske porazdelitve enaka momentno rodovni funkciji Gamma porazdelitve v točki $e^t - 1$. Tako lahko zapišemo:

$$\begin{aligned} M_Y(t) = M_\Lambda(e^t-1) &= \Big(\frac{c}{c-e^t+1}\Big)^M = \Big(\frac{\frac{c}{c+1}}{\frac{c}{c+1} - \frac{e^t}{c+1} + \frac{1}{c+1}}\Big)^M \\ &= \Big(\frac{p}{p-(1-p)e^t+1-p}\Big)^M = \Big(\frac{p}{1-(1-p)e^t}\Big)^M. \end{aligned} \tag{3.21}$$

Sedaj si poglejmo, kako izračunamo pričakovano vrednost in varianco negativne binomske porazdelitve, saj bomo le to v nadaljevanju še potrebovali. Označimo z X slučajno spremenljivko, ki ima verjetnostno funkcijo enako kot v (3.6). Potem po definicji velja

$$\begin{aligned} E(X) &= \sum_{k=0}^\infty k p_k = \sum_{k=0}^\infty k \binom{M+k-1}{k} p^M (1-p)^k \\ &= 0\binom{M+0-1}{0} p^M(1-p)^0 + \sum_{k=1}^\infty k \frac{(M+k-1)!}{k!(M-1)!} p^M(1-p)^k \\ &= 0 + \sum_{k=1}^\infty \frac{(M+k-1)!}{(k-1)!(M-1)!} p^M(1-p)^k \\ &= \frac{M(1-p)}{p} \sum_{k=1}^\infty \frac{(M+k-1)!}{(k-1)!M!} p^{M+1}(1-p)^{k-1}. \end{aligned} \tag{3.22}$$

Uvedemo novi spremenljivki $i = k-1$ ter $j = M+1$ in dobimo

$$\begin{aligned} E(X) &= \frac{M(1-p)}{p} \sum_{i=0}^\infty \frac{(j+i-1)!}{i!(j-1)!} p^j (1-p)^i \\ &= \frac{M(1-p)}{p} \sum_{i=0}^\infty \binom{j+i-1}{i} p^j (1-p)^i = \frac{M(1-p)}{p}. \end{aligned} \tag{3.23}$$

Tako smo prišli do matematičnega upanja negativne binomske porazdelitve. Ostal nam je še izračun variance. Ponovno si najprej poglejmo, kaj dobimo, če izračunamo $E(X^2)$:

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{\infty} k^2 p_k = \sum_{k=0}^{\infty} k^2 \binom{M+k-1}{k} p^M (1-p)^k \\ &= 0^2 \binom{M+0-1}{0} p^M (1-p)^0 + \sum_{k=1}^{\infty} k^2 \frac{(M+k-1)!}{k!(M-1)!} p^M (1-p)^k \\ &= \frac{M(1-p)}{p} \sum_{k=1}^{\infty} k \frac{(M+k-1)!}{(k-1)!M!} p^{M+1} (1-p)^{k-1}. \end{aligned} \tag{3.24}$$

Definirajmo novi spremenljivki $i = k-1$ ter $j = M+1$:

$$\begin{aligned} E(X^2) &= \frac{M(1-p)}{p} \sum_{i=0}^{\infty} (i+1) \frac{(j+i-1)!}{i!(j-1)!} p^j (1-p)^i \\ &= \frac{M(1-p)}{p} \sum_{i=0}^{\infty} (i+1) \binom{j+i-1}{i} p^j (1-p)^i \\ &= \frac{M(1-p)}{p} \Big(\sum_{i=0}^{\infty} i \binom{j+i-1}{i} p^j (1-p)^i + \sum_{i=0}^{\infty} \binom{j+i-1}{i} p^j (1-p)^i \Big) \\ &= \frac{M(1-p)}{p} \Big(E(X) + 1 \Big) = \frac{M(1-p)}{p} \Big(\frac{j(1-p)}{p} + 1 \Big) \\ &= \frac{M(1-p)}{p} \Big(\frac{(M+1)(1-p)}{p} + 1 \Big) \\ &= \frac{M(1-p)}{p} \Big(\frac{M(1-p)+1}{p} \Big) = \frac{M(1-p)\big(M(1-p)+1\big)}{p^2}. \end{aligned} \tag{3.25}$$

Iz tega lahko sedaj izračunamo varianco:

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 = \frac{M(1-p)\big(M(1-p)+1\big)}{p^2} - \frac{(M(1-p))^2}{p^2} \\ &= \frac{M(1-p)}{p^2}. \end{aligned} \tag{3.26}$$

Za negativno binomsko porazdelitev vidimo, da velja $E(X) < Var(X)$. To sledi iz dejstva, da je parameter $0 < p \leq 1$, kar pomeni, da je $p^2 \leq p$. Izraza za upanje in varianco imata v števcu isti vrednosti, v imenovalcu pa se razlikujeta ravno v vrednosti p^2. Ker torej pri enačbi za varianco števec delimo z manjšo vrednostjo, to pomeni, da je na koncu rezultat večji kot pri upanju. S tem smo res dobili bolj uporabno porazdelitev za ocenjevanje škodnih dogodkov v praksi. Na tem mestu velja omeniti, da bi lahko matematično upanje in varianco izračunali tudi s pomočjo momentno rodovne funkcije. Velja namreč, da je prvi odvod momentno rodovne funkcije enak matematičnemu upanju, drugi odvod pa varianci. S tem bi prišli do istega rezultata, kot če računamo matematično upanje in varianco po definiciji.

Obstaja pa še tretji tip parametrizacije, ki ga velja omeniti, saj je tudi lahko zelo uporaben. Pravzaprav gre za poseben primer, ki ga je možno uporabiti za

napovedovanje števila škod, ki se prenesejo na pozavarovalnico. Pri izpeljavi te oblike, si bomo pomagali z že prej definiranim kolektivnim modelom. Predstavljajmo si, da imamo XL pogodbo s pozavarovalnico. Z XL pogodbo mislimo na škodno presežkovno pozavarovanje, oznaka pa izhaja iz angleškega izraza excess of loss. Označimo ponovno z X slučajno spremenljivko, ki predstavlja originalni škodni zahtevek. Del, ki ga krije pozavarovalnica je potem enak

$$\max\{X - d, 0\} = (X - d)_+,$$

kjer vrednost d predstavlja odbitno franšizo. Če to sedaj posplošimo v obliko kolektivnega modela, dobimo, da v primeru N - tih škodnih zahtevkov pozavarovalnica plača naslednjo vsoto:

$$S^1 = \sum_{i=1}^{N} (X_i - d)_+.$$

Opazimo lahko, da pozavarovalnica krije škodo zgolj v primeru, ko je $X_i > d$ za $i = 1, ..., N$. Označimo s slučajno spremenljivko Z naslednjo indikatorsko funkcijo:

$$Z_i = \mathbf{1}_{\{X_i > d\}} \qquad i = 1, ..., N.$$

Predpostavimo, da so slučajne spremenljivke Z_i neodvisne in porazdeljene Bernoullijevo s parametrom π. Torej π označuje verjetnost dogodka, da je $X_i > d$ - $\pi = P(X_i > d)$. Velja pa naj še, da je število škod N porazdeljeno negativno binomsko s parametroma M in p. Z novo definiranimi dogodki lahko zapišemo:

$$N^1 = \sum_{i=1}^{N} \mathbf{1}_{\{X_i > d\}} = \sum_{i=1}^{N} Z_i. \tag{3.27}$$

Vidimo, da v vsoti sedaj ne gledamo več višine nastalih škod, temveč štejemo število dogodkov, pri katerih so višine škod večje od d. Vsota N^1 torej predstavlja število škod, ki se prenesejo na pozavarovalnico. Da bi ugotovili bolj natančno obnašanje slučajne spremenljivke N^1, moramo izračunati njeno porazdelitveno funkcijo. Pri tem si bomo pomagali z momentno rodovno funkcijo negativne binomske porazdelitve. Z izračunom smo že pokazali, da je le ta enaka:

$$M_N(t) = \Big(\frac{c}{c - (e^t - 1)}\Big)^M, \tag{3.28}$$

kjer je N porazdeljena negativno binomsko s parmetroma M in p ter $c = \frac{p}{1-p}$. S pomočjo te funkcije lahko sedaj izrazimo tudi momentno rodovno funkcijo slučajne

spremenljivke N^1:

$$\begin{aligned} M_{N^1}(t) &= E(e^{N^1\cdot t}) = E\Big(E(e^{(\sum_{i=1}^{N} Z_i)\cdot t}|N)\Big) \\ &= \sum_{k=0}^{\infty}\Big(E(e^{(\sum_{i=1}^{k} Z_i)\cdot t}|N=k)\Big)\cdot P(N=k) \\ &= \sum_{k=0}^{\infty}\Big(E(e^{(Z_1+Z_2+\ldots+Z_k)\cdot t})\Big)\cdot P(N=k) \\ &= \sum_{k=0}^{\infty}\Big(\prod_{i=1}^{k} E(e^{Z_i\cdot t})\Big)\cdot P(N=k) \\ &= \sum_{k=0}^{\infty}\Big(E\big(e^{Z\cdot t}\big)^k\Big)\cdot P(N=k) = \sum_{k=0}^{\infty}\big(M_Z(t)^k\big)P(N=k) \\ &= E\big((M_Z(t))^N\big). \end{aligned} \tag{3.29}$$

V tretjem koraku smo upoštevali, da so slučajne spremenljivke Z_i neodvisne od N, v četrti vrstici pa uporabimo neodvisnost Z_i med seboj, peto vrstico pa dobimo z uporabo lastnosti, da so Z_i enako porazdeljene. Dobljeni izraz lahko še malo poenostavimo:

$$M_{N^1}(t) = E\big(M_Z(t)^N\big) = E\big(e^{N\cdot\ln(M_Z(t))}\big) = M_N\big(\ln(M_Z(t))\big). \tag{3.30}$$

Dobili smo uporabno relacijo, s katero smo že blizu izračuna porazdelitve slučajne spremenljivke N^1. Vidimo, da moramo izračunati še momentno rodovno funkcijo Bernoullijeve slučajne spremenljivke:

$$M_Z(t) = E(e^{Z\cdot t}) = \pi\cdot e^t + (1-\pi). \tag{3.31}$$

Izračunajmo sedaj $M_{N^1}(t)$:

$$\begin{aligned} M_{N^1}(t) &= \Big(\frac{c}{c-(e^{\ln(M_Z(t))}-1)}\Big)^M = \Big(\frac{c}{c-(M_Z(t)-1)}\Big)^M \\ &= \Big(\frac{c}{c-(\pi\cdot e^t+1-\pi-1)}\Big)^M = \Big(\frac{c}{c-\pi\cdot(e^t-1)}\Big)^M \\ &= \Big(\frac{\frac{c}{\pi}}{\frac{c}{\pi}-(e^t-1)}\Big)^M. \end{aligned} \tag{3.32}$$

Opazimo lahko, da dobimo zelo podobno momentno rodovno funkcijo, kot jo ima prvotna negativna binomska porazdelitev s parametroma M in p, s to razliko, da je izraz c zamenjan z vrednostjo $\frac{c}{\pi}$. Lahko že ugotovimo, da je tudi novo dobljena slučajna spremenljivka N^1 porazdeljena negativno binomsko. Njen prvi parameter je prav tako enak M, drugi parameter pa moramo zgolj še izraziti s p. Iz izraza $p = \frac{c}{c+1}$ dobimo $c = \frac{p}{1-p}$. Upoštevamo to pri izračunu drugega parametra:

$$\frac{\frac{c}{\pi}}{\frac{c}{\pi}+1} = \frac{\frac{p}{\pi(1-p)}}{\frac{p}{\pi(1-p)}+1} = \frac{p}{p+\pi(1-p)}. \tag{3.33}$$

Dobili smo tudi drugi parameter in lahko zapišemo:

$$N^1 \sim NBin\Big(M, \frac{p}{p+\pi(1-p)}\Big). \tag{3.34}$$

Verjetnostna funkcija dobljene slučajne spremenljivke ima zato naslednjo obliko:
(3.35)

$$p_j = P(Y = j) = \binom{M+j-1}{j} \Big(\frac{p}{p+\pi(1-p)}\Big)^M \Big(\frac{\pi(1-p)}{p+\pi(1-p)}\Big)^j, \quad j = 0, 1, ...$$

Za slučajne spremenljivke, ki so porazdeljene negativno binomsko, pa velja še ena uporabna lastnost.

Trditev 3.2. *Naj bo $X \sim NBin(M, p)$ in $Y \sim NBin(L, p)$ ter naj velja, da sta slučajni spremenljivki X in Y neodvisni. Potem sledi:*

$$X + Y \sim NBin(M + L, p). \tag{3.36}$$

Dokaz. Dokažimo to trditev s pomočjo momentno rodovne funkcije negativne binomske porazdelitve. Ker sta slučajni spremenljivki X in Y neodvisni, velja, da je $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$:

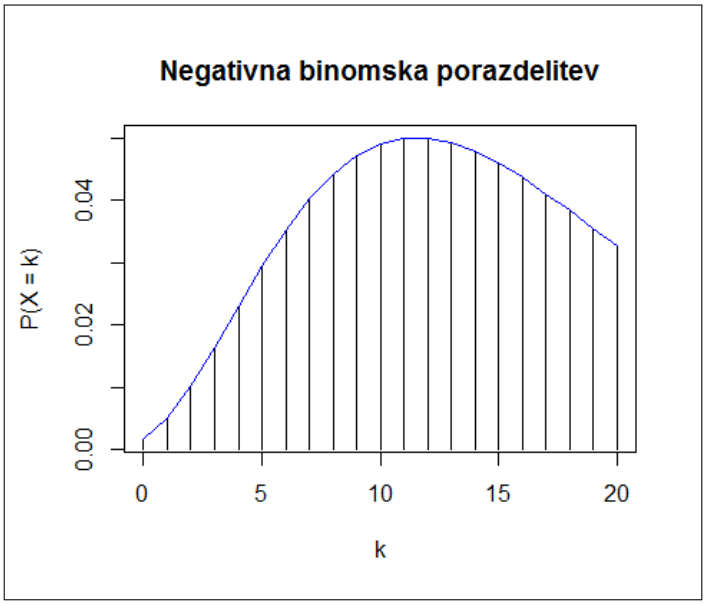
$$M_{X+Y}(t) = \Big(\tfrac{p}{1-(1-p)e^t}\Big)^M \cdot \Big(\tfrac{p}{1-(1-p)e^t}\Big)^L = \Big(\tfrac{p}{1-(1-p)e^t}\Big)^{M+L}.$$

Dobili smo točno momentno rodovno funkcijo negativne binomske porazdelitve s parametroma $M + L$, p. Torej lahko zapišemo:

$$X + Y \sim NBin(M + L, p).$$

□

Na spodnji sliki lahko vidimo še primer verjetnostne funkcije negativne binomske porazdelitve:



Slika 2. Primer negativne binomske binomske porazdelitve s parametroma $M = 4$ in $p = 0.2$

To so najbolj osnovne lastnosti teh dveh spremenljivk. V nadaljevanju bom sicer še sproti definiral nekatere značilnosti, ki nam bodo v pomoč pri sami anallizi. Zato si sedaj kar poglejmo bolj podrobno, kako se zavarovalnice lotijo ocenjevanja pričakovanih škod. Kot sem že omenil, se bom pri tem najbolj osredotočil na modeliranje s pomočjo negativne binomske porazdelitve, osnova vsemu pa predstavlja

Poissonova porazdelitev. Da bi lahko uporabili model z negativno binomsko porazdelitvijo, moramo najprej ugotoviti, ali je sploh smiselno izbrati tak model. Zato moramo poiskati test, ki bo povedal, ali je za slučajni vzorec, na katerem testiramo model, smiselno privzeti, da je porazdeljen negativno binomsko. Najprej si bomo pogledali različico Kolmogorov-Smirnovega testa, ki bolj nazorno prikaže sliko v teoretičnem smislu.

4.1. **Kolmogorov - Smirnov test za negativno binomsko porazdelitev.**

Naj bo $X_1, X_2, ..., X_n$ nek slučajni vzorec, ki ga dobimo iz neke porazdelitve s porazdelitveno funkcijo F. Najbolj osnovna in tudi najbolj poznana oblika Kolmogorov - Smirnovega testa ja naslednja:

$$H_0 : F = F_0, \qquad H_1 : F \neq F_0,$$

kjer je F_0 točno določena zvezna porazdelitvena funkcija. Ničelna hipoteza torej primerja enakost porazdelitve slučajnega vzorca z neko zvezno porazdelitveno funkcijo. Testna statistika primerja razlike med tema dvema porazdelitvama in je definirana kot

$$D_{KS} = \sup_x |\hat{F}(x) - F_0(x)|, \tag{4.1}$$

kjer je $\hat{F}(x)$ empirična porazdelitvena funkcija slučajnega vzorca $X_1, X_2, ..., X_n$. Z α označimo stopnjo značilnosti, pri kateri izvajamo test. Potem velja, da zavrnemo ničelno hipotezo H_0, če velja $D_{KS} \geq c_{n,\alpha}$, kjer $c_{n,\alpha}$ zadošča $P_G(D_{KS} \geq c_{n,\alpha}) = \alpha$ za vse zvezne porazdelitvene funkcije G. Točno takega testa ne moremo uporabiti na primeru negativne binomske porazdelitve, saj F_0 ne bi bila zvezna, ker ima negativna binomska porazdelitev diskretno porazdelitveno funkcijo. Zato v tem primeru definiramo Kolmogorov-Smirnov test na malo drugačen način.

Denimo, da je slučajni vzorec $X_1, X_2, ..., X_n$ dobljen iz negativne binomske porazdelitve s parametroma M in p. S T označimo vzorčno vsoto:

$$T = \sum_{i=1}^{n} X_i. \tag{4.2}$$

Predpostavimo, da poznamo parameter M, ne poznamo pa verjetnosti uspeha p. Videli bomo, da je T zadostna statistika za p. Test dobimo z uporabo pogojne porazdelitve slučajnega vzorca $(X_1, X_2, ..., X_n)$ glede na T. Pogojna testna statistika se zato v tem primeru zapiše kot:

$$D = \sup_x |\hat{F}(x) - E[\hat{F}(x)|T = t]|. \tag{4.3}$$

Da bi prišli do želenega testa, si moramo najprej pogledati, kakšna je pogojna porazdelitev slučajnega vzorca glede na T. To porazdelitev bomo označili z $F_t^{M,n}$ in jo izračunamo na naslednji način:

$$P\big((X_1, ..., X_n) = (x_1, ..., x_n)|T = t\big) = \frac{P\big(X_1 = x_1, ..., X_n = x_n, T = t\big)}{P(T = t)}, \tag{4.4}$$

kjer smo zgornji izraz zapisali po definiciji za pogojno verjetnost. Izraz v števcu je različen od 0 samo v primeru, ko je vsota slučajnih spremenljivk $X_1, ..., X_n$ enaka t. Slučajna spremenljvika T je po trditvi 3.2 porazdeljena negativno binomsko s parametroma nM, p, saj je vsota negativnih binomskih slučajnih spremenljivk s

parametroma M in p. Z upoštevanjem tega dejstva dobimo:
(4.5)

$$\begin{aligned}
P\big((X_1,...,X_n)=(x_1,...,x_n)|T=t\big) &= \frac{I_{(\sum_{i=1}^n x_i=t)}P(X_1=x_1)\cdot...\cdot P(X_n=x_n)}{\binom{nM+t-1}{t}p^{nM}(1-p)^t}\\
&= \frac{\prod_{i=1}^n\binom{M+x_i-1}{x_i}p^M(1-p)^{x_i}}{\binom{nM+t-1}{t}p^{nM}(1-p)^t}I_{(\sum_{i=1}^n x_i=t)}\\
&= \frac{p^{nM}(1-p)^{\sum_{i=1}^n x_i}\prod_{i=1}^n\binom{M+x_i-1}{x_i}}{\binom{nM+t-1}{t}p^{nM}(1-p)^t}I_{(\sum_{i=1}^n x_i=t)}\\
&= \frac{p^{nM}(1-p)^t\prod_{i=1}^n\binom{M+x_i-1}{x_i}}{p^{nM}(1-p)^t\binom{nM+t-1}{t}}I_{(\sum_{i=1}^n x_i=t)}\\
&= \frac{\prod_{i=1}^n\binom{M+x_i-1}{x_i}}{\binom{nM+t-1}{t}}I_{(\sum_{i=1}^n x_i=t)}.
\end{aligned}$$

V končni obliki lahko zapišemo:

$$F_t^{M,n}=P\big((X_1,...,X_n)=(x_1,...,x_n)|T=t\big)=\frac{\prod_{i=1}^n\binom{x_i+M-1}{M-1}}{\binom{nM+t-1}{t}}I_{(\sum_{i=1}^n x_i=t)}, \tag{4.6}$$

kjer z I_A označimo indikatorsko funkcijo dogodka A.

Imejmo sedaj slučajni vzorec $(Y_1,...,Y_n)$, ki je generiran s porazdelitvijo $F_t^{M,n}$. Za $i=1,...,n$ definiramo K_i kot število pojavov vrednosti i v zaporedju $Y_1,...,Y_n$. Ker imamo vzorec velikosti n, morata za vektor $(K_0,...,K_t)$ veljati spodnji enakosti:

$$\sum_{i=0}^t K_i=n, \qquad \sum_{i=0}^t iK_i=t. \tag{4.7}$$

Izračunajmo za $r=0,...,t$ pričakovano vrednost K_r:

$$E[K_r]=1\times P(Y_1=r)+1\times P(Y_2=r)+...+1\times P(Y_n=r). \tag{4.8}$$

Poglejmo si za primer $n=2$, kako bi izračunali $P(Y_1=r)$. V porazdelitveni funkciji (4.6) tega vzorca vidimo, da ima indikatorska funkcija pogoj, da je $\sum_{i=1}^n x_i=t$. To pomeni, da mora imeti Y_2 vrednost enako $t-r$, saj v nasprotnem primeru dobimo vedno verjetnost enako 0. Zato je vrednost v števcu porazdelitvene funkcije (4.6) enaka $\binom{r+M-1}{M-1}\binom{t-r+M-1}{M-1}$, kar lahko zapišemo tudi kot $\binom{r+M-1}{r}\binom{t-r+M-1}{t-r}$. Naš vzorec pa je velikosti n, zato mora imeti v tem primeru preostalih $n-1$ slučajnih spremenljivk vsoto enako $t-r$. S tem se izraz preoblikuje v $\binom{r+M-1}{r}\binom{t-r+(n-1)M-1}{t-r}$ Upoštevamo še, da so $Y_1....,Y_n$ enako porazdeljene, zato velja $P(Y_1=r)=...=P(Y_n=r)$ in dobimo

$$E[K_r]=nP(Y_1=r)=\frac{n\binom{r+M-1}{r}\binom{t-r+(n-1)M-1}{t-r}}{\binom{t+nM-1}{t}}. \tag{4.9}$$

Posledično je potem

$$E[K_0+...+K_r]=E[K_0]+...+E[K_r]=n\sum_{i=0}^r\frac{\binom{i+M-1}{i}\binom{t-i+(n-1)M-1}{t-i}}{\binom{t+nM-1}{t}}. \tag{4.10}$$

Definirajmo z $\hat{F}(x)$ empirično porazdelitveno funkcijo, dobljeno iz vzorca $Y_1, ..., Y_n$. Potem po definiciji velja, da je za $x = 0, ..., t$ le ta enaka

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x) = \frac{1}{n}\sum_{l=0}^{x} K_l. \tag{4.11}$$

Z uporabo (4.10) dobimo še

$$E[\hat{F}(x)|T = t] = \frac{1}{n}E[K_0 + ... + K_x] = \sum_{l=0}^{x} \frac{\binom{l+M-1}{l}\binom{t-l+(n-1)M-1}{t-l}}{\binom{t+nM-1}{t}}. \tag{4.12}$$

Tako imamo vse potrebne stvari, da lahko zapišemo Kolmogorov-Smirnovo testno statistiko:
(4.13)

$$D = \sup_x |\hat{F}(x) - E[\hat{F}(x)|T = t]| = \max_{x=0,...,t} |\hat{F}(x) - \sum_{l=0}^{x} \frac{\binom{l+M-1}{l}\binom{t-l+(n-1)M-1}{t-l}}{\binom{t+nM-1}{t}}|.$$

S pomočjo testne statistike lahko ugotovimo, ali je za opazovan vzorec mogoče trditi, da je porazdeljen negativno binomsko. To ugotovimo tako, da izračunamo $P(D \geq \nu)$ za neko vrednost $\nu > 0$. To je ekvivalentno, kot če izračunamo $1 - P(D < \nu)$. Dogodek, da je $D < \nu$ zahteva, da je razlika med $\hat{F}(x)$ in njeno pogojno pričakovano vrednostjo manjša od ν v vsaki točki $x = 0, ..., t$. Zato lahko to verjetnost zapišemo na naslednji način:

$$P(D < \nu) = P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t), \tag{4.14}$$

kjer je c_l najmanjše naravno število strogo večje od $E[K_0 + ... + K_l] - n\nu$ in d_l največje naravno število, ki je strogo manjše od $E[K_0 + ... + K_l] + n\nu$ za $l = 0, ..., t$. ν je torej opazovana vrednost testne statistike. Zato velja, da je $P(D \geq \nu|T = t)$ p-vrednost našega testa. Hipotezi testa pravita:

H_0: F je porazdeljena negativno binomsko z znanim parametrom M.
H_1: F ni porazdeljena negativno binomsko z znanim parametrom M.

Zopet naj α označuje stopnjo značilnosti. Potem lahko trdimo, da zavrnemo H_0, če velja, da je

$$P(D \geq \nu|T = t) \leq \alpha. \tag{4.15}$$

Da pridemo do p-vrednosti, moramo torej izračunati verjetnost (4.14). To računanje je kar precej zahtevno, zato bomo pri tem uporabili poseben algoritem, ki se imenuje Freyev algoritem.

4.1.1. *Računanje p-vrednosti.*

Kot v prejšnjem razdelku naj velja $(Y_1, ..., Y_n) \sim F_t^{M,n}$ in s K_i označimo število pojavov vrednosti i v zaporedju $Y_1, ..., Y_n$. Najprej si poglejmo, kako pridemo do $P((K_0, ..., K_t) = (k_0, ..., k_t))$, Verjetnost, da je $Y_1, ..., Y_n$ točno določeno zaporedje, v katerem se i ponovi K_i - krat za $i = 0, ..., t$, je

$$\frac{\prod_{i=0}^{t}\binom{i+M-1}{i}^{k_i}}{\binom{t+nM-1}{t}}. \tag{4.16}$$

Število zaporedij, v katerih se i ponovi K_i - krat za $i = 0, ..., t$, pa je

$$\frac{n!}{\prod_{i=0}^{t} k_i!}. \tag{4.17}$$

Torej velja, da je

$$P((K_0, ..., K_t) = (k_0, ..., k_t)) = \frac{\prod_{i=0}^{t} \binom{i+M-1}{i}^{k_i}}{\binom{t+nM-1}{t}} \cdot \frac{n!}{\prod_{i=0}^{t} k_i!}. \tag{4.18}$$

Ta izraz lahko malce preoblikujemo z upoštevanjem dejstev, da je $\sum_{i=0}^{t} k_i = n$ in $\sum_{i=0}^{t} ik_i = t$. Dobimo spodnje enakosti:

$$\begin{aligned}
n! &= n!^{\frac{1}{n}\sum_{i=0}^{t} k_i} = \prod_{i=0}^{t} (n!)^{\frac{k_i}{n}} = \prod_{i=0}^{t} ((n!)^{\frac{1}{n}})^{k_i}, \\
t! &= t!^{\frac{1}{t}\sum_{i=0}^{t} ik_i} = \prod_{i=0}^{t} (t!)^{\frac{ik_i}{t}} = \prod_{i=0}^{t} ((t!)^{\frac{1}{t}})^{ik_i}, \\
&\sum_{i=0}^{t} k_i(M - \frac{1}{n}) = nM - 1, \quad \sum_{i=0}^{t} k_i(i + M - \frac{1}{n}) = t + nM - 1, \\
(nM-1)! &= (nM-1)!^{\frac{1}{nM-1}\sum_{i=0}^{t} k_i(M-\frac{1}{n})} = \prod_{i=0}^{t} (((nM-1)!)^{\frac{1}{nM-1}})^{(M-\frac{1}{n})k_i}, \\
(t+nM-1)! &= (t+nM-1)!^{\frac{1}{t+nM-1}\sum_{i=0}^{t} k_i(i+M-\frac{1}{n})} \\
&= \prod_{i=0}^{t} (((t+nM-1)!)^{\frac{1}{t+nM-1}})^{(i+M-\frac{1}{n})k_i}.
\end{aligned} \tag{4.19}$$

Če sedaj zapišemo

$$\begin{aligned}
P((K_0, ..., K_t) = (k_0, ..., k_t)) &= \frac{\prod_{i=0}^{t} \binom{i+M-1}{i}^{k_i}}{\frac{(t+nM-1)!}{t!(nM-1)!}} \cdot \frac{n!}{\prod_{i=0}^{t} k_i!} \\
&= \prod_{i=0}^{t} \frac{\binom{i+M-1}{i}^{k_i} ((n!)^{\frac{k_i}{n}})((t!)^{\frac{ik_i}{t}})(((nM-1)!)^{\frac{(M-\frac{1}{n})k_i}{nM-1}})}{(k_i)!(((t+nM-1)!)^{\frac{(i+M-\frac{1}{n})k_i}{t+nM-1}})}
\end{aligned} \tag{4.20}$$

in definiramo

$$S(i,k) = \frac{\binom{i+M-1}{i}^{k} ((n!)^{\frac{1}{n}})^{k} ((t!)^{\frac{1}{t}})^{ik} (((nM-1)!)^{\frac{1}{nM-1}})^{(M-\frac{1}{n})k}}{k!(((t+nM-1)!)^{\frac{1}{t+nM-1}})^{(i+M-\frac{1}{n})k}}. \tag{4.21}$$

V tej obliki dobimo

$$P((K_0, ..., K_t) = (k_0, ..., k_t)) = \prod_{i=0}^{t} S(i, k_i). \tag{4.22}$$

Želeli pa bi si izračunati $P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t)$, saj le iz tega lahko dobimo p-vrednost in s tem veljavnost testa. Do te verjetnosti bomo prišli z uporabo nalsednjega algoritma in s pomočjo izraza $S(i,k)$, ki smo ga pravkar definirali.

4.1.2. *Freyev algoritem.*

Za $r = 0, ..., t$, $i = 0, ..., n$ in $j = 0, ..., t$ definirajmo z $\Gamma_r(i, j)$ množico vseh vektorjev $(K_0, ..., K_r)$, za katere veljajo naslednji pogoji:

- $K_l \geq 0$, za $l = 0, ..., r$,
- $c_l \leq K_0 + ... + K_l \leq d_l$, za $l = 0, ..., r$,
- $\sum_{l=0}^{r} k_l = i$,
- $\sum_{l=0}^{r} l k_l = j$.

Označimo še

$$Q_r(i, j) = \sum_{(k_0,...,k_r)\in\Gamma_r(i,j)} S(0, k_0) \cdot ... \cdot S(r, k_r). \tag{4.23}$$

Iz enačbe (4.22) potem sledi, da je $P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t) = Q_t(n, t)$, kjer lahko vrednost $Q_t(n, t)$ dobimo s pomočjo rekurzije. Ko je $t = 0$ velja, da je

$$\begin{aligned} Q_t(i, j) = Q_0(i, j) &= \sum_{(K_0)\in\Gamma_0(i,j)} S(0, K_0) \\ &= \begin{cases} S(0, i), & \text{če } i = c_0, ..., d_0; j = 0 \\ 0, & \text{sicer} \end{cases}. \end{aligned} \tag{4.24}$$

Če poznamo vrednosti $Q_{r-1}(i, j)$ za nek fiksen $r \geq 1$, kjer je $0 \leq i \leq n$ in $0 \leq j \leq t$, potem lahko s pomočjo tega pridemo do rešitve $Q_t(i, j)$ za $0 \leq i \leq n$ in $0 \leq j \leq t$. Da dobimo zaporedje $(k_0, ..., k_r)$, moramo začeti z zaporedjem $(k_0, ..., k_{r-1})$, ki pripada množici $\Gamma_{r-1}(i', j')$ za neka (i', j'). Temu zaporedju nato dodamo element k_r, da dobimo $(k_0, ..., k_r)$. Pri tem moramo biti pozorni na to, da $(k_0, ..., k_{r-1})$ ustreza pogojema $\sum_{l=0}^{r-1} k_l = i - k_r \geq 0$ in $\sum_{l=0}^{r-1} l k_l = j - r k_r \geq 0$. Velja še, da je $Q_r(i, j)$ različen od nič, kadar je $c_r \leq i \leq d_r$, posledično je tudi $Q_{r-1}(i - k_r, j - r k_r)$ različen od nič samo v primeru, ko je $c_{r-1} \leq i - k_r \leq d_{r-1}$. Z upoštevanjem tega lahko zapišemo

$$Q_r(i, j) = \begin{cases} \sum\limits_{k_r=max\{0,i-d_{r-1}\}}^{min\{i-c_{r-1}, \lfloor \frac{j}{r} \rfloor\}} Q_{r-1}(i - k_r, j - r k_r) S(r, k_r), & \text{če } c_r \leq i \leq d_r \\ 0, & \text{sicer} \end{cases}, \tag{4.25}$$

kjer smo z $\lfloor \frac{j}{r} \rfloor$ označili največje celo število, ki je manjše ali enako $\frac{j}{r}$. Če spodnja meja v vsoti preseže zgornjo mejo, potem je vsota enaka nič. Sedaj lahko uporabimo to rekurzijo za izračun $Q_t(n, t)$ in tako dobimo $P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t)$ na sledeči način:

(1) Izračunamo $\{Q_0(i, j), 0 \leq i \leq n, 0 \leq j \leq t\}$ z uporabo enačbe (4.24).
(2) Za $r = 1, ..., t$ izračunamo $\{Q_r(i, j), 0 \leq i \leq n, 0 \leq j \leq t\}$ iz $\{Q_{r-1}(i, j), 0 \leq i \leq n, 0 \leq j \leq t\}$ s pomočjo enačbe (4.25).
(3) Na koncu izrazimo $Q_t(n, t)$, ki predstavlja $P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t)$.

Zgornjemu algoritmu rečemo Freyev algoritem, s pomočjo katerega izračunamo p - vrednost, ki je enaka $1 - P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t)$.

Primer Freyevega algoritma sem napisal tudi sam v programu R. Najbolje bo, da si kar na primeru pogledamo, kako testiramo, ali so določeni podatki porazdeljeni negativno binomsko.

4.1.3. *Primeri Kolmogorov - Smirnovega testa.*

V tem razdelku si bomo ogledali na primeru, kako preučimo, ali je določen vzorec lahko porazdeljen negativno binomsko ali ne. Pri reševanju sem uporabljal program R, zato bodo vsi postopki opremljeni še s psevdokodo programa, ki sem ga uporabil. Denimo, da opazujemo 150 zavarovalniških polic, ki so bile vse sklenjene za obdobje enega leta. V spodnji tabeli so podatki o številu nastalih škod:

Število škod	Število polic
0	70
1	38
2	17
3	10
4	9
5	3
6	2
7	1

Menimo, da bi tem podatkom ustrezala negativna binomska porazdelitev s parametrom $M = 1$, ne vemo pa kakšen je parameter p. Iz podatkov lahko razberemo, da je $n = 150$ ter $t = 172$, kjer t izračunamo kot

$$t = \sum_{i=1}^{150} X_i = 0 \cdot 70 + 1 \cdot 38 + 2 \cdot 17 + 3 \cdot 10 + 4 \cdot 9 + 5 \cdot 3 + 6 \cdot 2 + 7 \cdot 1 = 172.$$

Postavimo hipotezo

H_0: Vzorec ima negativno binomsko porazdelitev s parametrom $M = 1$

in opravimo Kolmogorov-Smirnov test. Najprej moramo iz podatkov izračunati opazovano vrednost $\hat{F}$ in pričakovano vrednost $E[\hat{F}|T = t]$. Pri tem si pomagamo z naslednjim programom, kjer sem $\hat{F}$ in $E[\hat{F}|T = t]$ izračunal po definiciji, ki smo jo zapisali na začetku poglavja 3.1:

```
##Podatki:
vrednosti <- 0:7
stevilo <- c(70,38,17,10,9,3,2,1)
t = sum(vrednosti*stevilo)
n <- sum(stevilo)
M <- 1
#######################################################
##F streha:
F_streha <- c()
for (i in vrednosti){
  F_streha[i+1] <- 1/n * sum(stevilo[1:(i+1)])
}
#######################################################
##E(F|T=t):
pric_vr_F <- c()
for (i in vrednosti){
  j = 0
```

```
  upanje = 0
  while (j <= i){
    upanje <- upanje + (choose((j+M-1),j)*
              choose((t-j+(n-1)*M - 1),(t-j)))/(choose((t+n*M-1),t))
    j <- j+1
  }
  pric_vr_F[i+1] <- upanje
}
```

Ko imamo izračunano $\hat{F}$ in $E[\hat{F}|T = t]$, lahko sedaj dobimo tudi testno statistiko, ki zgolj primerja razlike med tema dvema vrednostima in vrne največjo absolutno razliko:

```
# max razlika - D:
abs_razlika <- abs(F_streha-pric_vr_F)
D <- max(abs_razlika)
```

Vse dobljene vrednosti so prikazane v spodnji tabeli:

Število škod	Število polic	$\hat{F}$	$E[\hat{F}\|T = t]$	Absolutna razlika
0	70	0.4667	0.4641745	0.002492212
1	38	0.72	0.7136682	0.006331776
2	17	0.833	0.8474094	0.014076065
3	10	0.90	0.9189063	0.018906253
4	9	0.96	0.9570229	0.002977128
5	3	0.98	0.9772874	0.002712596
6	2	0.993	0.9880308	0.005302511
7	1	1	0.9937105	0.006289536

Opazimo lahko, da je največja absolutna razlika prisotna pri vrednosti $X = 3$, zato zapišemo $D \approx 0.0189$. Ta vrednost je potem tudi enaka opazovani vrednosti testne statistike ν, kar bomo zdaj v nadaljevanju potrebovali pri računanju p - vrednosti. Računali bomo namreč $P(D < \nu)$, kar je pri tej izbiri ν zagotovljeno. Še prej pa moramo definirati funkcijo $S(i, k)$ za $i = 0, ..., n$ in $k = 0, ..., t$, ki jo uporabimo za izračun $P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t)$:

```
##Izračun za S(i,k)
S <- function(i,k){
  S <- k*log(choose((i+M-1),i)) + ((k/n)*lfactorial(n)) -
       lfactorial(k) - ((((i+M-1/n)*k)/(t+n*M-1))*
       lfactorial(t+n*M-1) - ((i*k)/t)*lfactorial(t) -
       (((M-1/n)*k)/(n*M-1))*lfactorial((n*M-1)))
  return (exp(S))
}
```

V psevdokodi lahko vidimo, da sem za izračun fakultete, npr. $k!$, uporabil funkcijo $lfactorial$ namesto $factorial$. V programu R je namreč funkcija $factorial$ definirana samo za vrednosti, ki so manjše ali enake 170 (lahko izračunamo $factorial(170)$, vrednost za $factorial(171)$ pa je že prevelika). Zato je bila bolj primerna uporaba funkcije $lfactorial$, ki omogoča izračune za veliko večja števila. Ta funkcija izračuna naravni logaritem fakultete določenega števila. Pri tem sem moral ustrezno preoblikovati izraz (4.21) in paziti na to, da na koncu funkcija vrne eksponentno vrednost. Definirati moramo še vektorja $c = (c_0, ..., c_t)$ in $d = (d_0, ..., d_t)$, kjer upoštevamo največjo absolutno razliko D, ki smo jo malo prej dobili.

```
povp <- c()
for (i in 0:t){
  j <- 0
  upanje <- 0
  while (j <= i){
    upanje <- upanje + (choose((j+M-1),j)*
              choose((t-j+(n-1)*M - 1),(t-j)))/
              (choose((t+n*M-1),t))
    j <- j+1
  }
  povp[i+1] <- n*upanje
}
##vrednosti c_0,...,c_t
c <- ceiling(povp-n*D)
##vrednosti d_0,...,d_t
d <- floor(povp+n*D)
```

Dobimo vektor c, ki je dolžine 173 in ima v našem primeru do c_{10} naslednje vrednosti

$$c = [67, 105, 125, 136, 141, 144, 146, 147, 147, 147, 148],$$

preostalih 162 vrednosti, od c_{11} do c_{172}, pa je enakih 148.
Prav tako je vektor d dolžine 173, njegove vrednosti do d_8 pa so enake

$$d = [72, 109, 129, 140, 146, 149, 151, 151, 152],$$

za ostalih 164 vrednosti pa dobimo $c_9 = c_{10} = ... = c_{172} = 152$.
Tako imamo pripravljene vse stvari, ki jih potrebujemo pri računanju $P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t)$. Napisal sem dve različici Frejevega algoritma, ki

izračuna to verjetnost. Najprej si poglejmo prvi primer, kjer pri računanju uporabimo rekurzivno funkcijo in sicer tako, kot smo ta algoritem definirali v razdelku 3.1.1. Psevdokoda tega algoritma je naslednja:

```
##Frey's Algorithm:
Q <- function(i,j,r){
  if (r == 0){
    c0 <- min(c[r+1],n)
    d0 <- min(d[r+1],n)
    if (c0 <= i & i <= d0 & j == 0){
      Q <- S(0,i)
    }
    else{
      Q <- 0
    }
  }
  else{
    c0 <- min(c[r],n)
    d0 <- min(d[r],n)
    c1 <- min(c[r+1],n)
    d1 <- min(d[r+1],n)
    if (c1 <= i & i <= d1){
      kmin <- max(0,i-d0)
      kmax <- min((i-c0),floor(j/r))
      if (kmin > kmax){
        Q <- 0
      }
      else{
        vsota <- 0
        for (k in kmin:kmax){
          vsota <- vsota + Q((i-k),(j-r*k),(r-1))*S(r,k)
        }
        Q <- vsota
      }
    }
    else{
      Q <- 0
    }
  }
  return (Q)
}
```

Izračunati hočemo $Q_t(n,t)$, kar dobimo z ukazom $Q(n,t,t)$. To nam vrne, da je $P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t) = 0.1583384$. Dobljena p - vrednost je potem enaka $1 - P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t) = 0.8416616$. Če vzamemo za stopnjo značilnosti α vrednost 0.05, potem na podlagi testa lahko trdimo, da je naš vzorec porazdeljen negativno binomsko s parametrom $M = 1$. Torej ne zavrnemo ničelne hipoteze, saj velja $P(D \geq \nu | T = t) > \alpha$ ($0.8416616 > 0.05$).

Vendar pa ta algoritem ni ravno najbolj uporaben, saj pri računanju potrebuje kar precej časa. Že v našem primeru, ko imamo vzorec velikosti 150, je za izračun verjetnosti potreboval okoli 11 ur. To je seveda veliko preveč, da bi opravljali s tem razne analize. Zato sem poskušal sprogramirati ta algoritem še na malo drugačen način. Rekurzivno funkcijo sem zamenjal z navadno for zanko, vrednosti pa sem shranjeval v matriko. Izkazalo se je, da je to veliko bolj učinkovito. Psevdokoda tega algoritma je prikazana spodaj:

```
##Frey's Algorithm 2 (veliko hitrejši):
Q_0 <- matrix(rep(0,(n+1)*(t+1)),nrow = n+1,ncol = t+1)
c0 <- min(c[1],n)
d0 <- min(d[1],n)
for (i in c0:d0){
  Q_0[i+1,1] <- S(0,i)
}
Q_matrika <- Q_0

for (r in 1:t){
  Q1 <- matrix(rep(0,(n+1)*(t+1)),nrow = n+1,ncol = t+1)
  c0 <- min(c[r],n)
  d0 <- min(d[r],n)
  c1 <- min(c[r+1],n)
  d1 <- min(d[r+1],n)
  for (i in c1:d1){
    kmin <- max(0,i-d0)
    for (j in 0:t){
      kmax <- min((i-c0),floor(j/r))
      if (kmax < kmin){
        Q1[i+1,j+1] <- 0
      }
      else{
        vsota <- 0
        for (k in kmin:kmax){
          vsota <- vsota + Q_matrika[(i-k+1),(j-r*k+1)]*S(r,k)
        }
        Q1[i+1,j+1] <- vsota
      }
    }
  }
  Q_matrika <- Q1
}
```

Na koncu dobimo matriko Q, ki je velikosti $(n+1)\times(t+1)$. $P(c_0 \leq K_0 \leq d_0, ..., c_t \leq K_0 + ... + K_t \leq d_t)$ predstavlja zadnji element v tej matriki. Dobimo ga z ukazom $Q\_matrika[n+1, t+1]$ in nam vrne isto vrednost kot v prejšnjem primeru, torej 0.1583384. To pomeni, da je tudi p - vrednost enaka 0.8416616, iz česar sledi, da ne zavrnemo ničelne hipoteze. Ta algoritem je bistveno hitrejši od prejšnjega, saj je to verjetnost izračunal v petih sekundah. Da bi bolje videli, kako učinkovit je ta

algoritem, sem opravil simulacije izračunov $P(0 \leq K_0 \leq n, ..., 0 \leq K_0 + ... + K_t \leq n)$ za različne vrednosti parametrov M,n,t. Ta verjetnost je seveda enaka 1, saj vedno velja $\sum_{i=0}^{r} K_i \leq n$ za vsak $r = 0, ..., t$, ker mora biti izpolnjeno $\sum_{i=0}^{t} K_i = n$. Prav zaradi tega je računanje te verjetnosti tudi najbolj časovno zahtevno, saj moramo opraviti največje možno število korakov. V spodnji tabeli so prikazani časi izračunov pri različnih vrednostih posameznih parametrov:

M = 1			
n	$t = \frac{n}{2}$	t = n	$t = 2n$
30	0.6	2.43	9.31
50	2.63	11.14	44.04
100	22.28	95.4	399.17
M = 5			
n	$t = \frac{n}{2}$	t = n	$t = 2n$
30	0.6	2.47	9.62
50	2.88	11.79	46.59
100	24.01	101.72	463.47
M = 20			
n	$t = \frac{n}{2}$	t = n	$t = 2n$
30	0.59	2.48	10.02
50	2.94	11.82	60.21
100	23.98	101.49	448.09

TABELA 1. Časi računanja $P(0 \leq K_0 \leq n, ..., 0 \leq K_0 + ... + K_t \leq n)$ v sekundah

Opazimo lahko, da čas za računanje verjetnosti $P(0 \leq K_0 \leq n, ..., 0 \leq K_0 + ... + K_t \leq n)$ ni odvisen od parametra M, saj so pri vseh vrednostih (1,5,20) časi dokaj podobni. Na čas izračuna torej najbolj vplivata parametra n in t. Vidimo, da s povečanjem teh dveh parametrov čas kar hitro narašča. Če vzamemo na primer $n = 100$, je že pri $t = 50$ algoritem porabil več kot 20 sekund, ko t povečamo na 200, pa čas že preseže 400 sekund.
Poglejmo si še en primer, kjer pa za vzorec ni mogoče trditi, da je porazdeljen negativno binomsko s parametrom $M = 1$. Imejmo isto primer iz zavarovalništva. Sedaj opazujemo 414 zavarovalniških polic. V tabeli so podana števila o nastalih škodah:

Število škod	Število polic
0	296
1	74
2	26
3	8
4	4
5	4
6	1
7	0
8	1

Ponovno najprej izračunamo opazovano porazdelitveno funkcijo $\hat{F}$ in pričakovano vrednost. Nato pogledamo še absolutne razlike in dobimo naslednje rezultate:

Število škod	Število polic	$\hat{F}$	$E[\hat{F}\|T=t]$	Absolutna razlika
0	296	0.7149758	0.6737357	0.04124012
1	74	0.8937198	0.8939108	0.000191
2	26	0.9565217	0.9656208	0.009099
3	8	0.9758454	0.9888972	0.0130518
4	4	0.9855072	0.9964267	0.0109195
5	4	0.9951691	0.9988540	0.0036849
6	1	0.9975845	0.9996337	0.0020492
7	0	0.9975845	0.9998833	0.0022988
8	1	1	0.9999630	0.0000370

Največja absolutna razlika je $D = 0.04124012$, ki se pojavi, ko je število škod enako 0. Vsota slučajnih spremenljivk t je v tem primeru enaka 200, za ν pa zopet vzamemo kar vrednost 0.04124012. Iz tega dobimo, da je $P(D < \nu)$ oziroma $P(0 \leq K_0 \leq n, ..., 0 \leq K_0 + ... + K_t \leq n)$ enaka 0.9987531. Posledično je potem p - vrednost enaka 0.001247, kar je manj kot stopnja zaupanja α, če zanjo ponovno vzamemo vrednost 0.05. Zaradi tega velja, da zavrnemo ničelno hipotezo, torej tukaj ne moremo sklepati, da je opazovan vzorec prišel iz negativne binomske porazdelitve z znanim parametrom $M = 1$. Pri računanju te verjetnosti je algoritem potreboval 59.84 sekund.
Kljub temu, da sem algoritem že izboljšal, lahko opazimo, da so časi izračunov še vedno kar precej visoki, ne glede na to, da vse izračune delamo na majhnih vzorcih. Če bi ta algoritem uporabili na veliko večjih vzorcih, ki jih zavarovalnice pri svojem delu pogosto srečujejo, bi porabili zelo veliko časa (verjetno kar nekaj ur). Tudi funkcija $lfactorial$, ki sem jo uporabil pri izračunu $S(i,k)$, ima pri velikih vzorcih omejitev in je program R ne more izračunati. Zato lahko ugotovimo, da v praksi ta algoritem ne bi bil ravno uporaben, s teoretičnega stališča pa nam lepo prikaže, na kakšen način lahko opravimo test. Poleg tega pa smo tudi privzeli, da poznamo parameter M negativne binomske porazdelitve, kar pa tudi v praksi ponavadi ne poznamo. Če imamo določene podatke, iz njih težko ocenimo, kakšen bi lahko bil ta parameter. V nadaljevanju si bomo sedaj ogledali, kako bi v praksi na boljši način preverili, ali se naš vzorec prilega negativni binomski porazdelitvi.

4.2. **Testiranje vzorcev v praksi.**

V praksi se testiranja določenega vzorca lotimo na malo drugačen način. Običajno imamo na voljo zgolj nekatere podatke, iz katerih težko sklepamo, za kateri tip verjetnostne porazdelitve točno gre, še težje pa lahko ocenimo pravilne parametre te porazdelitve. Zato si najprej narišemo graf podatkov, ki jih imamo na voljo (npr. histogram), in poskušamo sklepati, kateri znani porazdelitvi bi lahko najbolj ustrezal. Grafe verjetnostnih funkcij teh znanih porazdelitev namreč v splošnem poznamo tako, da hitro dobimo nekaj kandidatov. Nato kar predpostavimo, da podatki ustrezajo tej določeni porazdelitvi in potem poskušamo oceniti parametre predpostavljene porazdelitve. Iz opazovanih vrednosti, ki jih imamo na voljo, poskušamo razbrati čim več informacij, ki bi lahko koristile pri ocenjevanju teh parametrov. Na koncu,

ko dobimo ocene parametrov, pa poskušamo z raznimi statističnimi testi ugotoviti, ali se dobljena porazdelitev dovolj dobro ujema z vzorcem. Najbolj pogosti metodi, ki se uporabljata pri ocenjevanju parametrov sta metoda momentov in metoda največjega verjetja. Zato si bomo sedaj pogledali nekatere njune značilnosti in na kakšen način pridemo do ocen parametrov pri posamezni metodi.

4.2.1. *Metoda momentov.*

To metodo lahko uporabljamo za določanje neznanih parametrov tako zveznih, kot tudi diskretnih slučajnih spremenljivk. Glede na to, da se jaz v svojem delu bolj osredotočam na diskretne slučajne spremenljivke (negativna binomska porazdelitev), bom tudi v nadaljevanju definiral metodo momentov zgolj za diskretne slučajne spremenljivke. Če poznamo ta način, lahko potem hitro ugotovimo tudi ocene za zvezne slučajne spremenljivke (vsote zamenjajo integrali). Neznane parametre določene porazdelitvene funkcije p_k, za $k = 0, 1, ...,$ ponavadi označimo z grškimi črkami $\theta_1, \theta_2, ..., \theta_r$, kjer je $r > 0$. Ocenjujemo jih s pomočjo cenilk $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_r$, ki so pravzaprav funkcije slučajnih spremenljivk $X_1, ..., X_n$, pri dani velikosti vzorca n. Verjetnostna funkcija p_k je torej odvisna od neznanih parametrov $p_k = p_k(\theta_1, \theta_2, ..., \theta_r)$. Da to metodo lahko uporabimo, mora obstajati prvih r momentov. j - ti začetni moment porazdelitve je definiran kot

$$m_j = E[X^j]. \tag{4.26}$$

Momenti so tako odvisni od parametrov, zato je $m_j = m_j(\theta_1, \theta_2, ..., \theta_r)$. Po tej metodi moramo začetni moment porazdelitve primerjati z j - tim začetnim momentom vzorca, ki je definiran na sledeči način:

$$\hat{m}_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j. \tag{4.27}$$

Dobimo sistem enačb $m_j(\theta_1, \theta_2, ..., \theta_r) = \hat{m}_j$, ki ga moramo rešiti za $j = 1, 2, ..., r$. Do rešitve pa lahko pridemo tudi s centralnimi momenti. V tem primeru najprej izračunamo prvi vzorčni začetni moment:

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}, \tag{4.28}$$

nato pa rešujemo sistem enačb $\mu_j(\theta_1, \theta_2, ..., \theta_r) = \hat{\mu}_j$, za $j = 1, 2, ..., r$, kjer je $\hat{\mu}_j$ j - ti vzorčni centralni moment, definiran kot

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^j. \tag{4.29}$$

Denimo, da je zgornji sistem enačb mogoče rešiti, torej $\theta_j = \theta_j(m_1, m_2, ..., m_r)$ v prvem primeru in $\theta_j = \theta_j(\mu_1, \mu_2, ..., \mu_r)$ v primeru centralnih momentov, za $j = 1, 2, ..., r$. Če v prvi razrešeni sistem enačb namesto parametrov $m_1, m_2, ..., m_r$ vstavimo vzorčne začetne momente $\hat{m}_1, \hat{m}_2, ..., \hat{m}_r$ oziroma v drugi razrešeni sistem enačb namesto $\mu_1, \mu_2, ..., \mu_r$ centralne vzorčne momente $\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_r$, dobimo ocene oziroma cenilke $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_r$ neznanih parametrov $\theta_1, \theta_2, ..., \theta_r$. Pri tem moramo povedati, da eksplicitni zapis funkcij $\theta_j = \theta_j(m_1, m_2, ..., m_r)$ oz. $\theta_j = \theta_j(\mu_1, \mu_2, ..., \mu_r)$,

za $j = 1, 2, ..., r$, ni vedno mogoč. V tem primeru lahko sistem enačb rešimo tudi numerično, če ne gre analitično. Poglejmo si na primeru, kako ocenimo parametre negativne binomske porazdelitve po metodi momentov.

Primer 4.1. Denimo, da imamo vzorec slučajnih spremenljivk $X_1, ..., X_n$ z realizacijo (oziroma njihovimi vrednostmi) $x_1, ..., x_n$. Za oceno parametrov tega vzorca, za katerega menimo, da je porazdeljen negativno binomsko, moramo najprej rešiti sistem dveh enačb z dvema neznankama. Sistem dveh enačb imamo zato, ker iščemo oceno za dva parametra, to sta M in p. Iz enačb (3.23) in (3.26) iz razdelka 2.2 dobimo upanje in varianco negativne binomske porazdelitve. Izračunati moramo samo še vzorčna momenta. Prvi vzorčni začetni moment je enak $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, kar predstavlja povprečno vrednost vzorca. Za drugi vzorčni centralni moment pa posledično vzamemo varianco vzorca, ki jo označimo s $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Torej dobimo

$$\begin{aligned} \frac{M(1-p)}{p} &= \bar{x}, \\ \frac{M(1-p)}{p^2} &= \sigma^2. \end{aligned} \tag{4.30}$$

Iz prve enačbe izrazimo M in dobimo $M = \frac{p}{1-p}\bar{x}$, kar vstavimo v drugo enačbo. Oceni za parametra sta potem naslednji:

$$\begin{aligned} \hat{M} &= \frac{\bar{x}^2}{\sigma^2 - \bar{x}}, \\ \hat{p} &= \frac{\bar{x}}{\sigma^2}. \end{aligned} \tag{4.31}$$

Tako lahko izračunamo vrednosti parametrov, ki ustrezajo našemu vzorcu in dobimo cenilki $\hat{M}$ in $\hat{p}$. Dobljena izraza preprosto uporabimo v programu R, da izračunamo ocene parametrov za določen vzorec. Spodaj je primer vzorca velikosti 200, ki sem ga naključno generiral z negativno binomsko porazdelitvijo s parametroma $M = 3$ in $p = 0.2$.

```
vzorec <- rnbinom(200,3,0.2)
table(vzorec)
vzorec
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 21 28 41
10 12 14 13 18 14 14 15 11 16 11 10  6  7  5  4  7  5  1  2  2  2  1

upanje <- mean(vzorec)
varianca <- sd(vzorec)^2
#Metoda momentov:
ocena_M <- upanje^2/(varianca - upanje)
ocena_p <- upanje/varianca
```

Ko poženemo program, dobimo vrednosti za ocene parametrov enake:

$$\hat{M} = 2.778299, \qquad \hat{p} = 0.1823234.$$

Za cenilke je zaželjeno, da so nepristranske in imajo čim manjšo napako, ki jo merimo z izrazom MSE (mean squared error).

Definicija 4.2. $\hat{\theta}_j$ je nepristranska cenilka parametra θ_j, če velja

$$E[\hat{\theta}_j] = \theta_j. \tag{4.32}$$

Definicija 4.3. Izraz MSE za cenilko $\hat{\theta}_j$ parametra θ_j je definiran kot

$$MSE(\hat{\theta}_j) = E[(\hat{\theta}_j - \theta_j)^2] = Var(\hat{\theta}_j) + (E[\hat{\theta}_j] - \theta_j)^2. \tag{4.33}$$

Opazimo lahko, da za nepristranske cenilke velja $MSE(\hat{\theta}_j) = Var(\hat{\theta}_j)$. Torej nam $Var(\hat{\theta}_j)$ predstavlja neko mero za kvaliteto cenilke neznanega parametra.

4.2.2. *Metoda največjega verjetja.*

Tudi metoda največjega verjetja se uporablja tako za zvezne, kot za diskretne slučajne spremenljivke. Zopet označimo verjetnostno funkcijo diskretne slučajne spremenljivke s p_k, ki je odvisna od neznanih prametrov, torej $p_k = p_k(\theta_1, \theta_2, ..., \theta_r)$, za $k = 0, 1, ...$ in $r > 0$. S pomočjo vzorca $x_1, x_2, ..., x_n$, ki ga imamo na voljo, lahko ocenimo te neznane parametre. Pri tej metodi jih določimo tako, da je verjetnost pojava opazovanih vrednosti $x_1, x_2, ..., x_n$ neodvisnih in enako porazdeljenih diskretnih slučajnih spremenljivk $X_1, X_2, ..., X_n$ največja. Matematično je ta verjetnost, da se pojavijo dani $x_1, x_2, ..., x_n$, določena s funkcijo največjega verjetja, ki je definirana kot

$$L(\theta_1, \theta_2, ..., \theta_r) = \prod_{i=1}^{n} p_{x_i}(\theta_1, \theta_2, ..., \theta_r). \tag{4.34}$$

Ta izraz vsebuje neznane parametre. Vrednosti neznanih parametrov, ki maksimizirajo zgornjo funkcijo verjetja se imenujejo ocene po metodi največjega verjetja (velikokrat se uporabi kratica MLE, ki prihaja iz angleščine). Da se izognemo računanju s produktom, ki je precej zahtevno, ponavadi raje uporabimo logaritem funkcije verjetja in iščemo maksimum te dobljene funkcije. Produkt se nam v tem primeru poenostavi v vsoto:

$$\log(L(\theta_1, \theta_2, ..., \theta_r)) = l(\theta_1, \theta_2, ..., \theta_r) = \sum_{i=1}^{n} \log p_{x_i}(\theta_1, \theta_2, ..., \theta_r). \tag{4.35}$$

Da bi dobili maksimum zgornje funkcije, moramo le to odvajati in odvode izenačiti z 0. Dobimo sistem r nelinearnih enačb

$$0 = \frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial \log p_{x_i}(\theta_1, \theta_2, ..., \theta_r)}{\partial \theta_j}, \quad j = 1, 2, ..., r. \tag{4.36}$$

Ta sistem enačb v splošnem lahko nima rešitve, je pa verjetnost, da jo ima, zelo blizu 1, ko večamo velikost vzorca. Druga možnost pa je, da ima lahko tudi več rešitev. V tem primeru moramo biti pozorni, da res uporabimo pravo rešitev. Če izberemo napačno, lahko dosežemo zgolj lokalni maksimum funkcije verjetja, kar pa ne bi bilo pravilno. Ko enkrat dobimo rešitev, parametre označimo s $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_r$ in te vrednosti privzamemo za neznane parametre $\theta_1, \theta_2, ..., \theta_r$ predpostavljene porazdelitve. Kadar imamo opravka z diskretnimi slučajnimi spremenljivkami $X_1, X_2, ..., X_n$ z zalogo vrednosti v $\mathbb{N}$, ni nujno, da poznamo vse njihove vrednosti $x_1, x_2, ..., x_n$. V tem primeru namreč zadostuje že, če poznamo frekvence pojavljanja ν_k posameznih števil

iz zaloge vrednosti. Definirajmo $k_{max} = \max\{k|v_k \neq 0\}$. Potem se izraz v (4.35) poenostavi v

$$l(\theta_1, \theta_2, ..., \theta_r) = \sum_{k=0}^{k_{max}} \nu_k \cdot \log p_k(\theta_1, \theta_2, ..., \theta_r). \tag{4.37}$$

Ko ta izraz odvajamo, dobimo sistem naslednjih enačb, ki ga je potrebno rešiti:

$$0 = \frac{\partial l}{\partial \theta_j} = \sum_{k=0}^{k_{max}} \nu_k \frac{\partial \log p_k(\theta_1, \theta_2, ..., \theta_r)}{\partial \theta_j}, \quad j = 1, 2, ..., r. \tag{4.38}$$

Primer 4.4. Poglejmo si sedaj, kako izgleda funkcija verjetja negativne binomske porazdelitve. Za parametrizacijo verjetnostne funkcije vzamemo obliko iz enačbe (3.6) in po definiciji funkcije verjetja izračunamo

$$\begin{aligned} L(\theta_1, \theta_2, ..., \theta_r) &= \prod_{i=1}^{n} p_{x_i}(\theta_1, \theta_2, ..., \theta_r) = \prod_{i=1}^{n} \binom{M + x_i - 1}{x_i} p^M (1-p)^{x_i} \\ &= \Big(\frac{p^M}{(M-1)!}\Big)^n \prod_{i=1}^{n} \frac{(M + x_i - 1)!}{x_i!} (1-p)^{x_i}. \end{aligned} \tag{4.39}$$

Za računanje potrebujemo logaritemsko funkcijo verjetja, ki ima potem spodnjo obliko:

$$\begin{aligned} l(\theta_1, \theta_2, ..., \theta_r) = nM \log p - n \log((M-1)!) + \sum_{i=1}^{n} \log((M + x_i - 1)!) \\ - \sum_{i=1}^{n} \log(x_i!) + \log(1-p) \sum_{i=1}^{n} x_i. \end{aligned} \tag{4.40}$$

Ta izraz lahko zopet uporabimo v programu R za izračun ocen parametrov po metodi največjega verjetja. Poleg tega pa uporabimo še vgrajeno funkcijo *mle*, za katero moramo imeti vgrajen programski paket stats4. Ta funkcija zahteva izraz za $-l(\theta_1, \theta_2, ..., \theta_r))$, določiti pa moramo tudi začetno vrednost parametrov. Izbira za to je lahko povsem poljubna, velikokrat pa se uporabi za začetno vrednost kar ocena parametrov po metodi momentov.

```
#Metoda največjega verjetja:
vzorec <- rnbinom(200,3,0.2)
table(vzorec)
vzorec
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
 2  3  5  5  5  8 17  7 22 11 18 12  9  5  6  9  9  6  9  6  4  2  3
23 25 26 27 29 30 33 34 35 38
4  2  4  1  1  1  1  1  1  1

library(stats4)
#Logaritemska funkcija verjetja za negativno binomsko porazdelitev:
l <- function(M,p){
  n <- 200
  x <- vzorec
  - n*M*log(p) - sum(lfactorial(M+x-1)) + sum(lfactorial(x)) +
```

```
    n*lfactorial(M-1) - log(1-p)*sum(x)
}

ocenaMLE <- mle(l,start = list(M = 2,p = 0.5))
```

Jaz sem za začetni vrednosti izbral $M = 2$ in $p = 0,5$. Ko izvedemo test na našem vzorcu dobimo oceni $\hat{M} = 3,78$ ter $\hat{p} = 0,24$. Rezultati so predstavljeni spodaj, kjer lahko vidimo tudi standardno napako pri posamezni oceni parametra:

```
summary(ocenaMLE)
Maximum likelihood estimation

Call:
mle(minuslogl = l, start = list(M = 2, p = 0.5))

Coefficients:
   Estimate Std. Error
M 3.7846043 0.50606601
p 0.2397659 0.02553016

-2 log L: 1312.419
```

Obstaja pa še ena vgrajena funkcija v programu R znotraj paketa MASS, s pomočjo katere dobimo ocene parametrov po metodi največjega verjetja. Imenuje se *fitdistr* in je bolj preprosta kot funkcija *mle*, saj nam ni potrebno eksplicitno izračunati funkcije verjetja. Kar potrebujemo, je vzorec, ime porazdelitve h kateri želimo aproksimirati vzorec in začetne vrednosti parametrov. Za vzorec izberimo isti primer kot zgoraj in dobimo:

```
#Metoda največjega verjetja:
fitdistr(vzorec,"negative binomial", list(size = 2,prob = 0.5))
    size         prob
  3.78460432   0.23976593
 (0.50606604) (0.02553016)
```

Opazimo lahko, da smo dobili iste ocene za parametra M in p kot v prejšnjem primeru, pri čemer smo izbrali tudi isti vrednosti za začetne parametre kot pri funkciji *mle*.

To sta najbolj osnovna načina, ki nam ustrezno ocenita parametre predpostavljene porazdelitve. Seveda pa nas zanima, ali je ta ocena sploh smiselna in kako dobro aproksimira dejansko porazdelitev vzorca. Zato moramo ocene še stestirati. Pri tem se najbolj pogosto uporablja hi - kvadrat test (χ^2), ki nam pove kvaliteto aproksimacije.

4.2.3. *Testiranje kvalitete aproksimacije.*

Za najboljše ocene parametrov se štejejo cenilke, dobljene po metodi največjega verjetja. Lahko pa se zgodi, da pri določenem primeru to ni nujno res. Dobljene ocene za neznane parametre lahko najprej preučimo z nekaterimi enostavnimi absolutnimi in relativnimi merami. Med absolutnimi merami velja izpostaviti naslednji dve:

$$\epsilon_1 = \frac{\sum_{i=1}^{n} |\nu_i - \nu_i^*|}{n}, \qquad \epsilon_2 = \sqrt{\frac{\sum_{i=1}^{n} (\nu_i - \nu_i^*)^2}{n}}, \tag{4.41}$$

kjer ν_i predstavljajo empirične frekvence, z ν_i^* pa označimo frekvence aproksimiranega modela. Posledično dobimo za relativni meri naslednja dva izraza:

$$\delta_1 = \frac{\epsilon_1}{\sum_{i=1}^{n} \nu_i / n}, \qquad \delta_2 = \frac{\epsilon_2}{\sum_{i=1}^{n} \nu_i / n}. \tag{4.42}$$

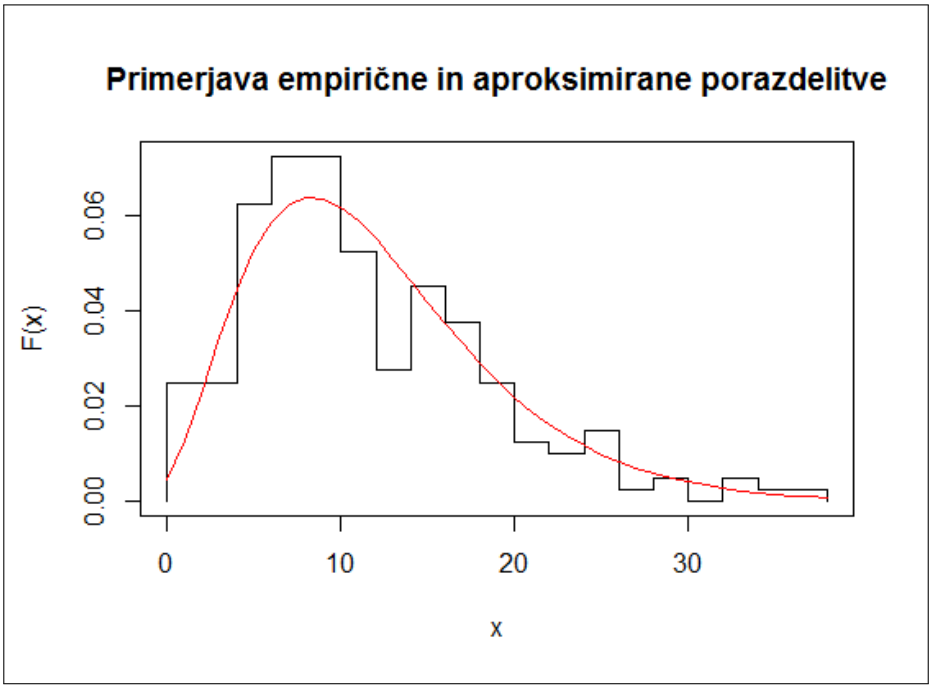
Z izračunom teh dveh mer že dobimo določen podatek o tem, ali je aproksimacija smiselna ali ne. Poglejmo si, kakšni sta ti dve meri v primeru zadnjega vzorca, kjer smo parametre ocenjevali po metodi največjega verjetja:

```
#empirične frekvence:
tabela <- table(vzorec)
frekv <- c()
for (i in 1:length(tabela)){
  frekv[i] <- tabela[[i]]
}
#aproksimirane frekvence:
MLE_M <- 3.7846043
MLE_p <- 0.2397659
x <- as.numeric(names(tabela))
frekv_model <- dnbinom(x,size = MLE_M,prob = MLE_p)*200

frekv
2  3  5  5  5  8 17  7 22 11 18 12  9  5  6  9  9  6  9  6  4  2  3
4  2  4  1  1  1  1  1  1  1
trunc(frekv_model)
0  2  4  6  8 10 11 12 12 12 12 11 11 10  9  8  7  6  5  5  4  3  3
2  1  1  1  0  0  0  0  0  0
#napake:
absolute <- mean(abs(frekv-trunc(frekv_model)))
relative <- absolute/mean(frekv)
```

Za absolutno mero ϵ_1 dobimo vrednost 2.121212, kar pomeni, da se v povprečju absolutne frekvence razlikujejo za približno 2. Relatvina mera δ_1 pa v tem primeru znaša 35%. Za bolj nazorni prikaz aproksimacije si poglejmo graf, na katerem je lepo prikazana razlika med empirično in aproksimirano porazdelitvijo. Psevdokoda za dobljeni graf je naslednja:

```
h <- hist(vzorec,breaks = 19)
b <- c(min(h$breaks),h$breaks)
y <- c(0,h$density,0)
plot(b,y,type ="s", xlab = "x", ylab = "F(x)",
     main = "Primerjava empirične in aproksimirane porazdelitve")

xfit <- seq(min(vzorec),max(vzorec))
yfit <- dnbinom(xfit,size = MLE_M,prob = MLE_p)
lines(xfit,yfit,col = "red")
```

Graf, ki ga dobimo, pa je prikazan spodaj:



Slika 3. Aproksimacija porazdelitve vzorca

Z rdečo barvo je narisana porazdelitvena funkcija negativne binomske porazdelitve z ocenjenima parametroma $M = 3.7846043$ in $p = 0.2397659$, s črno barvo pa je prikazana porazdelitev vzorca. Iz grafa lahko delno že razberemo, kako dober je model. Za bolj podrobno analizo pa uporabimo χ^2 - test, ki se uporablja za testiranje zveznih in tudi diskretnih slučajnih spremenljivk.

Označimo z O_i število ponovitev vrednosti i v vzorcu, ki ga testiramo. Temu številu bomo rekli opazovana frekvenca. Z E_i bomo označili število ponovitev vrednosti i v vzorcu generiranem s predpostavljeno porazdelitvijo, ki ima vrednosti parametrov enake ocenam, dobljenim bodisi z metodo momentov bodisi po metodi največjega verjetja. To število se imenuje pričakovana frekvenca vrednosti i. χ^2 test nam pove, ali vzorec ustreza določeni porazdelitvi. Jaz bom v večini primerov testiral, ali lahko za določen vzorec trdimo, da je porazdeljen negativno binomsko. Zato sta hipotezi definirani tako:

H_0 : Vzorec je porazdeljen negativno binomsko z ocenjenima parametroma $\hat{M}$ in $\hat{p}$,
H_A : Vzorec ni porazdeljen negativno binomsko z ocenjenima parametroma $\hat{M}$ in $\hat{p}$.

Da lahko do testa pridemo, moramo najprej izračunati testno statistiko, ki je v tem primeru naslednja:

$$\hat{\chi}^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{4.43}$$

kjer k predstavlja število različnih vrednosti v vzorcu oziroma število razredov, v katere razdelimo vrednosti vzorca. Ta statistika konvergira k porazdelitvenemu zakonu $\chi^2(d)$ s številom prostostnih stopenj d = k - s - 1, kjer s označuje število neznanih parametrov predpostavljene porazdelitve (v primeru negativne binomske porazdelitve je $s = 2$, saj ocenjujemo parametra M in p). Če je ničelna hipoteza pravilna, bo vrednost χ^2 blizu 0, saj si bodo števila O_i in E_i zelo blizu. V primeru, da ničelna hipoteza ne drži pa bo ta razlika večja in posledično bo tudi vrednost χ^2 večja. Spet moramo določiti, pri kakšni stopnji zaupanja α izvajamo test. Kritično vrednost $h_{d,\alpha}$ dobimo s pomočjo porazdelitvene tabele χ^2 porazdelitve in sicer mora veljati:

$$P(\chi^2(d) > h_{d,\alpha}) = \alpha. \tag{4.44}$$

Nato zavrnemo ničelno hipotezo v primeru, da je $\hat{\chi}^2 > h_{d,\alpha}$. Ta test lahko opravimo s pomočjo programa R. Preprosto izračunamo testno statistiko po definiciji in opravimo test. Če si spet pogledamo na zadnjem primeru, ki smo ga obravnavali pri metodi največjega verjetja, dobimo naslednje rezultate:

```
alpha <- 0.05
#testna statistika:
X2 <- sum((frekv-frekv_model)^2/frekv_model)
st <- length(frekv) - 2 - 1
#kritična vrednost:
h_st <- qchisq(1-alpha,st)
```

Program nam vrne za kritično vrednost $h_{d,\alpha}$ število 43.77297, testna statistika $\hat{\chi}^2$ pa je enaka 37.41099. Ker velja $\hat{\chi}^2 < h_{d,\alpha}$ ničelne hipoteze ne zavrnemo in lahko sklepamo, da je naš vzorec res porazdeljen negativno binomsko z ocenjenima parametroma $M = 3.7846043$ in $p = 0.2397659$. Ta test lahko opravimo še na bolj enostaven način, kjer nam ni potrebno eksplicitno računati testne statistike. To storimo s pomočjo vgrajene funkcije *chisq.test*, ki mora vsebovati izpis empiričnih frekvenc in vzorec aproksimiranih frekvenc z modelom:

```
chisq.test(frekv,p = frekv_model,rescale.p = TRUE)

        Chi-squared test for given probabilities

data:  frekv
X-squared = 36.1385, df = 32, p-value = 0.2812
```

Vidimo, da funkcija izpiše vrednost testne statistike, število prostostnih stopenj in p - vrednost. Ta je enaka $0,2812$, kar pomeni, da je večja od stopnje značilnosti $\alpha =$

$0,05$. Zato ne zavrnemo ničelne hipoteze in ponovno ugotovimo, da je smiselno privzeti, da je vzorec porazdeljen negativno binomsko s parametroma $M = 3.7846043$ in $p = 0.2397659$. Poleg tega pa obstaja za diskretne slučajne spremenljivke še ena posebna funkcija, ki preveri kakovost aproksimacije. Ta funkcija se imenuje *goodfit*, zanjo pa moramo imeti vgrajen paket *vcd*. V tej funkciji zgolj podamo vzorec, katero porazdelitev želimo aproksimirati in dodamo ocenjene vrednosti neznanih parametrov:

```
library(vcd)
test <- goodfit(vzorec,type = "nbinomial",method = "MinChisq",
                par = list(size = MLE_M,prob = MLE_p))
summary(test)
         Goodness-of-fit test for nbinomial distribution

                      X^2 df   P(> X^2)
Pearson          38.45368 38 0.44894550
Likelihood Ratio 44.23854 32 0.07347457
```

Spet ugotovimo, da je p - vrednost večja od stopnje značilnosti α, zato je negativna binomska porazdelitev s parametroma $M = 3.7846043$ in $p = 0.2397659$ dober izbor za aproksimacijo vzorca.

Opazimo, da je χ^2 - test v teh primerih lahko zelo uporaben, ima pa, tako kot velika večina ostalih testov, tudi določene pomanjkljivosti. Testna statistika in rezultati, ki jih dobimo, so odvisni od velikosti razredov, v katere razdelimo vzorec. Druga negativna lastnost tega testa je, da potrebujemo dovolj velik vzorec. Pri majhnih vzorcih se lahko zgodi, da ta način aproksimacije ne bi bil pravilen. V primeru, da na podlagi testa ugotovimo nepravilnost aproksimacije, ta model zavrnemo in poskusimo s kakšno drugo porazdelitvijo. Lahko se zgodi tudi, da ne moremo poiskati prave porazdelitve, ki bi dobro aproksimirala vzorec. Potem vseeno poskušamo izbrati tisto porazdelitev, ki se v resnici ne prilega najbolje, nam pa zagotavlja, da je test dokaj blizu uspeha.
V praksi zavarovalnice ta način ocenjevanja veliko uporabljajo, saj želijo oceniti kakšno bo pričakovano število škod v prihodnosti. Zato obstajajo modeli, s pomočjo katerih pridemo do želenih ocen. Sedaj si bomo pogledali, kako ravnamo v določenih primerih, da si zagotovimo čim boljše ocene.

Pri večini modelov, ki se uporabljajo pri ocenjevanju števila škod, igra pomembno vlogo Poissonova porazdelitev. Zanjo je značilno, da je uporabna pri procesih, kjer opazujemo število dogodkov v določenem časovnem intervalu. Prav zaradi tega je njena uporabnost v zavarovalništvu velika. Tipičen primer je uporaba pri avtomobilskem zavarovanju. Vemo, da se bodo nesreče zgodile in to neodvisno ena od druge. Tako dobimo idejo, da bi za začetni model izbrali Poissonovo porazdeltiev s parametrom $\lambda = \mu T$, kjer μ označuje intenzivnost pojavljanja škod, T pa je oznaka za časovni interval. Za začetek bi si želeli čim bolj splošen model. V ta namen razdelimo časovni interval T na K manjših časovnih obdobij enake dolžine $h = T/K$. To naredimo zato, ker je na manjših časovnih intervalih verjetnost pojava več kot enega dogodka zelo majhna. S tem dobimo, da je število dogodkov v k - tem časovnem intervalu I_k, za $k = 1, ..., K$, enako bodisi 1 bodisi 0. Število dogodkov, ki se zgodijo v celotnem obdobju T je potem enako:

$$N = I_1 + ... + I_K. \tag{5.1}$$

Če je verjetnost $p = P(I_k = 1)$ enaka za vse k in so dogodki neodvisni, potem je to navadno Bernoullijevo zaporedje z verjetnostjo uspeha p. Iz verjetnosti pa tudi vemo, da je vsota neodvisnih Bernoullijevih zaporedij porazdeljena binomsko, zato je porazdelitvena funkcija slučajne spremenljivke N enaka:

$$P(N = n) = \binom{K}{n} p^n (1-p)^{K-n}, \quad \text{za } n = 0, 1, ..., K. \tag{5.2}$$

Smiselno je privzeti, da je verjetnost p sorazmerna z dolžino časovnega intervala h. Torej lahko zapišemo, da je $p = \mu h$, kjer μ označuje intenzivnost pojava škod v časovni enoti. Model bomo zgradili tako, da bomo poiskali porazdelitev N v limiti, ko $K \to \infty$ in zapisali verjetnost dogodka kot $p = \mu h = \mu T/K$. Uporabimo to v (5.2) in dobimo:

$$P(N = n) = \frac{K!}{n!(K-n)!} \frac{(\mu T)^n}{K^n} (1 - \mu T/K)^{K-n}. \tag{5.3}$$

Izračunajmo sedaj limito tega izraza:

(5.4)

$$\begin{aligned} \lim_{K\to\infty} P(N = n) &= \frac{(\mu T)^n}{n!} \lim_{K\to\infty} \frac{K(K-1)...(K-n+1)(K-n)!}{K^n (K-n)!} \frac{(1-\mu T/K)^K}{(1-\mu T/K)^n} \\ &= \frac{(\mu T)^n}{n!} \lim_{K\to\infty} \frac{K(K-1)...(K-n+1)}{K^n} \frac{(1-\mu T/K)^K}{(1-\mu T/K)^n}. \end{aligned}$$

Prvi ulomek v limiti ima v števcu en koeficient enak K^n, vsi ostali koeficienti pa imajo stopnjo manjšo od n, zato bodo ti imeli v limiti vrednost enako 0. Ostane nam torej samo $\lim_{K\to\infty} \frac{K^n}{K^n}$, kar pa je enako 1. Za izraz $(1-\mu T/K)^K$ vemo, da ima v limiti, ko $K \to \infty$, vrednost enako $e^{-\mu T}$. Ostane nam samo še $(1 - \mu T/K)^n$, ki pa ima prav tako limito enako 1, saj je $\lim_{K\to\infty} \mu T/K = 0$. Če poenostavimo sedaj zgornji izraz dobimo:

$$\lim_{K\to\infty} P(N = n) = \frac{(\mu T)^n}{n!} e^{-\mu T}. \tag{5.5}$$

Opazimo lahko, da ima v limiti slučajna spremenljivka N znano Poissonovo porazdelitev s parametrom $\lambda = \mu T$. Ta model bi si želeli še malo bolj razširiti. Kot prvo bi radi, da ima lahko slučajna spremenljivka I_k še kakšne druge vrednosti poleg 0 in 1. Zato sedaj izberemo bolj primerno specifikacijo, ki izhaja iz zadnjega dobljenega rezultata za N:

$$P(I_k = 0) = 1 - \mu h + o(h), \quad P(I_k = 1) = \mu h + o(h), \quad P(I_k > 1) = o(h),$$

kjer $o(h)$ označuje izraz, za katerega velja

$$\frac{o(h)}{h} \to 0, \quad \text{ko } h \to 0. \tag{5.6}$$

Te majhne vrednosti $o(h)$ v limiti nimajo vpliva. Dokažimo to:

Dokaz. Izraz $\frac{o(h)}{h} \to 0$ lahko ekvivalentno zapišemo kot $Ko(1/K)$, ko $K \to \infty$, saj je $h = T/K$. Označimo z A dogodek, da so vsi $I_1, ..., I_K$ enaki 0 ali 1. Posledično potem komplement A^c označuje dogodek, da ima vsaj eden od $I_1, ..., I_K$ vrednost večjo od 1. Potem velja

$$\begin{aligned} P(A^c) = P(I_1 > 1 \cup ... \cup I_K > 1) &\leq \sum_{k=1}^{K} P(I_k > 1) = K \cdot o(h) \\ &= K \cdot o(1/K) \to 0, \end{aligned} \tag{5.7}$$

iz česar sledi, da $P(A^c) \to 0$ in po isti logiki dobimo $P(A) \to 1$. Po osnovni formuli o verjetnosti lahko zapišemo

$$P(N = n) = P(N = n|A)P(A) + P(N = n|A^c)P(A^c). \tag{5.8}$$

Z uporabo prejšnjega rezultata lahko ugotovimo, da velja $P(N = n) - P(N = n|A) \to 0$. Zadostuje, da izračunamo limito $P(N = n|A)$, za to verjetnost pa vemo, da ima binomsko porazdelitev z verjetnostjo uspeha:

$$\begin{aligned} p = \frac{P(I_k = 1)}{P(I_k = 0) + P(I_k = 1)} &= \frac{\mu h + o(h)}{1 - \mu h + o(h) + \mu h + o(h)} \\ &= \mu h + o(h) = \frac{\mu T}{K} + o(1/K). \end{aligned} \tag{5.9}$$

Če zapišemo verjetnostno funkcijo s tem parametrom p dobimo:

$$P(N = n|A) = \frac{K!}{n!(K-n)!} \frac{(\mu T + Ko(1/K))^n}{K^n} (1 - \mu T/K + o(1/K))^{K-n}. \tag{5.10}$$

Limito tega izraza dobimo na podoben način kot smo prišli do rezultata v (5.5). Izraz $\frac{(\mu T + Ko(1/K))^n}{n!}$ gre v limiti proti $\frac{(\mu T)^n}{n!}$, za izraz $\frac{K!}{(K-n)!K^n}$ pa smo že prej videli, da ima v limiti vrednost 1. Ostaneta nam samo še $(1 - \mu T/K + o(1/K))^K$ in $\frac{1}{(1-\mu T/K+o(1/K))^n}$. Poglejmo, kako bi izračunali limito prvega izraza. Tukaj si pomagamo s Taylorjevo vrsto logaritemske funkcije

$$\log(1 + x) = x - x^2/2 + x^3/3 + ... \tag{5.11}$$

V tej vrsti pišimo namesto x vrednost $-\mu T/K + o(1/K)$:

$$\begin{aligned} \log(1 - \mu T/K + o(1/K))^K &= K \log(1 - \mu T/K + o(1/K)) = -\mu T + Ko(1/K)) \\ &\to -\mu T, \quad \text{ko } K \to \infty, \end{aligned} \tag{5.12}$$

kar pomeni, da je limita $(1 - \mu T/K + o(1/K))^K$ enaka $e^{-\mu T}$. Za drugi izraz pa ugotovimo, da je v limiti njegova vrednost 1, saj gresta oba izraza $\mu T/K$ in $o(1/K)$ proti 0. Zapišemo lahko:

$$P(N = n|A) = \frac{(\mu T)^n}{n!} e^{-\mu T}. \tag{5.13}$$

Tako smo dobili povsem enake limite kot v primeru, kjer smo predpostavili za I_k vrednosti enake zgolj 1 ali 0. Vrednosti $o(h)$ torej v limit ne vplivajo na rezultat, saj vedno dobimo Poissonovo porazdelitev. □

To dejstvo, da majhne količine $o(h)$ ne vplivajo na rezultat, ni uporabno samo iz matematičnega stališča. Denimo, da imamo portfelj J zavarovalniških polic. Potem lahko na njih gledamo kot na J neodvisnih Poissonovih procesov. Označimo z μ_j intenzivnost pojavljanja škod v polici j in $\mathbf{I}_k$ naj označuje število vseh škod iz našega portfelja, ki so se zgodile v časovnem intervalu k. Ugotovimo naslednje:

$$P(\mathbf{I}_k = 0) = \prod_{j=1}^{J}(1 - \mu_j h) \quad \text{ter} \quad P(\mathbf{I}_k = 1) = \sum_{j=1}^{J}\Big(\mu_j h \prod_{i\neq j}(1 - \mu_i h)\Big). \tag{5.14}$$

Leva stran predstavlja verjetnost, da pri nobeni polici ni prišlo do škodnega zahtevka, zato imamo produkt verjetnosti preko vseh polic. Desna stran pa je verjetnost, da se je zgodila natanko ena škoda. To pomeni, da se je pri določeni polici zgodil škodni zahtevek, pri preostalih policah pa do tega ni smelo priti. Ker se ena škoda lahko pripeti v katerikoli izmed J - tih polic imamo vsoto od 1 do J. Ta dva izraza lahko še nekoliko poenostavimo. Poglejmo si, kaj dobimo, če izberemo število polic $J = 3$:

$$\begin{aligned} P(\mathbf{I}_k = 0) &= (1 - \mu_1 h) \cdot (1 - \mu_2 h) \cdot (1 - \mu_3 h) \\ &= (1 - \mu_2 h - \mu_1 h + \mu_1\mu_2 h^2)(1 - \mu_3 h) \\ &= 1 - \mu_3 h - \mu_2 h + \mu_2\mu_3 h^2 - \mu_1 h + \mu_1\mu_3 h^2 + \mu_1\mu_2 h^2 - \mu_1\mu_2\mu_3 h^3 \\ &= 1 - \Big(\sum_{j=1}^{3}\mu_j\Big)h + o(h). \end{aligned} \tag{5.15}$$

Člene s potenco h^2 in višje lahko izpustimo in jih zapišemo kot majhne količine $o(h)$. Podobno naredimo še za verjetnost, da se zgodi natančno en škodni dogodek in dobimo:

$$P(\mathbf{I}_k = 1) = \Big(\sum_{j=1}^{3}\mu_j\Big)h + o(h). \tag{5.16}$$

To ugotovitev lahko posplošimo na celoten portfelj J - tih polic:

$$P(\mathbf{I}_k = 0) = 1 - \Big(\sum_{j=1}^{J}\mu_j\Big)h + o(h) \quad \text{ter} \quad P(\mathbf{I}_k = 1) = \Big(\sum_{j=1}^{J}\mu_j\Big)h + o(h). \tag{5.17}$$

Opazimo, da dobimo za portfelj z zaporedjem $\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_K$ povsem enak model, kot smo ga dobili, ko smo opazovali samo posamezne police $I_1, I_2, ..., I_K$. Razlika je le v tem, da smo pri policah imeli zgolj eno intenzivnost μ, tukaj pa to intenzivnost predstavlja vsota intenzivnosti vseh polic v portfelju - $\mu_1 + \mu_2 + ... + \mu_J$. Iz tega sledi, da je število škod $\mathbf{N}$ v portfelju porazdeljeno po Poissonu s parametrom

$$\lambda = (\mu_1 + \mu_2 + ... + \mu_J)T = J\bar{\mu}T, \quad \text{kjer} \quad \bar{\mu} = (\mu_1 + \mu_2 + ... + \mu_J)/J.$$

Torej, kadar opazujemo določen portfelj, je pomembno samo povprečje intenzivnosti pojavljanja škod pri posameznih policah, ki ta portfelj sestavljajo.

Zaenkrat smo privzeli, da ima vsaka polica svojo konstantno intenzivnost. Temu pa vedno ni tako, saj se lahko s časom intenzivnost pojavljanja škod pri določeni polici tudi spreminja. V tem primeru moramo postaviti nekoliko drugačen model.

5.1. **Spreminanje intenzivnosti skozi čas.**

Intenzivnosti pojavljanja škod pogosto niso konstanta skozi celotno obdobje. To je lepo vidno na primeru avtomobilskega zavarovanja, kjer je število škodnih dogodkov povezano z letnim časom. Tako se več nesreč običajno zgodi pozimi, ko so spolzke ceste. Prav tako je večje število nesreč v deževnih obdobjih, kot pa poleti, ko je suho. V matematiki ta problem rešimo tako, da je intenzivnost pojavljanja škod sedaj neka funkcija časa $\mu = \mu(t)$. Števila škod v posameznih časovnih intervalih $I_1, ..., I_K$ so sedaj odvisna od različnih intenzivnosti $\mu_1, ..., \mu_K$, kjer je $\mu_k = \mu(t_k)$ za $k = 1, ..., K$. Celotno število škod $N = I_1 + ... + I_K$ je potem tudi odvisno od časa. Predstavljamo si lahko, da N sedaj dobimo na isti način, kot če bi imeli K različnih zavarovalnih polic v časovnem intervalu h in to potem seštejemo po vseh časih $k = 1, ..., K$. To nam ne spremeni porazdelitve slučajne spremenljivke N, zato ima še vedno Poissonovo porazdelitev s parametrom:

$$\lambda = h \sum_{k=1}^{K} \mu_k \to \int_0^T \mu(t)dt, \qquad \text{ko} \qquad h \to 0. \tag{5.18}$$

Zgoraj imamo limito Riemannove vsote, kar pa je po definiciji integral. Parameter slučajne spremenljivke N lahko zapišemo tudi kot

$$\lambda = T\bar{\mu}, \qquad \text{kjer} \qquad \bar{\mu} = \frac{1}{T} \int_0^T \mu(t)dt. \tag{5.19}$$

Vidimo, da je tudi v tem primeru parameter λ odvisen od povprečja intenzivnosti. Razlika od prejšnjega parametra pri konstantnih intenzivnostih je ta, da je tukaj povprečje odvisno od časa.

5.2. **Ocenjevanje intenzivnosti na podlagi zgodovinskih podatkov.**

V praksi za ocenjevanje parametra μ uporabimo zgodovinske podatke, ki nam povejo, koliko škod se je v preteklosti zgodilo. Označimo z $n_1, ..., n_J$ število škod, ki se nanašajo na J polic. Vsaka polica se sklene za določeno obdobje. Čase, ki predstavljajo to obdobje pri posameznih policah bomo označili s $T_1, ..., T_J$. Najbolj običajna in tudi najbolj preprosta ocena za intenzivnost pojavljanja škod je naslednja:

$$\hat{\mu} = \frac{n_1 + ... + n_J}{A}, \qquad \text{kjer} \qquad A = T_1 + ... + T_J. \tag{5.20}$$

Dobimo jo torej tako, da delimo celotno število škod z vsoto obdobij, za katera so bile sklenjene police. Za to cenilko velja, da je nepristranska, torej $E(\hat{\mu}) = \mu$. Standardni odklon te cenilke je odvisen od celotnega časovnega obdobja A in je enak $sd(\hat{\mu}) = \sqrt{\frac{\mu}{A}}$.

Tukaj velja omeniti še eno posebno lastnost. Kadar imamo opravka z zavarovalnimi policami, ki so bila sklenjena za ista obdobja - T_j so enaki za vse $j = 1, ..., J$, je dobro izračunati $D = s^2/\bar{n}$. V tem primeru s^2 označuje vzorčno varianco, $\bar{n}$ pa vzorčno upanje. Količnik D, ki ga dobimo, imenujemo koeficient razpršenosti in nam lahko veliko razkrije. Njegova uporabnost je v tem, da niha okoli ena, če so podatki neodvisni in enako porazdeljeni po Poissonu. To sledi iz tega, ker ima Poissonova porazdelitev enako upanje in varianco. V primeru, da je ta koeficient močno različen od ena, sta možnosti dve. Bodisi intenzivnosti pojavljanja škod niso enake, bodisi so podatki porazdeljeni drugače in ne po Poissonu.

Primer 5.1. Poglejmo si na primeru, kako bi dobili intenzivnost pojavljanja škod. Denimo, da opazujemo 23.589 zavarovalniških polic, ki so bile vse sklenjene z namenom avtomobilskega zavarovanja in za obdobje enega leta. To pomeni, da je v našem primeru $J = 23.589$, vredosti $T_1, ..., T_J$ pa so vse enake 1. V spodnji tabeli so predstavljeni podatki o številu škod:

Število škod	0	1	2	3	4	5	6
Število polic	20.592	2.651	297	41	7	0	1

Vsota vseh škod je torej

$$n_1 + ... + n_J = 0 \cdot 20.592 + 1 \cdot 2.651 + 2 \cdot 297 + 3 \cdot 41 + 4 \cdot 7 + 5 \cdot 0 + 6 \cdot 1 = 3.402.$$

Tako dobimo, da je cenilka intenzivnosti enaka

$$\hat{\mu} = \frac{3.402}{23.589} = 0,1442 = 14,4\%,$$

z ocenjenim standardnim odklonom

$$sd(\hat{\mu}) = \sqrt{\frac{0,1442}{23.589}} = 0,00247 = 0,25\%.$$

Rezultat nam pove, da približno en avto od sedmih zavarovanih vsako leto povzroči nesrečo. Na tem primeru preverimo še, koliko znaša količnik D:

$$D = \frac{s^2}{\bar{n}} = \frac{0.1638699}{0.1442198} = 1.136252.$$

Kljub temu, da je rezultat blizu 1, ne moremo z gotovostjo trditi, da gre za Poissonovo porazdelitev.

5.3. **Intenzivnost μ kot slučajna spremenljivka.**

V prejšnjem razdelku smo že omenili, da intenzivnosti niso vedno konstantne skozi čas. Takrat smo privzeli za intenzivnost funkcijo odvisno od časa. Predvsem pri avtomobilskem zavarovanju pa obstajajo še drugi dejavniki, ki vplivajo na pojave škod. Eden izmed njih je zagotovo starost voznika, poleg tega pa lahko izpostavimo še spol zavarovanca. Seveda pa nista to edini dve lastnosti. Vozniki se med seboj razlikujejo tudi po načinu vožnje. Nekateri so bolj agresivni, drugi bolj previdni. Enostavno lahko rečemo, da so nekateri vozniki boljši, drugi pa malo slabši. To so lastnosti, ki jih zavarovalnica težko pridobi od vsakega posameznika. Zato v tem primeru ravnamo tako, da se intenzivnost μ spremeni v slučajno spremenljivko. V matematiki se to odraža tako, da smo dobili stohastične procese. Za modele sedaj predpostavimo pogojne porazdelitve glede na intenzivnost in sicer

$$N|\mu \sim \text{Poisson}(\mu T) \quad \text{in} \quad N|\mu \sim \text{Poisson}(J\mu T). \tag{5.21}$$

Leva stran prikazuje porazdelitev v primeru določene zavarovalne police, desna stran pa porazdelitev na nivoju portfelja. Kot smo že velikokrat opazili, v primeru portfelja zgolj pomnožimo parameter μT s številom polic J v portfelju.

Ker je intenzivnost μ postala slučajna spremenljivka, moramo problem opredeliti nekoliko drugače. Sedaj ne moremo eksplicitno določiti, da bo intenzivnost enaka določeni vrednosti. Zato označimo povprečno vrednost $\xi = E[\mu]$ in standardni odklon $\sigma = sd(\mu)$. Za slučajno spremenljivko $N|\mu$ smo ugotovili, da ustreza Poissonovi porazdelitvi s parametrom μT, zato velja $E[N|\mu] = Var[N|\mu] = \mu T$. Z uporabo pravila pogojnega upanja lahko izrazimo $E[N]$:

$$E[N] = E\big[E[N|\mu]\big] = E[\mu T] = \xi T. \tag{5.22}$$

Na podoben način lahko z uporabo formule za pogojno varianco izračunamo $Var(N)$:

$$\begin{aligned} Var[N] &= E\big[Var[N|\mu]\big] + Var\big[E[N|\mu]\big] = E[\mu T] + Var[\mu T] \\ &= \xi T + \sigma^2 T^2. \end{aligned} \tag{5.23}$$

Opazimo, da v tem primeru velja $E[N] < Var[N]$, zato slučajna spremenljivka N ni več porazdeljena po Poissonu, saj bi moralo veljati $E[N] = Var[N]$. Model za intenzivnost μ poskušamo poiskati z mešanjem porazdelitev in z uporabo spodnje relacije:

$$P[N = n] = \int_0^\infty P(N = n|\mu) g(\mu) d\mu, \tag{5.24}$$

kjer $g(\mu)$ označuje gostoto porazdelitve slučajne spremenljivke μ. Zgornja enakost sledi iz osnovne formule o verjetnosti. Ko drobimo interval h na vedno manjše dele dobimo točke $\mu_i = ih$ za $i = 1, 2, ...$ in zgornji integral je približno enak vsoti $\sum_i P(N = n|\mu_i)P(\mu = \mu_i)$, kar pa vemo iz teorije verjetnosti, da je ravno $P(N = n)$. Zelo pogosta izbira za porazdelitveno gostoto $g(\mu)$ je gamma porazdelitev, kar bom v nadaljevanju tudi predstavil. Druga možnost, ki se še uporablja, pa je log - normalna porazdelitev. Vendar v tem primeru ne dobimo tako lepega rezultata za porazdelitev slučajne spremenljivke N, kot v primeru gamma porazdelitve.

5.4. **Ocenjevanje upanja in variance intenzivnosti μ.**

Oceni za ξ in σ lahko dobimo iz zgodovinskih podatkov, zato nam ni potrebno natančno poznati gostote porazdelitve $g(\mu)$ slučajne spremenljivke μ. Ponovno označimo z $n_1, ..., n_J$ število škod, ki ustrezajo J - tim policam, s $T_1, ..., T_J$ pa časovna obdobja, ki pripradajo določenim policam. Intenzivnost j - tega posameznika μ_j lahko ocenimo kot

$$\hat{\mu}_j = n_j / T_j. \tag{5.25}$$

S to oceno sicer veliko tvegamo, saj je lahko v resnici intenzivnost precej drugačna od te ocene. Vendar, če to uporabimo na nivoju portfelja, dobimo veliko bolj sigurno oceno. Ena izmed rešitev za oceno upanja intenzivnosti portfelja je naslednja

$$\hat{\xi} = \sum_{j=1}^{J} w_j \hat{\mu}_j, \qquad \text{kjer je} \qquad w_j = \frac{T_j}{\sum_{i=1}^{J} T_i}, \tag{5.26}$$

ocena za varianco pa je

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{J} w_j(\hat{\mu}_j - \hat{\xi})^2 - c}{1 - \sum_{j=1}^{J} w_j^2}, \qquad \text{kjer je} \qquad c = \frac{(J-1)\hat{\xi}}{\sum_{j=1}^{J} T_j}. \tag{5.27}$$

Trditev 5.2. *Cenilki* (5.26) *ter* (5.27) *za upanje in varianco sta nepristranski. Velja še:*

$$\lim_{J\to\infty} \hat{\xi} = \xi \qquad in \qquad \lim_{J\to\infty} \hat{\sigma}^2 = \sigma^2.$$

Dokaz. Oceni za $\xi = E(\mu)$ in $\sigma = sd(\mu)$ dobimo na podlagi $\hat{\mu_1}, ..., \hat{\mu_J}$, kjer je $\hat{\mu_j} = N_j/T_j$. Vemo, da je $N_j|\mu_j \sim Poisson(\mu_j T_j)$, iz česar dobimo:

$$E[\hat{\mu}_j|\mu_j] = E[N_j/T_j|\mu_j] = \tfrac{1}{T_j} \cdot \mu_j T_j = \mu_j,$$
$$Var[\hat{\mu}_j|\mu_j] = Var[N_j/T_j|\mu_j] = \tfrac{1}{T_j^2} \cdot \mu_j T_j = \mu_j/T_j.$$

Uporabimo pravilo za pogojno upanje in formulo za pogojno varianco, da dobimo:

$$E[\hat{\mu}_j] = E\big[E[\hat{\mu}_j|\mu_j]\big] = E[\mu_j] = \xi,$$
$$Var[\hat{\mu}_j] = E\big[Var[\hat{\mu}_j|\mu_j]\big] + Var\big[E[\hat{\mu}_j|\mu_j]\big] = E[\mu_j/T_j] + Var[\mu_j] = \tfrac{\xi}{T_j} + \sigma^2.$$

Ocena za upanje ξ je bila

$$\hat{\xi} = w_1\hat{\mu_1} + ... + w_J\hat{\mu_J}, \text{ kjer je } w_j = \tfrac{T_j}{T_1+...+T_J}.$$

Vidimo, da za uteži w_j velja, da je $w_1+...+w_J = 1$. Z upoštevanjem tega izračunajmo $E[\hat{\xi}]$:

$$E[\hat{\xi}] = w_1E[\hat{\mu}_1] + ... + w_JE[\hat{\mu}_J] = \xi(w_1 + ... + w_J) = \xi. \tag{5.28}$$

Dobili smo, da je $\hat{\xi}$ nepristranska cenilka za upanje ξ. Varianca te cenilke je

$$Var(\hat{\xi}) = \sum_{j=1}^{J} w_j^2 Var(\hat{\mu}_j) = \sum_{j=1}^{J} w_j^2(\sigma^2 + \frac{\xi}{T_j}). \tag{5.29}$$

Poglejmo si sedaj še cenilko za varianco. Njen izraz je

$$\hat{\sigma}^2 = \frac{Q - c}{1 - \sum_{j=1}^{J} w_j^2}, \qquad \text{kjer je} \qquad c = \frac{(J-1)\hat{\xi}}{\sum_{j=1}^{J} T_j} \qquad \text{in} \qquad Q = \sum_{j=1}^{J} w_j(\hat{\mu}_j - \hat{\xi})^2. \tag{5.30}$$

Izraz za Q lahko malo preoblikujemo:

$$\begin{aligned} Q &= \sum_{j=1}^{J} w_j(\hat{\mu}_j - \xi)^2 - 2\sum_{j=1}^{J} w_j(\hat{\mu}_j - \xi)(\hat{\xi} - \xi) + \sum_{j=1}^{J} w_j(\hat{\xi} - \xi)^2 \\ &= \sum_{j=1}^{J} w_j(\hat{\mu}_j - \xi)^2 - 2(\hat{\xi} - \xi)\Big(\sum_{j=1}^{J} w_j\hat{\mu}_j - \xi\sum_{j=1}^{J} w_j\Big) + (\hat{\xi} - \xi)^2\sum_{j=1}^{J} w_j \\ &= \sum_{j=1}^{J} w_j(\hat{\mu}_j - \xi)^2 - 2(\hat{\xi} - \xi)(\hat{\xi} - \xi) + (\hat{\xi} - \xi)^2 \\ &= \sum_{j=1}^{J} w_j(\hat{\mu}_j - \xi)^2 - (\hat{\xi} - \xi)^2. \end{aligned} \tag{5.31}$$

Tukaj smo upoštevali, da je $\sum_{j=1}^{J} w_j \hat{\mu}_j = \hat{\xi}$ in $\sum_{j=1}^{J} w_j = 1$. Iz tega sledi

$$\begin{aligned} E[Q] &= \sum_{j=1}^{J} w_j E(\hat{\mu}_j - \xi)^2 - E(\hat{\xi} - \xi)^2 \\ &= \sum_{j=1}^{J} w_j Var(\hat{\mu}_j) - Var(\hat{\xi}) \\ &= \sum_{j=1}^{J} w_j (\sigma^2 + \frac{\xi}{T_j}) - \sum_{j=1}^{J} w_j^2 (\sigma^2 + \frac{\xi}{T_j}) \\ &= \sigma^2 \sum_{j=1}^{J} w_j + \xi \sum_{j=1}^{J} \frac{w_j}{T_j} - \sigma^2 \sum_{j=1}^{J} w_j^2 - \xi \sum_{j=1}^{J} \frac{w_j^2}{T_j} \\ &= \sigma^2 \big(1 - \sum_{j=1}^{J} w_j^2\big) + \xi \big(\sum_{j=1}^{J} \frac{w_j - w_j^2}{T_j}\big) \\ &= \sigma^2 \big(1 - \sum_{j=1}^{J} w_j^2\big) + \frac{(J-1)\xi}{\sum_{j=1}^{J} T_j}. \end{aligned} \tag{5.32}$$

Opazimo lahko, da je drugi člen v zadnjem izrazu ravno $E(c)$. Upoštevamo to in zapišemo:

$$E(\hat{\sigma}^2) = \frac{E(Q) - E(c)}{1 - \sum_{j=1}^{J} w_j^2} = \sigma^2. \tag{5.33}$$

Torej je $\hat{\sigma}$ res nepristranska cenilka za varianco σ. Pokazati moramo še, da sta ξ in σ limiti naših cenilk. To naredimo tako, da gledamo na časovna obdobja $T_1, ..., T_J$ kot na slučajni vzorec neodvisnih in enako porazdeljenih slučajnih spremenljivk. Dokazali smo že, da je $E[\hat{\mu}_j | T_j] = \xi$, s to razliko, da tukaj pogojujemo glede na slučajne spremenljivke T_j. Velja še $E[\hat{\mu}_j T_j | T_j] = \xi T_j$ in z uporabo pravila o pogojnem upanju dobimo:

$$E[\hat{\mu}_j T_j] = E\big[E[\hat{\mu}_j T_j | T_j]\big] = E[\xi T_j] = \xi T_j.$$

Po enačbi (5.26) je potem

$$\hat{\xi} = \frac{\frac{1}{J}\sum_{j=1}^{J} \hat{\mu}_j T_j}{\frac{1}{J}\sum_{j=1}^{J} T_j} \to \frac{E[\hat{\mu}_1 T_1]}{E[T_1]} = \frac{\xi E[T_1]}{E[T_1]} = \xi.$$

S tem smo dokazali $\hat{\xi} \to \xi$, kjer smo pri izračunu limite upoštevali šibki zakon velikih števil. Na podoben način sledi izračun za $\hat{\sigma}^2$:

$$Q = \frac{\frac{1}{J}\sum_{j=1}^{J} T_j (\hat{\mu}_j - \xi)^2}{\frac{1}{J}\sum_{j=1}^{J} T_j} - (\hat{\xi} - \xi)^2 \to \frac{(\hat{\mu}_1 - \xi)^2 E[T_1]}{E[T_1]} - 0,$$

kjer smo zopet upoštevali šibki zakon velikih števil. To limito lahko še poenostavimo:

$$E[T_1(\hat{\mu}_1 - \xi)^2 | T_1] = T_1 Var[\hat{\mu}_1 | T_1] = T_1(\sigma^2 + \xi/T_1) = T_1\sigma^2 + \xi.$$

Z uporabo pravila o pogojnem upanju izračunamo

$$E\big[E[T_1(\hat{\mu}_1 - \xi)^2 | T_1]\big] = E[T_1(\hat{\mu}_1 - \xi)^2] = \sigma^2 E[T_1] + \xi.$$

Iz tega sledi, da $Q \to \sigma^2 + \xi/E[T_1]$ in opazimo lahko, da je:

$$c = \frac{(J-1)\xi}{\sum_{j=1}^{J} T_j} = \frac{\frac{J-1}{J}\xi}{\frac{1}{J}\sum_{j=1}^{J} T_j} \to \frac{\xi}{E[T_1]}$$

in

$$J \sum_{j=1}^{J} w_j^2 = \frac{\frac{1}{J}\sum_{j=1}^{J} T_j^2}{(\frac{1}{J}\sum_{j=1}^{J} T_j)^2} \to \frac{E[T_1^2]}{E[T_1]^2},$$

kar pomeni, da $\sum_j w_j^2 \to 0$. Torej lahko zapišemo:

$$\hat{\sigma}^2 = \frac{Q-c}{1-\sum_{j=1}^{J} w_j^2} \to \frac{\sigma^2 + \xi/E[T_1] - \xi/E[T_1]}{1-0} = \sigma^2. \tag{5.34}$$

Vdimo, da res velja $\hat{\sigma}^2 \to \sigma^2$, ko $J \to \infty$. □

Pri tem je potrebno omeniti, da cenilka za varianco ni nujno pozitivna. Če se zgodi, da za $\hat{\sigma}^2$ dobimo negativno vrednost, lahko sklepamo, da nihanje intenzivnosti μ v portfelju ni pomembno in velja $\hat{\sigma} = 0$. Te dobljene ocene predstavljajo osnovo in jih lahko uporabimo pri različnih modelih. Jaz bom sedaj predstavil model, ki temelji na negativni binomski porazdelitvi.

5.5. **Negativni binomski model.**

Negatvni binomski model se običajno uporablja takrat, ko imamo opravka s prerazpršenostjo podatkov. To pomeni, da je vzorčna varianca večja od vzorčnega upanja. Najbolj pogost način, s katerim poskušamo oceniti intenzivnost pojavljanja škod μ je, da zanjo predpostavimo naslednji izraz:

$$\mu = \xi G, \qquad \text{kjer je} \qquad G \sim Gamma(\alpha). \tag{5.35}$$

Tukaj G označuje standardno Gamma porazdelitev z upanjem 1. Parameter α nam pove podatek o tem, s kakšno negotovostjo μ niha okoli upanja ξ. Bolj nazorno to pomeni:

$$E[\mu] = \xi E[G] = \xi, \qquad \text{in} \qquad sd(\mu) = \xi/\sqrt{\alpha}. \tag{5.36}$$

V primeru, ko $\alpha \to \infty$ opazimo, da $sd(\mu) \to 0$. V limiti torej dobimo Poissonov model s konstanto intenzivnostjo μ. V naslednjem koraku nas zanima, kakšna je porazdelitev števila škod N. V razdelku 4.3 smo predpostavili, da za slučajno spremenljivko $N|\mu$ velja, da je porazdeljena po Poissonu s parametrom μT. Porazdelitev za N dobimo s pomočjo mešanja porazdelitev. Ker smo v (5.36) privzeli, da μ vsebuje Gamma porazdelitev, moramo za funkcijo $g(\mu)$ izbrati porazdelitveno gostoto Gamma porazdelitve. Ti dve dejstvi nam data:

$$P(N = n|\mu) = \frac{(\mu T)^n}{n!} e^{-\mu T} \qquad \text{in} \qquad g(\mu) = \frac{(\alpha/\xi)^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\mu\alpha/\xi} \ .$$

Z uporabo enačbe (5.24) in izpeljave v (3.14) vemo, da ima potem slučajna spremenljivka N negativno binomsko porazdelitev. V (3.14) moramo parameter M zamenjati z α, parameter p pa je v tem primeru enak:

$$p = \frac{\alpha}{\alpha + \xi T}.$$

Ta model je velikokrat zelo uporaben ravno zaradi tega, ker smo za N dobili znano porazdelitev. V (3.23) in (3.26) smo izpeljali upanje in varianco negativne binomske

porazdelitve. Upoštevamo, da je v tem primeru $M = \alpha$ in dobimo izraza za $E[N]$ ter $sd[N]$:

$$
\begin{aligned}
E[N] &= \frac{\alpha(1-p)}{p} = \frac{\alpha\xi T/(\alpha+\xi T)}{\alpha/(\alpha+\xi T)} = \xi T, \\
sd[N] &= \sqrt{Var[N]} = \sqrt{\frac{\alpha(1-p)}{p^2}} = \sqrt{\frac{\alpha\xi T/(\alpha+\xi T)}{\alpha^2/(\alpha+\xi T)^2}} = \sqrt{\frac{\xi T(\alpha+\xi T)}{\alpha}}.
\end{aligned} \tag{5.37}
$$

Za negativno binomsko porazdelitev pa velja še ena lastnost, ki je zelo uporabna.

Trditev 5.3. *Naj bodo $N_1, ..., N_J$ neodvisne slučajne spremenljivke porazdeljene negativno binomsko s parametroma ξ in α. Potem velja:*

$$N_1 + ... + N_J \sim nbin(J\xi, J\alpha). \tag{5.38}$$

Ta lastnost velja prav tako za Poissonove slučajne spremenljivke. Dokažimo sedaj to trditev.

Dokaz. Dokazati moramo, da ima vsota negativno binomsko prazdeljenih slučajnih spremenljivk tudi negativno binomsko porazdelitev. Označimo $\mathbf{N} = N_1 + ... + N_J$, kjer so $N_1, ...N_J$ neodvisne in porazdeljene $nbin(\xi, \alpha)$. Za vsak N_j pa vemo, da velja

$$N_j|\mu_j \sim Poisson(\mu_j T), \qquad \text{kjer je} \qquad \mu_j = \xi G_j, \qquad G_j \sim Gamma(\alpha).$$

Ker je vsota neodvisnih Poissonivh slučajnih spremenljvik porazdeljena po Poissonu, lahko ugotovimo, da je $\mathbf{N} = N_1 + ... + N_J$ pri pogoju $\mu_1, ...\mu_J$ porazdeljena po Poissonu s parametrom:

$$\theta = \mu_1 + ... + \mu_J = \xi(G_1 + ... + G_J) = J\xi\bar{G}, \qquad \text{kjer} \qquad \bar{G} = (G_1 + ... + G_J)/J \tag{5.39}$$

$\bar{G}$ označuje povprečje J - tih slučajnih spremenljivk, porazdeljenih $Gamma(\alpha)$. Iz tega sledi, da je $\bar{G}$ porazdeljena $Gamma(J\alpha)$. To nam da naslednje rezultate:

$$\mathbf{N}|\theta \sim Poisson(\theta T), \qquad \text{kjer je} \qquad \theta = (J\xi)\bar{G} \qquad \text{in} \qquad \bar{G} \sim Gamma(J\alpha). \tag{5.40}$$

Porazdelitev za $\mathbf{N}$ dobimo z mešanjem Poissonove porazdelitve s parametrom θT in Gamma porazdelitve s parametrom $J\alpha$. Pri tem si pomagamo z uporabo formule (5.24) in na koncu pridemo do že dokazane ugotovitve:

$$\mathbf{N} \sim NBin(J\xi, J\alpha). \tag{5.41}$$

S tem smo dokazali trditev. □

5.6. **Aproksimiranje negativne binomske porazdelitve.**

Sedaj, ko imamo postavljen model in smo dobili, da je število škodnih zahtevkov porazdeljeno negativno binomsko, lahko ocenimo neznana parametra ξ in α. S pomočjo teh ocen bomo lahko ugotovili, kakšna je pričakovana intenzivnost pojavljanja škod. Pri ocenah parametrov si pomagamo s cenilkama definiranima v (5.26) in (5.27). Za obe cenilki smo že dokazali, da sta nepristranski. Za oceno parametra ξ tako vzamemo kar vrednost (5.26). Pri oceni za varianco σ^2, pa si pomagamo še z izrazom (5.36) in dobimo

$$\hat{\sigma} = \hat{\xi}/\sqrt{\hat{\alpha}}, \qquad \text{oziroma} \qquad \hat{\alpha} = \hat{\xi}^2/\hat{\sigma}^2. \tag{5.42}$$

Če se zgodi, da je $\hat{\sigma} = 0$, lahko to interpretiramo kot posledica neskončnosti parametra α. V tem primeru dobimo zopet Poissonov model.

Druga možnost pa je, da do ocen pridemo s pomočjo funkcije verjetja. V primeru 3.4 smo že izpeljali funkcijo verjetja za negativno binomsko porazdelitev. Takrat smo imeli parametra M in p, sedaj pa je M zamenjan z α. Pri tem moramo biti pozorni tudi na to, da je parameter $p = \alpha/(\alpha + \xi T_j)$ odvisen od posameznega časa T_j, za katerega je bila sklenjena j - ta polica. Uporabimo verjetnostno funkcijo v (3.14), kjer število n zamenjamo s posameznim številom škod n_j. Zato se izraz za funkcijo verjetja malo spremeni, ker nekaterih členov ne moremo izpostaviti pred vsoto:

$$\begin{aligned} L(\xi, \alpha) &= \prod_{j=1}^{J} \binom{\alpha + n_j - 1}{n_j} (\frac{\alpha}{\alpha + \xi T_j})^{\alpha} (\frac{\xi T_j}{\alpha + \xi T_j})^{n_j} \\ &= \Big(\frac{\alpha^{\alpha}}{(\alpha - 1)!}\Big)^{J} \prod_{i=1}^{J} \frac{(\alpha + n_j - 1)!}{n_j!} (\frac{1}{\alpha + \xi T_j})^{\alpha} (\frac{\xi T_j}{\alpha + \xi T_j})^{n_j}. \end{aligned} \tag{5.43}$$

Logaritemska funkcija se potem glasi:

$$\begin{aligned} l(\xi, \alpha) &= J\Big(\alpha \log(\alpha) - \log(\Gamma(\alpha))\Big) + \sum_{j=1}^{J} \log(\Gamma(\alpha + n_j)) \\ &\quad + \sum_{j=1}^{J} \Big[n_j \log(\xi T_j) - (n_j + \alpha) \log(\alpha + \xi T_j)\Big] - \sum_{j=1}^{J} \log(n_j). \end{aligned} \tag{5.44}$$

Zadnja vsota v logaritemski funkciji verjetja je konstanta, zato bi jo lahko tudi izpustili, saj ne vsebuje parametra α ali ξ in na rezultat tako nima vpliva.
Predstavljene metode lahko uporabimo v praksi, da ocenimo pričakovano število škod. Tudi jaz bom sedaj uporabil negativni binomski model pri analizi podatkov iz zavarovalništva.

5.7. **Primer iz zavarovalništva.**

Dobljeni negativni binomski model bomo sedaj uporabili na realnih podatkih iz zavarovalništva. Žal mi ni uspelo pridobiti točno določenih podatkov od neke konkretne zavarovalnice, zato sem moral podatke dobiti na drugačen način. Program R vsebuje veliko uporabnih paketov, med drugim tudi paket *insuranceData*. V njem so zbrani primeri zavarovalnih polic, s pripadajočim število škod in še nekaterimi ostalimi podatki, ki so zelo uporabni pri modeliranju števila škod. Jaz bom predstavil samo en del tega paketa, ki vsebuje podatke o avtomobilskem zavarovanju in se imenuje *dataCar*. Podatki so sestavljeni na podlagi enoletnih zavarovalniških polic, ki so bile sklenjene med letoma 2004 in 2005. Skupno število polic je 67.856 in vsebujejo naslednje podatke:

- *veh_value*: vrednost vozila - izražena v 10.000 - ih ameriških dolarjev,
- *exposure*: izpostavljenost škodi - vrednosti med 0 in 1,
- *clm*: indikator pojava škode - ima vrednost 0, v primeru, da do škode ni prišlo in 1 v nasprotnem primeru,
- *numclaims*: število nastalih škod,
- *claimcst*0: višina nastale škode,
- *veh_body*: tip vozila,
- *veh_age*: starost vozila, razdeljena v štiri starostne kategorije (od 1 - 4),
- *gender*: spol zavarovanca - F - ženska, M - moški,
- *area*: območje, ki je razdeljeno v 6 razredov (A,B,C,D,E,F),
- *agecat*: starost zavarovanca, ki vsebuje vrednosti od 1 do 6 - 1 pomeni najmlajšo starostno skupino, 6 pa najstarejšo.

Spodaj je primer nekaterih vrstic iz podatkov:

	veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat
15147	2.200	0.85420945	1	4	6521.550	STNWG	3	F	B	3
54370	2.950	0.91444216	1	4	2356.410	STNWG	3	F	E	5
12616	0.720	0.70362765	1	3	2677.664	HBACK	4	M	A	1
67674	2.830	0.90896646	1	3	1173.270	PANVN	2	M	B	2
9855	1.330	0.66803559	1	3	3097.860	TRUCK	2	M	C	2

Slika 4. Primeri podatkov iz paketa *insuranceData* (*dataCar*)

Vidimo, da imamo vse podatke, ki jih potrebujemo za analizo s pomočjo negativnega binomskega modela. Poglejmo si, do kakšnega rezultata pridemo. Največje število škod, ki se je pripetilo posameznemu zavarovancu je bilo 4. Razporeditev števila škod je prikazana v spodnji tabeli:

Število škod	0	1	2	3	4
Število polic	63232	4333	271	18	2

Definirajmo sedaj naše podatke tako, kot smo jih v razdelku 4.4. Število zavarovalnih polic je $J = 67.856$, števila škod $n_1, ..., n_J$ ustrezajo stolpcu *numclaims*, časi $T_1, ..., T_J$ pa so vsi enaki 1, ker vzorec vsebuje samo police, sklenjene za obdobje enega leta. Iz teh podatkov lahko najprej izračunamo posamezne ocene za intenzivnosti $\hat{\mu}_j$, $j = 1, ..., 67.856$:

```
#podatki
library(insuranceData)
podatki <- read.table("C:/Šola/Magistrska/avt.txt")

st_skod <- podatki[,4]
J <- length(podatki[,1])
#imamo samo enoletne police:
T_ji <- rep(1,J)
u_streha <- st_skod/T_ji
```

Da bi dobili pričakovano intenzivnost in njeno varianco glede na celoten portfelj, moramo izračunati parametra $\hat{\xi}$ in $\hat{\sigma}^2$ tako, kot smo definirali v (5.26) ter (5.27). Psevdokoda je predstavljena spodaj:

```
#oceni za upanje in varianco:
utezi_w <- 1/(sum(T_ji))
ksi_streha <- sum(utezi_w*u_streha)

Q <- sum(utezi_w*(u_streha - ksi_streha)^2)
c <- (J-1)*ksi_streha/(sum(T_ji))
sigma_streha <- (Q - c)/(1 - sum(utezi_w^2))
```

Videli smo, da si pri negativnem binomskem modelu pomagamo s temi ocenami. Ocena za ξ je kar zgornji $\hat{\xi}$, pri oceni za α pa si pomagamo s $\hat{\sigma}^2$. Dobimo:

```
#negativni binomski model:
ksi <- ksi_streha
alfa_streha <- ksi_streha^2/sigma_streha

pr <- alfa_streha/(alfa_streha + ksi)
```

Dobljeni oceni sta naslednji:

$$\hat{\xi} = 0,07275701 = 7,28\%, \qquad \hat{\alpha} = 1,14.$$

Povprečna intenzivnost pojavljanja škod je torej približno 7%, kar pomeni, da eden izmed 14 avtov vsako leto povzroči škodo. Dobljeni rezultat nam ne pomeni veliko dokler ne preverimo, kakšna je njegova natančnost. Zato naredimo sedaj χ^2 test. Na podlagi ocen $\hat{\xi}$ in $\hat{\alpha}$ lahko izračunamo pričakovane frekvence dobljenega modela:

```
##izračun frekvenc:
st_skod_tabela <- table(st_skod)
vrednosti_skod <- as.numeric(names(st_skod_tabela))
#frekvence modela:
pric_frekvence <- dnbinom(vrednosti_skod,size=alfa_streha,prob=pr)*J
```

Dobimo naslednje vrednosti:

Število škod	0	1	2	3	4
Število polic	63235	4325	278	17	1

Že na podlagi teh rezultatov lahko sklepamo, da smo na dobri poti. Kljub temu izračunajmo še testno statistiko $\hat{\chi}^2$:

```
##test:
emp_frekvence <- c()
for (i in 1:length(st_skod_tabela)){
  emp_frekvence[i] <- st_skod_tabela[[i]]
}

X2_test <- sum((emp_frekvence - pric_frekvence)^2 / pric_frekvence)
st_test <- length(emp_frekvence) - 2 - 1
h_st_test <- qchisq(1-0.05,st_test)
```

Test sem opravil pri stopnji zaupanja $\alpha = 0,05$. V tem primeru dobimo $\hat{\chi}^2 = 0,9696926$ in $h_{2,0,05} = 5.991465$. Ker velja $\hat{\chi}^2 < h_{2,0,05}$, ne zavrnemo ničelne hipoteze in lahko sklepamo, da je opazovan vzorec res porazdeljen negativno binomsko s parametroma $\hat{\alpha}$ in $p = \hat{\alpha}/\hat{\alpha} + \hat{\xi}$. Tukaj smo test izvedli povsem po definiciji. Lahko bi uporabili tudi vgrajeno funkcijo *chisq.test* in bi dobili:

```
chisq.test(emp_frekvence,p = pric_frekvence,rescale.p = TRUE)

        Chi-squared test for given probabilities

data:  emp_frekvence
X-squared = 0.9697, df = 4, p-value = 0.9144
```

Vidimo, da pridemo do istega zaključka in ne zavrnemo ničelne hipoteze, ker je p - vrednost $0,9144 > 0.05$. Intenzivnost $\hat{\mu}$ je tako slučajna spremenljivka, ki ustreza izrazu:

$$\hat{\mu} = 0,073 \cdot Gamma(1, 14).$$

Sedaj lahko izračunamo J intenzivnosti na podlagi zgornjega izraza. Dobimo:

```
#intenzivnosti:
u_model <- ksi*pgamma(1:J,alfa_streha))
mean(u_model)
0.07275626
```

Vidimo, da imajo generirane intenzivnosti upanje enako $0,073$ in res velja $E(\hat{\mu}) = \hat{\xi}$. Poskušajmo sedaj pridobiti ocene parametrov ξ in α še s pomočjo metode največjega verjetja. Pri tem uporabimo logaritemsko funkcijo verjetja, ki smo jo definirali v

(5.44). Paziti moramo na to, da funkcija *mle* zahteva negativno vrednost logaritemske funkcije verjetja, torej $-l(\xi, \alpha)$:

```
#metoda največjega verjetja:
maxlike <- function(oce_alfa,oce_ksi){
  x_nov <- st_skod
  maxlike <- -sum(lfactorial(oce_alfa+x_nov-1)) +
            J*(lfactorial(oce_alfa-1) - oce_alfa*log(oce_alfa))
            - log(oce_ksi)*sum(x_nov) + log(oce_alfa + oce_ksi)
            *sum(x_nov+oce_alfa) + sum(log(x_nov))
  return (maxlike)
}

ocenaMLE_primer <- mle(minuslog = maxlike, start =
                         list(oce_alfa = 2, oce_ksi = 0.5))
summary(ocenaMLE_primer)
Maximum likelihood estimation

Call:
mle(minuslogl = maxlike, start = list(oce_alfa = 2, oce_ksi = 0.5))

Coefficients:
           Estimate  Std. Error
oce_alfa 1.15906307 0.143417676
oce_ksi  0.07276176 0.001067348

-2 log L: 35646.46
```

Za začetni vrednosti parametrov sem izbral $\alpha = 2$ in $\xi = 0,5$. Pri teh izbranih vrednostih dobimo naslednji oceni:

$$\hat{\xi} = 0.07276176 = 7,28\%, \qquad \hat{\alpha} = 1.15906307.$$

Vidimo, da dobimo oceno za ξ skoraj identično kot pri prejšnji oceni, parametra α pa se čisto malo razlikujeta. Vprašanje, ki se nam pojavi je, katera ocena je boljša. Poglejmo, kaj v tem primeru vrne χ^2 test. Pričakovane frekvence, dobljene na podlagi teh dveh ocen so naslednje:

```
alfa_MLE <- 1.15906307
ksi_MLE <- 0.07276176
p_MLE <- alfa_MLE/(alfa_MLE + ksi_MLE)
pric_frekvence_MLE <- dnbinom(vrednosti_skod,size = alfa_MLE,
                              prob = p_MLE)*J
round(pric_frekvence_MLE)
63233  4329   276    17     1
```

Dobimo zelo podobne vrednosti za števila škod kot pri prejšnji oceni. Poglejmo si, kakšno p - vrednost vrne χ^2 test v tem primeru:

```
chisq.test(emp_frekvence,p = pric_frekvence_MLE,rescale.p = T)

        Chi-squared test for given probabilities

data:  emp_frekvence
X-squared = 0.9836, df = 4, p-value = 0.9123
```

Vidimo, da je p - vrednost enaka $0,9123$ in je malo manjša od prejšnje, ki je bila $0,9144$. Zato bi se lahko na podlagi tega odločili, da izberemo tiste ocene z višjo p - vrednostjo. V tem primeru torej tiste ocene, ki smo jih dobili v prvem primeru. Ker sta si obe oceni zelo blizu, lahko sklepamo, da je aproksimacija z negativnim binomskim modelom smiselna. To potrdita tudi visoki p - vrednosti v obeh primerih. Tako lahko sedaj uporabimo dobljene rezultate tudi za napovedi pričakovanega števila škod v prihodnje. Denimo, da pričakujemo v naslednjem letu 50.000 sklenjenih zavarovalnih polic. Potem preprosto generiramo naključni vzorec velikosti 50.000, ki je porazdeljen negativno binomsko s pripadajočima parametroma $\hat{\alpha} = 0,073$ in $p = \hat{\alpha}/(\hat{\alpha} + \hat{\xi}) = 0,94$:

```
st_polic <- 50000
pric_stevilo_skod <- rnbinom(st_polic,alfa_streha,pr)
table(pric_stevilo_skod)
```

Pričakovano število škod, ki ga dobimo, je prikazano v spodnji tabeli:

Število škod	0	1	2	3	4
Število polic	46602	3197	193	7	1

Na koncu smo dobili ocene za pričakovano število škod, ki dajejo zavarovalnici uporabno informacijo, saj se lahko malo pripravijo na bodoče nastale škode in si oblikujejo rezerve. V svojem delu sem predstavil postopke, s katerimi si pomagamo pri ocenjevanju števila škod. Seveda to ni edini dejavnik, ki zavarovalnice zanima. Ko enkrat poznamo pričakovano število škod, je naravno vprašanje, koliko znaša višina teh nastalih škod. To je prav tako zelo pomemben podatek, za katerega pa moramo zopet uporabiti nove metode, ki zahtevajo podrobnejšo analizo. Zato jaz v svojem delu tega področja nisem posebej predstavljal.

Literatura

[1] E. Bolviken, *Computation and Modelling in Insurance and Finance*, (2014) strani 279–313.

[2] A. Hazra, *An Exact Kolmogorov-Smirnov Test for the Negative Binomial Distribution with Unknown Probability of Success*, Journal of Statistics, Vol.**1** (2013) strani 1–14.

[3] *Negative binomial regression — R data analysis examples*, [ogled 22. 6. 2018], dostopno na `https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/`.

[4] F. Herzog, *Statistical methods*, (2013).

[5] V. Ricci, *Fitting distributions with R*, (2005).

[6] V. Pacáková, D. Zapletal, *Mixture Distributions in Modelling of Insurance Losses*, (2013).