UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ZOVŠEOBECNENÉ GEOMETRICKÉ ROZDELENIE

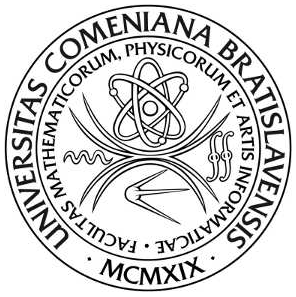
Bakalárska práca

2021

Tatiana Hrabovská

UNIVERZITA KOMENSKÉHO V BRATISLAVE

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



ZOVŠEOBECNENÉ GEOMETRICKÉ ROZDELENIE

Bakalárska práca

Študijný program:	Poistná matematika
Študijný odbor:	1113 Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Školiteľ:	Mgr. Lívia Rosová, PhD.

Bratislava, 2021

Tatiana Hrabovská

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Tatiana Hrabovská
Študijný program: poistná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Zovšeobecnené geometrické rozdelenie
Generalized geometric distribution

Anotácia: Diskrétne dáta sa objavujú v mnohých oblastiach ako napríklad v poisťovníctve, medicíne, ekonómii, lingvistike a iných.
Pri ich modelovaní sa často vyžaduje, aby modely vedeli zohľadniť aj prípady, kedy je napríklad disperzia väčšia ako stredná hodnota
alebo sa v dátach vyskytuje nadmerné množtvo núl.
Jeden zo spôsobov, ako sa vysporiadať s takýmito dátami, je prispôsobiť a rozšíriť už existujúce modely.
Táto práca sa bude venovať zovšeobecneným geometrickým rozdeleniam,
ktoré môžu byť interpretované ako diskrétne analógie zovšeobecnených exponenciálnych rozdelení.

Vedúci: Mgr. Lívia Rosová, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Dátum zadania: 08.10.2020

Dátum schválenia: 14.10.2020 doc. RNDr. Katarína Janková, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Abstrakt

Diskrétne náhodné premenné majú svoje využitie v rôznych oblastiach, konkrétne nápomocné môžu byť napríklad v oblasti poisťovníctva pre modelovanie počtov poistných nárokov. Pri výbere rozdelenia, ktorým môžeme tieto počty modelovať, môže byť užitočné pozrieť sa na zovšeobecnenia už známych rozdelení. Bakalárska práca je zameraná na zovšeobecnené geometrické rozdelenia, pričom konkrétne sa venuje trom zovšeobecneniam. Práca ukazuje prepojenie medzi danými zovšeobecnenými geometrickými rozdeleniami a im prislúchajúcimi zovšeobecneniami exponenciálneho rozdelenia. Ďalej pre každé z rozdelení uvádza jeho základné vlastnosti a kladie dôraz na odhad parametrov spomínaných rozdelení. Pre odhadovanie parametrov je používaná metóda maximálnej vierohodnosti. Získané teoretické poznatky a samotné rozdelenia sú následne demonštrované na reálnych sadách dát ako možné aplikácie daných rozdelení.

Kľúčové slová: geometrické rozdelenie, zovšeobecnené geometrické rozdelenie, diskrétne rozdelenie pravdepodobnosti, odhad parametrov, modelovanie.

Abstract

Discrete random variables have their uses in different areas, for example they can be helpful in the insurance for modelling counts of insurance claims. When we are choosing the distribution, which we want to use in our model, it might be useful to take into account the generalizations of already known distribution. This bachelor thesis is focused on generalized geometric distributions, while specifically, it deals with three generalizations. This thesis shows the connection between defined generalized geometric distributions and their corresponding generalizations of the exponential distribution. Further for each of the distributions, it states their basic properties and emphasis is on the estimation of parameters of the mentioned distributions. Maximum likelihood estimation method is used to estimate the parameters. In the end are all the generalized distributions applied on the real data sets and they are compared compared with each other.

Keywords: geometric distribution, generalized geometric distribution, discrete probability distribution, estimation of parameters, modelling.

Predhovor

Som rada, že som si vybrala práve *Zovšeobecnené geometrické rozdelenie,* nakoľko som seba aj čitateľov chcela oboznámiť s novými typmi rozdelení a táto téma mi priniesla mnoho nových poznatkov z oblasti teórie pravdepodobnosti, konkrétne o diskrétnych náhodných premenných. Samotné geometrické rozdelenie som mala možnosť spoznať už na predmetoch, ktoré sú vyučované na fakulte, avšak všetky zovšeobecnenia tohto rozdelenia boli pre mňa nové a veľmi prínosné. Všetky ciele, ktoré som si stanovila na začiatku tvorby bakalárskej práce, sa mi aj podarilo v dostatočnej miere naplniť. Som veľmi vďačná, že mi práca pomohla ujasniť si diskretizáciu náhodných premenných, a v neposlednom rade určite posilnila moje myslenie a praktické schopnosti programovania v softvéri R. Dúfam, že moja bakalárska práca bude prínosná aj pre ostatných čitateľov.

Touto cestou by som tiež chcela vyjadriť vďaku vedúcej bakalárskej práce Mgr. Lívii Rosovej, PhD. za jej ochotu, podporu, cenné rady a pripomienky počas tvorby práce. Taktiež by som sa chcela poďakovať všetkým pedagógom za ich vedomosti, ktoré mi doposiaľ odovzdali. Ďakujem aj rodine a priateľom za ich podporu počas celého doterajšieho štúdia.

Obsah

Úvod

Geometrické rozdelenie je diskrétne rozdelenie pravdepodobnosti, ktoré má svoje vyžitie v mnohých oblastiach, ako napríklad v aktuárstve. Je to síce pomerne jednoduché rozdelenie, avšak v rôznych oblastiach môže poslúžiť pri tvorbe počiatočných jednoduchých modelov. Problém pri odhadovaní parametrov a modelovaní dát môže nastať, keď sa hodnota nula nadobúda s vysokou pravdepodobnosťou. Práve vtedy má zmysel uvažovať nad geometrickým rozdelením, ktoré sa s tým vie dobre vysporiadať. Pri riešení komplexnejších problémoch nemusí byť geometrické rozdelenie postačujúce a má zmysel uvažovať nad rôznymi jeho zovšeobecneniami. Využitie týchto zovšeobecnených rozdelení môžeme nájsť napríklad v oblasti neživotného poistenia pri modelovaní počtu nárokov, resp. počtu poistných plnení.

V prvej kapitole bakalárskej práce najskôr v krátkosti opisujeme klasické geometrické rozdelenie a následne predstavujeme odhadovanie parametrov pomocou metódy maximálnej vierohodnosti. Ďalšie tri kapitoly, ktoré tvoria hlavnú časť teoretickej časti, sú venované definovaniu zovšeobecnených geometrických rozdelení. Postupne prezentujeme tri zovšeobecnenia geometrického rozdelenia. Pre každé zo zovšeobecnených rozdelení ukazujeme prepojenie daného geometrického zovšeobecnenia s prislúchajúcim exponenciálnym zovšeobecneným rozdelením. Okrem toho spomíname aj strednú hodnotu a disperziu konkrétneho zovšeobecnenia. Pri každom rozdelení uvádzame odhady parametrov, ktoré sme získali pomocou metódy maximálnej vierohodnosti z vygenerovaných hodnôt z daného rozdelenia.

Cieľom bakalárskej práce je nielen vytvoriť prehľad vlastností konkrétnych zovšeobecnených rozdelení, ale aj poukázať na využitie spomínaných rozdelení v rôznych oblastiach. Práve preto je posledná piata kapitola zameraná na praktické využitie jednotlivých rozdelení. V tejto časti sú všetky rozdelenia porovnané na dvoch reálnych sadách dát. Ako spomíname vyššie, často sú tieto rozdelenia používané v oblasti poistnej matematiky, preto ako jednu možnú aplikáciu sme vybrali modelovanie dát práve z tohto prostredia. Aby sme ukázali využitie daných rozdelení v rôznych a celkom odlišných oblastiach, ako druhú možnú aplikáciu sme zvolili modelovanie dát z prostredia medicíny. Pri modelovaní dát opäť odhadujeme parametre jednotlivých rozdelení a pomocou chí-kvadrát hodnoty porovnávame, ako rozdelenia sedia na konkrétnu sadu dát.

1 Geometrické rozdelenie

Úvodná kapitola je venovaná klasickému geometrickému rozdeleniu s jedným parametrom. Je v nej tiež zahrnutý princíp odhadovania parametrov pomocou metódy maximálnej vierohodnosti.

Definícia 1.1 *Nech náhodná premenná X má pravdepodobnostnú funkciu*

$$p_X(x) = \mathrm{P}(X = x) = pq^x,$$

kde $x = 0, 1, 2, \ldots$, $0 < q < 1$ a $q = 1 - p$. Potom X je náhodná premenná, ktorá má geometrické rozdelenie s parametrom q. Používame označenie $X \sim Geo(q)$.

Veta 1.1 *Nech X je náhodná premenná, ktorá má geometrické rozdelenie, potom distribučná funkcia X zodpovedá tvaru*

$$F_X(x) = \mathrm{P}(X < x) = \begin{cases} 0 & ak\ x \leq 0, \\ \sum_{i=0}^{\lfloor x \rfloor} pq^i & ak\ x > 0, \end{cases}$$

ak $0 < q < 1$ a $p = 1 - q$.

Dôkaz.

Dôkaz vety vyplýva priamo z definície distribučnej funkcie diskrétnej náhodnej premennej. □

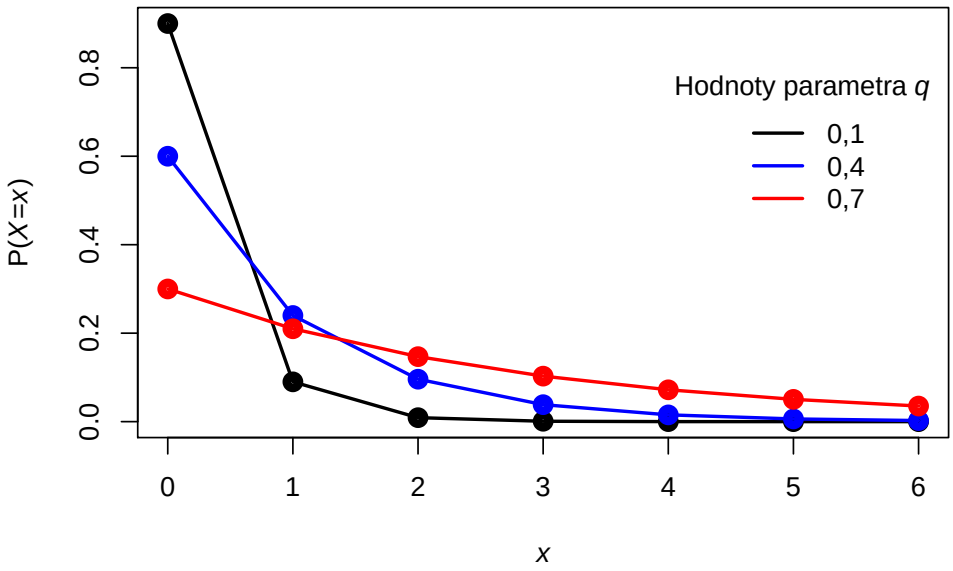
Na Obr. 1 môžeme vidieť pravdepodobnostnú funkciu geometrického rozdelenia pre rôzne hodnoty parametra q, ktorá má vo všetkých prípadoch klesajúci tvar. Toto pozorovanie nie je dielom náhody, pretože pravdepodobnostná funkcia geometrického rozdelenia má vždy klesajúci charakter.

Ďalej pre strednú hodnotu a disperziu tohto rozdelenia platí

$$\mathrm{E}(X) = \frac{q}{1-q},$$

$$\mathrm{D}(X) = \frac{q}{(1-q)^2},$$

kde $0 < q < 1$.



Obr. 1: Pravdepodobnostná funkcia $Geo(q)$ vykreslená pre rôzne hodnoty parametra q.

(zdroj: vlastné spracovanie)

1.1 Metóda maximálnej vierohodnosti

Známou a často používanou metódou odhadovania parametrov je práve metóda maximálnej vierohodnosti, ktorú označujeme aj pomocou skratky MLE z anglického názvu *maximum likelihood estimation*. Princíp metódy je založený na vierohodnostnej funkcii, ktorú v našej práci označujeme $L(\boldsymbol{x}, \boldsymbol{\theta})$.

Definícia 1.2 *Majme náhodný výber $X_1, X_2, ..., X_n$, ktorý má rozdelenie pravdepodobnosti s hustotou $f(x_i, \boldsymbol{\theta})$, kde $\boldsymbol{\theta}$ je neznámy parameter, alebo vektor parametrov tohto rozdelenia. Taktiež uvažujme jeho realizáciu $x_1, x_2, ..., x_n$. Potom funkciu vierohodnosti definujeme nasledovným vzťahom*

$$L(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}).$$

Metóda maximálnej vierohodnosti má za úlohu nájsť maximum vierohodnostnej funkcie, pričom za maximálne vierohodný odhad parametrov zoberie argument, v ktorom sa nadobúda maximum vierohodnostnej funkcie. To znamená

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{x}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}),$$

kde Θ je parametrický priestor pre $\boldsymbol{\theta}$.

Keďže vierohodnostná funkcia je definovaná ako súčin, spočítanie jej derivácie, pomocou ktorej môže byť nájdené maximum, je vo veľa prípadoch pomerne zložitá úloha. Preto sa často používa logaritmická vierohodnostná funkcia, ktorú ďalej označujeme $l(\boldsymbol{x}, \boldsymbol{\theta})$.

Definícia 1.3 *Za platnosti vyššie spomínaných predpokladov, logaritmus funkcie vierohodnosti je daný nasledujúcou rovnosťou*

$$l(\boldsymbol{x}, \boldsymbol{\theta}) = \ln L(\boldsymbol{x}, \boldsymbol{\theta}) = \ln \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i, \boldsymbol{\theta}).$$

Po použití logaritmickej vierohodnostnej funkcie sa vo veľa prípadoch zjednoduší úloha spočítania derivácie. Ak existujú parciálne derivácie danej funkcie, odhad parametrov je možné dostať ako riešenie sústavy rovníc, kedy sa derivácie rovnajú nule.

Pre odhadovanie parametrov je potrebné mať realizáciu náhodného výberu. Preto pre potreby tejto bakalárskej práce sme naprogramovali generátor náhodných hodnôt pre jednotlivé rozdelenia, ktorý sa zakladá na metóde inverznej transformácie a algoritmus pre túto funkciu môžeme nájsť napríklad v publikácií [6].

2 Transmutované geometrické rozdelenie

Transmutované geometrické rozdelenie (*transmuted geometric distribution*) autori článku [4] odvodili pomocou metódy transmutácie kvadratického rádu (*quadratic rank transmutation*). My toto odvodenie neuvádzame, zamerali sme sa na samotné rozdelenie pravdepodobnosti a ukázali sme vzťah medzi diskrétnym transmutovaným geometrickým rozdelením a jeho spojitou analógiou- zovšeobecneným exponenciálnym rozdelením. V neposlednom rade sme sa sústredili na generovanie realizácií a následný odhad parametrov, od ktorých toto rozdelenie závisí. Poznatky o spomínanom rozdelení sme čerpali z článku [4].

2.1 Definícia a vlastnosti transmutovaného geometrického rozdelenia

Transmutované geometrické rozdelenie je zovšeobecnením $Geo(q)$, takže je to diskrétne rozdelenie závislé od dvoch parametrov.

Definícia 2.1 *Nech Y je diskrétna náhodná premenná, ktorej pravdepodobnostná funkcia má tvar*

$$p_Y(y) = (1-\alpha)q^y(1-q) + \alpha(1-q^2)q^{2y}, \tag{1}$$

kde y = 0, 1, 2, ..., $0 < q < 1$ a $-1 < \alpha < 1$. Potom hovoríme, že Y má transmutované geometrické rozdelenie pravdepodobnosti.

Veta 2.1 *Nech Y je náhodná premenná, ktorá má transmutované geometrické rozdelenie, potom distribučná funkcia tohto rozdelenia pravdepodobnosti vyzerá*

$$F_Y(y) = \begin{cases} 0 & ak\ y \leq 0, \\ \sum_{i=0}^{\lfloor y \rfloor} (1-\alpha)q^i(1-q) + \alpha(1-q^2)q^{2i} & ak\ y > 0, \end{cases}$$

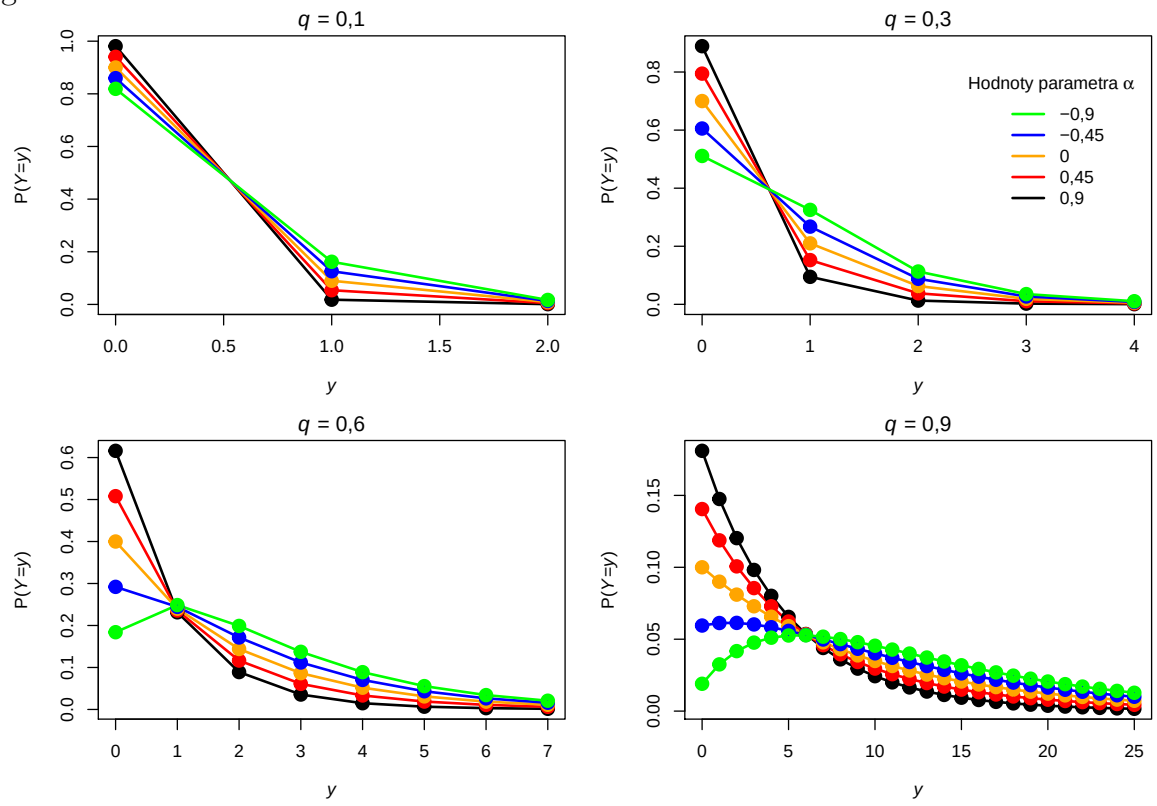
pre $0 < q < 1$ a $-1 < \alpha < 1$.

Dôkaz.

Dôkaz vety vyplýva priamo z definície distribučnej funkcie diskrétnej náhodnej premennej. □

Poznámka. Ak by sme parameter α položili rovný nule, dostali by sme klasické $Geo(q)$, ako ho už poznáme.

Pravdepodobnostnú funkciu tohto rozdelenia sme vykreslili pre rôzne kombinácie parametrov q a α na Obr. 2. Poznamenajme, že pre transmutované geometrické rozdelenie sme zachovali označenie, ktoré zaviedli autori článku [4]. Takže transmutované geometrické rozdelenie označujeme $TGD(q, \alpha)$ z anglického názvu *transmuted geometric distribution.*



Obr. 2: Grafické porovnanie pravdepodobnostnej funkcie $TGD(q, \alpha)$ pri rôznych parametroch q a α.
(*zdroj: vlastné spracovanie na základe* [4]).

Shaw a Buckley v roku 2009 v článku [7] uviedli zovšeobecnenie exponenciálneho rozdelenia pomocou hustoty

$$f_X(x) = (1-\alpha)\beta \mathrm{e}^{-\beta x} + 2\alpha\beta \mathrm{e}^{-2\beta x},$$

pre $x > 0$ s parametrami $\beta > 0$ a $-1 < \alpha < 1$. Distribučná funkcia takto definovaného rozdelenia je určená rovnicou

$$F_X(x) = (1-\alpha)(1-\mathrm{e}^{-\beta x}) + \alpha(1-\mathrm{e}^{-2\beta x}), \tag{2}$$

kde $x > 0$, $\beta > 0$ a $-1 < \alpha < 1$.

Veta 2.2 *Nech X je kladná spojitá náhodná premenná. Ak X má Shawovo-Buckleyho exponenciálne rozdelenie s parametrami $\beta > 0$ a $-1 < \alpha < 1$, tak jeho celá dolná časť má transmutované geometrické rozdelenie s parametrami $\mathrm{e}^{-\beta}$ a α.*

Dôkaz.

Nech $Y = \lfloor X \rfloor$, kde $\lfloor x \rfloor$ je funkcia celej dolnej časti reálneho čísla x, potom náhodnú premennú Y chápeme, ako diskrétnu analógiu spojitej náhodnej premennej X. Následne aplikujme vzťah $\mathrm{P}(Y = y) = F_X(y+1) - F_X(y)$ na Shawovo-Buckleyho exponenciálne rozdelenie.

$$\begin{aligned}
\mathrm{P}(Y = y) =& F_X(y+1) - F_X(y) \\
=& (1-\alpha)(1-\mathrm{e}^{-\beta(y+1)}) + \alpha(1-\mathrm{e}^{-2\beta(y+1)}) - (1-\alpha)(1-\mathrm{e}^{-\beta y}) \\
& - \alpha(1-\mathrm{e}^{-2\beta y}) \\
=& -\mathrm{e}^{-\beta y-\beta} + \alpha\mathrm{e}^{-\beta y-\beta} - \alpha\mathrm{e}^{-2\beta y-2\beta} + \mathrm{e}^{-\beta y} - \alpha\mathrm{e}^{-\beta y} + \alpha\mathrm{e}^{-2\beta y} \\
=& (-\mathrm{e}^{-\beta y}\mathrm{e}^{-\beta} + \alpha\mathrm{e}^{-\beta y}\mathrm{e}^{-\beta} + \mathrm{e}^{-\beta y} - \alpha\mathrm{e}^{-\beta y}) + (-\alpha\mathrm{e}^{-2\beta y}\mathrm{e}^{-2\beta} + \alpha\mathrm{e}^{-2\beta y}) \\
=& \mathrm{e}^{-\beta y}(1-\alpha)(1-\mathrm{e}^{-\beta}) + \alpha(1-\mathrm{e}^{-2\beta})\mathrm{e}^{-2\beta y}.
\end{aligned}$$

Dostali sme

$$\mathrm{P}(Y = y) = (1-\alpha)\mathrm{e}^{-\beta y}(1-\mathrm{e}^{-\beta}) + \alpha(1-\mathrm{e}^{-2\beta})\mathrm{e}^{-2\beta y},$$

pre $y = 0, 1, 2 ..., 0 < q < 1$ a $-1 < \alpha < 1$. Daná rovnica je pravdepodobnostná funkcia $TGD(q, \alpha)$, kde parameter $\alpha = \alpha$ a parameter q je rovný $\mathrm{e}^{-\beta}$, takže $Y \sim TGD(\mathrm{e}^{-\beta}, \alpha)$. □

Poznámka. Veta 2.2 dokázaná v článku [4] nie je, počas dokazovania sme však natrafili na nezrovnalosť v tomto článku. Autori definovali distribučnú funkciu Shawovho-Buckleyho exponenciálneho rozdelenia vzťahom

$$F_X(x) = (1+\alpha)(1-\mathrm{e}^{-\beta x}) - \alpha(1-\mathrm{e}^{-2\beta x})^2. \tag{3}$$

Keď sme však do vzorca $\mathrm{P}(Y = y) = F_X(y+1) - F_X(y)$ dosadili vzťah (3), zistili sme, že rovnosť neplatí, a preto sme sa pokúsili zrátať distribučnú funkciu Shawovho-Buckleyho exponenciálneho rozdelenia pomocou definície, a to

$$F_X(x) = \int_{-\infty}^{x} (1-\alpha)\beta\mathrm{e}^{-\beta u} + 2\alpha\beta\mathrm{e}^{-2\beta u}\mathrm{d}u.$$

Po zrátaní daného integrálu sme zistili, že distribučná funkcia Shawovho-Buckleyho exponenciálneho rozdelenia má tvar (2). Keď sme následne vzťah (2) aplikovali v dôkaze Vety 2.2, podarilo sa nám danú vetu dokázať.

Nech Y je náhodná premenná, ktorá má $TGD(q, \alpha)$ rozdelenie pravdepodobnosti. Pre strednú hodnotu náhodnej premennej Y platí, že

$$\mathrm{E}(Y) = \frac{q(1-\alpha)+q^2}{1-q^2}$$

a pre disperziu náhodnej premennej Y platí vzťah

$$\mathrm{D}(Y) = \frac{q(1-\alpha^2+q(1-\alpha^2+q(1-\alpha)+2))}{(1-q^2)^2},$$

pre dané parametre $0 < q < 1$ a $-1 < \alpha < 1$.

2.2 Odhad parametrov

V rámci tejto podkapitoly sme sa sústredili na odhady parametrov $TGD(q, \alpha)$. Najskôr sme spravili jednoduché prvotné odhady pre parametre tohto rozdelenia, a následne sme sa snažili použiť presnejšiu metódu odhadovania, konkrétne sme aplikovali metódu MLE, ktorej princíp sme popísali v podkapitole 1.1.

Pri prvotných odhadoch sme najskôr uvažovali rovnosti

$$p_Y(0) = (1-\alpha)(1-q) + \alpha(1-q^2),$$
$$p_Y(1) = (1-\alpha)q(1-q) + \alpha(1-q^2)q^2,$$

z ktorých sme dostali nasledujúcu sústavu rovníc

$$\alpha = \frac{qp_0 - p_1}{q(1-q)^2(1+q)},$$
$$q^3 + (p_0-1)q^2 + (p_0-1)q + 1 - p_0 - p_1 = 0,$$

kde označenie $\hat{p}_0$ zodpovedá odhadu $p_Y(0)$, ktorý sme odhadli ako pomer medzi početnosťou čísla nula v realizácii a celkovým rozsahom daného súboru, pre ktorý sme odhadovali parametre. Analogický význam má označenie $\hat{p}_1$. Prvotné odhady sme následne zrátali z danej sústavy rovníc pomocou softvéru R [5].

Pomocou vzťahu (1) odvoďme vierohodnostnú funkciu, ktorá vyzerá

$$L(\boldsymbol{y}, q, \alpha) = \prod_{i=1}^{n}((1-\alpha)q^{y_i}(1-q) + \alpha q^{2y_i}(1-q^2)).$$

Logaritmus funkcie vierohodnosti je určený nasledovným tvarom

$$l(\boldsymbol{y}, q, \alpha) = n\ln(1-q) + n\bar{y}\ln(q) + \sum_{i=1}^{n}\ln((1-\alpha) + \alpha q^{y_i}(1+q)),$$

ktorý sme prevzali z článku [4]. V oboch prípadoch $0 < q < 1$, $-1 < \alpha < 1$.

2.2.1 Simulácie

Pre odhadovanie parametrov sme si najskôr vygenerovali viacero dátových sád z $TGD(q, \alpha)$ s rôznymi parametrami. Prvotné odhady sme spočítali pre súbory veľkosti 10 000 hodnôt a zapísali sme ich do Tabuľky 1. Prvý stĺpec zaznamenáva veľkosť súboru, nasledujúce dva stĺpce zaznamenávajú skutočné hodnoty parametrov a v posledných štyroch stĺpcoch sú zapísané odhady. Prvotné odhady boli spočítané pomocou softvéru R [5], kde pre výpočet odhadu q sme aplikovali funkciu `cubic()` z balíka [9].

Tabuľka 1: Prvotné odhady parametrov $TGD(q, \alpha)$.

n	q	α	$\hat{q}$	$\hat{\alpha}$	$\hat{q}$	$\hat{\alpha}$
10 000	0,9	−0,9	0,9086	−0,8585	0,8903	−0,9471
10 000	0,4	0,7	0,3044	0,3597	0,3400	0,4968
10 000	0,6	0,3	0,5992	0,2850	0,6036	0,3082
10 000	0,1	−0,2	0,1037	−0,1551	0,0882	−0,3776

(*zdroj: vlastné spracovanie*)

Po spočítaní prvotných odhadov sme sa mohli pokúsiť aplikovať nejakú rafinovanejšiu metódu. Ako sme už spomínali vyššie, použili sme metódu MLE, samotné výsledky však boli spočítané numericky. Presnejšie v softvéri R sme aplikovali funkciu `optim()`, aby sme našli argument, v ktorom sa nadobúda maximum logaritmu funkcie vierohodnosti. Ako štartovacie body vo funkcii `optim()` sme zvolili vyššie spomínané prvotné odhady. Najskôr sme vygenerovali 10 hodnôt a vypočítali odhady $\hat{q}$ a $\hat{\alpha}$. Následne sme vygenerovali 10 000 hodnôt, pre ktoré sme takisto odhadli dané parametre, aby sme dostali informáciu o tom, aký presný je náš odhad pri veľkom súbore dát. Odhadnuté parametre q a α $TGD(q, \alpha)$ rozdelenia sú zapísané v Tabuľke 2. Prvý stĺpec Tabuľky 2 zaznamenáva, z akého veľkého súboru bol náš odhad urobený, v druhom a treťom stĺpci sú zaznamenané reálne hodnoty q a α. Do posledných troch blokov stĺpcov sme zapísali naše odhady pomocou MLE metódy, vždy pre daný rozsah súboru a pre rovnaké parametre, ktoré sú určené v predchádzajúcich stĺpcoch.

Pre hodnotu n rovnú 10 sme zistili, že naše odhady $\hat{q}$ a $\hat{\alpha}$ sú výrazne rozdielne od skutočných hodnôt q a α. Je to zapríčinené tým, že máme vygenerovaných len 10 hodnôt, čo je veľmi málo na to, aby sme dostali odhad blízky reálnym para-

metrom. Môžeme si všimnúť, že pre $n = 10\ 000$ sme s našimi odhadmi oveľa presnejší. Hlavne odhad $\hat{q}$ je vo väčšine prípadov veľmi blízky reálnemu q. Na druhej strane odhady $\hat{\alpha}$ sa v mnohých prípadoch výrazne líšia od skutočnej hodnoty α, aj keď máme súbor s veľkosťou $n = 10\ 000$ hodnôt.

Tabuľka 2: Odhady parametrov $TGD(q, \alpha)$ pomocou metódy MLE.

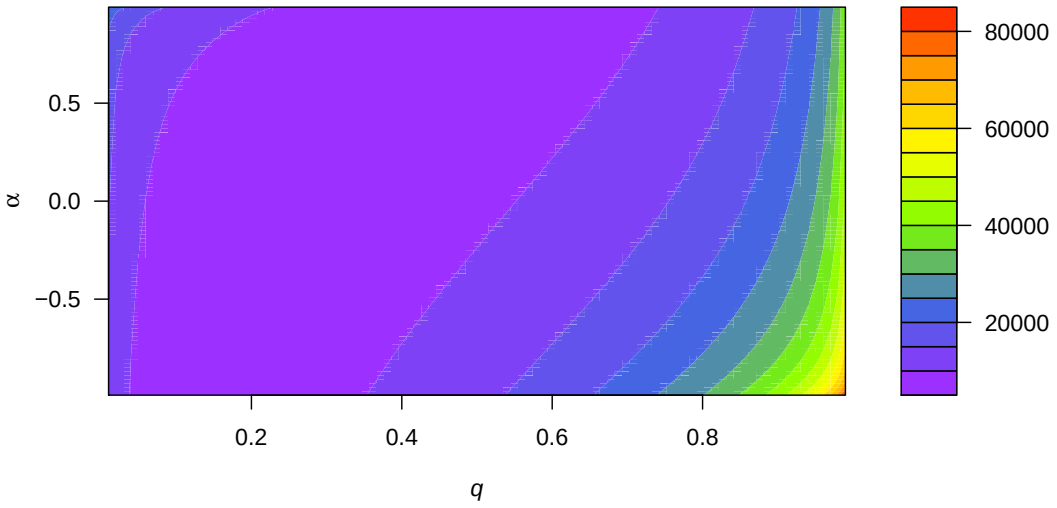
n	q	α	$\hat{q}$	$\hat{\alpha}$	$\hat{q}$	$\hat{\alpha}$	$\hat{q}$	$\hat{\alpha}$
10	0,9	−0,9	0,9081	−0,9900	0,8703	−0,9900	0,8739	−0,6921
10 000	0,9	−0,9	0,9017	−0,8974	0,8998	−0,9122	0,9006	−0,8854
10	0,4	0,7	0,4319	0,6063	0,2875	0,3435	0,5745	0,4095
10 000	0,4	0,7	0,4329	0,7935	0,4012	0,7128	0,3616	0,5653
10	0,6	0,3	0,6691	0,3060	0,6868	0,4935	0,6622	0,4330
10 000	0,6	0,3	0,6040	0,3410	0,5944	0,2593	0,6198	0,3643
10	0,1	−0,2	0,0951	−0,9900	0,2337	−0,3889	0,2337	−0,3889
10 000	0,1	−0,2	0,0871	−0,3254	0,1018	−0,2160	0,1071	−0,1044

(*zdroj: vlastné spracovanie*)

Nakoľko pre súbor veľkosti 10 000 hodnôt by sme očakávali presnejšie odhady, pokúšali sme sa zistiť, prečo sa nám v niektorých prípadoch nedarí parametere odhadovať podľa našich očakávaní. V softvéri R sme vykreslili logaritmus funkcie vierohodnosti prenásobený mínus jednotkou, presnejšie úrovňové množiny (takzvané vrstevnice) tejto funkcie. Logaritmus funkcie vierohodnosti sme prenásobili mínus jednotkou kvôli tomu, aby vykreslená funkcia presne zodpovedala funkcii, z ktorej sme počítali odhady parametrov. Realizácia, ktorá vstupovala do vykreslenej funkcie bola generovaná pri parametroch $q = 0{,}4$ a $\alpha = 0{,}7$ a obsahovala 10 000 hodnôt. Tieto parametre sme zvolili podľa Tabuľky 2, pretože sa zdá, že odhady pre tieto parametre sa najviac líšia od skutočných hodnôt.

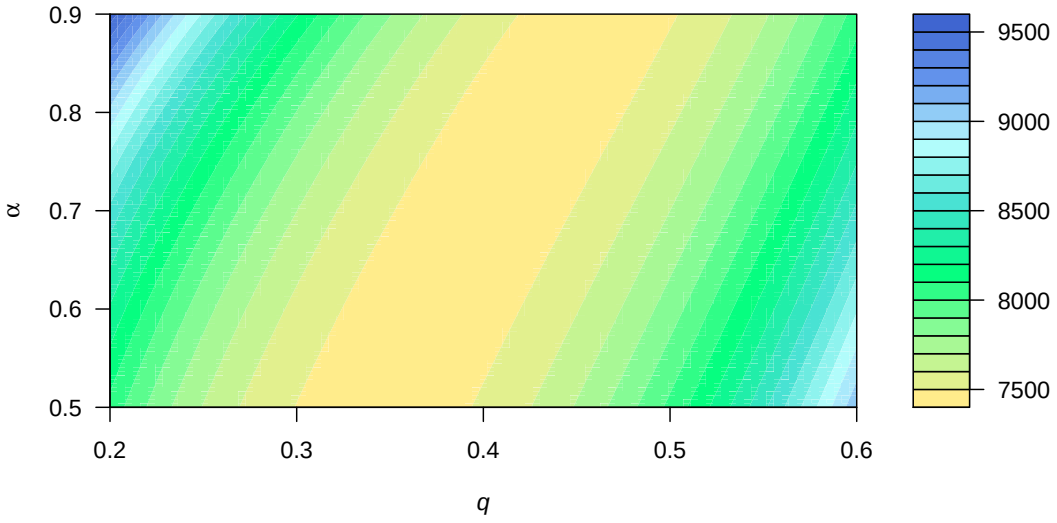
Ako prvé sme vykreslili vrstevnice danej funkcie na celom definičnom obore, konkrétne pre $0 < q < 1$ a $-1 < \alpha < 1$, zobrazené na Obr. 3. Podľa tohto obrázka sa môže zdať, že pre konkrétnu hodnotu parametra q sa parameter α môže rovnať rôznym hodnotám, pričom funkčná hodnota zostane takmer nemenná. Preto sme sa rozhodli pozrieť na túto funkciu bližšie a vykreslili sme ju pre určité okolie parametrov q a α na Obr. 4. Podľa tohto obrázka sme mohli potvrdiť, že keď $q = 0{,}4$ parameter α môže nadobúdať

rôzne hodnoty a funkčná hodnota zostane prakticky nemenná, preto môžu vzniknúť viditeľné odchýlky pri odhadovaní parametrov, aj keď veľkosť súboru je až 10 000 hodnôt.



Obr. 3: Graf úrovňových množín logaritmu funkcie vierohodnosti prenásobeného mínus jednotkou $TGD(q, \alpha)$ rozdelenia, kde realizácia bola generovaná pri parametroch $q = 0{,}4$ a $\alpha = 0{,}7$, vykreslený pre všetky hodnoty parametrov.

(*zdroj: vlastné spracovanie*)



Obr. 4: Graf úrovňových množín logaritmu funkcie vierohodnosti prenásobeného mínus jednotkou $TGD(q, \alpha)$ rozdelenia, kde realizácia bola generovaná pri parametroch $q = 0{,}4$ a $\alpha = 0{,}7$, vykreslený pre okolie skutočných parametrov.

(*zdroj: vlastné spracovanie*)

3 Gómez-Dénizovo geometrické rozdelenie

V tejto kapitole sme sa pozreli na ďalšie zovšeobecnené geometrické rozdelenie. Našu pozornosť sme nesústredili na to, ako presne toto rozdelenie vzniklo, ale zamerali sme sa na podobné vlastnosti ako v prípade $TGD(q, \alpha)$. Všetky potrebné informácie a vlastnosti sme čerpali z článku [3].

3.1 Definícia a vlastnosti Gómez-Dénizovho geometrického rozdelenia

Gómez-Dénizovo geometrické rozdelenie je ďalším zovšeobecnením $Geo(q)$, takže patrí k diskrétnym rozdeleniam pravdepodobnosti, je však závislé od dvoch parametrov.

Definícia 3.1 *Majme diskrétnu náhodnú premennú Y. Nech pre pravdepodobnostnú funkciu Y platí*

$$p_Y(y) = \frac{\alpha q^y (1-q)}{(1-\bar{\alpha}q^{y+1})(1-\bar{\alpha}q^y)},$$

kde $y = 0, 1, 2, \ldots$, $0 < q < 1$, $\alpha > 0$ a $\bar{\alpha} = 1 - \alpha$. Potom hovoríme, že Y má Gómez-Dénizovo geometrické rozdelenie pravdepodobnosti.

Veta 3.1 *Nech Y je náhodná premenná, ktorá má Gómez-Dénizovo geometrické rozdelenie. Distribučná funkcia tohto rozdelenia má tvar*

$$F_Y(y) = \begin{cases} 0 & ak\ y \leq 0, \\ \sum_{i=0}^{\lfloor y \rfloor} \frac{\alpha q^i (1-q)}{(1-\bar{\alpha}q^{i+1})(1-\bar{\alpha}q^i)} & ak\ y > 0, \end{cases}$$

pričom $0 < q < 1$, $\alpha > 0$ a $\bar{\alpha} = 1 - \alpha$.

Dôkaz.

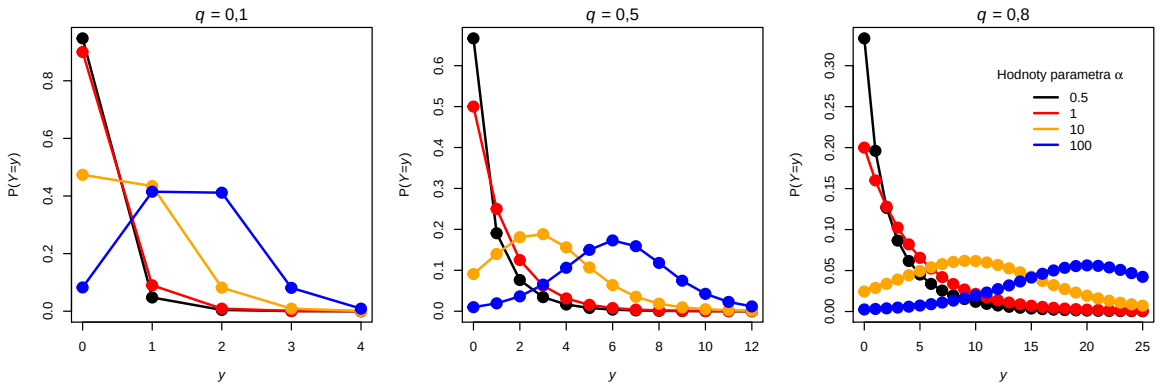
Dôkaz vety vyplýva priamo z definície distribučnej funkcie diskrétnej náhodnej premennej. □

Autor v článku [3] neuvádza špecifický názov daného rozdelenia, nazýva ho len ako zovšeobecnené geometrické rozdelenie. Aby sme toto rozdelenie v našej práci jednoznačne odlíšili od ostatných, nazývame ho Gómez-Dénizovo geometrické rozdelenie.

Pre tento názov sme sa rozhodli práve kvôli jeho autorovi, avšak pre rýchlejšiu orientáciu sme dodržali označenie, ktoré zaviedol sám autor. Preto využívame označenie $GG(q, \alpha)$ z anglického názvu *generalized geometric distribution.*

Poznámka. Môžeme si všimnúť, že ak by sme parameter α položili rovný jednej, tak z distribučnej funkcie $GG(q, \alpha)$ by sme dostali distribučnú funkciu $Geo(q)$.

Aby sme videli, aký tvar môže mať pravdepodobnostná funkcia $GG(q, \alpha)$ rozdelenia, vykreslili sme ju na Obr. 5 pre rôzne kombinácie parametrov q a α.



Obr. 5: Grafické porovnanie pravdepodobnostnej funkcie $GG(q, \alpha)$ pri rôznych parametroch q a α.
(*zdroj: vlastné spracovanie*)

Ako uvádza článok [3], v roku 1997 Marshall a Olkin predstavili zovšeobecnené exponenciálne rozdelenie, ktoré opísali pomocou nasledujúcej hustoty

$$f_X(x) = \frac{\alpha\beta \mathrm{e}^{-\beta x}}{(1 - \bar{\alpha}\mathrm{e}^{-\beta x})^2}$$

pre $x > 0$ a parametre $\beta > 0$, $\alpha > 0$ a $\bar{\alpha} = 1 - \alpha$.

Veta 3.2 *Nech X je kladná náhodná premenná, ktorá má Marshallovo-Olkinovo exponenciálne rozdelenie s parametrami $\beta > 0$, $\alpha > 0$ a $\bar{\alpha} = 1 - \alpha$, potom jeho celá dolná časť má Gómez-Dénizovo geometrické rozdelenie s parametrami $\mathrm{e}^{-\beta}$ a α.*

Dôkaz.

Nech $Y = \lfloor X \rfloor$, kde $\lfloor x \rfloor$ je funkcia dolnej celej časti reálneho čísla x, potom platí

$$p_Y(y) = \int_y^{y+1} \frac{\alpha\beta \mathrm{e}^{-\beta x}}{(1 - \bar{\alpha}\mathrm{e}^{-\beta x})^2} \mathrm{d}x.$$

Zaveďme substitúciu $u = 1 - \bar{\alpha}\mathrm{e}^{-\beta x}$. Po aplikovaní substitúcie dostávame

$$p_Y(y) = \int_y^{y+1} \frac{\alpha\beta \mathrm{e}^{-\beta x}}{(1 - \bar{\alpha}\mathrm{e}^{-\beta x})^2} \mathrm{d}x = \frac{\alpha}{\bar{\alpha}} \int_{1-\bar{\alpha}\mathrm{e}^{-\beta y}}^{1-\bar{\alpha}\mathrm{e}^{-\beta(y+1)}} \frac{1}{u^2} \mathrm{d}u =$$

$$= -\frac{\alpha}{\bar{\alpha}}\left[\frac{1}{1-\bar{\alpha}\mathrm{e}^{-\beta(y+1)}} - \frac{1}{1-\bar{\alpha}\mathrm{e}^{-\beta y}}\right] = -\frac{\alpha}{\bar{\alpha}}\left[\frac{1-\bar{\alpha}\mathrm{e}^{-\beta y}-1+\bar{\alpha}\mathrm{e}^{-\beta(y+1)}}{(1-\bar{\alpha}\mathrm{e}^{-\beta(y+1)})(1-\bar{\alpha}\mathrm{e}^{-\beta y})}\right].$$

Označme $q = \mathrm{e}^{-\beta}$, potom

$$p_Y(y) = -\alpha\frac{-q^y + q^{y+1}}{(1-\bar{\alpha}q^{(y+1)})(1-\bar{\alpha}q^y)} = \frac{\alpha q^y(1-q)}{(1-\bar{\alpha}q^{(y+1)})(1-\bar{\alpha}q^y)},$$

kde $y = 0, 1, 2, ..., 0 < q < 1, \alpha > 0$ a $\bar{\alpha} = 1 - \alpha$. Daná rovnosť je pravdepodobnostná funkcia $GG(q, \alpha)$, pričom $q = \mathrm{e}^{-\beta}$, takže $Y \sim GG(\mathrm{e}^{-\beta}, \alpha)$. □

Nech náhodná premenná $Y \sim GG(q, \alpha)$. Stredná hodnota Y zodpovedá tvaru

$$\mathrm{E}(Y) = \sum_{y=1}^{\infty}\frac{\alpha q^y}{1-\bar{\alpha}q^y}.$$

Následne pomocou strednej hodnoty môžeme vyjadriť vzťah pre disperziu, ktorá vyzerá

$$\mathrm{D}(Y) = \sum_{y=1}^{\infty}\frac{(2y-1)\alpha q^y}{1-\bar{\alpha}q^y} - \mathrm{E}(Y)^2.$$

Oba tieto vzťahy platia pre $y = 0, 1, 2, ..., 0 < q < 1, \alpha > 0$ a $\bar{\alpha} = 1 - \alpha$.

3.2 Odhad parametrov

Nasledujúcu podkapitolu sme venovali odhadom parametrov $GG(q, \alpha)$. Ako prvé sme spočítali prvotné odhady parametrov daného rozdelenia. Pre presnejšie odhady sme použili metódu MLE, ktorej princíp sme popísali v podkapitole 1.1.

V prípade $GG(q, \alpha)$ boli prvotné odhady spomenuté v článku, ktorý sa venuje tomuto rozdeleniu, a to konkrétne článok [3]. Z pravdepodobnostnej funkcie $GG(q, \alpha)$ vyplýva, že

$$p_Y(0) = \frac{(1-q)}{(1-\bar{\alpha}q)},$$

a zároveň

$$p_Y(1) = \frac{\alpha q(1-q)}{(1-\bar{\alpha}q^2)(1-\bar{\alpha}q)}.$$

Pomocou týchto vzťahov autori vyjadrili odhady požadovaných parametrov. Platia pre ne nasledovné vzťahy

$$\hat{q} = \frac{\hat{p}_0 - \hat{p}_0^2 - \hat{p}_0\hat{p}_1}{\hat{p}_1},$$

$$\hat{\alpha} = \frac{\hat{p}_0 - 2\hat{p}_0^2 + \hat{p}_0^3 - \hat{p}_1 + \hat{p}_0^2\hat{p}_1}{\hat{p}_0^2(\hat{p}_0 + \hat{p}_1 - 1)},$$

kde označenie $\hat{p}_0$ zodpovedá odhadu $p_Y(0)$, ktorý bol odhadnutý pomocou realizácie ako pomer medzi početnosťou čísla nula v danom súbore a celkovým rozsahom súboru, pre ktorý sme práve parametre odhadovali. Analogický význam má označenie $\hat{p}_1$.

Aby sme mohli odhadnúť parametre pomocou metódy MLE, je potrebné si definovať vierohodnostnú funkciu pre $GG(q, \alpha)$, ktorá má tvar

$$L(\boldsymbol{y}, q, \alpha) = \prod_{i=1}^{n} \frac{\alpha q^{y_i}(1-q)}{(1-\bar{\alpha}q^{y_i+1})(1-\bar{\alpha}q^{y_i})},$$

kde $y = 0,\ 1,\ 2,\ ...,\ 0 < q < 1,\ \alpha > 0$ a $\bar{\alpha} = 1 - \alpha$. Ďalej pre logaritmus funkcie vierohodnosti platí vzťah

$$l(\boldsymbol{y}, q, \alpha) = n[\ln \alpha + \ln(1-q) + \bar{y} \ln q] - \sum_{i=1}^{n} [\ln(1-\bar{\alpha}q^{y_i+1}) + \ln(1-\bar{\alpha}q^{y_i})],$$

keď $\alpha > 0$ a $\bar{\alpha} = 1 - \alpha$. Tento tvar uvádza článok [3].

3.2.1 Simulácie

Pre odhadovanie parametrov sme si ako prvé vygenerovali viaceré sady dát z rozdelenia $GG(q, \alpha)$ pre rôzne parametre q a α. Počiatočné odhady sme pre ilustráciu spravili na súbore veľkosti 10 000 hodnôt. Na spočítanie odhadov sme využili softvér R [5] a po spočítaní odhadov sme ich zapísali do Tabuľky 3. Prvý stĺpec zaznamenáva spomínaný rozsah súboru, v nasledujúcej dvojici stĺpcov sú zaznamenané reálne hodnoty parametrov a zvyšné stĺpce znázorňujú ich odhady.

Tabuľka 3: Prvotné odhady parametrov $GG(q, \alpha)$.

n	q	α	$\hat{q}$	$\hat{\alpha}$	$\hat{q}$	$\hat{\alpha}$
10 000	0,3	0,6	0,3107	0,5743	0,2789	0,6806
10 000	0,2	19	0,1904	21,3059	0,2017	18,0312
10 000	0,4	3	0,3919	3,1417	0,4068	2,9013
10 000	0,6	0,2	0,5757	0,2169	0,5602	0,2381

(*zdroj: vlastné spracovanie*)

Po vypočítaní prvotných odhadov sme aplikovali metódu MLE. Samotné odhady sme zrátali numericky v softvéri R aplikovaním funkcie `optim()` na logaritmus funkcie vierohodnosti prenásobený mínus jednotkou, aby sme našli bod, v ktorom sa nadobúda

jej maximum. Tu sme využili aj naše prvotné odhady, ktoré sme doplnili do funkcie `optim()` ako štartovacie body. Aby sme mohli porovnať odhady aj navzájom medzi rozdeleniami, odhady sme vykonali na súboroch o rovnakej veľkosti. Odhady q a α sme najskôr vykonali pre súbor o veľkosti n rovné 10 a následne sme tento súbor podstatne zväčšili na 10 000 hodnôt. Odhady parametrov sme zapísali do Tabuľky 4, konkrétne do prvého stĺpca tejto tabuľky sme zaznamenali veľkosť súboru. V nasledujúcom stĺpci sme zaznamenali reálne hodnoty q a α, pre ktoré sme spravili realizáciu. Do posledných troch blokov stĺpcov sme zapísali naše odhady $\hat{q}$ a $\hat{\alpha}$, vždy pre inú realizáciu.

Tabuľka 4: Odhady parametrov $GG(q, \alpha)$ pomocou metódy MLE.

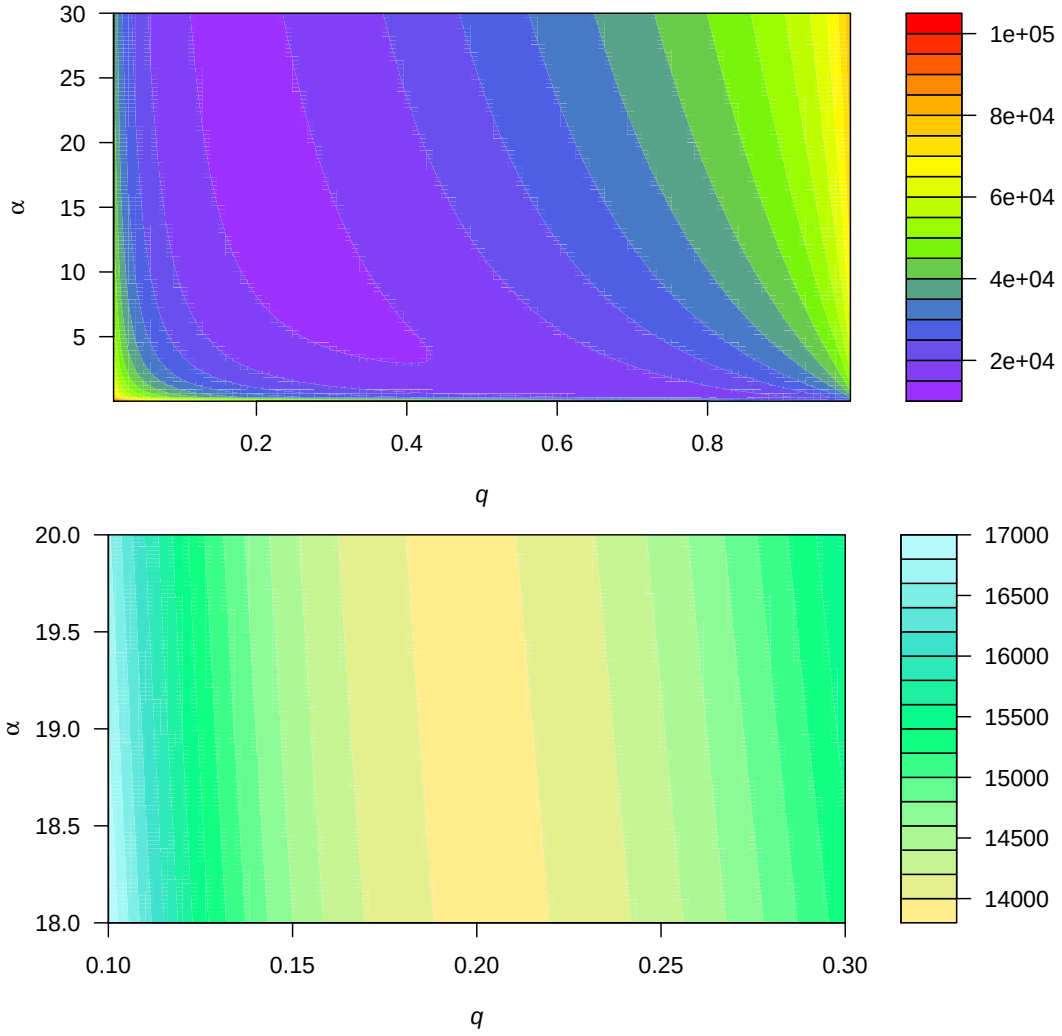
n	q	α	$\hat{q}$	$\hat{\alpha}$	$\hat{q}$	$\hat{\alpha}$	$\hat{q}$	$\hat{\alpha}$
10	0,3	0,6	0,3602	0,8473	0,5250	0,6215	0,4330	0,5621
10 000	0,3	0,6	0,2935	0,6129	0,2990	0,6185	0,3095	0,5521
10	0,2	19,0	0,2295	13,4233	0,1672	22,0287	0,2380	11,2018
10 000	0,2	19,0	0,1996	19,3211	0,1998	19,1482	0,2050	18,7215
10	0,4	3,0	0,4030	1,0538	0,5581	0,9272	0,3699	4,2089
10 000	0,4	3,0	0,3956	3,0660	0,3976	3,0161	0,3995	3,0248
10	0,6	0,2	0,5638	0,2003	0,2196	1,5736	0,4030	1,0538
10 000	0,6	0,2	0,5984	0,2049	0,6071	0,1890	0,6088	0,1952

(*zdroj: vlastné spracovanie*)

V prípade $n = 10$ naše odhady neboli presné ani pri tomto type rozdelenia. Je to spôsobené tým, že 10 hodnôt je naozaj veľmi málo na to, aby sme dostali nejaký odhad, ktorý by bol dostatočne blízky reálnym hodnotám parametrov. Keď sme si však porovnali Tabuľku 4 s Tabuľkou 2 a pozreli sme sa na súbor o rozsahu 10 000 hodnôt, všimli sme si, že odhady parametra α sú pri $GG(q, \alpha)$ o niečo presnejšie ako pri $TGD(q, \alpha)$. Avšak musíme uviesť, že pre väčšie hodnoty parametra α sa nám odhady nepodarili vykonať. Taktiež bol problém s odhadovaním parametra q pre hodnoty bližšie číslu 1.

Tak ako pri $TGD(q, \alpha)$, opäť sme sa bližšie pozreli na funkciu, pomocou ktorej sme parametre odhadovali. Vykreslili sme úrovňové množiny (tzv. vrstevnice) logaritmu funkcie vierohodnosti prenásobeného mínus jednotkou, kde realizácia náhodného výberu, ktorá vstupovala do vykreslenej funkcie, bola generovaná pre $q = 0{,}2$ a $\alpha = 19$

a bola veľkosti 10 000 hodnôt. Na Obr. 6 je najskôr funkcia vykreslená pre široké rozpätie parametrov, konkrétne $0 < q < 1$ a $0 < \alpha \leq 30$. Z obrázka môžeme vidieť, že funkcia má podobný tvar ako v prípade rozdelenia $TGD(q, \alpha)$. Na Obr. 6 sme následne funkciu vykreslili aj pre okolie reálnych parametrov. Môžeme si všimnúť, že okolo skutočnej hodnoty je vytvorený pás, kde sa síce parametre líšia od skutočných hodnôt, ale funkčná hodnota zostáva takmer nemenná, kvôli čomu mohlo nastať, že odhadnuté parametre sa líšia od reálnych hodnôt parametrov napriek veľkosti súboru 10 000 hodnôt.



Obr. 6: Graf úrovňových množín logaritmu funkcie vierohodnosti prenásobeného mínus jednotkou $GG(q, \alpha)$ rozdelenia, kde realizácia bola generovaná pri parametroch $q = 0{,}2$ a $\alpha = 19$.
(zdroj: vlastné spracovanie)

4 Umocnené zovšeobecnené geometrické rozdelenie

V rámci tejto kapitoly sme sa venovali ďalšiemu zovšeobecnenému geometrickému rozdeleniu. Ako pri predchádzajúcich dvoch rozdeleniach, nezamerali sme sa na vznik tohto rozdelenia, skôr sme sa venovali jeho vlastnostiam a odhadu parametrov. Odvodenie tohto rozdelenia a všetky vlastnosti môžeme nájsť v [2].

4.1 Definícia a vlastnosti umocneného zovšeobecneného geometrického rozdelenia

Umocnené zovšeobecnené geometrické rozdelenie je zovšeobecnením $Geo(q)$, takže opäť je to diskrétne rozdelenie pravdepodobnosti, avšak v tomto prípade je závislé od troch parametrov.

Definícia 4.1 *Nech Y je diskrétna náhodná premenná, ktorej pravdepodobnostná funkcia vyzerá*

$$p_Y(y) = \left(\frac{1-q^{y+1}}{1-\bar{\alpha}q^{y+1}}\right)^{\gamma} - \left(\frac{1-q^{y}}{1-\bar{\alpha}q^{y}}\right)^{\gamma} \tag{4}$$

pre $y = 0, 1, 2, \ldots$, $0 < q < 1$, $\alpha > 0$, $\bar{\alpha} = 1 - \alpha$ a $\gamma > 0$. Potom hovoríme, že Y má umocnené zovšeobecnené geometrické rozdelenie.

Veta 4.1 *Nech Y je náhodná premenná, ktorá má umocnené zovšeobecnené geometrické rozdelenie, potom distribučná funkcia tohto rozdelenia má tvar*

$$F_Y(y) = \begin{cases} 0 & ak\ y \leq 0, \\ \sum\limits_{i=0}^{\lfloor y \rfloor} \left(\frac{1-q^{y+1}}{1-\bar{\alpha}q^{y+1}}\right)^{\gamma} - \left(\frac{1-q^{y}}{1-\bar{\alpha}q^{y}}\right)^{\gamma} & ak\ y > 0, \end{cases}$$

kde $0 < q < 1$, $\alpha > 0$, $\bar{\alpha} = 1 - \alpha$ a $\gamma > 0$.

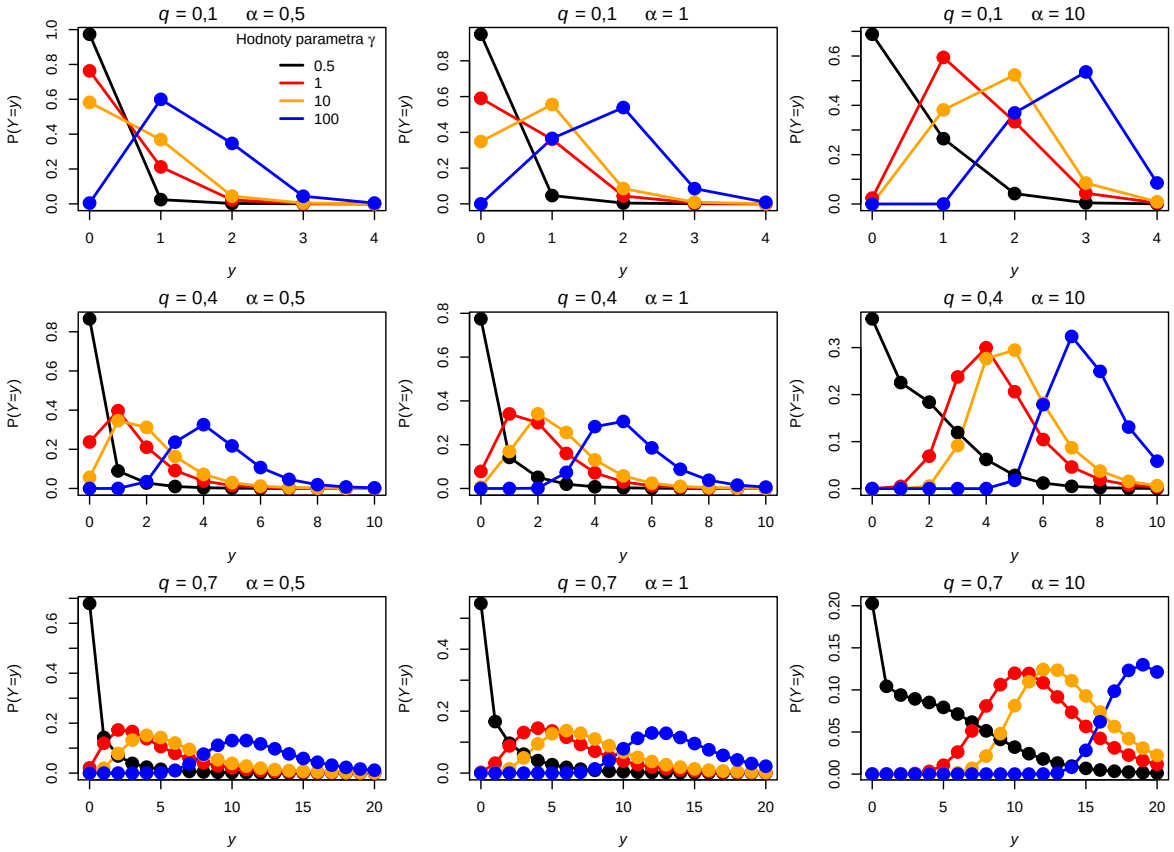
Dôkaz.

Dôkaz vety vyplýva priamo z definície distribučnej funkcie diskrétnej náhodnej premennej. □

Ďalej budeme umocnené zovšeobecnené geometrické rozdelenie označovať skratkou, ktorú zaviedli samotní autori, a to $EGG(q, \alpha, \gamma)$ podľa anglického názvu *exponentiated generalized geometric distribution*.

Poznámka. Ak by sme parameter γ v distribučnej funkcii $EGG(q, \alpha, \gamma)$ položili rovný jednej, dostali by sme rozdelenie $GG(q, \alpha)$. Následne, ak by sa aj parameter α rovnal jednej, dostali by sme tvar distribučnej funkcie $Geo(q)$.

Aby sme získali predstavu o tom, ako vyzerá pravdepodobnostná funkcia tohto rozdelenia, vykreslili sme ju pre rôzne kombinácie parametrov na Obr. 7.



Obr. 7: Grafické porovnanie pravdepodobnostnej funkcie $EGG(q, \alpha, \gamma)$ pri rôznych parametroch q, α a γ.
(*zdroj: vlastné spracovanie*).

Článok [8] uvádza nový typ zovšeobecneného exponenciálneho rozdelenia. Autori ho zadefinovali nasledujúcou distribučnou funkciou

$$F_X(x) = \left(\frac{1 - \mathrm{e}^{-\beta x}}{1 - \bar{\alpha}\mathrm{e}^{-\beta x}} \right)^{\gamma},$$

kde $x > 0$, $0 < \alpha < 1$, $\bar{\alpha} = 1 - \alpha$, $\beta > 0$ a $\gamma > 0$. Nasledujúca veta ukázala vzťah medzi týmto exponenciálnym rozdelením a $EGG(q, \alpha, \gamma)$.

Veta 4.2 *Nech náhodná premenná X má vyššie spomenuté zovšeobecnené exponenciálne rozdelenie s uvedenými parametrami. Potom náhodná premenná Y, ktorá je dolnou celou časťou náhodnej premennej X, má rozdelenie $EGG(\mathrm{e}^{-\beta}, \alpha, \gamma)$.*

Dôkaz.

Nech $Y = \lfloor X \rfloor$, kde $\lfloor x \rfloor$ je funkcia dolnej celej časti reálneho čísla x. Potom platí

$$\mathrm{P}(Y = y) = F_X(y+1) - F_X(y) = \left(\frac{1 - \mathrm{e}^{-\beta(y+1)}}{1 - \bar{\alpha}\mathrm{e}^{-\beta(y+1)}}\right)^{\gamma} - \left(\frac{1 - \mathrm{e}^{-\beta y}}{1 - \bar{\alpha}\mathrm{e}^{-\beta y}}\right)^{\gamma}$$

pre $y = 0, 1, 2, ..., 0 < \alpha < 1$, $\bar{\alpha} = 1 - \alpha$, $\beta > 0$ a $\gamma > 0$. Priamo vzťah (4) ukazuje, že keď sa parameter q rovná $\mathrm{e}^{-\beta}$, tak $Y \sim EGG(q = \mathrm{e}^{-\beta}, \alpha, \gamma)$. □

Nech Y je náhodná premenná, ktorá má $EGG(q, \alpha, \gamma)$ rozdelenie. Potom stredná hodnota Y vyzerá

$$\mathrm{E}(Y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \omega_{i,j}(\alpha, \gamma)\left(\frac{q^{i+j}}{1 - q^{i+j}}\right),$$

kde

$$\omega_{i,j}(\alpha, \gamma) = (-1)^{i+1}\frac{\gamma\bar{\alpha}}{i!j!}\frac{\Gamma(\gamma + j)}{\Gamma(\gamma + 1 - i)}.$$

Následne pre disperziu náhodnej premennej $Y \sim EGG(q, \alpha, \gamma)$ platí vzťah

$$\mathrm{D}(Y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \omega_{i,j}(\alpha, \gamma)\left(\frac{q^{i+j}}{1 - q^{i+j}}\right)\left(1 + \frac{q^{i+j}}{1 - q^{i+j}}\right) - \mathrm{E}(Y)^2.$$

Vo všetkých týchto vzťahoch platí, že $0 < q < 1$, $\alpha > 0$, $\bar{\alpha} = 1 - \alpha$ a $\gamma > 0$.

4.2 Odhad parametrov

V tejto podkapitole sme sa sústredili na odhad parametrov $EGG(q, \alpha, \gamma)$. V prípade tohto rozdelenia sme prvotné odhady neodvádzali. Keď sme sa snažili spraviť prvotné odhady pre toto trojparametrické rozdelenie, zistili sme, že odvodiť ich by bolo veľmi náročné oproti tomu, aký by to malo prínos, preto sme sa rozhodli prvotné odhady voliť náhodne. Pre odhady parametrov sme aplikovali metódu MLE, jej princíp sme popísali v podkapitole 1.1.

Funkcia vierohodnosti $EGG(q, \alpha, \gamma)$ má tvar

$$L(\boldsymbol{y}, q, \alpha, \gamma) = \prod_{i=1}^{n}\left[\left(\frac{1 - q^{y_i+1}}{1 - \bar{\alpha}q^{y_i+1}}\right)^{\gamma} - \left(\frac{1 - q^{y_i}}{1 - \bar{\alpha}q^{y_i}}\right)^{\gamma}\right].$$

Autori článku [2] neuviedli upravený vzťah pre logaritmus funkcie vierohodnosti. Logaritmus funkcie vierohodnosti vyzerá

$$l(\boldsymbol{y}, q, \alpha, \gamma) = \sum_{i=1}^{n}\ln\left[\left(\frac{1 - q^{y_i+1}}{1 - \bar{\alpha}q^{y_i+1}}\right)^{\gamma} - \left(\frac{1 - q^{y_i}}{1 - \bar{\alpha}q^{y_i}}\right)^{\gamma}\right].$$

Obe tieto funkcie platia pre $0 < q < 1$, $\alpha > 0$, $\bar{\alpha} = 1 - \alpha$ a $\gamma > 0$.

4.2.1 Simulácie

Ako prvé sme si vygenerovali sady dát z $EGG(q, \alpha, \gamma)$ pre rôzne parametre. Rozsah menšieho zo súborov sme v tomto prípade pozmenili oproti prvým dvom rozdeleniam, pretože pri súbore o veľkosti 10 hodnôt sa nedarilo odhadnúť parametre daného rozdelenia. To znamená, že najskôr sme odhady parametrov q, α, γ spravili na súbore veľkosti 50 hodnôt a následne na súbore veľkosti 10 000 hodnôt. Odhady boli vypočítané v softvéri R [5], presnejšie boli zrátané numericky pomocou funkcie `optim()`, v ktorej štartovacie body boli v prípade tohto rozdelenia zvolené náhodne. Všetky spravené odhady sme zapísali do Tabuľky 5.

V prvom stĺpci Tabuľky 5 sú zapísané veľkosti súborov, na ktorých boli jednotlivé odhady vykonané. V druhom bloku stĺpcov sú zaznamenané reálne hodnoty parametrov, ktoré sme odhadovali. Zvyšné stĺpce ukazujú odhady jednotlivých parametrov pre danú veľkosť súboru. Kvôli veľkosti tabuľky sme všetky odhady parametrov zaokrúhľovali na dve desatinné miesta.

Tabuľka 5: Odhady parametrov $EGG(q, \alpha, \gamma)$ pomocou metódy MLE.

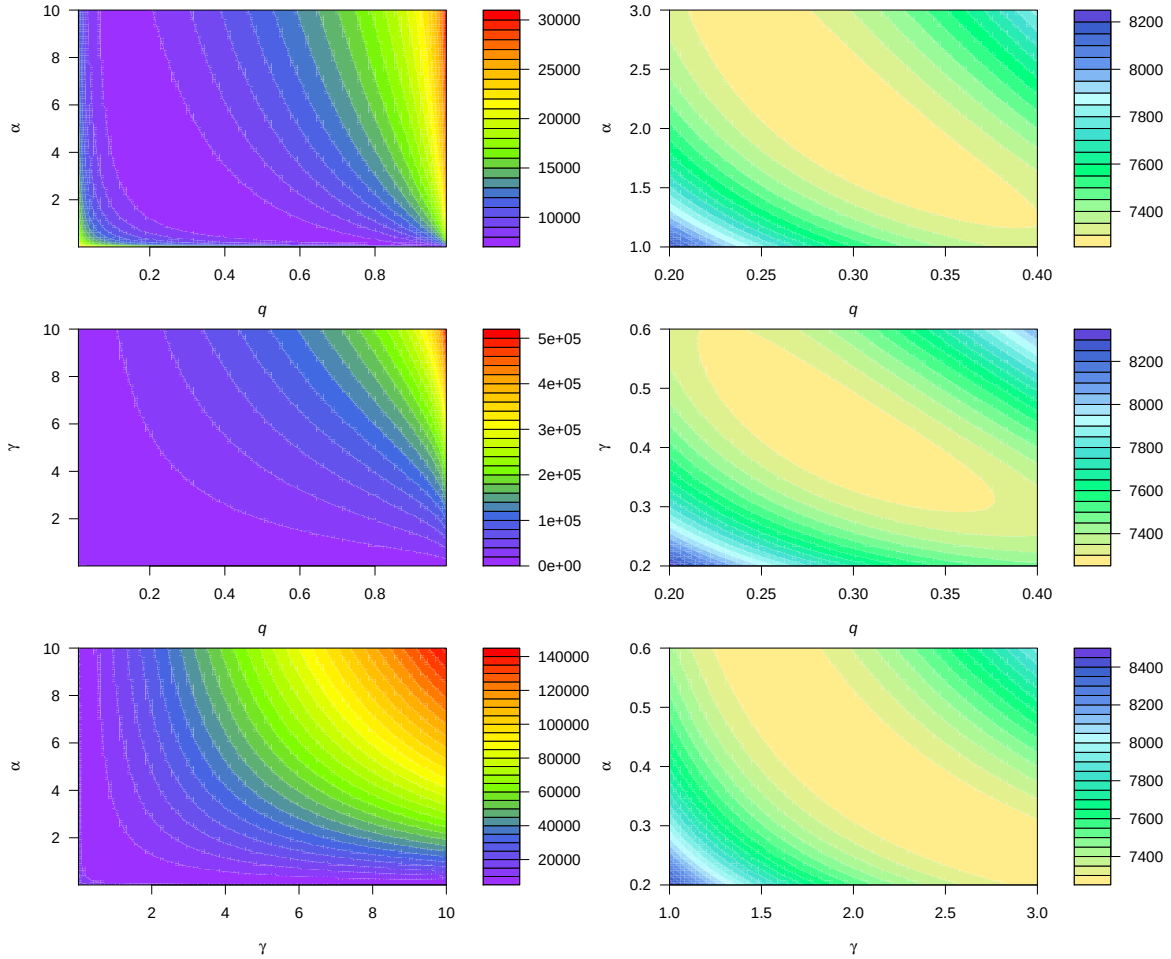
n	q	α	γ	$\hat{q}$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{q}$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{q}$	$\hat{\alpha}$	$\hat{\gamma}$
50	0,5	0,2	0,8	0,41	0,12	3,42	0,24	7,30	0,07	0,49	3,07	0,20
10 000	0,5	0,2	0,8	0,51	0,10	1,48	0,49	0,17	0,94	0,53	0,06	2,29
50	0,3	2,0	0,4	0,35	0,24	2,03	0,23	11,54	0,12	0,24	14,62	0,16
10 000	0,3	2,0	0,4	0,31	1,60	0,44	0,37	0,05	8,34	0,26	5,15	0,23
50	0,6	0,8	5,0	0,49	9,97	1,14	0,60	2,30	2,05	0,60	3,90	1,60
10 000	0,6	0,8	5,0	0,60	0,77	5,07	0,59	0,87	4,77	0,60	0,88	4,63
50	0,8	9,0	3,0	0,81	6,64	3, 43	0,82	1,94	7,33	0,77	6,06	6,84
10 000	0,8	9,0	3,0	0,81	4,72	4,08	0,81	4,68	4,30	0,79	13,37	2,50

(*zdroj: vlastné spracovanie*)

Ako sa ukázalo, odhady parametrov $EGG(q, \alpha, \gamma)$ sa v niektorých prípadoch výrazne líšia od skutočných hodnôt aj pre súbor, ktorý má veľkosť až 10 000 hodnôt. V prípade parametra q sú odhady blízke, ale parametre α a γ sa v mnohých prípadoch nepodarilo dobre odhadnúť. Preto sme sa rozhodli pozrieť na to, ako vyzerá funkcia, pomocou ktorej boli parametre odhadované. Vykreslili sme úrovňové množiny (tzv. vrs-

tevnice) logaritmu funkcie vierohodnosti $EGG(q, \alpha, \gamma)$ rozdelenia prenásobeného mínus jednotkou. Realizácia, ktorá vstupovala do vykreslenej funkcie bola generovaná pre $q = 0{,}3$, $\alpha = 2$ a $\gamma = 0{,}4$. Keďže dané rozdelenie závisí až od troch parametrov a vykresľovali sme len dvojrozmerné grafy, rozhodli sme sa funkciu vykresliť postupne pre kombináciu q, α, následne pre q, γ a na záver pre α, γ, pričom zvyšný parameter sa vo všetkých prípadoch rovnal vyššie spomínaným hodnotám.

Na Obr. 8 sme zobrazili vrstevnice danej funkcie pre všetky kombinácie parametrov, pričom pre každú kombináciu sme najskôr vykreslili graf pre široký rozsah parametrov a následne sme sa pozreli na blízke okolie parametrov.



Obr. 8: Graf úrovňových množín logaritmu funkcie vierohodnosti prenásobeného mínus jednotkou $EGG(q, \alpha, \gamma)$ rozdelenia, vykreslený pre všetky kombinácie parametrov, kde realizácia bola generovaná pri parametroch $q = 0{,}3$, $\alpha = 2$ a $\gamma = 0{,}4$.

(zdroj: vlastné spracovanie)

5 Aplikácie geometrických rozdelení na reálnych dátach

V predchádzajúcich kapitolách sme sa oboznámili s viacerými zovšeobecneniami geometrických rozdelení. Nasledujúcu kapitolu sme venovali aplikáciam a porovnaniu týchto rozdelení na reálnych sadách dát z dvoch odlišných prostredí.

5.1 Chí-kvadrát štatistika

Na porovnanie jednotlivých typov rozdelení sme použili chí-kvadrát štatistiku. Vypočítali sme ju pomocou vzorca

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_{m_i} - E_{m_i})^2}{E_{m_i}},$$

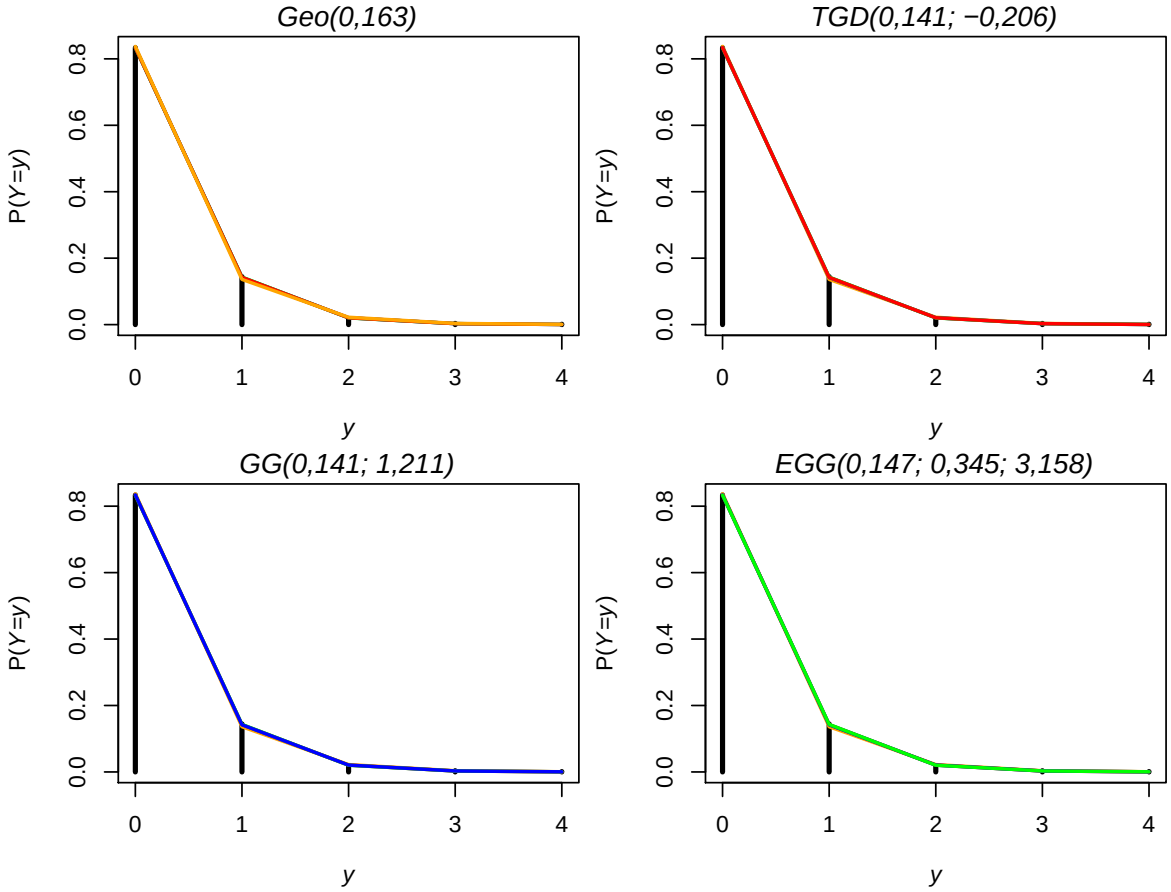
kde k označuje počet kategórií, do ktorých sú jednotlivé hodnoty rozdelené, resp. počet hodnôt, ktoré sú nadobúdané v konkrétnej sade dát a m_i označuje konkrétnu kategóriu. O_{m_i}, z anglického *observed*, je označenie pre početnosť konkrétnej hodnoty v kategórii m_i. Takže napríklad O_{m_1} nám hovorí, presne koľkokrát sa konkrétna hodnota (ako napríklad v našom prípade 0) nachádza v danom súbore v prvej kategórii. E_{m_i}, z anglického *estimated*, je odhad počtu konkrétnych hodnôt, spravený na základe nášho modelu v danej sade dát. Napríklad E_{m_1} hovorí o tom, koľkokrát sa nadobúda konkrétna hodnota (v našom prípade 0) v danom rozdelení v prvej kategórii.

Po implementovaní získaných dát sme ako prvé odhadli parametre rozdelenia, ktorého zhodu sme práve skúmali. Pre odhad parametrov sme použili metódu MLE, pre každý typ rozdelenia máme definované funkcie vierohodnosti v kapitolách, v ktorých sa venujeme jednotlivým rozdeleniam. Po odhadnutí parametrov sme pomocou pravdepodobnostnej funkcie odhadli početnosti hodnôt v daných rozdeleniach pre odhadnuté $\hat{q}$, $\hat{\alpha}$, resp. $\hat{\gamma}$. Následne po odhadnutí početností sme vypočítali vyššie spomenutú chí-kvadrát štatistiku a porovnali sme, ktoré z rozdelení $Geo(q)$, $TGD(q, \alpha)$, $GG(q, \alpha)$ a $EGG(q, \alpha, \gamma)$ je to najlepšie pre danú sadu dát.

5.2 Aplikácia v aktuárstve

Ako prvé sme skúmali ako dobre tieto rozdelenia opisujú sadu dát, ktorá je z prostredia poisťovní. Dáta sme prevzali z článku [1], konkrétne táto sada dát zaznamenáva

počty poistných udalostí automobilov poistencov. V Tabuľke 7 sú v prvých dvoch stĺpcoch zaznamenané tieto poistné udalosti. Presnejšie v prvom stĺpci sú zapísané hodnoty, ktoré sú nadobúdané v tejto sade dát, konkrétne 0, 1, 2, 3, 4, pričom hodnota 4 zaznamenáva 4 a viac poistných plnení. V druhom stĺpci sú zaznamenané početnosti jednotlivých hodnôt. To znamená, koľkokrát je v sade dát nadobudnutá 0, koľkokrát je v sade dát obsiahnutá 1, a tak ďalej.



Obr. 9: Pravdepodobnosti, podľa ktorých sú nadobúdané jednotlivé hodnoty v reálnej sade dát z poisťovníctva v porovnaní s pravdepodobnosťami pre odhadnuté parametre rozdelení $Geo(q)$, $TGD(q, \alpha)$, $GG(q, \alpha)$ a $EGG(q, \alpha, \gamma)$.
(*zdroj: vlastné spracovanie*)

Na Obr. 9 sme vykreslili pravdepodobnosti, s ktorými sa nadobúdajú jednotlivé hodnoty, aby sme mohli porovnať a predpokladať, ktoré rozdelenie sa najlepšie hodí na opis našich dát. Čierne stĺpiky zaznamenávajú odhadnuté pravdepodobnosti $\hat{p}_Y(0)$, $\hat{p}_Y(1)$, $\hat{p}_Y(2)$, ..., ktoré sme odhadli podľa nadobudnutých hodnôt v sade dát. Do grafu sú zakreslené aj odhady pravdepodobností, s ktorými sa nadobúdajú hodnoty pri jednotlivých typoch rozdelení s odhadnutými parametrami $\hat{q}$, $\hat{\alpha}$, resp. $\hat{\gamma}$.

Podľa Obr. 9 nie je jasné, ktoré rozdelenie je najvhodnejšie na modelovanie tohto typu dát. Vidíme, že pre všetky rozdelenia, sú pravdepodobnosti $\hat{p}_Y(0)$, $\hat{p}_Y(1)$, $\hat{p}_Y(2)$, ... prakticky rovnaké. Keby sme zakreslili do jedného obrázka všetky typy rozdelení, všetky rozdelenia by sa doslovne prekryli, a preto sme zostrojili Tabuľku 6, kde sme zapísali odhady pravdepodobností, s ktorými sa nadobúdajú jednotlivé hodnoty v každom type rozdelenia, ako aj v reálnej sade dát. Keď sme sa pozreli na tieto odhady, všimli sme si, že odchýlka medzi pravdepodobnosťami, ktoré sme získali z reálnej sady dát a odhadnutými pravdepodobnosťami je v prípade $Geo(q)$ najväčšia. Na základe tohto pozorovania môžeme predpokladať, že $Geo(q)$ bude spomedzi našich rozdelení najslabšie modelovať reálnu sadu dát.

Tabuľka 6: Porovnanie odhadnutých pravdepodobností nadobúdania hodnôt pri jednotlivých rozdeleniach.

	Dáta	$Geo(q)$	$TGD(q,\alpha)$	$GG(q,\alpha)$	$EGG(q,\alpha,\gamma)$
$\hat{p}_Y(0)$	0,8336	0,8374	0,8338	0,8339	0,8332
$\hat{p}_Y(1)$	0,1445	0,1362	0,1422	0,1420	0,1431
$\hat{p}_Y(2)$	0,0171	0,0221	0,0206	0,0206	0,0202
$\hat{p}_Y(3)$	0,0037	0,0036	0,0029	0,0029	0,0030
$1-\sum_{i=0}^{3}\hat{p}_Y(i)$	0,0011	0,0007	0,0005	0,0005	0,0005
Odhadnuté parametre		$\hat{q}=0{,}163$	$\hat{q}=0{,}141$ $\hat{\alpha}=-0{,}206$	$\hat{q}=0{,}141$ $\hat{\alpha}=1{,}211$	$\hat{q}=0{,}147$ $\hat{\alpha}=0{,}345$ $\hat{\gamma}=3{,}158$

(*zdroj: vlastné spracovanie*)

Pre lepšie porovnanie, ktoré rozdelenie je najvhodnejšie na modelovanie týchto dát sme vytvorili Tabuľku 7. V prvej časti tabuľky sú zaznamenané nadobúdané hodnoty. Ako sme už spomínali, prvé dva stĺpce sú venované skutočným hodnotám. Zvyšné stĺpce tabuľky zaznamenávajú odhadnuté početnosti pre konkrétne typy rozdelenia, ktoré boli spočítané po odhadnutí parametrov. V riadku s názvom „Spolu" sme zaznamenali rozsah súboru a v nasledujúcom riadku sú pre konkrétne rozdelenie zapísané odhadnuté parametre. Do posledného riadku Tabuľky 7 sme zapísali vypočítané chí-kvadrát hodnoty pre jednotlivé rozdelenia.

Tabuľka 7: Modelovanie aktuárskych dát pomocou geometrického rozdelenia a jeho zovšeobecnení.

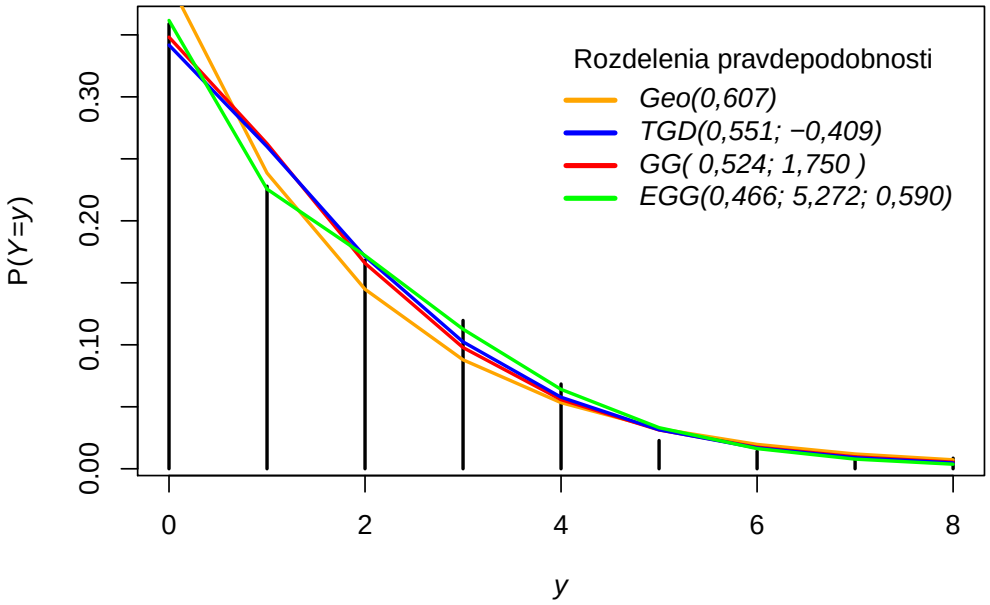
Nadobúdané hodnoty	Početnosť	$Geo(q)$	$TGD(q,\alpha)$	$GG(q,\alpha)$	$EGG(q,\alpha,\gamma)$
0	1563	1570,16	1563,42	1563,60	1562,30
1	271	255,28	266,65	266,31	268,26
2	32	41,50	38,57	38,70	37,94
3	7	6,75	5,46	5,49	5,55
4	2	1,31	0,90	0,90	0,96
Spolu	1875	1875	1875	1875	1875
Odhadnuté parametre		$\hat{q} = 0{,}163$	$\hat{q} = 0{,}141$ $\hat{\alpha} = -0{,}206$	$\hat{q} = 0{,}141$ $\hat{\alpha} = 1{,}211$	$\hat{q} = 0{,}147$ $\hat{\alpha} = 0{,}345$ $\hat{\gamma} = 3{,}158$
χ^2-hodnota		3,93	3,58	3,60	3,06

(*zdroj: vlastné spracovaniena základe* [1])

Chí-kvadrát hodnota nám potvrdila to, čo sme si všimli v Tabuľke 6. Rozdelenie $Geo(q)$ spomedzi definovaných rozdelení najslabšie modeluje danú sadu dát, avšak nemôžeme tvrdiť, že by bolo pre túto sadu úplne nepoužiteľné, pretože aj pri tomto rozdelení chí-kvadrát hodnota vyšla pomerne malá a veľmi porovnateľná s hodnotami pre ostatné tri rozdelenia. V prípade rozdelení $TGD(q,\alpha)$ a $GG(q,\alpha)$ môžeme vidieť isté zlepšenie od $Geo(q)$. Naviac si môžeme všimnúť, že tieto rozdelenia majú medzi sebou veľmi blízke chí-kvadrát hodnoty. Môže to byť zapríčinené tým, že obe tieto rozdelenia sú závislé práve od dvoch parametrov. Rozdelenie, ktoré najlepšie modeluje dáta spomedzi rozdelení, ktoré sme porovnávali je $EGG(q,\alpha,\gamma)$. Nie je to prekvapujúce, pretože závisí až od troch parametrov, čo dodáva rozdeleniu lepšiu pružnosť, avšak toto zlepšenie v chí-kvadrát hodnote nie je veľmi výrazné. Ako ukazuje Tabuľka 7, odhadnuté početnosti pre dvojparametrové rozdelenia ($TGD(q,\alpha)$, $GG(q,\alpha)$) nám vyšli veľmi podobné v porovnaní s trojparametrových rozdelením ($EGG(q,\alpha,\gamma)$). Podľa týchto faktov sme si dovolili skonštatovať, že na modelovanie týchto dát by sme použili jedno z dvojparametrových rozdelení ($TGD(q,\alpha)$, $GG(q,\alpha)$), nakoľko práca s dvomi parametrami je jednoduchšia ako s tromi a tieto rozdelenia dobre modelujú danú sadu.

5.3 Aplikácia v medicíne

Ako druhé sme skúmali dáta z lekárskeho prostredia, ktoré sme opäť získali z článku [1]. Dáta v druhej sade zaznamenávajú počty epileptických záchvatov a sú zapísané v Tabuľke 9. Presnejšie v prvom stĺpci sú zapísané hodnoty, ktoré sa nadobúdajú v tejto sade dát, konkrétne 0, 1, ..., 8, kde hodnota 8 zaznamenáva 8 a viac epileptických záchvatov. V druhom stĺpci sú zaznamenané početnosti jednotlivých hodnôt, to znamená, koľkokrát je v sade dát nadobudnutá 0, koľkokrát je v sade dát obsiahnutá 1, a tak ďalej.



Obr. 10: Pravdepodobnosti, podľa ktorých sú nadobúdané jednotlivé hodnoty v reálnej sade dát z medicíny v porovnaní s pravdepodobnosťami pre odhadnuté parametre rozdelení $Geo(q)$, $TGD(q, \alpha)$, $GG(q, \alpha)$ a $EGG(q, \alpha, \gamma)$.
(*zdroj: vlastné spracovanie*)

Opäť sme si ako prvý vykreslili obrázok pravdepodobností, s ktorými sa nadobúdajú jednotlivé hodnoty. Čierne stĺpiky zaznamenávajú odhadnuté pravdepodobnosti, ktoré sme získali z reálnej sady dát. Čiarovými grafmi sú znázornené odhadnuté pravdepodobnosti pre každé rozdelenie. Presné farebné odlíšenie a odhadnuté parametre pre každé rozdelenie môžeme vidieť v legende na Obr 10.

Na Obr. 10 môžeme vidieť, že $EGG(q, \alpha, \gamma)$ najlepšie modeluje danú sadu dát z lekárskeho prostredia. Na druhej strane vidíme, že odhadnuté pravdepodobnosti $Geo(q)$ rozdelenia najslabšie kopírujú čierne stĺpiky, čo znamená, že $Geo(q)$ spomedzi rozdelení,

ktoré sme porovnávali, najhoršie sedí na danú sadu dát z lekárskeho prostredia. Aby sme lepšie videli tieto pravdepodobnosti a odhady pravdepodobností pre jednotlivé rozdelenia zostrojili sme Tabuľku 8.

Tabuľka 8: Porovnanie odhadnutých pravdepodobností nadobúdania hodnôt pri jednotlivých rozdeleniach.

	Dáta	$Geo(q)$	$TGD(q,\alpha)$	$GG(q,\alpha)$	$EGG(q,\alpha,\gamma)$
$\hat{p}_Y(0)$	0,3590	0,3931	0,3482	0,3418	0,3616
$\hat{p}_Y(1)$	0,2279	0,2386	0,2623	0,2599	0,2256
$\hat{p}_Y(2)$	0,1681	0,1448	0,1658	0,1712	0,1720
$\hat{p}_Y(3)$	0,1197	0,0879	0,0978	0,1024	0,1126
$\hat{p}_Y(4)$	0,0684	0,05334	0,0558	0,0577	0,0640
$\hat{p}_Y(5)$	0,0228	0,0324	0,0313	0,0315	0,0332
$\hat{p}_Y(6)$	0,0142	0,0197	0,0174	0,0168	0,0163
$\hat{p}_Y(7)$	0,0114	0,0119	0,0096	0,0089	0,0078
$1-\sum_{i=0}^{7}\hat{p}_Y(i)$	0,0085	0,0184	0,0119	0,0099	0,0069
Odhadnuté parametre		$\hat{q}=0{,}607$	$\hat{q}=0{,}551$ $\hat{\alpha}=-0{,}409$	$\hat{q}=0{,}524$ $\hat{\alpha}=1{,}750$	$\hat{q}=0{,}466$ $\hat{\alpha}=5{,}273$ $\hat{\gamma}=0{,}590$

(*zdroj: vlastné spracovanie*)

Tabuľka 8 ukazuje, že odchýlka medzi pravdepodobnosťami nadobúdania pre reálne hodnoty a odhadnutými pravdepodobnosťami je pre $Geo(q)$ najväčšia. Na druhej strane môžeme vidieť, že v prípade $EGG(q,\alpha,\gamma)$ je odchýlka odhadnutých pravdepodobností a pravdepodobností nadobúdania z reálnych dát najmenšia. Takže na základe obrázka aj odhadov pravdepodobností sme mohli predpokladať, že najlepšie rozdelenie na modelovanie tejto sady dát je $EGG(q,\alpha,\gamma)$. Preto sme zostrojili Tabuľku 9, kde sme zapísali vypočítané chí-kvadrát hodnoty, aby sme sa utvrdili, ktoré rozdelenie je pre danú sadu dát najvhodnejšie.

V prvej časti Tabuľky 9 sú v prvých dvoch stĺpcoch zaznamenané reálne dáta. Zvyšné stĺpce sú venované jednotlivým rozdeleniam, konkrétne odhadom početností, ktoré sme spočítali po odhadnutí parametrov vždy pre konkrétne rozdelenie. Samotné odhady parametrov sú zapísané v riadku hneď pod riadkom s názvom „Spolu", ktorý

zaznamenáva celkový rozsah súboru. V poslednom riadku Tabuľky 9 sú pre každé rozdelenie zapísané chí-kvadrát hodnoty.

Tabuľka 9: Modelovanie lekárskych dát pomocou geometrického rozdelenia a jeho zovšeobecnení.

Nadobúdané hodnoty	Početnosť	$Geo(q)$	$TGD(q,\alpha)$	$GG(q,\alpha)$	$EGG(q,\alpha,\gamma)$
0	126	137,96	122,21	119,96	126,94
1	80	83,74	92,05	91,22	79,19
2	59	50,82	58,19	60,09	60,36
3	42	30,85	34,31	35,93	39,52
4	24	18,72	19,58	20,25	22,48
5	8	11,36	10,99	11,04	11,64
6	5	6,90	6,11	5,91	5,72
7	4	4,19	3,39	3,13	2,73
8	3	6,46	4,17	3,47	2,43
Spolu	351	351	351	351	351
Odhadnuté parametre		$\hat{q} = 0{,}607$	$\hat{q} = 0{,}551$ $\hat{\alpha} = -0{,}409$	$\hat{q} = 0{,}524$ $\hat{\alpha} = 1{,}750$	$\hat{q} = 0{,}466$ $\hat{\alpha} = 5{,}273$ $\hat{\gamma} = 0{,}590$
χ^2-hodnota		9,65	6,24	5,75	4,39

(*zdroj: vlastné spracovaniena základe* [1])

Po porovnaní všetkých chí-kvadrát hodnôt sme videli, že rozdelenie, ktoré najslabšie modeluje túto sadu dát je naozaj $Geo(q)$, tak ako sme predpokladali. Dôvodom je, že toto rozdelenie závisí len od jedného parametra. Pri rozdeleniach $TGD(q,\alpha)$ a $GG(q,\alpha)$ vidíme isté zlepšenie v chí-kvadrát hodnote, je to spôsobené tým, že obe tieto rozdelenia závisia od dvoch parametrov. Práve tento fakt spôsobil to, že v prípade týchto dvoch rozdelení sú si podobné aj samotné chí-kvadrát hodnoty. Najmenšia hodnota vyšla pre $EGG(q,\alpha,\gamma)$, to znamená, že toto rozdelenie je spomedzi porovnávaných najvhodnejšie na modelovanie konkrétnej sady dát. Tu jasne zavážil fakt, že $EGG(q,\alpha,\gamma)$ závisí od troch parametrov. Takže na modelovanie tejto sady dát by sme zvolili práve toto rozdelenie.

Záver

V rôznych oblastiach je často potrebné modelovať dáta, ktoré nadobúdajú len celočíselné hodnoty. Práve v týchto oblastiach majú využitie diskrétne rozdelenia pravdepodobnosti. Táto bakalárska práca bola venovaná zovšeobecneniam geometrického rozdelenia. Postupne sme prezentovali tri zovšeobecnenia, odhady ich parametrov a využitie na reálnych sadách dát. Práca mala priniesť akýsi prehľad základných charakteristík týchto rozdelení a napríklad aj ukázať, ako rôzne môžu vyzerať pravdepodobnostné funkcie.

Prvá kapitola sa zameriavala na geometrické rozdelenie. Keďže toto rozdelenie je dobre známe, nevenovali sme mu veľa pozornosti. Uvádzame ho pre úplnosť informácií, nakoľko práca sa zaoberá jeho zovšeobecneniami. Okrem samotného geometrického rozdelenia je v prvej kapitole popísaný aj princíp odhadovania parametrov pomocou metódy maximálnej vierohodnosti, ktorý je následne používaný v priebehu celej práce.

Hlavná časť bakalárskej práce postupne prezentovala tri zovšeobecnenia geometrického rozdelenia pravdepodobnosti. Najskôr sme každé z rozdelení definovali, ďalej sme ukázali analógiu medzi danými zovšeobecnenými geometrickými rozdeleniami a zodpovedajúcimi spojitými zovšeobecneniami exponenciálneho rozdelenia, a následne sme uviedli strednú hodnotu a disperziu každého z rozdelení. Veľká časť práce bola venovaná odhadovaniu parametrov jednotlivých rozdelení. Pre odhadovanie parametrov sme používali metódu maximálnej vierohodnosti, a preto súčasťou každej kapitoly, ktorá sa venovala daným zovšeobecneniam, bola funkcia vierohodnosti. Spomínané boli aj logaritmické verzie funkcií vierohodnosti, pre ktoré sme vykreslili aj grafy úrovňových množín, aby sme priblížili správanie sa odhadnutých parametrov.

Posledná piata kapitola bola venovaná aplikácií geometrického rozdelenia ako aj jeho zovšeobecnení na reálnych sadách dát. Ukázali sme využitie spomínaných rozdelení pri reálnych problémoch v prostredí poisťovní a v oblasti medicíny. Dané rozdelenia sú medzi sebou porovnávané pomocou chí-kvadrát štatistiky. Na reálnych sadách dát bol ukázaný prínos ďalších parametrov do geometrického rozdelenia. Možno teda konštatovať, že má zmysel v daných oblastiach uvažovať aj nad rôznymi zovšeobecneniami už známeho geometrického rozdelenia.

Ciele, ktoré sme si na začiatku stanovili, boli naplnené v plnom rozsahu. Podarilo sa nám opísať základné vlastnosti daných rozdelení a aplikáciu samotných rozdelení demonštrovať na reálnych dátach, čím sme ukázali, že dané rozdelenia môžu byť nápomocné v oblasti poisťovníctva. Samozrejme v danej tématike by sa dalo ďalej pokračovať. Teoretická časť by sa dala rozšíriť o mnoho ďalších zovšeobecnení geometrického rozdelenia, alebo by sme mohli zahrnúť iné rôzne vlastnosti, ktoré už spomínané rozdelenia majú. Praktická časť by sa dala obohatiť napríklad testovaním daných rozdelení a mohli by sme skúmať ďalšie sady dát z odlišných oblastí. V tom prípade by však práca prekračovala štandardný rozsah bakalárskych prác, a preto táto práca môže byť motiváciou k hlbšiemu skúmaniu danej problematiky.

Zoznam použitej literatúry

[1] Bhati, D., Sastry, D. V. S., Qadri PZ M.: A New Generalized Poisson-Lindley Distribution: Applications and Properties. *Austrian Journal of Statistics*, 44, 35-51, 2015.

[2] Bidram, H., Roozegar, R., Nekoukhou, V.: Exponentiated generalized geometric distribution: A new discrete distribution. *Hacettepe Journal of Mathematics and Statistics*, 45(6), 1767-1779, 2016.

[3] Gómez-Déniz, E.: Another generalization of the geometric distribution, 19, 399-415, 2010. [cit. 30.4.2021] Dostupné na adrese: `https://doi.org/10.1007/s11749-009-0169-3`.

[4] Chakraborty, S., Bhati, D.: Transmuted geometric distribution with applications in modelling and regression analysis of count data. *Statistics and Operations Research Transactions*, 40(1), 153-176, 2016.

[5] R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. [cit. 30.4.2021] Dostupné na adrese: `https://www.R-project.org/`.

[6] Ross, S. M.: *Simulation*. Fifth Edition, Academic Press, United States, 2013, ISBN 978-0-12-415825-2.

[7] Shaw, W. T., Buckley, I. R. C.: The alchemy of probability distributions: beyond Gram-Charlier expansions, and a skew-kurtotic-normal distribution from a rank transmutation map, 2009. [cit. 30.4.2021] Dostupné na adrese: `https://arxiv.org/abs/0901.0434`.

[8] Silva, R. B., Barreto-Souza, W., Cordeiro, G. M.: A new distribution with decreasing, increasing and upside-down bathtub failure rate. *Computational Statistics and Data Analysis*, 54, 935-944, 2010.

[9] Wolfertz J.: *AlgebraicHaploPackage: Haplotype Two Snips Out of a Paired Group of Patients*. R package version 1.2, 2015. [cit. 06.04.2021] Dostupné na adrese: `https://CRAN.R-project.org/package=AlgebraicHaploPackage`.

Prílohy

Príloha A

Programový kód odhadovania parametrov $TGD(q, \alpha)$ rozdelenia, ktorému sme venovali kapitolu 2, vybraný z prostredia R [5]. Pri počítaní prvotných parametrov $TGD(q, \alpha)$ sme použili potrebný balík, konkrétne [9].

```
#DEFINOVANIE POTREBNYCH FUNKCII
#PRAVDEPODOBNOSTNA FUNKCIA
pmf.TGD<- function(x,q,alfa)
{
  p<- c()
  p<- (1-alfa)*q^x*(1-q)+alfa*(1-q^2)*q^(2*x)
  return(p)
}

#DISTRIBUCNA FUNKCIA
F.TGD<- function(x,q,alfa)
{
  f<- c()
  f<- 1-(1-alfa)*q^(x+1)-alfa*q^(2*(x+1))
  return(f)
}

#GENERATOR PSEUDO-NAHODNYCH HODNOT
rTGD<- function(n, q, alfa)
{
  u<- runif(n, 0, 1)
  r<- c()
  d<- F.TGD(0:200, q, alfa)
  i<- 1

  for (i in 1:n)
  {
    for (j in 1:200)
    {
      if (u[i]< d[j])
      {
        r[i]<- j-1
        break
      }
    }
  }

  return(r)
}
```

```
#LOG-VIEROHODNOSTNA FUNKCIA
l.TGD<- function(parametre)
{
  x<- nv
  q<- parametre[1]
  alfa<- parametre[2]
  return( -sum( log(pmf.TGD(x, q, alfa)) ) )
}

#ODHADOVANIE PARAMETROV
#ODHAD PRAVDEPODOBNOSTI p_0 A p_1
hat_p<- function(nv)
{
  counter0 <- 0
  counter1 <- 0

  for(i in 1:length(nv))
  {
    if (nv[i] == 0)
    {
      counter0 <- counter0 + 1
    }

    if (nv[i] == 1)
    {
      counter1 <- counter1 + 1
    }
  }

  p0<- counter0 / length(nv)
  p1<- counter1 / length(nv)

  return( c(p0, p1) )
}

#FUNKCIA, KTORA VRATI PRVOTNE ODHADY
#instalacia a nacitanie potrebneho balika [9]
install.packages("AlgebraicHaploPackage"); library(AlgebraicHaploPackage);

prvotne_odhadyTGD<- function(nv)
{
  p0<- hat_p(nv)[1]
  p1<- hat_p(nv)[2]

  q<- cubic( 1, p0-1, p0-1, 1-p0-p1 )[3]
  alfa<- (q*p0-p1)/(q*(1-q)^2*(1+q))

  return(c(q, alfa))
}
```

```
#PARAMETRE A GENEROVANIE HODNOT
q<- 0.9
alfa<- -0.9
set.seed(540)
nv<- rTGD(10000, q, alfa)

#ODHAD PARAMETROV ZALOZENY NA PRINCIPE MLE
prvotne_odhadyTGD(nv)          #vypisanie prvotnych odhadov
optim( par= c( prvotne_odhadyTGD(nv)[1], prvotne_odhadyTGD(nv)[2] ),
       l.TGD, method =  "L-BFGS-B", lower = -0.99, upper = 0.99 )
```