Univerzita Komenského v Bratislave
Fakulta Matematiky, Fyziky a Informatiky

Diplomová Práca

Bratislava 2013

Bc. Barbora Litvajová

UNIVERZITA KOMENSKÉHO V BRATISLAVE
Fakulta matematiky, fyziky a informatiky

NOVÉ ZOVŠEOBECNENIE GEOMETRICKÉHO ROZDELENIA AKO MODEL PRE FREKVENCIE GRAFÉM

Diplomová práca

Evidenčné číslo: 70156c51-f6f1-4bde-aaea-c7683adaf15e

Študijný program: Ekonomická a finančná matematika
Študijný odbor: 9.1.9. Aplikovaná matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky

Vedúci diplomovej práce: Mgr. Ján Mačutek, PhD.

Bratislava 2013

Bc. Barbora Litvajová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Barbora Litvajová
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov: Nové zovšeobecnenie geometrického rozdelenia ako model pre frekvencie grafém

Cieľ: Pokus modelovať usporiadane frekvencie grafém novým zovšeobecnením geometrického rozdelenia. Testovanie zhody medzi modelom a dátami.

Vedúci: Mgr. Ján Mačutek, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 25.01.2012

Dátum schválenia: 26.01.2012 prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Čestné prehlásenie

Prehlasujem, že som túto diplomovú prácu vypracovala samostatne s použitím uvedenej odbornej literatúry a ďalších zdrojov.

Bratislava 22.4.2013 .....

Barbora Litvajová

Poďakovanie

Chcem sa poďakovať Mgr. Jánovi Mačutkovi, PhD. za rady a pomoc pri vypracovaní diplomovej práce. Takisto sa chcem poďakovať môjmu manželovi Lukášovi Litvajovi za podporu.

Autorka

Abstrakt

LITVAJOVÁ, Barbora : Nové zovšeobecnenie geometrického rozdelenia ako model pre frekvencie grafém [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Mačutek, PhD., Bratislava, 2013, 48 s.

Cieľom práce je aplikácia nového zovšeobecnenia geometrického rozdelenia v matematickom modelovaní v lingvistike, konkrétne ide o model frekvencie grafém. Model a jeho modifikáciu aplikujeme na dáta z ruštiny, slovenčiny, slovinčiny, ukrajinčiny a tamilčiny. Vhodnosť modelu testujeme pomocou χ^2 testu dobrej zhody.

Kľúčové slová

teória pravdepodobnosti, geometrické rozdelenie, frekvencie grafém, odhady parametrov, χ^2 test dobrej zhody

Abstract

LITVAJOVÁ, Barbora : A new generalization of the geometric distribution as a model for grapheme frequencies [Master thesis], Comenius University in Bratislava, Faculty of mathematics, physics and informatics, Department of applied mathematics and statistics; thesis supervisor: : Mgr. Ján Mačutek, PhD., Bratislava, 2013, 48 p.

The goal of this master thesis is an application of the new generalized geometric distribution in the mathematical modelling in linguistics, specifically it is a model for grapheme frequencies. Model and its modification is applied on the data from Russian, Slovak, Slovene, Ukrainian and Tamil language. The criterion of the good fit is χ^2 goodness of fit test.

Key words

probability theory, geometric distribution, grapheme frequencies, estimation of parameters, χ^2 goodness of fit test

Obsah

Zoznam obrázkov

Zoznam tabuliek

1 Úvod

V diplomovej práci budeme skúmať zovšeobecnenie geometrického rozdelenia predstavené v článku [1] a aplikovať ho na dáta z oblasti lingvistiky - konkrétne na usporiadané frekvencie grafém. Graféma je najmenšia funkčná jednotka písma. V článku [5] bolo ukázané, že geometrické rozdelenie nie je vhodným rozdelením pre tento typ dát. Programy budú napísané v programovacom prostredí R. Výstupom našich programov budú jednotlivé odhady parametrov rozdelení a hodnota χ^2 štatistík pre štyri slovanské jazyky - ruštinu, slovenčinu, slovinčinu a ukrajinčinu.

V druhej kapitole, v ktorej popíšeme teoretický základ diplomovej práce, spomenieme základné pojmy, s ktorými neskôr v článku budeme pracovať. Kapitola je rozdelená na pojmy z oblasti pravdepodobnosti a pojmy z oblasti matematickej štatistiky. Okrem základných pojmov z oblasti pravdepodobnosti sa zameriame najmä na detailnejší popis geometrického rozdelenia. Pri pojmoch z oblasti štatistiky sa pozrieme na všeobecný postup testovania štatistických hypotéz a najmä na to, na akom princípe je založený χ^2 test.

V ďalšej kapitole predstavíme zovšeobecnené geometrické rozdelenie. Toto rozdelenie je pružnejšie, pretože namiesto jedného parametra má dva, $\alpha > 0$ a $0 < \theta < 1$. V ďalších podkapitolách tejto kapitoly o novom modeli spomenieme niektoré vlastnosti nového rozdelenia, ako sú stredná hodnota, disperzia, unimodálnosť. Potom sa pozrieme na generovanie náhodných čísel. Pri zovšeobecnenom geometrickom rozdelení budeme potrebovať počiatočné odhady parametrov kvôli tomu, že optimalizujeme cez dva neznáme parametre iteračnou metódou z oblasti nelineárneho programovania. R vyžaduje pri optimalizácii cez jeden parameter len jeho ohraničenie, pri dvoch ohraničenie nestačí - potrebujeme aj počiatočné odhady. Tie odvodíme tiež v teoretickej časti.

V štvrtej kapitole aplikujeme teoretické poznatky na konkrétne dáta z oblasti frekvencie grafém štyroch spomínaných slovanských jazykov. Cieľom je porovnanie zhody dát a modelu pre geometrické rozdelenie a pre jeho zovšeobecnenie. Keďže skúmané zovšeobecnenie geometrického rozdelenia má dva parametre (teda o jeden viac ako geometrické), dá sa predpokladať, že bude lepším modelom.

Skúmame aj iný typ jazyka - tamilčinu. Tamilčina patrí do skupiny jazykov, v

ktorých sa jedna graféma vyskytuje výrazne častejšie ako ostatné, preto pre ňu prirodzene tento model nevykazuje dobré výsledky. Preto modifikujeme pravdepodobnosť najpočetnejšej grafémy.

Okrem oblasti frekvencie grafém máme k dispozícii aj dáta reprezentujúce dĺžku slov (meranú počtom slabík) v angličtine a litovčine. Znova sú to dáta z kvantitatívnej lingvistiky, len iného typu. Aj na tento iný typ dát sa pokúsime aplikovať náš nový model a urobiť obdobné štatistiky ako v prechádzajúcich prípadoch.

V závere aplikačnej časti práce spomenieme aj simulované p-hodnoty pre dáta z litovčiny. P-hodnota je v oblasti štatistiky kritériom toho, či skúmanú hypotézu zamietneme alebo prijmeme. Nebudeme používať vstavané funkcie, p-hodnotu vypočítame pomocou generovania náhodných čísel.

Piata kapitola je záver - teda zhrnutie výsledkov a zhodnotenie toho, či sa všetky naše hypotézy ukázali ako pravdivé alebo nie.

V prílohe uvádzame zdrojové kódy všetkých použitých programov v programovacom prostredí R.

2 Základné definície

2.1 Základné pojmy z teórie pravdepodobnosti

V tejto podkapitole budeme čerpať pojmy z kníh [9] a [7].

Množina výsledkov náhodného pokusu sa v teórii pravdepodobnosti nazýva priestorom elementárnych udalostí, zvyčajne sa označuje Ω. Jednotlivé výsledky pokusu sa označujú ω. Udalosť A nastáva, ak výsledkom experimentu je elementárna udalosť ω taká, že $\omega \in A$. Množinu všetkých reálnych čísel značíme $\mathbf{R}$.

Definícia 2.1 (σ algebra) *Nech $\Omega \neq \emptyset$. Neprázdny systém S podmnožín množiny Ω budeme nazývať σ algebrou, ak platí*

- *Ak $A \in S$, tak $A^C \in S$.*
- *Ak $A_i \in S; i = 1, 2, ..., tak \cup_{i=1}^{\infty} A_i \in S$.*

Definícia 2.2 (Systém borelovských množín) *Najmenšiu σ algebru nad systémom*

$$C = \{G \subset \boldsymbol{R}; G \text{ je otvorená v } \boldsymbol{R}\}$$

podmnožín množiny $\boldsymbol{R}$ budeme nazývať systémom borelovských množín v $\boldsymbol{R}$ a značiť β.

Definícia 2.3 (Axiomatická definícia pravdepodobnosti) *Pravdepodobnosť je zobrazenie $P : S \to \boldsymbol{R}$ definované na σ - algebre S podmnožín Ω, pričom platí:*

1. *$\Omega \in S$*
2. *Ak $A \in S$, tak aj $A^C = \Omega - A \in S$*
3. *Ak $A_n \in S$, n=1, 2, ..., tak aj $\cup_{i=1}^{\infty} A_n \in S$*
4. *$P(A) \geq 0$ pre všetky $A \in S$*
5. *P(Ω) = 1*
6. *Ak $A_n \in S$, n=1, 2, ... a $A_i \cap A_j = \emptyset, i \neq j$, tak $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$*

Trojica (Ω, S, P) sa nazýva pravdepodobnostný priestor.

Definícia 2.4 (Náhodná premenná) *Nech je daný pravdepodobnostný priestor* (Ω, S, P). *Zobrazenie* $X : \Omega \to R$ *nazývame náhodnou premennou (náhodnou veličinou), ak pre všetky* $x \in R$ *platí*

$$\{\omega; X(\omega) < x\} \in S.$$

Definícia 2.5 (Rozdelenie pravdepodobnosti) *Nech* (Ω, S, P) *je pravdepodobnostný priestor,* X *náhodná premenná a* β *systém všetkých borelovských množín. Rozdelením pravdepodobnosti náhodnej premennej nazývame funkciu* $P_X : M \to \boldsymbol{R}$ *definovanú vzťahom* $P_X(B) = P(X^{-1}(B)), B \in \beta$, *kde* $X^{-1}(B) = \{\omega; X(\omega) \in B\}$ *je vzor množiny* B *pri zobrazení* X.

Definícia 2.6 (Diskrétna náhodná premenná) *Náhodná premenná* X *sa nazýva diskrétna, ak existuje postupnosť nezáporných reálnych čísel* $\{x_i\}$ *a postupnosť nezáporných reálnych čísel* $\{p_i\}$ *taká, že platí*

$$P(X = x_i) = p_i \text{ a zároveň } \sum_i P(X = x_i) = \sum_i p_i = 1.$$

Rozdelenie diskrétnej náhodnej premennej je dané množinou dvojíc $(x_i, p_i), i = 1, 2, ..$ a podmienkou $\sum_{i=1}^{\infty} p_i = 1$.

Definícia 2.7 (Stredná hodnota náhodnej premennej) *Nech* X *je diskrétna náhodná premenná, ktorá nadobúda hodnoty* x_i *s pravdepodobnosťami* p_i *podľa definície diskrétneho rozdelenia. Potom stredná hodnota náhodnej premennej* X *je*

$$E(X) = \sum_{i=1}^{n} x_i p_i,$$

ak tento rad konverguje absolútne.

Definícia 2.8 (Disperzia náhodnej premennej) *Disperzia diskrétnej náhodnej premennej* X *so strednou hodnotou E(X) je číslo*

$$D(X) = E[(X - E(X))^2],$$

ak táto stredná hodnota existuje.

Disperziu vieme napísať aj ako

$$D(X) = E(X^2) - E^2(X), \tag{1}$$

ak existuje $E(X)$ a $E(X^2)$.

2.2 Geometrické rozdelenie

Geometrické rozdelenie je diskrétne rozdelenie pravdepodobnosti. Náhodná premenná X má geometrické rozdelenie s parametrom $p \in (0, 1)$, ak platí

$$P(X = k) = p(1 - p)^k \quad k = 0, 1, 2, 3... \tag{2}$$

Parameter p voláme aj pravdepodobnosť úspechu. Ako príklad si môžeme predstaviť hod mincou - tieto hody predstavujú postupnosť nezávislých pokusov. Máme dve možnosti - buď padne hlava alebo znak. Obe tieto udalosti nastávajú s rovnakou pravdepodobnosťou - $\frac{1}{2}$. My chceme, aby padla hlava, hádžeme teda mincou, až kým nepadne. Premenná X potom predstavuje počet neúspešných hodov. Pre náš konkrétny príklad platí $P(X = k) = \frac{1}{2}(\frac{1}{2})^k$. Ak sme museli mincou hodiť dvakrát a na tretí raz padla hlava, potom platí $P(X = 2) = \frac{1}{2}(\frac{1}{2})^2 = \frac{1}{8}$.

2.3 Základné pojmy z oblasti matematickej štatistiky

V tejto podkapitole uvedieme základnú terminológiu z oblasti testovania štatistických hypotéz a bližšie sa pozrieme na χ^2 test dobrej zhody. Znova čerpáme z knihy [9].

Definícia 2.9 (Hypotéza) *Hypotézou v štatistike nazývame určité tvrdenie, ktoré sa týka rozdelenia pravdepodobnosti náhodných premenných alebo hodnôt parametrov týchto rozdelení.*

Overovanie správnosti tvrdenia voláme testovaním štatistických hypotéz. Hypotézu, ktorej platnosť overujeme, nazývame testovanou alebo nulovou hypotézou a označujeme ju symbolom H_0. Proti nej kladieme alternatívnu hypotézu H_1. Výsledkom testovania je rozhodnutie, ktorú hypotézu prijmeme a ktorú zamietneme.

2.4 χ^2 test dobrej zhody

χ^2 test dobrej zhody testuje pre nejaké rozdelenie pravdepodobnosti nasledovnú dvojicu hypotéz:

H_0: početnosti v jednotlivých kategóriách sa rovnajú teoretickým početnostiam,

vs.

H_1: početnosti v jednotlivých kategóriách sa nerovnajú teoretickým početnostiam.

Máme teda nejaké početnosti, v našom prípade usporiadané frekvencie grafém a na druhej strane sú tu teoretické početnosti vyplývajúce z definície rozdelenia, ktorého vhodnosť budeme testovať. Ak rozdiel nie je štatisticky významný, teda je malý (v závislosti od rôznych vstupných parametrov), tak nulovú hypotézu nezamietame. Testovacia štatistika má tvar

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}, \tag{3}$$

kde

- f_i sú empirické početnosti,
- p_i sú teoretické pravdepodobnosti vychádzajúce z teoretického pravdepodobnostného rozdelenia,
- n je veľkosť súboru, ktorého prvkov početnosť skúmame ,
- k je počet kategórii, do ktorých prvky rozdeľujeme - v tomto prípade počet rôznych druhov grafém.

Potom táto testovacia štatistika má pre $n \to \infty$ asymptoticky rozdelenie $\chi^2(k-1)$, ak poznáme všetky parametre hypotetického rozdelenia. Od početností žiadame, aby platilo $np_i \geq 5, i = 1, 2, ..., k$. Na tento problém sa môžeme pozrieť aj trochu inak: Chceme testovať, či sú pravdepodobnosti výskytu grafémy v texte $p_i^0, i = 1, 2, ..., N$ (získame ich ako $\frac{f_i}{N}$) rovné teoretickým pravdepodobnostiam daných rozdelením $p_i, i = 1, 2, ..., k$. Naša hypotéza môže mať aj tvar

$$H_0 : (p_1 = p_1^0, p_2 = p_2^0, ..., p_N = p_k^0)$$

Test takejto hypotézy pomocou štatistiky (3) sa nazýva test dobrej zhody.

Trochu odlišný prípad nastáva, keď sú jednotlivé pravdepodobnosti závislé od neznámych parametrov, nazvime ich vo všeobecnosti $u_1, u_2, ..., u_m$. Potom testovacia štatistika vyzerá

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i(u))^2}{np_i(u)} \tag{4}$$

a má rozdelenie χ^2 s $k - m - 1$ stupňami voľnosti, kde m značí počet neznámych parametrov.

Najprv budeme musieť parametre odhadnúť, aby sme získali odhady pre pravdepodobnosti, ktoré potom dosadíme do vzorca pre testovaciu štatistiku (4). Teda platí $p_i = p_i(u), i = 1, 2, ..., k$. Takýto problém budeme riešiť aj v tejto diplomovej práci.

Jednou z možností, ako odhadnúť neznámy parameter u je, že za odhad zoberieme také u, ktoré minimalizuje premennú χ^2. Takýmto odhadom potom hovoríme odhady získané metódou minimálneho χ^2.

Z praktického hľadiska, v tejto práci budeme pracovať v prostredí R so vstavanými funkciami *optim* a *optimize* za účelom nájdenia parametra, ktorý minimalizuje štatistiku. Funkciu *optimize* použijeme pri minimalizácii parametra p v geometrickom rozdelení a funkciu *optim* pri minimalizácii parametrov α a θ v zovšeobecnenom geometrickom rozdelení.

Definícia 2.10 (Hladina významnosti) *Chyba, ktorá spočíva v zamietnutí nulovej hypotézy, hoci je správna, sa nazýva chybou prvého druhu. Pravdepodobnosť chyby prvého druhu nazývame aj hladinou významnosti, označujeme ju α. Zvyčajne volíme $\alpha = 0.01$ alebo $\alpha = 0.05$.*

Definícia 2.11 (P-hodnota) *P-hodnota je pravdepodobnosť, že testovacia štatistika nadobudne väčšiu hodnotu ako tá práve pozorovaná, za podmienky, že H_0 platí.*

Nulovú hypotézu neprijímame, ak je p-hodnota menšia ako požadovaná hladina významnosti.

3 Zovšeobecnené geometrické rozdelenie

3.1 Definícia a vlastnosti nového rozdelenia

Vychádzajúc z článku [1] môžeme zadefinovať pravdepodobnostnú funkciu zovšeobecneného geometrického rozdelenia

$$P[X = x] = \frac{\alpha\theta^x(1-\theta)}{[1-(1-\alpha)\theta^{x+1}][1-(1-\alpha)\theta^x]}, \tag{5}$$

kde pre jeho parametre platí $0 < \theta < 1$ a $0 < \alpha$. Ak $\alpha = 1$, dostávame geometrické rozdelenie.

Základnú charakteristiku modelu tvorí stredná hodnota a disperzia rozdelenia. V tejto práci ich nebudeme odvádzať, keďže pre našu tému nie sú kľúčové, ale ukážeme ich všeobecný tvar, vychádzajúc taktiež zo základného článku [1].

Stredná hodnota tohto rozdelenia je daná ako

$$\mu(\alpha,\theta) = \alpha\frac{d}{d\alpha}log_0\Phi_0(;\frac{1}{\theta},1-\alpha)$$

kde $(a;q)_\infty$ je takzvaný q-Pochhammerov symbol (bližšie vysvetlený v článku [4]), ktorý je definovaný ako

$$(a;q)_\infty = \prod_{k=0}^{\infty}(1-aq^k), \quad 0 < q < 1$$

a ${}_A\Phi_B(:,:)$, je dané ako

$${}_A\Phi_B(a_1,....,a_A;b_1,....b_B;q,z) = \sum_{j=0}^{\infty}\frac{(a_1;q)_j...(a_A;q)_j z^j}{(b_1;q)_j...(b_B;q)_j}.$$

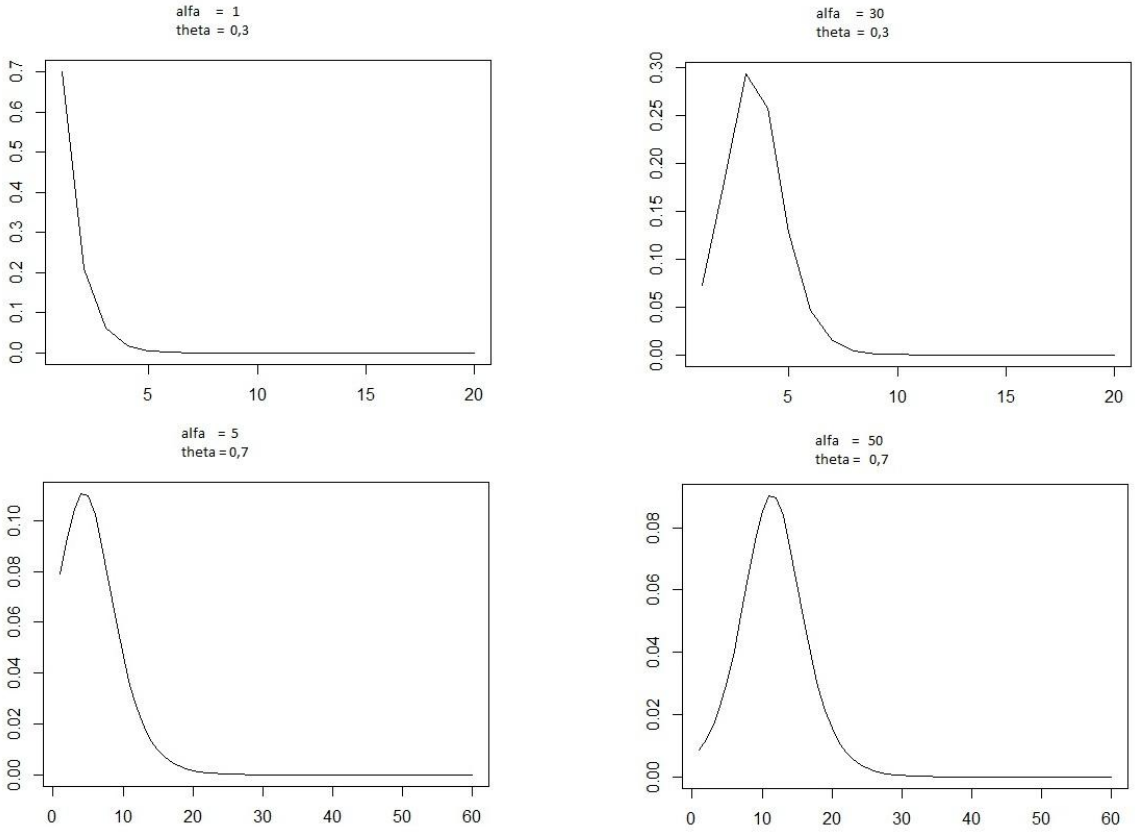
Pomocou vzťahu (1) a iných vzťahov, podľa postupu uvedenom v článku, **disperzia** rozdelenia je

$$D(X) = \sum_{x=1}^{\infty}\frac{(2x-1)\alpha\theta^x}{1-(1-\alpha)\theta^x} - (\mu(\alpha,\theta))^2.$$

Jednou z vlastností rozdelenia je **unimodálnosť**. Znamená to, že určité rozdelenie pravdepodobnosti má len jeden modus. Modus je hodnota, v ktorej pravdepodobnostná

funkcia nadobúda maximum. To, že toto rozdelenie je unimodálne je takisto ukázané v článku [1].

Na Obr.1 vidíme pravdepodobnostnú funkciu zovšeobecneného geometrického rozdelenia pre rôzne parametre. Vo všetkých prípadoch sa môžeme presvedčiť, že funkcia je unimodálna - má práve jedno maximum.



Obr. 1: Pravdepodobnostná funkcia zovšeobecneného geometrického rozdelenia pre rôzne parametre.

3.2 Minimalizačná úloha pre nové rozdelenie a počiatočné odhady jeho parametrov

Pri odhade parametrov zovšeobecneného geometrického rozdelenia, definovaného vzťahom (5), metódou minimálneho χ^2 riešime minimalizačnú úlohu

$$\min_{\alpha,\theta}\{\sum_{i=1}^{k}\frac{(f_i - Np_i(\alpha,\theta))^2}{Np_i(\alpha,\theta)}; 0 < \theta < 1, 0 < \alpha\} \tag{6}$$

Na optimalizáciu použijeme kvázinewtonovskú BFGS (Broyden-Fletcher-Goldfarb-Shanno) metódu, ktorá sa používa pri nelineárnej optimalizácii. Je to funkcia, ktorú má prostredie R predprogramovanú a zo všetkých možností sme ju vybrali práve preto, že je možné vybrať jej modifikáciu L-BFGS-R s ohraničeniami na parametre.

Keďže táto metóda je iteračná, potrebujeme počiatočné odhady parametrov, od ktorých sa budú odvíjať ďalšie iterácie. Tie získame z dát tak, že si vyjadríme prvé dva členy vygenerovaného súboru teoretických pravdepodobností. Keďže v ideálnom prípade platí vzťah $p_i = \frac{f_i}{N}$, tak počiatočné odhady parametrov môžeme získať z nasledujúcich vzťahov:

$$p_1 := \frac{f_1}{N} = \frac{\alpha\theta^0(1-\theta)}{[1-(1-\alpha)\theta^1][1-(1-\alpha)\theta^0]} = \frac{(1-\theta)}{[1-(1-\alpha)\theta]}$$

$$p_2 := \frac{f_2}{N} = \frac{\alpha\theta^1(1-\theta)}{[1-(1-\alpha)\theta^2][1-(1-\alpha)\theta^1]} = \frac{\alpha\theta(1-\theta)}{[1-(1-\alpha)\theta^2][1-(1-\alpha)\theta]}$$

Dostávame dve rovnice o dvoch neznámych, ktorých výsledkom sú nasledujúce dva počiatočné odhady parametrov, ktoré použijeme v aplikačnej časti práce:

$$\begin{aligned} \theta_0 &= \frac{(1-p_1-p_2)p_1}{p_2} \\ \alpha_0 &= \frac{(1-\theta_0)(1-p_1)}{\theta_0 p_1} \end{aligned} \tag{7}$$

3.3 Generovanie náhodných čísel z nového rozdelenia

V skriptách [8] je uvedená všeobecná metóda generovania náhodných čísel z diskrétnych rozdelení s konečnou strednou hodnotou, ktorú využijeme pri programovaní v prostredí R. Náhodné čísla z určitého rozdelenia sú realizácie náhodnej premennej z daného rozdelenia. Existujú rôzne generátory náhodných čísel založené na odlišných princípoch. Avšak, čo majú spoločné je, že najprv generujeme náhodné čísla z rovnomerného rozdelenia na intervale $(0, 1)$ a potom tieto náhodné čísla transformujeme do požadovaného rozdelenia. Celá vstupná informácia, ktorú pri tom potrebujeme, je tvar pravdepodobnostnej funkcie - čo je v prípade geometrického rozdelenia (2) a v prípade nového rozdelenia (5).

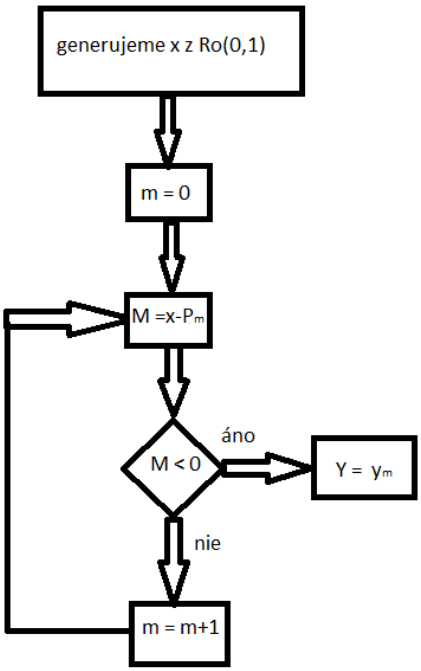
Označme Y náhodnú premennú, ktorá je výstupom nášho generátora náhodných čísel - táto premenná nadobúda hodnoty y_k s pravdepodobnosťami p_k, kde $k = 0, 1, 2, ...$ Ak má premenná X rovnomerné rozdelenie na intervale $(0, 1)$, potom platí $P(a \leq X < b) = b - a$ pre $0 \leq a < b \leq 1$. Potom analogicky platí

$$P\left(\sum_{j=0}^{m-1} p_j \leq X < \sum_{j=0}^{m} p_j\right) = p_m, \quad m = 0, 1, 2, ..$$

Inými slovami, generujeme náhodné číslo $x \sim Ro(0, 1)$ a potrebujeme nájsť index m, pri ktorom je splnená nasledujúca nerovnosť :

$$\sum_{j=0}^{m-1} p_j \leq x < \sum_{j=0}^{m} p_j$$

Pre zjednodušenie zavedieme $P_m = \sum_{j=0}^{m} p_j$ a skúmame, či $x - P_m < 0$. Ak je táto nerovnica splnená, tak sme našli hľadaný index m. Schéma je zobrazená na Obr.2.

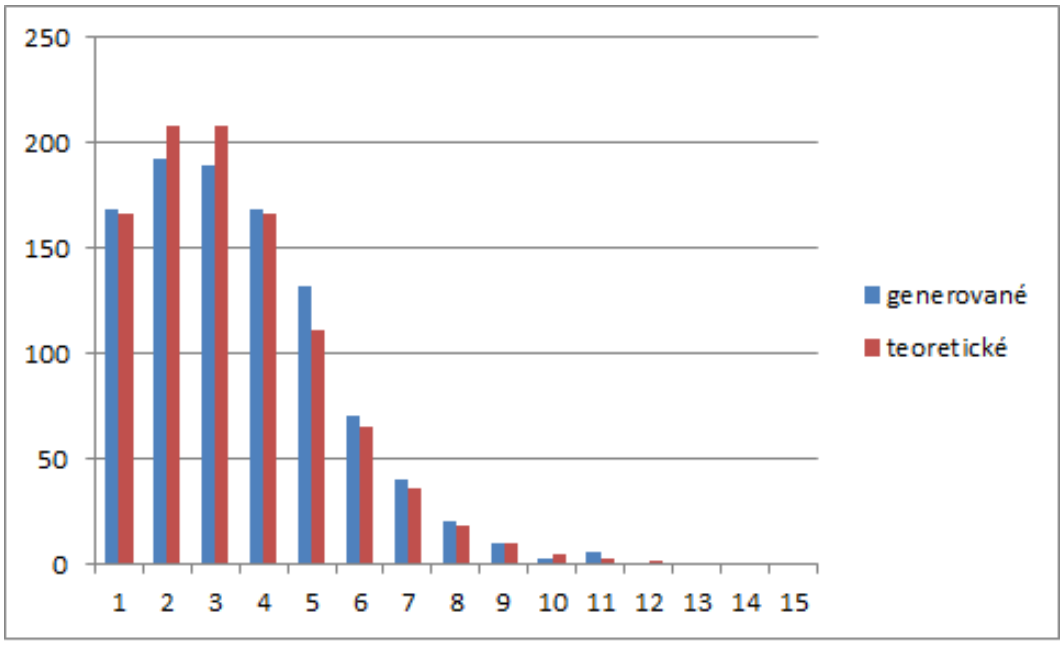


Obr. 2: Schéma generovania náhodných čísel z diskrétnych rozdelení

Táto schéma je teda všeobecná a aplikovateľná na generovanie náhodných čísel z akéhokoľvek rozdelenia s konečnou strednou hodnotou, ak poznáme predpis pravdepodobnostnej funkcie. V tejto práci sme ju využili dvakrát. V prvom prípade sme ju

použili na testovanie toho, či je odhad parametrov z dát dobrý (vygenerovali sme náhodné čísla s vopred známym parametrom a porovnávali s parametrom, ktorý nám odhadol náš program). Tento program sme taktiež využili v kapitole o simulovaných p-hodnotách.

Na ukážku toho, ako dobre tento program funguje, vygenerujeme $N = 1000$ náhodných čísel z nového rozdelenia s parametrami $\alpha = 5$ a $\theta = 0.5$ a zrátame ich frekvencie, v tomto prípade frekvencie výskytu vygenerovaných čísel počnúc od čísla 0. To porovnáme s teoretickými frekvenciami s tými istými parametrami. Tieto frekvencie dostaneme ako $f_i = Np_i$. Porovnanie môžeme vidieť na Obr. 3.



Obr. 3: Porovnanie frekvencií vygenerovaných náhodných čísel a teoretických frekvencií pre parametre $\alpha = 5$ a $\theta = 0.5$.

4 Aplikácie

4.1 Aplikácia základného modelu na frekvencie grafém

K dispozícii máme dáta z nasledujúcich slovanských jazykov: ruský, slovinský, slovenský a ukrajinský. Dáta sme prebrali z článku [6]. Počet znakov (písmen) v jednotlivých abecedách, ako aj celkový počet znakov pozorovaných štatistických súborov sa líši. Našim cieľom je posúdiť, či nové, zovšeobecnené geometrické rozdelenie s dvomi parametrami popisuje tento typ dát lepšie ako geometrické rozdelenie.

V Tabuľkách 2 až 5 môžeme vidieť dáta - v prvom stĺpci je poradové číslo znaku (prvý v poradí je znak s najvyššou frekvenciou), v druhom sa nachádzajú frekvencie grafém z textov v jednotlivých jazykoch, v treťom teoretické frekvencie z geometrického rozdelenia a v štvrtom stĺpci sú teoretické frekvencie zo zovšeobecneného geometrického rozdelenia. Ďalšie štyri stĺpce sú pokračovaním. N udáva celkový počet znakov v pozorovanom štatistickom súbore. Pod tabuľkami môžeme vidieť výstupy.

Vidíme, že v prípade všetkých pozorovaných jazykov je hľadaná minimálna χ^2 štatistika menšia pre nové rozdelenie. Štatistika sa nám môže zdať pomerne vysoká, čo je čiastočne spôsobené tým, ako sme konštruovali jednotlivé teoretické pravdepodobnosti - keďže dát je konečný počet, teoretické pravdepodobnosti sme museli generovať v počte o jednu menej ako je počet znakov v pozorovanom súbore a poslednú zložku sme určili kumulatívne ako

$$p_k = 1 - \sum_{i=1}^{k-1} p_i.$$

Tento postup sa používa veľmi často, keď potrebujeme len konečný počet pravdepodobností, ktorých je prirodzene nekonečný počet. Tento postup sa odzrkadľuje v tabuľkách - frekvencie písmena s najmenším výskytom sa veľmi líšia.

V podkapitole 3.2 sme spomínali, že pri novom rozdelení potrebujeme počiatočné odhady parametrov (7), najprv teda uvádzame tabuľku týchto počiatočných odhadov pre naše štyri súbory dát - Tabuľku 1.

V prípade ruského jazyka sa naše počiatočné odhady vymykali ohraničeniam, ktoré sú dané modelom, $0 < \theta < 1$ a $0 < \alpha$. Pôvodné parametre vyšli ako $\theta = 0.10280$ a $\alpha = -0.2138$. V tomto prípade sme nastavili θ_0 na najvyššiu možnú hodnotu, pre ktorú

Jazyk	θ_0	α_0
ruský	0.9000	0.8729
slovenský	0.8351	1.8531
slovinský	0.7999	2.2002
ukrajinský	0.9311	0.6938

Tabuľka 1: Počiatočné odhady parametrov pre nové rozdelenie.

metóda ziterovala ku konečnej hodnote χ^2 štatistiky, teda $\theta_0 = 0.9$ a α_0 sme dopočítali podľa pôvodného vzorca, t.j. $\alpha_0 = 0.8729$. Keď tieto odhady, ktoré sú jedným zo vstupov do programu, máme, môžeme spustiť iteračný proces pre nové rozdelenie.

Na Obr. 4, 5, 6 a 7 vidíme príslušné histogramy k týmto frekvenčným tabuľkám.

V tejto práci porovnávame to, či je nové rozdelenie vhodnejšie na popis nášho typu dát ako geometrické. Avšak to nám nehovorí o tom, či sú dáta s modelom dostatočne zhodné. χ^2 štatistika rastie spolu s rastúcim N - počtom dát. Ak máme mať univerzálne kritérium dobrej zhody, musí byť indiferentné voči menlivému počtu dát. Preto sa často používa tzv. koeficient diskrepancie (nesúladu) C, ktorý vyrátame ako

$$C = \frac{\chi^2}{N}. \tag{8}$$

Dobrú zhodu dát s modelom dosiahneme, ak je tento koeficient $C < 0.02$. Táto hranica sa zaviedla v článku [3]. Vo výstupoch preto okrem celkového počtu dát, jednotlivých χ^2 štatistík a odhadov parametrov uvádzame aj tento koeficient.

RUSKÝ JAZYK							
i	f(i)	Np(i)-geom.	Np(i)-zovš.	i	f(i)	Np(i)-geom.	Np(i)-zovš.
1	982048	918392.47	806191.19	17	181684	154022.84	164556.59
2	763584	821415.92	756080.02	18	163449	137758.98	145485.01
3	701891	734679.49	704374.31	19	156929	123212.49	128458.50
4	593949	657101.89	652142.62	20	151944	110202.00	113295.95
5	563591	587716.00	600322.17	21	146832	98565.36	99823.11
6	52783	525656.84	549701.37	22	138459	88157.47	87874.84
7	456610	470150.70	500914.17	23	101994	78848.59	77296.66
8	423657	420505.73	454443.99	24	93156	70522.67	67945.41
9	403285	376102.93	410634.73	25	77999	63075.91	59689.60
10	353818	336388.80	369706.30	26	69870	56415.49	52409.26
11	295548	300868.22	331772.61	27	54464	50458.36	45995.59
12	272216	269098.41	296860.18	28	30584	45130.79	35385.42
13	262459	240683.28	264926.14	29	24421	40364.79	35385.42
14	248196	215268.62	235874.74	30	22314	36102.52	31021.66
15	222221	192537.59	209571.96	31	9578	32290.32	27188.59
16	195629	172206.81	185857.81	32	2257	273507.18	191258.11

N=8697409

p=0.1056 χ^2_{geom}=444772.2 C_{geom}=0.0511

α =1.4016 θ=0.8747 χ^2_{zovs}=372540.3 C_{zovs}=0.0428

Tabuľka 2: Frekvencie pre dáta z ruského jazyka.

SLOVENSKÝ JAZYK							
i	f(i)	Np(i)-geom.	Np(i)-zovš.	i	f(i)	Np(i)-geom.	Np(i)-zovš.
1	14194	14117.36	13052.78	24	1611	1393.31	1404.30
2	13772	12765.18	12107.57	25	1593	1259.85	1259.65
3	12701	11542.52	11193.61	26	1465	11.1839	1129.51
4	9285	10436.97	10316.95	27	1422	1030.07	1012.51
5	8323	9437.30	9482.09	28	1395	931.41	907.37
6	7099	8533.39	8692.1720	29	1294	842.20	812.95
7	6562	7716.05	7949.0942	30	1073	761.53	728.19
8	6534	6977.00	7253.72	31	947	688.59	652.14
9	6164	6308.74	6606.02	32	719	622.64	583.92
10	6091	5704.48	6005.26	33	402	563.00	522.76
11	5731	5158.10	5450.16	34	346	509.08	467.94
12	5659	4664.05	4938.98	35	297	460.32	418.81
13	5103	4217.32	4469.68	36	270	416.23	374.80
14	4121	3813.38	4040.03	37	253	376.36	335.37
15	3845	3448.13	3647.62	38	172	340.31	300.07
16	3135	3117.87	3290.04	39	131	307.72	268.46
17	2676	2819.24	2964.84	40	124	278.24	240.16
18	2660	2549.21	2669.60	41	47	251.59	214.84
19	2408	2305.04	2402.01	42	27	227.50	192.17
20	2262	2084.26	2159.82	43	10	205.71	171.88
21	1954	1884.63	1940.89	44	3	186.00	153.73
22	1825	1704.12	1743.23	45	2	168.19	137.49
23	1685	1540.90	1564.95	46	0	1587.77	1161.88
N=147392							
p=0.0958			χ^2_{geom}=5571.448		C_{geom}=0.0378		
α =1.2198		θ=0.8940	χ^2_{zovs}=5140.743		C_{zovs}=0.0349		

Tabuľka 3: Frekvencie pre dáta zo slovenského jazyka.

SLOVINSKÝ JAZYK							
i	f(i)	Np(i)-geom.	Np(i)-zovš.	i	f(i)	Np(i)-geom.	Np(i)-zovš.
1	32036	37409.82	29980.71	13	10514	8152.21	9396.18
2	31891	32949.07	28923.9	14	10216	7180.14	8123.07
3	31122	29020.22	27520.72	15	9568	6323.98	6995.07
4	27150	25559.84	25842.79	16	7446	5569.91	6003.46
5	22905	22512.08	23968.60	17	6413	4905.75	5137.56
6	16088	19827.73	21976.45	18	5361	4320.79	4385.69
7	16084	17463.47	19938.67	19	5055	3805.58	3735.95
8	15221	15381.12	17917.65	20	4608	3351.80	3176.76
9	14668	13547.08	15963.62	21	2606	2952.13	2697.13
10	14043	11931.72	14114.09	22	2554	2600.12	2286.95
11	13034	10508.98	12394.53	23	2463	2290.08	1937.01
12	10517	9255.89	10819.82	24	1675	2017.01	1639.08
				25	497	14898.50	8859.51
N=313745							
p=0.1192				χ^2_{geom}=23016.67		C_{geom}=0.0734	
α =1.7770		θ=0.8419		χ^2_{zovs}=15586.23		C_{zovs}=0.0497	

Tabuľka 4: Frekvencie pre dáta zo slovinského jazyka.

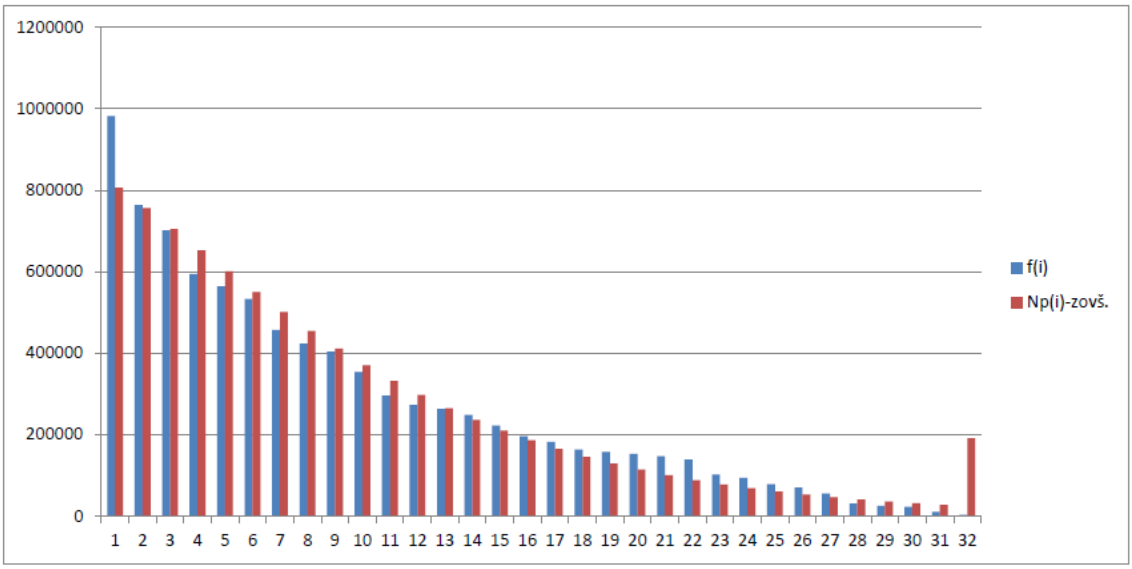
UKRAJINSKÝ JAZYK							
i	f(i)	Np(i)-geom.	Np(i)-zovš.	i	f(i)	Np(i)-geom.	Np(i)-zovš.
1	37267	36185.86	29797.47	17	8944	7510.78	8453.21
2	32774	32798.99	28719.55	18	8877	6807.80	7571.38
3	25080	29729.13	27480.18	19	7487	6170.61	6768.52
4	24639	26946.59	26112.05	20	6888	5593.07	6040.42
5	21053	24424.48	24648.56	21	6406	5069.57	5382.38
6	20941	22138.44	23122.35	22	5850	4595.08	4789.48
7	20075	20066.36	21564.00	23	5074	4165.00	4256.71
8	19171	18188.22	20001.13	24	4625	3775.17	3779.12
9	16296	16485.87	18457.78	25	3876	3421.83	3351.90
10	16240	14942.85	16954.03	26	3843	3101.56	2970.45
11	13936	13544.26	15506.02	27	3565	2811.26	2630.42
12	13697	12276.56	14126.06	28	2857	2548.14	2327.76
13	12959	11127.52	12822.90	29	2790	2309.64	2058.71
14	12949	10086.03	11602.17	30	2484	2093.47	1819.80
15	12398	9142.01	10466.81	31	2407	1897.53	1607.87
16	10584	8286.35	9417.53	32	506	1719.93	1420.05
				33	78	16656.05	10589.24

N=386616

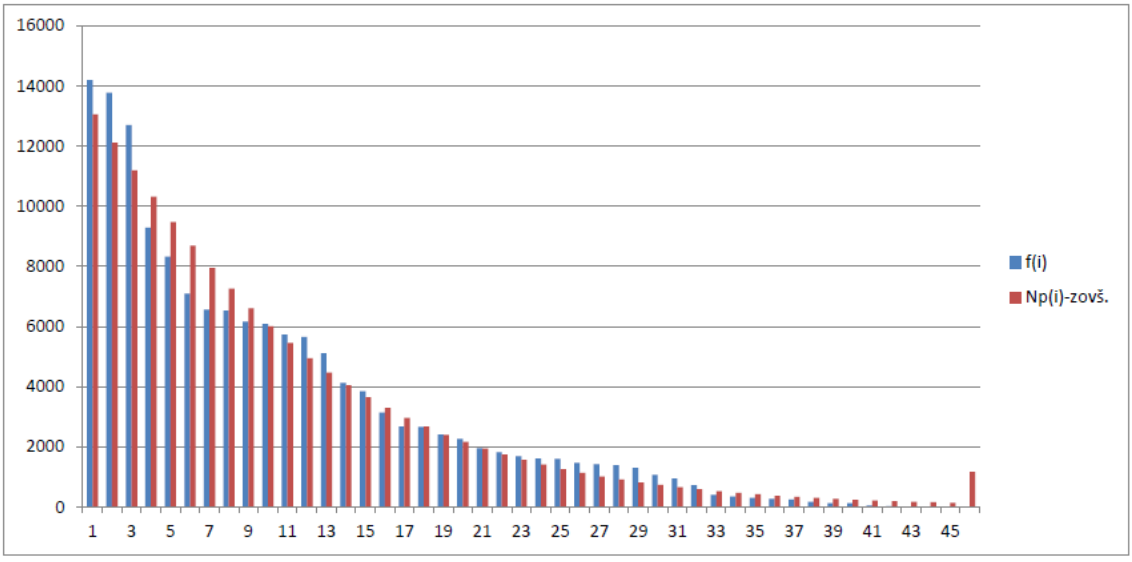
p=0.0936 χ^2_{geom}=25453.14 C_{geom}=0.0658

α =1.6225 θ=0.8805 χ^2_{zovs}=18655.71 C_{zovs}=0.0483

Tabuľka 5: Frekvencie pre dáta z ukrajinského jazyka.

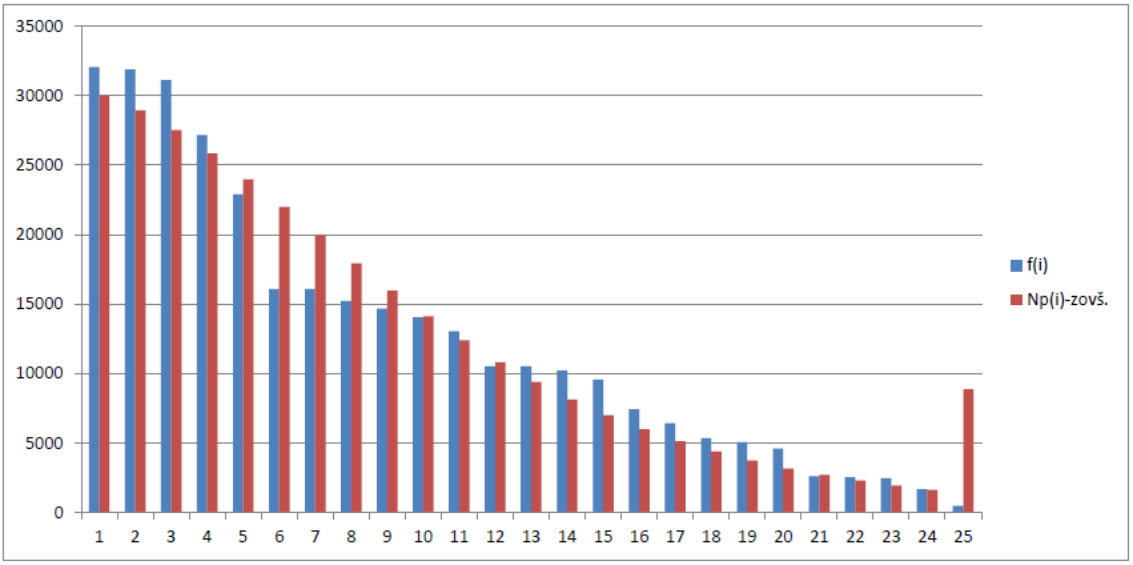


Obr. 4: Porovnanie empirických a teoretických frekvencií (pre zovšeobecnené geometrické rozdelenie) pre dáta z ruštiny.

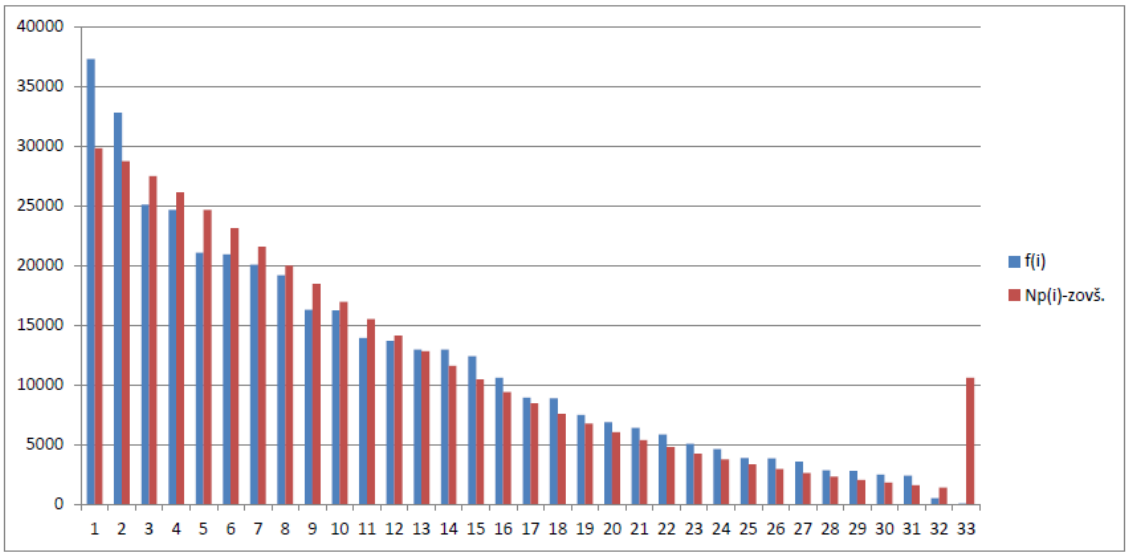


Obr. 5: Porovnanie empirických a teoretických frekvencií (pre zovšeobecnené geometrické rozdelenie) pre dáta zo slovenčiny.

Vidíme, že vo všetkých štyroch prípadoch predstavuje nové rozdelenie zlepšenie oproti geometrickému. Ďalšou možnosťou je použiť vpravo useknutú verziu zovšeobecneného geometrického rozdelenia - použijeme len toľko pravdepodobností, koľko je tried v našom štatistickom súbore. Potom každú pravdepodobnosť vydelíme ich súč-



Obr. 6: Porovnanie empirických a teoretických frekvencií (pre zovšeobecnené geometrické rozdelenie) pre dáta zo slovinčiny.



Obr. 7: Porovnanie empirických a teoretických frekvencií (pre zovšeobecnené geometrické rozdelenie) pre dáta z ukrajinčiny.

tom - normalizujeme ju. Nové pravdepodobnosti majú tvar

$$p_i' = \frac{p_i}{\sum_{i=1}^{k} p_i}, \quad i = 1, 2, ..., k.$$

RUSKÝ JAZYK							
i	f(i)	Np(i)-geom.	Np(i)-zovš.	i	f(i)	Np(i)-geom.	Np(i)-zovš.
1	982048	880763.51	866241.77	17	181684	170728.11	171575.62
2	763584	794922.28	787023.50	18	163449	154088.55	154511.12
3	701891	717447.34	714296.10	19	156929	139070.73	139114.99
4	593949	647523.28	647669.09	20	151944	125516.58	125229.64
5	563581	584414.17	586747.33	21	146832	113283.45	112711.29
6	532783	527455.83	531138.25	22	138459	102242.58	101428.98
7	456610	476048.78	480457.41	23	101994	92277.79	91263.62
8	423657	429651.97	434332.74	24	93156	83284.18	82107.00
9	403285	387777.11	392407.61	25	77999	75167.11	73860.94
10	353818	349983.46	354343.00	26	69870	67841.15	66436.47
11	295548	315873.27	319818.83	27	54464	61229.20	59752.98
12	272216	285087.54	288534.69	28	30584	55261.66	53737.54
13	262459	257302.26	260210.06	29	24421	49875.73	48324.20
14	248196	232224.99	234584.14	30	22314	45014.73	43453.37
15	222221	209591.81	211415.41	31	9578	40627.49	39071.22
16	195629	189164.51	190480.95	32	2257	36667.84	35129.15

N=8697409			
p=0.0975		χ^2_{geom}=150543.9	C_{geom}=0.0173
α =1.0598	θ=0.8987	χ^2_{zovs}=149458.4	C_{zovs}= 0.0172

Tabuľka 6: Frekvencie pre dáta z ruského jazyka s použitím sprava useknutých pravdepodobností.

Takýmto spôsobom zostáva zaručené, že $\sum_{i=1}^{k} p'_i = 1$. Vpravo useknuté zovšeobecnené geometrické rozdelenie opäť aplikujeme na dáta z Tabuliek 2 až 5. Uvidíme, že táto zmena predstavuje zlepšenie výsledkov, t.j. zníženie χ^2 štatistiky. Výsledky vidíme v Tabuľkách 6 až 9.

SLOVENSKÝ JAZYK							
i	f(i)	Np(i)-geom.	Np(i)-zovš.	i	f(i)	Np(i)-geom.	Np(i)-zovš.
1	14194	13873.98	13433.74	24	1611	1466.59	1462.94
2	13772	12582.62	12318.01	25	1593	1330.09	1321.54
3	12701	11411.45	11277.89	26	1465	1206.28	1193.60
4	9285	10349.30	10311.32	27	1422	1094.01	1077.90
5	8323	9386.01	9415.67	28	1395	992.18	973.28
6	7099	8512.38	8587.88	29	1294	899.83	878.71
7	6562	7720.06	7824.62	30	1073	816.07	793.24
8	6534	7001.49	7122.35	31	947	740.12	716.02
9	6164	6349.81	6477.45	32	719	671.23	646.26
10	6091	5758.78	5886.26	33	402	608.75	583.24
11	5731	5222.77	5345.16	34	346	552.09	526.34
12	5659	4736.64	4850.62	35	297	500.70	474.95
13	5103	4295.76	4399.22	36	270	454.10	428.56
14	4121	3895.92	3987.67	37	253	411.83	386.65
15	3845	3533.30	3612.85	38	172	373.50	348.88
16	3135	3204.43	3271.81	39	131	338.73	314.75
17	2676	2906.17	2961.78	40	124	307.21	283.96
18	2660	2635.67	2680.14	41	47	278.61	256.17
19	2408	2390.34	2424.49	42	27	252.68	231.09
20	2262	2167.85	2192.58	43	10	229.16	208.46
21	1954	1966.07	1982.31	44	3	207.83	188.04
22	1825	1783.08	1791.77	45	2	188.49	169.62
23	1685	1617.11	1619.19	46	0	170.94	153.00

N=1473929

p=0.0931 χ^2_{geom}=3979.568 C_{geom}=0.0270

α =1.0968 θ=0.9018 χ^2_{zovs}=3910.018 C_{zovs}=0.0265

Tabuľka 7: Frekvencie pre dáta zo slovenského jazyka s použitím sprava useknutých pravdepodobností.

SLOVINSKÝ JAZYK							
i	f(i)	Np(i)-geom.	Np(i)-zovš.	i	f(i)	Np(i)-geom.	Np(i)-zovš.
1	32036	35136.44	32768.00	13	10514	9281.95	9588.12
2	31891	31447.14	30281.28	14	10216	8307.36	8495.51
3	31122	28145.22	27825.40	15	9568	7435.09	7515.39
4	27150	25190.00	25436.08	15	7446	6654.41	6639.00
5	22905	22545.07	23141.63	17	6413	5955.71	5857.52
6	16088	20177.86	20963.25	618	5361	5330.36	5162.36
7	16084	18059.20	18915.63	19	5055	4770.68	4545.31
8	15221	16163.00	17007.73	20	4608	4269.76	3998.61
9	14668	14465.90	15243.69	21	2606	3821.44	3515.03
10	14043	12947.00	13623.70	22	2554	3420.19	3087.89
11	13034	11587.57	12144.88	23	2463	3061.08	2711.10
12	10517	10370.89	10801.99	24	1675	2739.67	2379.07
				25	497	2452.00	2086.78
N=313745							
p=0.1050		χ^2_{geom}=6213.464				C_{geom}=0.0198	
α =1.2962		θ=0.8744	χ^2_{zovs}=5539.527			C_{zovs}= 0.0177	

Tabuľka 8: Frekvencie pre dáta zo slovinského jazyka s použitím sprava useknutých pravdepodobností.

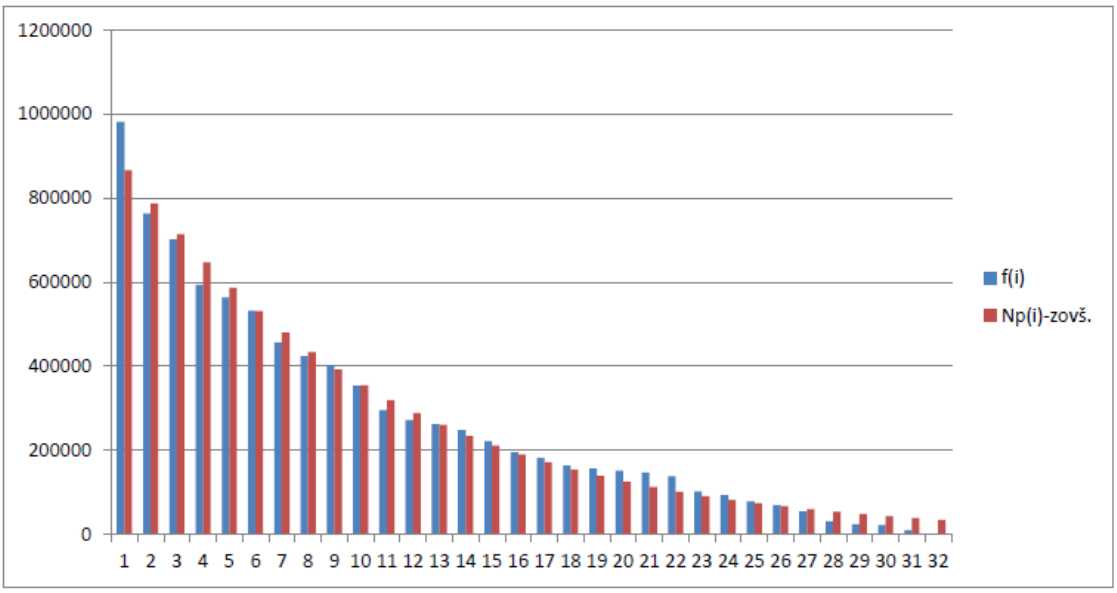
UKRAJINSKÝ JAZYK							
i	f(i)	Np(i)-geom.	Np(i)-zovš.	i	f(i)	Np(i)-geom.	Np(i)-zovš.
1	37267	34124.51	32528.45	17	8944	8493.04	8670.33
2	32774	31283.56	30327.51	18	8877	7785.97	7908.96
3	25080	28679.13	28209.99	19	7487	7137.77	7210.08
4	24639	26291.52	26183.76	20	6888	6543.53	6569.33
5	21053	24102.69	24254.45	21	6406	5998.77	5982.51
6	20941	22096.09	22425.64	22	5850	5499.36	5445.62
7	20075	20256.53	20699.19	23	5074	5041.52	4954.84
8	19171	18570.13	19075.42	24	4625	4621.80	4506.59
9	16296	17024.12	17553.39	25	3876	4237.03	4097.47
10	16240	15606.82	16131.13	26	3843	3884.28	3724.33
11	13936	14307.52	14805.81	27	3565	3560.91	3384.20
12	13697	13116.38	13574.00	28	2857	3264.45	3074.34
13	12959	12024.41	12431.75	29	2790	2992.68	2792.19
14	12949	11023.35	11374.79	30	2484	2743.53	2535.39
15	12398	10105.63	10398.64	31	2407	2515.13	2301.77
16	10584	9264.31	9498.70	32	506	2305.74	2089.30
				33	78	2113.79	1896.14

N=386616

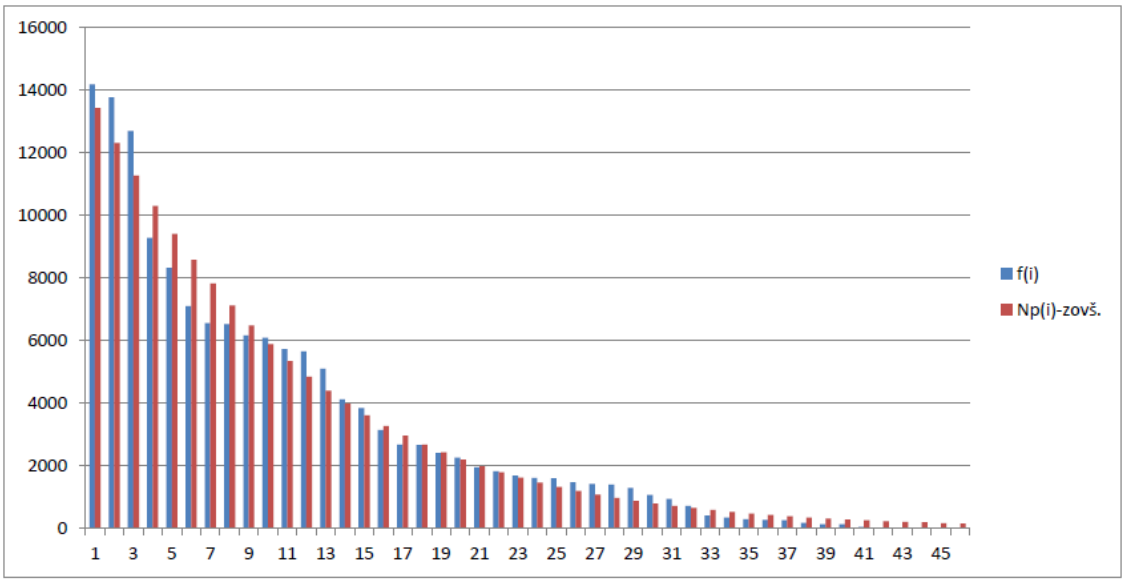
p=0.0833 χ^2_{geom}=6345.157 C_{geom}=0.0164

α =1.1863 θ=0.9061 χ^2_{zovs}=5966.918 C_{zovs}=0.0154

Tabuľka 9: Frekvencie pre dáta z ukrajinského jazyka s použitím sprava useknutých pravdepodobností.

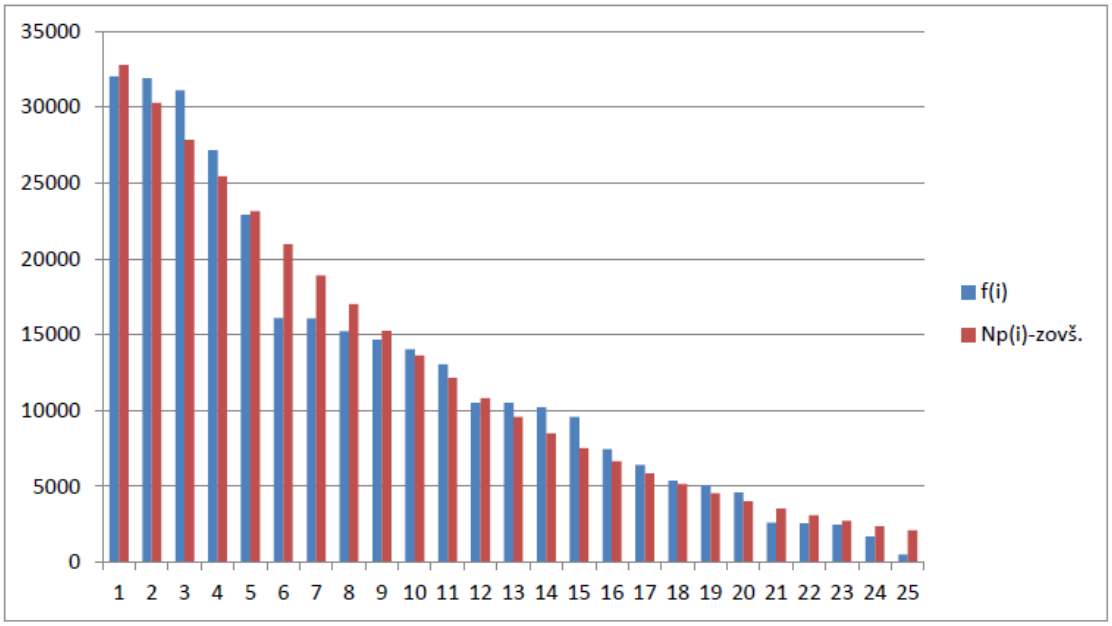


Obr. 8: Porovnanie empirických a teoretických frekvencií (pre zovšeobecnené geometrické rozdelenie) pre dáta z ruštiny s použitím sprava useknutých pravdepodobností.

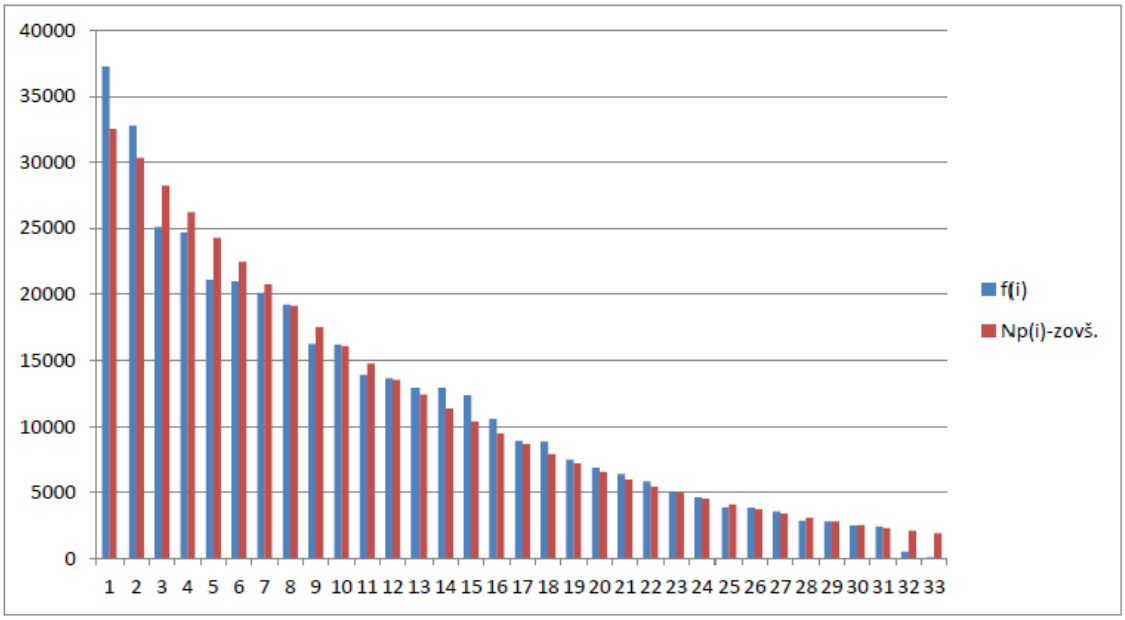


Obr. 9: Porovnanie empirických a teoretických frekvencií (pre zovšeobecnené geometrické rozdelenie) pre dáta zo slovenčiny s použitím sprava useknutých pravdepodobností.

Vidíme, že pre všetky skúmané jazyky okrem slovenčiny platí, že $C < 0.02$, teda model popisuje dáta dostatočne dobre.



Obr. 10: Porovnanie empirických a teoretických frekvencií (pre zovšeobecnené geometrické rozdelenie) pre dáta zo slovinčiny s použitím sprava useknutých pravdepodobností.



Obr. 11: Porovnanie empirických a teoretických frekvencií (pre zovšeobecnené geometrické rozdelenie) pre dáta z ukrajinčiny s použitím sprava useknutých pravdepodobností.

4.2 Modifikácia

Sú jazyky, v ktorých sa určité písmeno (graféma) vyskytuje omnoho častejšie ako zvyšné. Takýmto jazykom je napríklad tamilčina, dáta sú uvedené v Tabuľke 10. Tamilčina je jazyk, ktorým sa rozpráva na juhu Indie a na severo-východe Srí Lanky. Má 12 samohlások a 18 spoluhlások. Pochopiteľne, potom náš model v jeho pôvodnej forme nemôžeme na dáta v tomto jazyku aplikovať, musíme vykonať určitú zmenu. Musíme modifikovať, podľa článku [11], jednotlivé pravdepodobnosti výskytu jednotlivých písmen v texte. Modifikácií sa v článku uvádza niekoľko druhov, napr. modifikácia prvej triedy, ktorú použijeme aj my, sa vykonáva vtedy, ak má výrazne vyššiu frekvenciu len jedno písmeno. Pravdepodobnosti teda zmeníme nasledovným spôsobom:

$$\begin{aligned} \pi_1 &= 1 - \gamma(1 - P_1) \\ \pi_i &= \gamma P_i; \qquad i = 2, 3, ...k \end{aligned} \tag{9}$$

kde $P_i; i = 1, ..., k$ sú jednotlivé teoretické pravdepodobnosti výskytu písmen v texte, ktoré vstupujú do modelu a parameter $\gamma \in (0, [1 - P_1]^{-1})$. Platí, že pre $\gamma = 1$ dostávame pôvodné pravdepodobnosti. Parameter γ je neznámy a potrebujeme ho odhadnúť. π_i je znova rozdelenie pravdepodobnosti.

Dáta z tamilčiny sme prevzali z článku [6]. Ak sa pokúsime na dáta v tamilčine aplikovať náš program v prostredí R, ktorý v slovanských jazykoch dosahoval relatívne uspokojivé výsledky, dostaneme $C = 0.0354 > 0.02$. V tejto modifikovanej forme nášho problému riešime nasledujúcu minimalizačnú úlohu:

$$\min_{\alpha,\theta,\gamma}\{\sum_{i=1}^{k} \frac{(f_i - N\pi_i(\alpha, \theta, \gamma))^2}{N\pi_i(\alpha, \theta, \gamma)}; 0 < \theta < 1, 0 < \alpha, \gamma \in (0, [1 - P_1]^{-1})\} \tag{10}$$

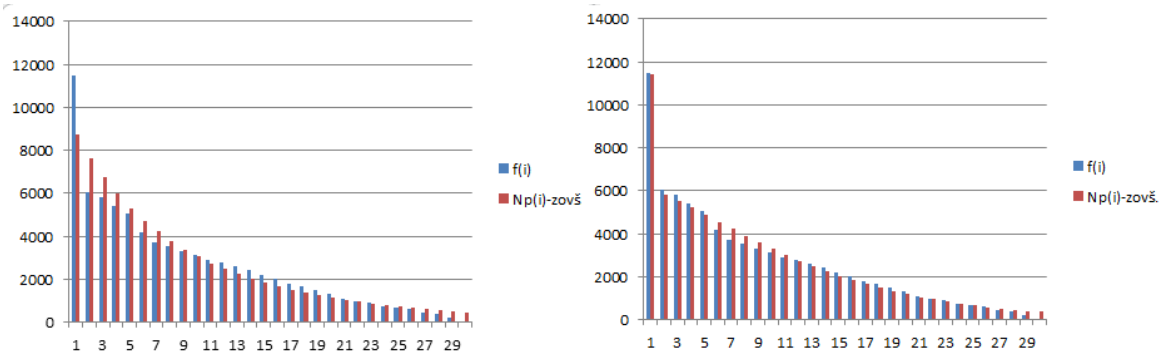
Počiatočné odhady parametrov α a θ sme nechali rovnaké, ako v pôvodnom programe, avšak výstupy nevyšli konečné, preto sme vstupné parametre museli ručne upravovať, aby proces dokonvergoval ku konečným výsledkom. Pre tieto dáta z tamilčiny máme po úprave počiatočné parametre $\theta = 0.90$ a $\alpha = 0.75$. Počiatočný odhad parametra γ sme pre jednoduchosť nastavili na $\gamma_0 = 1$.

Samotný program, ktorý vykonal minimalizáciu, môžeme nájsť v prílohe práce.

TAMILSKÝ JAZYK					
1	11462	11399.42	16	2019	1829.35
2	6050	5805.73	17	1763	1643.84
3	5844	5514.09	18	1669	1474.42
4	5423	5204.95	19	1510	1320.29
5	5083	4884.67	20	1329	1180.52
6	4157	4559.16	21	1087	1054.16
7	3729	4233.73	22	970	940.21
8	3526	3912.97	23	899	837.70
9	3331	3600.74	24	755	745.66
10	3119	3300.11	25	674	663.19
11	2902	3013.45	26	623	589.39
12	2807	2742.43	27	420	523.47
13	2599	2488.14	28	382	464.64
14	2440	2251.16	29	201	412.21
15	2214	2031.63	30	0	365.53
N=78987		$\gamma = 0.9349$			α =1.5270
θ=0.8836		χ^2=871.876			C = 0.0110

Tabuľka 10: Frekvencie pre dáta z tamilčiny.

Výsledok našej modifikácie modelu je pre tento štatistický súbor uspokojivý - koeficient $C < 0.02$, teda pozmenený model dobre vystihuje naše dáta. Na Obr. 12 môžeme vidieť dva histogramy. Vľavo je porovnanie dát s teoretickými frekvenciami pred modifikáciou a vpravo porovnanie dát s teoretickými frekvenciami po modifikácii.



Obr. 12: Porovnanie frekvencií pred a po modifikácii pre tamilčinu.

4.3 Modelovanie dĺžky slov

Ťažiskom tejto diplomovej práce je modelovať frekvencie grafém, ale môžeme tiež skúmať, či zovšeobecnené geometrické rozdelenie nemodeluje dobre aj štatistické súbory iného typu. Vybrali sme dĺžku slov meranú počtom slabík, dáta sme prevzali z článku [2], dáta sú tentokrát z anglického a litovského jazyka. Opäť porovnávame mieru zhody dát s geometrickým a zovšeobecneným geometrickým rozdelením. Budeme pracovať len s useknutými verziami týchto rozdelení.

V Tabuľkách 11 a 12 uvidíme znova dáta a následne aj výstupy usporiadané tak, ako doteraz.

Koeficient diskrepancie C naznačuje v oboch prípadoch akceptovateľnú mieru zhody medzi modelom a dátami.

ANGLICKÝ JAZYK			
počet slabík	f(i)	Np(i)-geom.	Np(i)-zovš
1	3987	3845.66	3970.54
2	831	1022.69	857.98
3	281	271.97	269.08
4	121	72.33	93.51
5	15	19.23	33.64
6	2	5.12	12.248
N=5237			
p=0.7341		χ^2_{geom}=77.0111	C_{geom}=0.0147
α =0.5539	θ=0.3667	χ^2_{zovs}=28.4268	C_{zovs}=0.0054

Tabuľka 11: Frekvencie pre dáta z angličtiny.

LITOVSKÝ JAZYK			
počet slabík	f(i)	Np(i)-geom.	Np(i)-zovš
1	525	514.12	521.49
2	116	136.52	125.00
3	48	36.25	37.56
4	9	9.63	12.02
5	2	3.48	3.92
N=700			
p=0.7345		χ^2_{geom}=7.7923	C_{geom}=0.0111
α =0.7043	θ=0.3294	χ^2_{zovs}=5.2714	C_{zovs}=0.0075

Tabuľka 12: Frekvencie pre dáta z litovčiny.

4.4 Simulované p-hodnoty

Podľa článku [10], ak je súbor dát príliš veľký (čo je často v lingvistike pravdou), nulovú hypotézu, uvedenú v podkapitole 2.4, test zamieta. Preto sa v lingvistike používa práve koeficient diskrepancie C, nie p - hodnota. Z tohto dôvodu sa budeme simulovanou p-hodnotou zaoberať len pre dáta z litovčiny, kde $N = 700$.

V tejto podkapitole budeme p-hodnotu simulovať, pričom budeme vychádzať z jej definície uvedenej v podkapitole 2.4. Pre pripomenutie, p-hodnota je pravdepodobnosť toho, že testovacia štatistika nadobudne väčšiu hodnotu ako je tá práve pozorovaná.

Postup je nasledovný:

1. Nájdeme parametre, ktoré minimalizujú χ^2 štatistiku. Tieto parametre pre obe rozdelenia môžeme vidieť v Tabuľke 12.

2. Vygenerujeme N náhodných čísel z príslušného rozdelenia. V tomto prípade je $N = 700$.

3. Tento postup zopakujeme dostatočne veľa ráz - v našom programe to je 2000.

4. Zrátame, koľkokrát boli χ^2 štatistiky väčšie ako štatistika z Tabuľky 12. Simulovaná p-hodnota je potom pomer počtu týchto štatistík, ktoré boli väčšie ako štatistika z Tabuľky 12 k číslu 2000, čo je počet všetkých vygenerovaných štatistík.

Keď program spustíme, pri geometrickom rozdelení vyjde p-hodnota okolo 0.2605 (ak spúšťame program opakovane, čísla sa stále pohybujú okolo tejto hodnoty) a pri zovšeobecnenom 0.3775. Zvyčajne sa ako kritická hodnota berie 0.05, čiže v oboch prípadoch nezamietame nulovú hypotézu, že dáta pochádzajú z daného rozdelenia, resp. že sa frekvencie grafém rovnajú teoretickým frekvenciám.

5 Záver

V tejto práci sme prebrali vzorec pre pravdepodobnostnú funkciu zovšeobecneného geometrického rozdelenia. Naším cieľom bolo porovnať mieru zhody dát s geometrickým a zovšeobecneným geometrickým rozdelením. Dáta boli z oblasti kvantitatívnej lingvistiky - frekvencie grafém z ruského, slovenského, slovinského a ukrajinského jazyka. Ako kritérium zhody nám poslúžil χ^2 test dobrej zhody, kde výstupmi boli optimálne odhady parametrov rozdelení, samotné hodnoty testovacích štatistík a koeficient diskrepancie C. Tento koeficient sme použili ako konečné kritérium toho, či je model prijateľný alebo nie. Model je prijateľný, ak jeho koeficient diskrepancie $C < 0.02$. Čím nižší je tento koeficient, tým je zhoda dát s modelom lepšia.

Naším primárnym cieľom bolo porovnať dve spomenuté rozdelenia a overiť, že zovšeobecnené geometrické rozdelenie popisuje tento typ dát lepšie ako geometrické, čo sa nám aj podarilo. Nezávisle od výberu rôznych dát v tejto práci sa vždy nové rozdelenie ukázalo ako vhodnejšie, čo čitateľ môže vidieť v jednotlivých tabuľkách porovnávajúcich dané frekvencie grafém a teoretické frekvencie z oboch rozdelení. Sekundárnym cieľom bolo zhodnotiť v každom súbore dát, či sú jednotlivé rozdelenia vhodné na popis dát (to, že je nejaké rozdelenie lepšie ako iné ešte neznamená, že je vhodné), na čo nám poslúžil koeficient diskrepancie. Uplatnili sme dve metódy generovania pravdepodobností. V Tabuľke 13 uvádzame koeficienty diskrepancie pre štyri slovanské jazyky a pre dve metódy generovania pravdepodobností kvôli prehľadu. Druhá metóda generovania pravdepodobností, tzv. sprava useknutých, sa ukazuje byť lepšia ako prvá metóda, kde posledný člen radu pravdepodobností určujeme kumulatívne ako $p_k = 1 - \sum_{i=1}^{k-1} p_i$. GR v Tabuľke 13 znamená geometrické rozdelenie a ZGR zovšeobecnené geometrické rozdelenie.

Z Tabuľky 13 sa dá vyčítať záver, ktorý bol v práci niekoľkokrát spomenutý - ZGR vždy dosahuje lepšie výsledky ako GR.

Najlepšie výsledky dosahujeme pri kombinácii druhej metódy generovania pravdepodobností a použitia ZGR. Vtedy je koeficient diskrepancie pre ruštinu, slovinčinu a ukrajinčinu vždy menší ako 0.02.

Okrem štyroch slovanských jazykov sme skúmali aj tamilčinu špecifickú tým, že jedna graféma sa vyskytovala oveľa častejšie ako ostatné. Z tohto dôvodu sme program museli upraviť, čo sa nám aj úspešne podarilo. Po upravení bol opäť koeficient diskrepancie $C < 0.02$.

	1.metóda - GR	1.metóda ZGR	2.metóda - GR	2.metóda ZGR
ruština	0.0511	0.0428	0.0173	0.0172
slovenčina	0.0378	0.0349	0.0270	0.0265
slovinčina	0.0734	0.0497	0.0198	0.0177
ukrajinčina	0.0658	0.0483	0.0164	0.0154

Tabuľka 13: Porovnanie modelov.

Posledné dva súbory dát, ktoré sme skúmali, boli dáta trochu iného typu - boli to merania počtu slabík slov v textoch z angličtiny a litovčiny. Opäť sme prišli k rovnakému výsledku - ZGR vykazovalo lepšie výsledky ako GR. K litovčine sme vygenerovali aj simulované p-hodnoty.

Výsledkom tejto práce je potvrdenie, že pre naše dáta z oblasti kvantitatívnej lingvistiky je naozaj zovšeobecnené geometrické rozdelenie vhodnejším rozdelením ako pôvodné, geometrické.

Literatúra

[1] Gómez - Déniz, E. Another generalization of the geometric distribution, In: *Test 19 (2010)*, pp. 399-415.

[2] Grzybek, P. (ed.) *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, pp. 15-99. Dordrecht: Springer, 2007.

[3] Grzybek, P., Stadlober, E., Antić, G. Mathematical aspects and modifications of Fucks' generalized Poisson distribution, In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook*, pp 158-205. Berlin: de Gruyter, 2005.

[4] Johnson, N.L., Kemp, A.W., Kotz, S. *Univariate Discrete Distributions.* Hoboken: Wiley, 2005.

[5] Kelih, E., Altmann, G., Grzybek, P. Graphemhäufigkeiten, Teil II: Modelle der Häufigkeitsverteilung, In: *Anzeiger fur Slavische Philologie XXIII(2004)*, pp. 25–54.

[6] Mačutek, J. A generalization of the geometric distribution and its application in quantitative linguistics, In: *Romanian Reports in Physics, Vol. 60 (2008)*, pp. 501–509.

[7] Pázman, A., Janková, K., *Pravdepodobnosť a štatistika*, Univerzita Komenského, 2011.

[8] Pekár, J. Kalas, J. *Simulačné metódy*, Univerzita Komenského, 1991.

[9] Potocký, R., Lamoš, F. *Pravdepodobnosť a matematická štatistika*, Univerzita Komenského, 1998.

[10] Wimmer, G., Mačutek, J. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics, In: *Journal of Quantitative Linguistics*, (to appear) 2013.

[11] Witkovský, V., Altmann, G., Wimmer, G. Modification of probability distributions applied to word length research, In: *Journal of Quantitative Linguistics (1999)*, pp. 257–268.

Prílohy

Zdrojové kódy programov v R.

Generovanie náhodných čísel z nového rozdelenia pre vopred určené parametre

```
#generujeme pravdepodobnosti pre zovseobecnene geometricke
N<-1000
alpha<-5
alpha2<-(1-alpha)
lambda<-0.59
#theta<-exp(-lambda)
theta<-0.5
Pr<-rep(0,times=N)

for (i in 0:(N-2))
  Pr[i+1]<-(alpha*theta^i*(1-theta))/((1-alpha2*theta^(i+1))*(1-alpha2*theta^i))
  Pr[N]<-1-sum(Pr)

#V Pr su pravdepodobnosti pre ZGR
#teraz generujeme u~Ro(0,1)

u <- runif(N)
Y <- rep(0,times=length(u))
M<-0

for (i in 1:length(u))  {
  xx <- u[i]
  for (m in 0:99){
    Pm <- sum(Pr[1:(m+1)])
    M <- xx - Pm
    if (M < 0){
      Y[i] <- m
      break
    }
  }
}

max<-0

for (i in 1:N)
    if (Y[i]>max) max<-Y[i]

frek<-rep(0,times=max+1)
```

```
for (i in 1: (N+1))
   frek[Y[i]+1] <- (frek[Y[i]+1]+1)
```

χ^2 test pre geometrické rozdelenie

```
frek1 <- read.table("cesta ku súboru")
attach(frek1)
frek<-t(frek1)
N<-sum(frek)
length(frek)
#mame načítané dáta
TeorPrav<-function(p) {
teor<-rep(0,times=length(frek))
sucet<-0
for (k in 0:(length(frek)-2)) {
teor[k+1] <- (1-p)^k*p
sucet<-sucet+teor[k+1]
}
teor[length(frek)]<- 1-sucet
vysledne<-teor*N
 for (i in 1:(length(vysledne)))
   if(vysledne[i]==0) vysledne[i]<-0.00000001
Statistika <- sum( (frek-vysledne)^2 / vysledne )
return(Statistika)
}

optimize(TeorPrav,c(0,1))
```

χ^2 test pre zovšeobecnené geometrické rozdelenie

```
frek1 <- read.table("cesta ku súboru")
attach(frek1)
frek<-t(frek1)
N<-sum(frek)
length(frek)
P1<-frek[1]/N
P2<-frek[2]/N
pocThe<-((1-P1-P2)*P1)/P2
if (pocThe>1) pocThe<-0.9
if (pocThe<0) pocThe<-0.1
pocAlp<-((1-pocThe)*(1-P1))/(pocThe*P1)
if (pocAlp<0) pocAlp<-0.1
#alpha=x[1], theta=x[2]
TeorPrav<-function(x) {
teor<-rep(0,times=length(frek))
```

```
sucet<-0
for (k in 0:(length(frek)-2)) {
teor[k+1] <- (x[1]*x[2]^k*(1-x[2]))/((1-(1-x[1])*x[2]^(k+1))*
(1-(1-x[1])*x[2]^k))
sucet<-sucet+teor[k+1]
}
teor[length(frek)]<- 1-sucet
vysledne<-teor*N
Statistika <- sum( (frek-vysledne)^2 / vysledne )
return(Statistika)
}

optim(par=c(pocAlp,pocThe),fn=TeorPrav,method="L-BFGS-B",
lower=c(0.00001,0.00001),upper=c(99.9999,0.999999),
 control=list(maxit=2000))
```

χ^2 test pre geometrické rozdelenie, použitie sprava useknutých pravdepodobností

```
frek1 <- read.table("cesta ku súboru")
attach(frek1)
frek<-t(frek1)
N<-sum(frek)
length(frek)
TeorPrav<-function(p) {
teoret<-rep(0,times=length(frek))
teor<-rep(0,times=length(frek))
sucet<-0
for (k in 0:(length(frek)-1)) {
teoret[k+1] <- (1-p)^k*p
sucet<-sucet+teoret[k+1]
}
teor<-teoret/sucet
vysledne<-teor*N
for (i in 1:(length(vysledne)))
   if(vysledne[i]==0) vysledne[i]<-0.00000001
Statistika <- sum( (frek-vysledne)^2 / vysledne )
return(Statistika)
}
optimize(TeorPrav,c(0,1))
```

χ^2 test pre zovšeobecnené geometrické rozdelenie, použitie sprava useknutých pravdepodobností

```
frek1 <- read.table("cesta ku súboru")
attach(frek1)
```

```
frek<-t(frek1)
N<-sum(frek)
length(frek)
P1<-frek[1]/N
P2<-frek[2]/N
pocThe<-((1-P1-P2)*P1)/P2
if (pocThe>1) pocThe<-0.9
if (pocThe<0) pocThe<-0.1
pocAlp<-((1-pocThe)*(1-P1))/(pocThe*P1)
if (pocAlp<0) pocAlp<-0.1
#alpha=x[1], theta=x[2]
TeorPrav<-function(x) {
teoret<-rep(0,times=length(frek))
teor<-rep(0,times=length(frek))
sucet<-0
for (k in 0:(length(frek)-1)) {
teoret[k+1] <- (x[1]*x[2]^k*(1-x[2]))/((1-(1-x[1])*x[2]^(k+1))*
(1-(1-x[1])*x[2]^k))
sucet<-sucet+teoret[k+1]
}
teor<-teoret/sucet
vysledne<-teor*N
Statistika <- sum( (frek-vysledne)^2 / vysledne )
return(Statistika)
}

optim(par=c(pocAlp,pocThe),fn=TeorPrav,method="L-BFGS-B",
lower=c(0.00001,0.00001),upper=c(99.9999,0.999999),
 control=list(maxit=2000))
```

Modifikácia prvého stupňa

```
frek1 <- read.table("cesta ku súboru")
attach(frek1)
frek<-t(frek1)
N<-sum(frek)
length(frek)

P1<-frek[1]/N
P2<-frek[2]/N

pocThe<-((1-P1-P2)*P1)/P2
pocAlp<-((1-pocThe)*(1-P1))/(pocThe*P1)

if (pocThe>1) pocThe<-0.9
if (pocThe<0) pocThe<-0.1
```

```
#pocAlp<-((1-pocThe)*(1-P1))/(pocThe*P1)
pocAlp<-0.75
pocGama<-1

#alpha=x[1], theta=x[2],  gama=x[3]

TeorPrav<-function(x) {
teoret<-rep(0,times=length(frek))
teor<-rep(0,times=length(frek))
new<-rep(0,times=length(frek))

sucet<-0
for (k in 0:(length(frek)-1)) {
teoret[k+1] <- (x[1]*x[2]^k*(1-x[2]))/((1-(1-x[1])*x[2]^(k+1))*
(1-(1-x[1])*x[2]^k))

 if (k==0) new[k+1]<-(1-x[3]*(1-teoret[k+1]))
           else    new[k+1]<-x[3]*teoret[k+1]

sucet<-sucet+new[k+1]
}
teor<-new/sucet
vysledne<-teor*N
Statistika <- sum( (frek-vysledne)^2 / vysledne )
return(Statistika)
}
optim(par=c(pocAlp,pocThe,pocGama),fn=TeorPrav,method="L-BFGS-B",
lower=c(0.00001,0.00001,0.00001),upper=c(99.9999,0.999999,4.9999),
 control=list(maxit=2000))
```

Simulované p-hodnoty pre geometrické rozdelenie

```
count<-0
for (h in 1:2000)  {
dlzka<-100
N<-700
u <- runif(N)
prav<-rep(0,times=dlzka)
pravdep<-rep(0,times=dlzka)
p<-0.7345

#generujeme  pravdepodobnosti p pre geometricke rozdelenie

sucet<-0
for (k in 0:(dlzka-1)) {
```

```
prav[k+1] <- (1-p)^k*p
sucet<-sucet+prav[k+1]
}

pravdep<-prav/sucet

Y <- rep(0,times=length(u))
M<-0

for (i in 1:length(u))  {
  xx <- u[i]
  for (m in 0:(dlzka-1)){
    Pm <- sum(pravdep[1:(m+1)])
    M <- xx - Pm
    if (M < 0){
      Y[i] <- m
      break
    }
  }
}


max<-0

for (i in 1:N)
    if (Y[i]>max) max<-Y[i]

frek<-rep(0,times=max+1)

for (i in 1: (N+1))
   frek[Y[i]+1] <- (frek[Y[i]+1]+1)

teor<-rep(0,times=max+1)
teoret<-rep(0,times=max+1)

sucet<-0
for (k in 0:(max)) {
teor[k+1] <- (1-p)^k*p
sucet<-sucet+teor[k+1]
}

teoret<-teor/sucet
vysledne<-teoret*N

Statistika[h] <- sum( (frek-vysledne)^2 / vysledne )
if (Statistika[h] > 7.7923)  count<-count+1
```

```
}

p.value<-count/2000
```

Simulované p-hodnoty pre nové rozdelenie

```
count<-0

for (h in 1:2000)  {
dlzka<-100
N<-700
u <- runif(N)
prav<-rep(0,times=dlzka)
pravdep<-rep(0,times=dlzka)
x[1]<-0.7043
x[2]<-0.3294

teoret<-rep(0,times=length(frek))
pravdep<-rep(0,times=length(frek))

sucet<-0
for (k in 0:(length(frek)-1)) {
teoret[k+1] <- (x[1]*x[2]^k*(1-x[2]))/((1-(1-x[1])*x[2]^(k+1))*(1-(1-x[1])*x[2]^k
sucet<-sucet+teoret[k+1]
}

pravdep<-teoret/sucet

Y <- rep(0,times=length(u))
M<-0

for (i in 1:length(u))  {
  xx <- u[i]
  for (m in 0:(dlzka-1)){
    Pm <- sum(pravdep[1:(m+1)])
    M <- xx - Pm
    if (M < 0){
      Y[i] <- m
      break
    }
  }
}

max<-0

for (i in 1:N)
```

```
    if (Y[i]>max) max<-Y[i]

frek<-rep(0,times=max+1)

for (i in 1: (N+1))
   frek[Y[i]+1] <- (frek[Y[i]+1]+1)

teor<-rep(0,times=max+1)
teoret<-rep(0,times=max+1)

sucet<-0
for (k in 0:(max)) {
teor[k+1] <- (1-p)^k*p
sucet<-sucet+teor[k+1]
}

teoret<-teor/sucet
vysledne<-teoret*N

Statistika[h] <- sum( (frek-vysledne)^2 / vysledne )
if (Statistika[h] > 5.2714)  count<-count+1
}
p.value<-count/2000
```