U n i v e r z i t a K o m e n s k é h o v B r a t i s l a v e

Matematické modely pre frekvencie grafém

Tomáš Kopilec

2009

Matematické modely pre frekvencie grafém

DIPLOMOVÁ PRÁCA

Tomáš Kopilec

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
KATEDRA APLIKOVANEJ MATEMATIKY A ŠTATISTIKY

Študijný odbor: 9.1.9 Aplikovaná matematika
Študijný program: Ekonomická a finančná matematika

Vedúci práce: Mgr. Ján Mačutek, PhD.

BRATISLAVA 2009

Čestné vyhlásenie

Vyhlasujem, že diplomovú prácu som vypracoval samostatne a čerpal som len z uvedenej literatúry a informačných zdrojov.

V Bratislave, 27. apríl 2009 .....

podpis

Poďakovanie

Chcel by som sa úprimne poďakovať Mgr. Jánovi Mačutkovi, PhD. za cenné rady a pomoc pri vypracovaní diplomovej práce.

Matematické modely pre frekvencie grafém
[diplomová práca]

Fakulta matematiky, fyziky a informatiky

autor: Tomáš Kopilec
vedúci diplomovej práce: Mgr. Ján Mačutek, PhD.
Apríl 2009

Abstrakt

V súčasnosti jednou z oblasti, ktorou sa zaoberá kvantitatívna lingvistika, je hľadanie modelu pre frekvencie grafém. Doposiaľ adekvátnym modelom sa ukazuje negatívne hypergeometrické rozdelenie, no stále sa hľadajú aj iné alternatívne modely. V tejto práci je predstavený nový model (nové diskrétne rozdelenie pravdepodobnosti). Sú odvodené jeho základné charakteristiky. Tento model je potom aplikovaný na modelovanie frekvencii grafém niektorých vybraných jazykov. V závere je model testovaný.

Kľúčové slová: graféma, frekvencie grafém, model, test dobrej zhody

Mathematical models for grapheme frequencies
[master thesis]

Faculty of Mathematics, Physics and Informatics

Author: Tomáš Kopilec
Supervisor: Mgr. Ján Mačutek, PhD.
April 2009

Abstract

At present quantitative linguistics is focused – among others - on finding a model for grapheme frequencies. Till now the negative hypergeometric distribution is an adequate model, but linguists are still looking for other alternative models. In this paper a new model (a new discrete probability distribution) is discussed. Its basic characteristics are studied. This model is then applied to modelling rank frequencies of graphemes of some selected languages. Finally, the model is tested.

Keywords: grapheme, grapheme frequencies, model, chi-square goodness of fit test

Obsah

Úvod

V súčasnosti dochádza k prieniku kvantitatívnych metód aj do takých oblasti, kde ešte relatívne nedávno neboli predstaviteľné – medzi nimi aj do lingvistiky. Kvantifikácia v nich nie je cieľom, ale prostriedkom. Umožňuje exaktnejšie formulovať hypotézy a najmä testovať ich adekvátnosť, rozdiely medzi nimi a podobne.

Tato práca analyzuje jeden model pre usporiadane frekvencie grafém (existuje mnoho definícii grafém, zhruba sa dá povedať, že graféma = písmeno). Ako model navrhujeme nove diskrétne rozdelenie. Odvádzame jeho základne vlastnosti a aplikujeme ho na dáta z niekoľkých jazykov.

1 Základné pojmy

1.1 Diskrétna premenná a jej charakteristiky

V prvej časti definujeme niektoré základné pojmy z pravdepodobnosti a štatistiky, ktorých definície možno nájsť v [7]. Tieto pojmy budú neskôr použité v tejto práci.

Definícia. Náhodná premenná X je *diskrétna*, ak nadobúda konečne alebo spočítateľne veľa hodnôt x_i s pravdepodobnosťami p_i

$$P(X = x_i) = p_i \quad \text{pre } i = 0, 1, \dots, n \ , n \leq \infty \tag{1.1}$$

a navyše musí platiť

$$\sum_{i=1}^{n} p_i = 1 \ . \tag{1.2}$$

Pod symbolom $\lceil . \rceil$ budeme rozumieť hornú celú časť čísla (napr. $\lceil 3.7 \rceil = 4$).

Definícia. Nech X je diskrétna náhodná premenná. Potom *distribučná funkcia* $F: \mathbb{R} \to \mathbb{R}$ náhodnej premennej X je definovaná vzťahom

$$F(x) = P(X < x) = \sum_{i=1}^{\lceil x-1 \rceil} p_i \ . \tag{1.3}$$

Definícia. Nech X je diskrétna náhodná premenná, ktorá nadobúda hodnoty x_i s pravdepodobnosťami p_i podľa (1.1). Potom *stredná hodnota* náhodnej premennej X je číslo

$$E(X) = \sum_{i=1}^{n} x_i \, p_i \ , \tag{1.4}$$

ak tento rad konverguje absolútne.

Vlastnosti strednej hodnoty

Nech X je diskrétna náhodná premenná a $g: \mathbb{R} \to \mathbb{R}$ je spojitá funkcia. Nech X má strednú hodnotu. Potom

$$E\big(g(X)\big) = \sum_{i=1}^{n} g(x_i) \, p_i \tag{1.5}$$

ak táto existuje.

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

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

ak táto stredná hodnota existuje.

Vlastnosti disperzie

Nech X je diskrétna náhodná premenná. Nech existuje stredná hodnota $E(X)$ a $E(X^2)$. Potom disperzia náhodnej premennej X je daná vzťahom

$$D(X) = E(X^2) - E^2(X) \ . \tag{1.6}$$

Definícia. Nech X a $\left(X - E(X)\right)^k$ pre $k \in \mathbb{N}$ sú náhodné premenné. Potom číslo

$$\mu_k = E\left[\left(X - E(X)\right)^k\right] \tag{1.7}$$

sa nazýva k - *tým centrálnym momentom*, ak táto stredná hodnota existuje.

Definícia. Nech X je diskrétna náhodná premenná, ktorá nadobúda nezáporné hodnoty $0,1,2,\dots,n$ $(n \leq \infty)$. Potom *vytvárajúca funkcia* $G\colon [0,1] \to \mathbb{R}$ náhodnej premennej X je

$$G(t) = \sum_{i=1}^{n} t^i \, p_i \ . \tag{1.8}$$

Z vytvárajúcej funkcie diskrétnej náhodnej premennej, ktorej definícia sa dá nájsť v [12], vieme ľahko spočítať niektoré základné charakteristiky náhodnej premennej a tiež pravdepodobnosti $P(X = i)$ $(i = 0, 1, 2, \dots, n)$.

Použitím (1.5) a (1.8) dostávame $G(t) = E(t^X)$. Spočítaním prvej derivácie a jej následným vyčíslením v bode 1 dostávame strednú hodnotu $E(X)$.

$$G'(t) = E(X\, t^{X-1})$$
$$G'(1) = E(X)$$

Analogicky ako strednú hodnotu môžeme spočítať aj disperziu. Najskôr spočítame druhú deriváciu vytvárajúcej funkcie (1.8) v bode 1

$$G''(t) = E(X(X-1)\, t^{X-2}),$$
$$G''(1) = E(X^2) - E(X).$$

Potom použitím (1.6) vieme vyjadriť disperziu $D(X)$ pomocou vytvárajúcej funkcie

$$D(X) = E(X^2) - E(X) = G''(1) + G'(1) - (G'(1))^2.$$

Z vytvárajúcej funkcie vieme spočítať aj pravdepodobnosti $p_i = P(X = i)$.

$$\begin{aligned} G(t) &= p_0 + tp_1 + t^2p_2 + \cdots + t^np_n \\ G'(t) &= p_1 + 2tp_2 + \cdots + nt^{n-1}p_n \\ &\vdots \\ G^{(i)}(t) &= i!\,p_i + (i+1)!\,t\,p_{i+1} + \cdots + n(n-1)\ldots(n-i+1)t^{n-i}p_n \end{aligned}$$

Potom dostávame $p_i = P(X = i) = \frac{G^{(i)}(0)}{i!}$.

1.2 Odhad parametrov metódou minimálneho χ^2

Častokrát sa v teórii pravdepodobnosti stretávame s takými rozdeleniami pravdepodobnosti, v ktorých nám vystupuje neznámy parameter $\boldsymbol{\theta}$. Tento parameter musíme potom z dát odhadovať. Metód na odhad neznámeho parametra je viacero (napr. metóda maximálnej vierohodnosti, momentová metóda, odhad metódou minimálneho χ^2). Keďže v práci bude použitý test dobrej zhody, preto je vhodné ako metódu odhadu parametra použiť metódu minimálneho χ^2, ktorá minimalizuje χ^2 testovaciu štatistiku.

Uvažujme náhodný výber z diskrétneho rozdelenia, pričom náhodná premenná X nadobúda iba celočíselné nezáporné hodnoty $i = 1,2,\ldots,n$. Tieto výberové hodnoty rozdelíme do k tried. Do prvej triedy zaradíme tie výberové hodnoty, ktoré sú menšie alebo sa rovnajú číslu s ($s \in \mathbb{N}$). Do ďalších tried budeme postupne zaraďovať hodnoty $s+1, s+2, \ldots, s+k-2$. Posledná k -ta trieda bude obsahovať hodnoty väčšie alebo rovnajúce sa číslu $s+k-1$. Početnosti v triedach označíme $M_1, M_2, \ldots, M_k$, pričom musí platiť, že $N = M_1 + M_2 + \cdots + M_k$, kde N je rozsah náhodného výberu.

Potrebujeme ešte poznať pravdepodobnosti, s akými náhodná premenná X padne do i – tej triedy pre $i = 1,2,\ldots,k$. Tieto pravdepodobnosti sú závislé od parametra $\boldsymbol{\theta}$. Nech tieto pravdepodobnosti sú

$$p_i = p_i(\boldsymbol{\theta}) \quad \text{pre } i = 1,2,\ldots,k\,,$$

kde $\boldsymbol{\theta} \in \mathbb{R}^m$ a musí platiť $p_1(\boldsymbol{\theta}) + p_2(\boldsymbol{\theta}) + \cdots + p_k(\boldsymbol{\theta}) = 1$ pre všetky $\boldsymbol{\theta} \in \mathbb{R}^m$. To znamená, že číslo $p_i(\boldsymbol{\theta})$ nám udáva pravdepodobnosť, s akou padne vybrané číslo

do i -tej triedy ($i = 1, \dots k$). Na získanie odhadu parametra $\boldsymbol{\theta}$ musíme mať predpoklad o rozdelení pravdepodobnosti $p_i(\boldsymbol{\theta})$ pre $i = 1,2, \dots , k$ (musíme mať predpoklad, z akého typu rozdelenia pravdepodobnosti pochádzajú dáta).

Za odhad parametra $\boldsymbol{\theta}$ sa berie tá hodnota, ktorá pri daných hodnotách $M_1, M_2, \dots , M_k$ minimalizuje premennú

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{\left(M_i - Np_i(\boldsymbol{\theta})\right)^2}{Np_i(\boldsymbol{\theta})} . \qquad (1.9)$$

Takémuto odhadu neznámeho parametra, uvedenom v [7], hovoríme odhad získaný metódou minimálneho χ^2 . Hodnoty $M_1, M_2, \dots , M_k$ sú empirické početnosti, ktoré získame z experimentálnych pozorovaní a $Np_1(\widehat{\boldsymbol{\theta}}), Np_2(\widehat{\boldsymbol{\theta}}), \dots , Np_n(\widehat{\boldsymbol{\theta}})$ sú teoretické početnosti, kde $\widehat{\boldsymbol{\theta}}$ je hodnota parametra, ktorá minimalizuje výraz (1.9), pričom za dostatočne veľký rozsah výberu N sa považuje také N, pre ktoré platí

$$Np_i(\boldsymbol{\theta}) \geq 5 \text{ pre } i = 1,2, \dots , n. \qquad (1.10)$$

Podmienka (1.10) je stanovená empiricky, teda nie je nutné, aby bola splnená. Ak podmienka (1.10) nie je splnená, tak nemusíme odhad touto metódou zavrhnúť (najmä pri malých vzorkách dát, pri ktorých sa nedajú vhodne zvoliť triedy). Vtedy môžeme buď podmienku (1.10) zjemniť, teda môžeme napr. požadovať iba $Np_i(\boldsymbol{\theta}) \geq 3$ pre $i = 1,2, \dots , n$ alebo pri väčších vzorkách môžeme zmeniť počet tried (napr. znížiť počet tried k).

Pri odhade neznámych parametrov musíme počítať optimálnu hodnotu výrazu (1.9) (minimalizovať χ^2 testovaciu štatistiku). Minimálna hodnota (1.9) $\chi^2(\widehat{\boldsymbol{\theta}})$ predstavuje testovaciu štatistiku hypotézy (1.11)

$$H_0\colon \textit{náhodný výber pochádza zo zvoleného rozdelenia} , \qquad (1.11)$$

pričom $\chi^2(\widehat{\boldsymbol{\theta}})$ má χ - kvadrát rozdelenie s $k - m - 1$ stupňami voľnosti, teda $\chi^2(k - m - 1)$, kde m je počet neznámych parametrov rozdelenia pravdepodobnosti p_i . Tomuto testu, uvedenom v [7] sa hovorí χ - kvadrát test dobrej zhody.

Teda najskôr musíme mať predpoklad o rozdelení pravdepodobnosti náhodného výberu, z ktorého odhadneme neznámy parameter, a potom testujeme či tento náhodný výber pochádza z predpokladaného rozdelenia. Ak testujeme túto hypotézu na hladine významnosti $\alpha = 0.05$, tak nulovú hypotézu (1.11) nezamietame ak dostaneme

p - hodnotu väčšiu ako 0.05, teda test nám potvrdil, že náhodný výber pochádza zo zvoleného rozdelenia.

Nevýhodou testu dobrej zhody je, že χ^2 testovacia štatistika rastie približne lineárne z rozsahom náhodného výberu *N*. Keďže v kvantitatívnej lingvistike sa pracuje s veľkými vzorkami (niekoľko stoviek tisíc), pre tieto veľké rozsahy vzoriek test dobrej zhody zamieta skoro vždy nulovú hypotézu (1.11).

Z tohto dôvodu sa zaviedol v lingvistike ako testovacie kritérium koeficient C, spomenutý v [2] a [8], ktorý získame vynormovaním χ^2 štatistiky rozsahom náhodného výberu

$$C = \frac{\chi^2}{N}. \qquad (1.12)$$

Empiricky určenou hraničnou hodnotou je $C = 0.02$. Ak dostaneme hodnotu koeficientu (1.12) menšiu ako 0.02, tak zhoda modelu s dátami je dobrá.

2 Negatívne hypergeometrické rozdelenie ako model pre usporiadané frekvencie grafém

Kvantitatívna lingvistika sa už dlhú dobu zaoberá modelovaním grafém. Donedávna bolo negatívne hypergeometrické rozdelenie jediným platným modelom pre frekvencie grafém. Toto rozdelenie bolo aplikované na všetkých dostupných dátach z rôznych jazykov a súhrne dalo najlepšie výsledky zo všetkých použitých modelov, teda sa ukázalo ako vhodné pre väčšinu jazykov (pre niektoré konkrétne jazyky sú vhodnejšie iné modely).

2.1 Odvodenie negatívneho hypergeometrického rozdelenia

Odvodenie negatívneho hypergeometrického rozdelenia sa dá nájsť v [12]. Uvažujme urnu s bielymi a čiernymi guľkami. Čiernych guliek máme *C* a bielych *K*-*C*. Guľky vyberáme bez návratu. Pýtame sa aká je pravdepodobnosť P_x, že vytiahneme práve *x* bielych guliek predtým ako vytiahneme $n-$ tú čiernu.

Najskôr spočítame pravdepodobnosť, že vytiahneme $n-1$ čiernych guliek, potom vytiahneme x bielych a nakoniec $n-$ tú čiernu.

$$\frac{C}{K}\frac{C-1}{K-1}\cdots\frac{C-[n-1-1]}{K-[n-1-1]}\frac{K-C}{K-(n-1)}\frac{K-C-1}{K-(n-1)-1}\cdots\frac{K-C-(x-1)}{K-(n-1)-(x-1)}\frac{C-(n-1)}{K-(n-1)-x}=$$

$$=\frac{C!\,(K-C)!\,(K-n-x)!}{K!\,(C-n)!\,(K-C-x)!}$$

Máme spočítanú pravdepodobnosť jednej možnosti ako vybrať *x* bielych guliek, predtým ako vytiahneme $n-$ tú čiernu. Ďalšia možnosť by bola napr.: vytiahneme prvú čiernu, potom x bielych a potom $n-1$ čiernych guliek. Pravdepodobnosť by vyšla aj v tomto prípade rovnaká ako pre vyššie uvedený prípad.

Teda všetkých kombinácii ako vybrať $n-1$ čiernych guliek a *x* bielych guliek a nakoniec čiernu guľku je

$$\binom{n+x-1}{x}=\binom{n+x-1}{n-1}$$

a pravdepodobnosť tohto výberu je vo všetkých prípadoch rovnaká. Potom vyššie spomínaná pravdepodobnosť P_x bude

$$P_x = \binom{n+x-1}{n-1}\frac{C!\,(K-C)!\,(K-n-x)!}{K!\,(C-n)!\,(K-C-x)!} = \frac{\binom{n+x-1}{n-1}\binom{K-n-x}{C-n}}{\binom{K}{C}},$$

pre $x = 0, 1, \ldots, n$ a $K, C, n \in \mathbb{N}$, $K \geq C \geq n$. Takto dané pravdepodobnostné rozdelenie nazývame *negatívne hypergeometrické rozdelenie*.

Pri modelovaní grafém požadujeme, aby $x \geq 1$, preto musíme x posunúť o jednotku vyššie. Potom dostávame

$$P_x = \frac{\binom{n+x-2}{n-1}\binom{K-n-x-1}{C-n}}{\binom{K}{C}},$$

pre $x = 1, \ldots, n+1$, kde $n+1$ predstavuje počet grafém v jazyku.

Zmení sa aj pôvodná podmienka na neznáme parametre n, K, C. Stále požadujeme aby $n \in \mathbb{N}$, ale zmení sa požiadavka na parametre K a C, teda $K, C \in \mathbb{R}$.

Na jednej strane dáva tento model po aplikácií na modelovanie usporiadaných frekvencii grafém dobré výsledky, ale na druhej strane je toto rozdelenie odvodené z binárnej urnovej schémy, teda vyberáme buď bielu alebo čiernu guľku. Pomocou alternatívnej možnosti modelovať frekvencie grafém, ktorých je niekoľko desiatok, je nereálne v lingvistike. Vzniká tu problém pre lingvistov s interpretáciou parametrov *C* a *K* (v matematike je interpretácia *C* a *K* jednoduchá, parametre predstavujú počty guliek).

Teda naskytuje sa tu možnosť isť dvoma smermi. Buď sa ponúka snaha o inú interpretáciu parametrov negatívneho hypergeometrickeho rozdelenia, ako jediného doposiaľ platného modelu pre frekvencie grafém alebo hľadať nové modely, ktoré nie sú odvodené z binárnej urnovej schémy.

V súčasnosti sa v lingvistike rieši aj interpretácia parametrov negatívneho hypergeometrického rozdelenia.

3 Nový model pre frekvencie grafém

V tejto práci sa budeme zaoberať druhým smerom, teda budeme hľadať nový model, ktorý by modeloval čo najlepšie usporiadané frekvencie grafém v jazyku a nebude mať korene v binárnej urnovej schéme.

Uvažujme diskrétne rozdelenie pravdepodobnosti

$$P_i = c\left[\left(1-\frac{i}{n}\right)p_1^{i-1} + \frac{i}{n}\,p_2^{i-1}\right] \qquad \text{pre } i = 1,2,\dots,n\,, \tag{3.1}$$

s parametrami $0 \le p_1 \le 1$, $0 \le p_2 \le 1$, $n \in \mathbb{N}$ a c je normalizačná konštanta.

Diskrétne rozdelenie pravdepodobnosti (3.1) je vytvorené ako „lineárna kombinácia" dvoch geometrických rozdelení, kde váhy pri týchto geometrických rozdeleniach nie sú konštantné (závisia od i). Keďže je to nové rozdelenie, v tejto časti sa ním budeme podrobnejšie zaoberať. Odvodíme jeho základné matematické vlastnosti a charakteristiky.

Pre normalizačnú konštantu c musí platiť podmienka (1.2).

$$c^{-1} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} p_1^{i-1} - \frac{1}{n}\sum_{i=1}^{n} i\,p_2^{i-1} + \frac{1}{n}\sum_{i=1}^{n} i\,p_2^{i-1}$$

Pripomenieme už známe vzťahy:

$$\sum_{i=1}^{n} p^{i-1} = \frac{1-p^n}{1-p} \tag{3.2}$$

$$\sum_{i=1}^{n} i\,p^{i-1} = \frac{1-p^{n+1}}{(1-p)^2} - \frac{(n+1)p^n}{1-p} \tag{3.3}$$

Potom dosadením (3.2) a (3.3) dostávame pre normalizačnú konštantu c

$$c = \left[\frac{1-p_1^n}{1-p_1} - \frac{1}{n}\left[\frac{1-p_1^{n+1}}{(1-p_1)^2} - \frac{(n+1)p_1^n}{1-p_1} - \frac{1-p_2^{n+1}}{(1-p_2)^2} + \frac{(n+1)p_2^n}{1-p_2}\right]\right]^{-1} =$$

$$= n\left[\frac{n+p_1^n}{1-p_1} - \frac{1-p_1^{n+1}}{(1-p_1)^2} - \frac{(n+1)p_2^n}{1-p_2} + \frac{1-p_2^{n+1}}{(1-p_2)^2}\right]^{-1}.$$

Výpočet konštanty c sa dá zjednodušiť, keď do (3.1) dosadíme za $i = 1$, potom dostávame

$$P_1 = c \ .$$

Distribučnú funkciu rozdelenia (3.1) vypočítame podľa (1.3)

$$F(x) = P(X < x) = \sum_{i=1}^{\lceil x-1 \rceil} P_i = c \sum_{i=1}^{\lceil x-1 \rceil} p_1^{i-1} - \frac{c}{n} \sum_{i=1}^{\lceil x-1 \rceil} i\, p_2^{i-1} + \frac{c}{n} \sum_{i=1}^{\lceil x-1 \rceil} i\, p_2^{i-1} \ .$$

Opäť dosadením (3.2) a (3.3) do predošlého výrazu dostávame nasledovný vzťah pre distribučnú funkciu

$$F(x) = c \left[\frac{1 - p_1^{\lceil x-1 \rceil}}{1 - p_1} - \frac{1}{n} \left[\frac{1 - p_1^{\lceil x-1 \rceil + 1}}{(1 - p_1)^2} - \frac{(\lceil x - 1 \rceil + 1) p_1^{\lceil x-1 \rceil}}{1 - p_1} - \frac{1 - p_2^{\lceil x-1 \rceil + 1}}{(1 - p_2)^2} + \frac{(\lceil x - 1 \rceil + 1) p_2^{\lceil x-1 \rceil}}{1 - p_2} \right] \right].$$

Vytvárajúca funkcia pravdepodobnostného rozdelenia (3.1) počítaná podľa (1.8)

$$G(t) = \sum_{i=1}^{n} t^i P_i = c \sum_{i=1}^{n} t^i \left[\left(1 - \frac{i}{n}\right) p_1^{i-1} + \frac{i}{n} p_2^{i-1} \right] =$$

$$= ct \sum_{i=1}^{n} (p_1 t)^{i-1} - \frac{ct}{n} \sum_{i=1}^{n} i (p_1 t)^{i-1} + \frac{ct}{n} \sum_{i=1}^{n} i (p_2 t)^{i-1} =$$

$$= \frac{ct}{n} \left[\frac{n + (p_1 t)^n}{1 - p_1 t} - \frac{1 - (p_1 t)^{n+1}}{(1 - p_1 t)^2} - \frac{(n+1)(p_2 t)^n}{1 - p_2 t} + \frac{1 - (p_2 t)^{n+1}}{(1 - p_2 t)^2} \right].$$

Strednú hodnotu rozdelenia (3.1) počítame podľa (1.4)

$$\mu = \sum_{i=1}^{n} i\, P_i = c \sum_{i=1}^{n} i \left[\left(1 - \frac{i}{n}\right) p_1^{i-1} + \frac{i}{n} p_2^{i-1} \right] =$$

$$= c \sum_{i=1}^{n} i\, {p_1}^{i-1} - \frac{c}{n} \sum_{i=1}^{n} i^2 {p_1}^{i-1} + \frac{c}{n} \sum_{i=1}^{n} i^2 {p_2}^{i-1}$$

Uvedieme ďalšie známe vzťahy:

$$\sum_{i=1}^{n} i^2 p^{i-1} = \frac{1}{(1-p)^3}\left[\, 1 + p - (n+1)^2 p^n + (2n^2+2n-1)p^{n+1} - n^2 p^{n+2} \,\right] \tag{3.4}$$

$$\sum_{i=1}^{n} i^3\, p^{i-1} = \frac{p^2 - (n+1)^3 p^n + (3n^3+6n^2-4)p^{n+1} - (3n^3+3n^2-3n+1)p^{n+2} + n^3 p^{n+3}}{(1-p)^4} +$$

$$+\ \frac{1+4p}{(1-p)^4} \tag{3.5}$$

Potom môžeme strednú hodnotu pomocou (3.3) a (3.4) prepísať na tvar

$$\mu = \frac{c}{n}\left[-\frac{(n^2+n)p_1^n}{1-p_1} + \frac{n(1-p_1^{n+1})}{(1-p_1)^2} - \frac{1+p_1-(n+1)^2 {p_1}^n + (2n^2+2n-1){p_1}^{n+1} - n^2 {p_1}^{n+2}}{(1-p_1)^3} + \right.$$

$$\left. + \frac{1+p_2-(n+1)^2 {p_2}^n + (2n^2+2n-1){p_2}^{n+1} - n^2 {p_2}^{n+2}}{(1-p_2)^3} \right].$$

Disperziu náhodnej premennej počítame podľa vzťahu (1.6), teda

$$D(X) = E(X^2) - \mu^2 . \tag{3.6}$$

Počítajme $E(X^2)$ pomocou (1.5)

$$E(X^2) = \sum_{i=1}^{n} i^2 P_i = c \sum_{i=1}^{n} i^2 \left[\left(1 - \frac{i}{n}\right) p_1^{i-1} + \frac{i}{n} p_2^{i-1} \right] =$$

$$= c \sum_{i=1}^{n} i^2 {p_1}^{i-1} - \frac{c}{n} \sum_{i=1}^{n} i^3 {p_1}^{i-1} + \frac{c}{n} \sum_{i=1}^{n} i^3 {p_2}^{i-1}$$

Potom použitím (3.4), (3.5) a dosadením do (3.6) dostaneme nasledovný vzťah pre disperziu

$$D(X) = \frac{c}{n}\left[n \frac{1+p_1-(n+1)^2 {p_1}^n + (2n^2+2n-1){p_1}^{n+1} - n^2 {p_1}^{n+2}}{(1-p_1)^3} - \right.$$

$$- \frac{1+4p_1+{p_1}^2-(n+1)^3 {p_1}^n + (3n^3+6n^2-4){p_1}^{n+1} - (3n^3+3n^2-3n+1){p_1}^{n+2} + n^3 {p_1}^{n+3}}{(1-p_1)^4} +$$

$$\left. + \frac{1+4p_2+{p_2}^2-(n+1)^3 {p_2}^n + (3n^3+6n^2-4){p_2}^{n+1} - (3n^3+3n^2-3n+1){p_2}^{n+2} + n^3 {p_2}^{n+3}}{(1-p_2)^4} \right] -$$

$$- \mu^2 \quad .$$

Ďalšou charakteristikou diskrétnej náhodnej premennej je miera opakovania (repeat rate). Vznik tejto miery má pôvod v ekonómii ako Herfindahlov index (ozn. *HI*). Uvažujme, že máme na trhu n firiem s rovnakým podielom $s_i = \frac{1}{n}$ $(i = 1, \dots, n)$. Spočítame hodnotu *HI* definovanú ako

$$HI = \sum_{i=1}^{n} s_i^2 = \sum_{i=1}^{n} \frac{1}{n^2} = \frac{1}{n}.$$

Toto je jeden extrémny prípad, ktorý predstavuje dokonalú konkurenciu na trhu. Druhým extrémnym prípadom je jedna firma na trhu, teda monopol ($s_1 = 1$ a $s_i = 0$ pre $i = 2, \dots, n$). V tomto prípade $HI = 1$. Teda tento index nadobúda hodnoty z intervalu $(\frac{1}{n}, 1)$.

Tento index sa začal používať aj v lingvistike. Ak je miera opakovania (rr) malé číslo (blízko $\frac{1}{n}$, kde táto hodnota predstavuje diskrétne rovnomerne rozdelenie, teda $P(X = i) = \frac{1}{n}$ pre všetky *i*), tak všetky grafémy majú približne rovnaké rozloženie. Ak je miera opakovania blízka jednotke (teda $P(X = 1) = 1$ a pre ostatné i sú pravdepodobnosti nulové), tak niektoré grafémy majú väčšie zastúpenie oproti ostatným, teda majú vyššiu mieru opakovania.

Miera opakovania pre rozdelenie pravdepodobnosti (3.1).

$$rr = \sum_{i=1}^{n} P_i^2 = c^2 \sum_{i=1}^{n} \left[\left(1 - \frac{i}{n}\right) p_1^{i-1} + \frac{i}{n} p_2^{i-1} \right]^2 =$$

$$= c^2 \sum_{i=1}^{n} \left[\left(1 - 2\frac{i}{n} + \frac{i^2}{n^2}\right) (p_1^2)^{i-1} + 2\left(\frac{i}{n} - \frac{i^2}{n^2}\right) (p_1 p_2)^{i-1} + \frac{i^2}{n^2} (p_2^2)^{i-1} \right] =$$

$$= c^2 \left[\sum_{i=1}^{n} (p_1^2)^{i-1} - \frac{2}{n} \left[\sum_{i=1}^{n} i (p_1^2)^{i-1} - \sum_{i=1}^{n} i (p_1 p_2)^{i-1} \right] + \frac{1}{n^2} \left[\sum_{i=1}^{n} i^2 (p_1^2)^{i-1} - 2 \sum_{i=1}^{n} i^2 (p_1 p_2)^{i-1} + \sum_{i=1}^{n} i^2 (p_1^2)^{i-1} \right] \right]$$

$$= c^2 \left\{ \frac{1 - p_1^{2n}}{1 - p_1^2} - \frac{2}{n} \left[\frac{1 - p_1^{2n+2}}{(1 - p_1^2)^2} - \frac{(n+1) p_1^{2n}}{1 - p_1^2} - \frac{1 - p_1 {p_2}^{n+1}}{(1 - p_1 p_2)^2} + \frac{(n+1)(p_1 p_2)^n}{1 - p_1 p_2} \right] + \right.$$

$$+ \frac{1}{n^2} \left[\frac{1 + p_1^2 - (n+1)^2 p_1^{2n} + (2n^2 + 2n - 1) p_1^{2n+2} - n^2 p_1^{2n+4}}{(1 - p_1^2)^3} - \right.$$

$$- \frac{1 + p_1 p_2 - (n+1)^2 p_1 {p_2}^n + (2n^2 + 2n - 1) p_1 {p_2}^{n+1} - n^2 p_1 {p_2}^{n+2}}{(1 - p_1 p_2)^3} +$$

$$\left. \left. + \frac{1 + p_2^2 - (n+1)^2 p_2^{2n} + (2n^2 + 2n - 1) p_2^{2n+2} - n^2 p_2^{2n+4}}{(1 - p_2^2)^3} \right] \right\}$$

3.1 Aplikácia nového modelu

Chceme aplikovať nový model (3.1) na modelovanie usporiadaných frekvencii grafém. Najskôr potrebujeme získať dáta o frekvenciách grafém. Zoberieme text zo zvoleného jazyka a budeme počítať výskyt každej grafémy v texte. Získané frekvencie potom usporiadame od najväčšej frekvencie po najmenšiu, lebo našou snahou je modelovať usporiadané grafémy. Môže vzniknúť otázka, prečo nenecháme grafémy v abecednom poradí. Dôvod je, že abeceda je len konvencia v usporiadaní grafém v jazyku (napr. slovenčina a ruština sú príbuzné jazyky, ale majú rôzne usporiadanie písmen v abecede).

Z vybraného textu sme získali dáta. Teraz aplikujeme model (3.1) na tieto frekvencie. Potrebujeme odhadnúť neznáme parametre p_1, p_2, n rozdelenia pravdepodobnosti (3.1). Parameter n je pevne daný, predstavuje počet grafém v danom jazyku. Neznámymi parametrami sú už len p_1, p_2.

Tieto parametre budeme odhadovať metódou minimálneho χ^2 opísanou v kapitole 1.2. Triedy vytvoríme tak, že položíme $s = 1$ a $k = n$. Neznámy parameter je $\boldsymbol{\theta} = (p_1, p_2)^T$. Pravdepodobnosť toho, že náhodná premenná padne do *i* - tej triedy bude

$$P_i = P(X = i) = P_i(p_1, p_2) = c\left[\left(1 - \frac{i}{n}\right)p_1^{i-1} + \frac{i}{n}p_2^{i-1}\right] \quad \text{pre } i = 1,2,\dots,n\,.$$

Početnosť M_i v každej *i* - tej triede bude reprezentovaná frekvenciou *i* - tej grafémy f_i. Celkový rozsah *N* je daný ako

$$N = \sum_{i=1}^{n} f_i\ .$$

Odhady parametrov p_1, p_2 nájdeme minimalizáciou (1.9). Musíme ešte zohľadniť podmienku pre neznámy parameter $0 \le p_1 \le 1$, $0 \le p_2 \le 1$. Dostávame nasledovnú minimalizačnú úlohu

$$\min\left\{\sum_{i=1}^{k}\frac{(f_i - NP_i(p_1,p_2))^2}{NP_i(p_1,p_2)}, 0 \le p_1 \le 1, 0 \le p_2 \le 1\right\}. \qquad (3.7)$$

Táto úloha nemá explicitné riešenie, preto ju budeme riešiť numericky. Aby sme mohli použiť nejakú numerickú metódu, potrebujeme určiť vhodné počiatočné odhady p_1^0, p_2^0 (pracujeme so štatistickým softvérom *R*, v ktorom vhodnou numerickou metódou je *L-BFGS-B*, ktorá vyžaduje, aby boli počiatočné odhady z ohraničení úlohy). Ako najjednoduchšia možnosť výpočtu počiatočných odhadov sa ponúka porovnanie v prvej až tretej teoretickej a nameranej frekvencie.

Teda vyriešime systém rovníc

$$f_1 = NP_1 = Nc$$

$$f_2 = NP_2 = Nc\left[\left(1-\frac{2}{n}\right)p_1 + \frac{2}{n}p_2\right]$$

$$f_3 = NP_3 = Nc\left[\left(1-\frac{3}{n}\right){p_1}^2 + \frac{3}{n}{p_2}^2\right] \qquad (3.8)$$

Z prvej rovnice dostaneme odhad normalizačného koeficientu $\hat{c} = \frac{f_1}{N}$, potom pre jednoduchosť zavedieme substitúcie do rovníc (3.8): $F_2 = \frac{f_2}{Nc}$ a $F_3 = \frac{f_3}{Nc}$. Následne dostávame počiatočné odhady parametrov ako riešenie systému rovníc (3.8)

$$p_{1\,1}^0 = \frac{3F_2(n-2) + 2\sqrt{3F_2^2(3-n) + F_3(3n-8)}}{3n-8}$$

$$p_{2\,1}^0 = \frac{2F_2(n-3) - (n-2)\sqrt{3F_2^2(3-n) + F_3(3n-8)}}{3n-8}$$

$$p_{1\,2}^0 = \frac{3F_2(n-2) - 2\sqrt{3F_2^2(3-n) + F_3(3n-8)}}{3n-8}$$

$$p_{2\,2}^0 = \frac{2F_2(n-3) + (n-2)\sqrt{3F_2^2(3-n) + F_3(3n-8)}}{3n-8}.$$

Lenže po aplikovaní tohto spôsobu výpočtu počiatočných odhadov na konkrétnych dátach sa tento spôsob ukázal ako nevhodný, lebo napr. pre ruské frekvencie grafém (tab. 4) sme dostali riešenie systému (3.8)

$$p_{1\,1}^0 = 0.8722221$$

$$p_{2\,1}^0 = -0.5479728$$

$$p_{1\,2}^0 = 0.7207917$$

$$p_{2\,2}^0 = 1.572053\,.$$

Sme mimo požadovaného intervalu pre neznáme parametre $p_1, p_2 \in (0,1)$. Ak by sme týmto istým spôsobom vypočítali počiatočné odhady pre slovenské frekvencie (tab. 1), tak by sme dostali komplexné riešenie systému (3.8)

$$p_{1\,1}^0 = 0.985196 - 0.0348031\,\mathrm{i}$$

$$p_{2\,1}^0 = 0.64187 + 0.765667\,\mathrm{i}$$

$$p^0_{1\,2} = \ 0.985196 \ + 0.0348031 \text{ i}$$

$$p^0_{2\,2} = 0.64187 \ - 0.765667 \text{ i} \, .$$

V obidvoch prípadoch nemôžeme odštartovať riešenie úlohy (3.7). Preto sme zvolili iný postup na výpočet počiatočných odhadov. Rozdelenie pravdepodobnosti (3.1) je vytvorené ako „lineárna kombinácia“ dvoch geometrických rozdelení. Tak budeme na chvíľu predpokladať, že frekvencie grafém sa správajú podľa geometrického rozdelenia (bližšie geometrické rozdelenie pozri v [12])

$$P_i = P(X = i) = (1-p)^{i-1}p \quad \text{pre } \ i = 1,2, \ldots \ .$$

Potom odhad parametra *p* geometrického rozdelenia metódou maximálnej vierohodnosti (viac k metóde maximálnej vierohodnosti pozri v [7]) je

$$\hat{p} = \frac{N}{\sum_{i=1}^{n} i \, f_i} .$$

Zvolíme počiatočné odhady $p^0_1 = \ p^0_2 = \hat{p}$.

Teraz môžeme odštartovať numerickú metódu na riešenie úlohy (3.7). Na riešenie použijeme software *R* (pozri program v prílohe), konkrétne optimalizačnú metódu *L-BFGS-B*. Tá umožňuje optimalizáciu na ohraničenom intervale (v našom prípade $p_1, p_2 \in \ (0,1)$).

Po nájdení odhadov parametrov rozdelenia musíme ešte overiť, či je splnená podmienku (1.10), teda či

$$NP_i(\hat{p}_1, \hat{p}_2) \geq 5 \ \text{ pre } i = 1,2, \ldots , n \, .$$

Keďže frekvencie sú usporiadané o najväčšej po najmenšiu, tak podmienka (1.10) bude nesplnená najskôr na posledných frekvenciách. Teda ak nájdeme také *i*, pre ktoré bude ešte splnená podmienka (1.10), ale pre $i + 1$ už splnená nebude. V takomto prípade zmeníme $k = i$ a opäť budeme riešiť úlohu (3.7), pričom sa zmení početnosť v *k* - tej triede

$$M_k = \sum_{j=k}^{n} M_j \ ,$$

zmení sa aj pravdepodobnosť, že náhodná premenná padne do k - tej triedy

$$P_k = \sum_{j=k}^{n} P_j(p_1,\ p_2)\ .$$

3.2 Ordov graf

Ord ukázal grafické zobrazenie niektorých vybraných rozdelení pravdepodobnosti, ktoré možno nájsť v [9]. Na grafe zobrazil závislosť S od I, kde

$$S = \frac{\mu_3}{\mu_2} \quad \text{a} \quad I = \frac{\mu_2}{\mu}\ . \tag{3.9}$$

Pre $k = 1,2,3$ počítame hodnotu centrálneho k - teho momentu μ_k podľa (1.7). *Ordov graf* sa dá zostrojiť pre teoretické, simulované alebo pozorované náhodné čísla zo zvoleného rozdelenia.

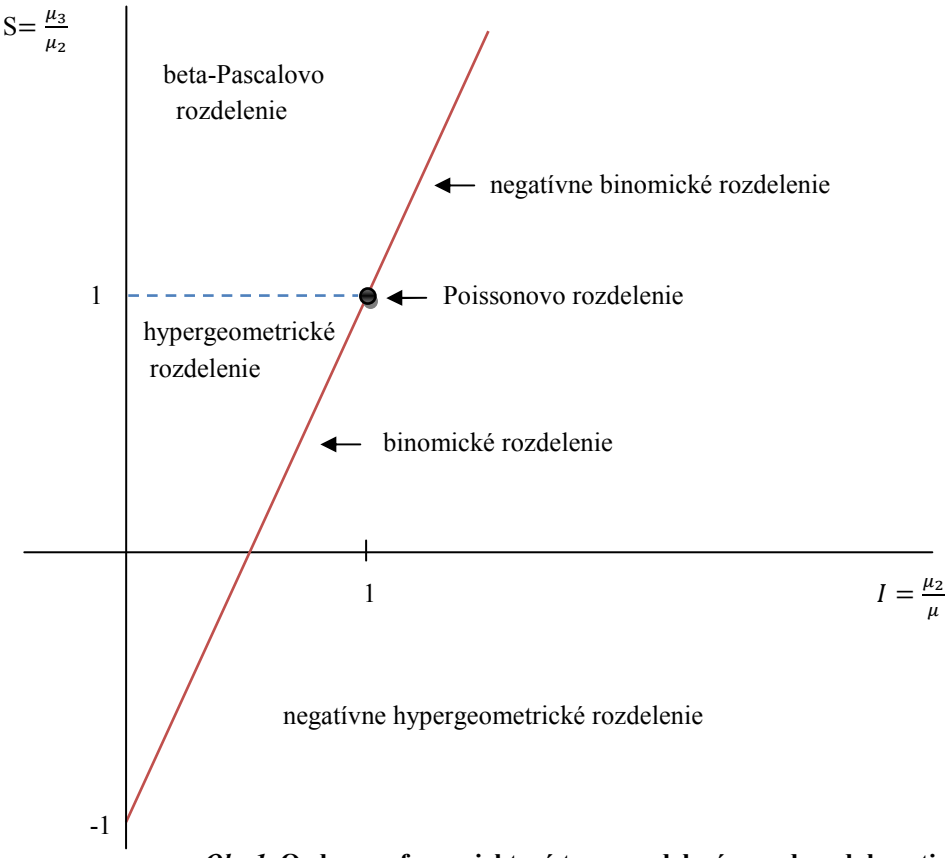
Existuje viacero možnosti na konštrukciu *Ordovho grafu* pre teoretické pravdepodobnosti. Jedna z možnosti je použiť vzťahy pre strednú hodnotu (μ), disperziu (μ_2) a šikmosť (μ_3) zvoleného rozdelenia, tie potom dosadiť do (3.9) a dostaneme vzťahy na výpočet S a I. Pre rozdelenie (3.1) vyšli tieto vzťahy príliš komplikované (viď kapitola 3), preto je tento spôsob nevhodný. Ďalším spôsobom je simulácia. Predpokladajme, že pravdepodobnostné rozdelenie obsahuje parameter $\theta \in \Theta$. Zoberieme nejakú konkrétnu hodnotu parametra θ a nasimulujeme pravdepodobnosti

$$P(X = i) = P_i(\theta) \quad \text{pre } i = 1,2,\dots,n \ (n \leq \infty).$$

Keď už poznáme pravdepodobnosti s akými náhodná premenná nadobúda hodnoty $i = 1,2,\dots,n$, vieme ľahko spočítať stredné hodnoty (1.7) potrebné na výpočet centrálnych momentov μ_k , $k = 1,2,3$. Takto vypočítané hodnoty S a I potom zobrazíme do grafu. Ak by sme tento postup opakovali pre všetky hodnoty parametra $\theta \in \Theta$, jednotlivé body zobrazené na grafe nám vytvoria oblasť príznačnú pre zvolené rozdelenie. Na obrázku (obr. 1) sú zobrazené oblasti pre niektoré známe rozdelenia pravdepodobnosti (napr. pre Poissonovo rozdelenie dostávame pre ľubovoľný parameter vždy bod $[1, 1]$).

Ak máme k dispozícii pozorované dáta, tak pre tieto dáta vieme spočítať pozorované centrálne momenty (1.7) a tým aj vypočítať hodnoty S a I podľa (3.9), ktoré nanesieme do *Ordovho grafu*. Ak máme k dispozícii viac pozorovaných dát, tak rovnakým postupom zostrojíme body pre pozorované dáta.

Teda *Ordov graf* slúži ako „optické kritérium“, či pozorované dáta pochádzajú z predpokladaného rozdelenia. Ak sa pozorovaný bod (resp. pozorované body) leží (resp. ležia) „blízko“ teoretickej oblasti zvoleného (= predpokladaného) rozdelenia, nezavrhujeme predpoklad, že pozorované dáta pochádzajú zo zvoleného rozdelenia. *Ordov graf* ale nie je dôkaz toho, že pozorované dáta pochádzajú zo zvoleného rozdelenia, lebo oblasti v *Ordovom grafe* sa pre viaceré rozdelenia prekrývajú (niektoré typy rozdelení majú prvé tri momenty rovnaké alebo majú rovnaké pomery momentov). Preto v takomto prípade musíme použiť ešte iný spôsob na potvrdenie, či dáta pochádzajú zo zvoleného rozdelenia (napr. test dobrej zhody). Ak by sa pozorovaný (resp. pozorované body) nachádzal (resp. nachádzali) „ďaleko“ od teoretickej oblasti, tak zavrhujeme predpoklad, že dáta pochádzajú zo zvoleného rozdelenia a nemusíme už použiť test dobrej zhody.



***Obr.1.* Ordov graf pre niektoré typy rozdelení pravdepodobnosti**

3.3 Testovanie nového modelu

Model (3.1) aplikujeme na modelovanie frekvencii grafém (bližšie v kapitole 3.4) a chceme ho testovať. Keďže využívame štatistický aparát, môžeme testovať hypotézy.

Ako jeden test použijeme *Ordov graf*. Zostrojíme teoretické a pozorované oblasti (bližšie opísané v časti 3.2), ktoré následne porovnáme. Ak sú „opticky blízko“ seba, tak nezavrhneme predpoklad, že frekvencie grafém pochádzajú zo zvoleného rozdelenia (z nového modelu (3.1)) a prejdeme k exaktným metódam. Na druhej strane ak sú „opticky ďaleko“ od seba, tak zavrhneme predpoklad zhody modelu s dátami o frekvenciách grafém.

Ďalším použitím testom je χ – *kvadrát test dobrej zhody*, ktorý je bližšie popísaný v závere časti 1.2. Teda testujeme nulovú hypotézu (1.11). Testovacia štatistika tohto testu má χ – kvadrát rozdelenie $\chi^2(k-m-1)$, kde *m* je pre model (3.1) rovné 2, lebo máme dva neznáme parametre p_1, p_2 . Keďže test dobrej zhody skoro vždy pre veľké rozsahy náhodných výberov zamieta hypotézu (1.11), použijeme ako testovacie kritérium modelu koeficient *C* (spomenutý v závere časti 1.2).

3.4 Aplikácia modelu na niektorých frekvenciách grafém

Máme dostupné pozorované usporiadané frekvencie grafém[1], na ktoré aplikujeme nový model (3.1). Teda sa snažíme odhadnúť neznáme parametre rozdelenia (3.1). Parameter *n* bude predstavovať počet grafém v zvolenom jazyku a parametre p_1 a p_2 odhadneme spôsobom uvedeným v časti 3.1.

V tabuľke (tab. 1) sú pozorované usporiadané frekvencie grafém slovenského jazyka. Teraz chceme porovnať pozorované frekvencie s frekvenciami grafém nového modelu. Teoretické frekvencie modelu (3.1) vyrátame tak, že celkový počet grafém N vynásobíme pravdepodobnosťou výskytu *i* – tej grafémy P_i

$$f_i^{teor} = N\, P_i(\hat{p}_1, \hat{p}_2)\,.$$

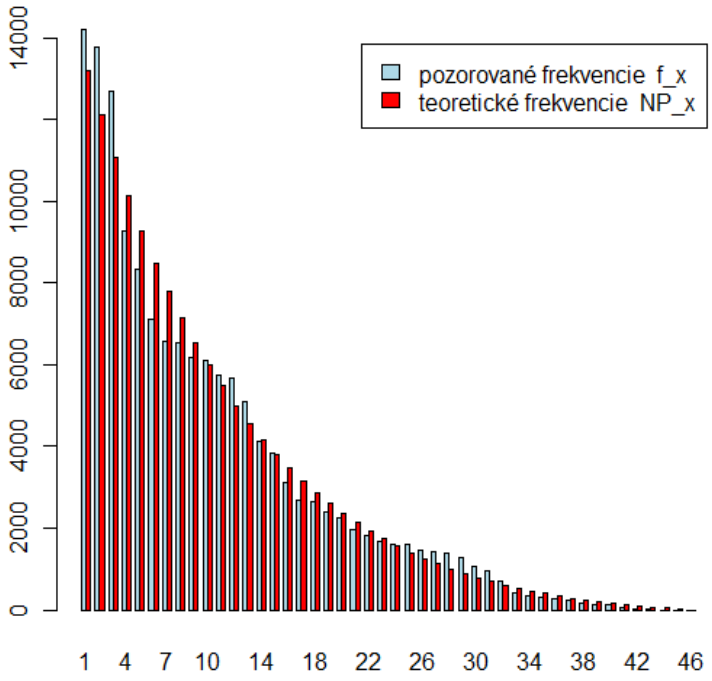
Porovnanie pozorovaných frekvencii grafém z textu slovenského jazyka a teoretických z modelu (3.1) je zobrazené na obrázku (obr. 2). Modré stĺpce predstavujú pozorované frekvencie grafém a červené stĺpce predstavujú teoretické frekvencie určené z nového modelu (3.1). Z obrázku je vidieť „dobrú“ zhodu, ale to nie je postačujúce. Musíme to potvrdiť exaktnejším spôsobom. Použijeme spomínaný test pomocou *Ordovho grafu* a ak sa ten ukáže ako pozitívny, tak použijeme ešte *test dobrej zhody*.

[1] Slovenské frekvencie grafém sú prevzaté z [3], ukrajinské frekvencie z [4], slovinské frekvencie z [6], ruské frekvencie grafém z [5], nemecké frekvencie z [1] a nakoniec tamilské frekvencie grafém sú prevzaté z [10], všetky tieto usporiadané frekvencie sú v danom poradí uvedené v tab. 1-6.

Tab. 1. Pozorované frekvencie grafém v slovenskom jazyku

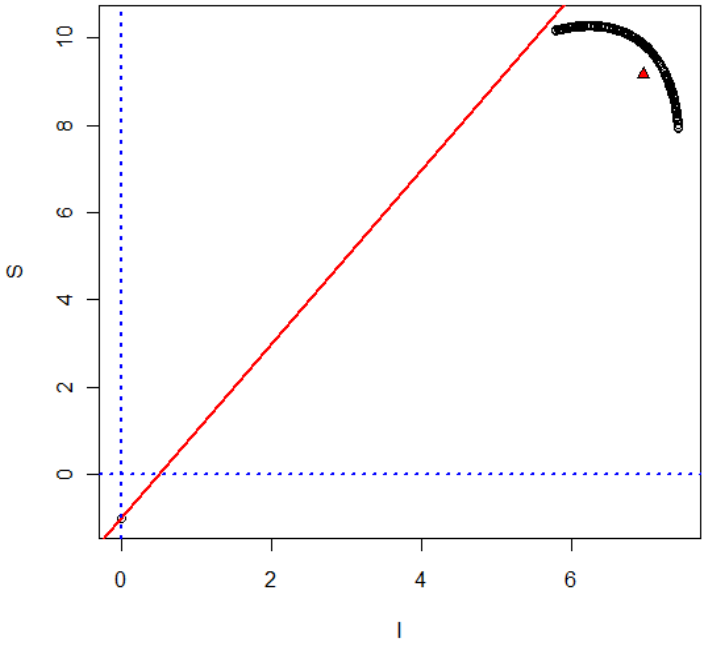
i	f_i	i	f_i	i	f_i	i	f_i
1	14194	***13***	5103	***25***	1593	***37***	253
2	13772	***14***	4121	***26***	1465	***38***	172
3	12701	***15***	3845	***27***	1422	***39***	131
4	9285	***16***	3135	***28***	1395	***40***	124
5	8323	***17***	2676	***29***	1294	***41***	47
6	7099	***18***	2660	***30***	1073	***42***	27
7	6562	***19***	2408	***31***	947	***43***	10
8	6534	***20***	2262	***32***	719	***44***	3
9	6164	***21***	1954	***33***	402	***45***	2
10	6091	***22***	1825	***34***	346	***46***	0
11	5731	***23***	1685	***35***	297		
12	5659	***24***	1611	***36***	270		
n = 46				N = 147392			
p1 = 0.9412				C = 0.0174			
p2 = 0.4288				p_hodnota = 0.000			

Ako testovacie kritérium modelu (3.1) pre slovenské frekvencie najskôr použijeme *Ordov graf*. Vytvoríme teoretickú oblasť pre hodnoty neznámych parametrov ($p_1 \in (0.90, 0.98)$, $p_2 \in (0.39, 0.47)$, $n = 46$) v blízkosti odhadov parametrov $\hat{p}_1$ a $\hat{p}_2$, ktoré sú uvedené dole v tabuľke (tab. 1).



Obr. 2. **Porovnanie pozorovaných a teoretických frekvencii grafém slovenského jazyka**

Ako je vidieť na obrázku (obr. 3), tak pozorovaný bod (červený trojuholník) vypočítaný z frekvencii grafém uvedených v tabuľke (tab. 1) je „blízko“ teoretickej oblasti (čierna krivka). Teda nezamietame, že frekvencie grafém slovenského jazyka pochádzajú z pravdepodobnostného rozdelenia (3.1).



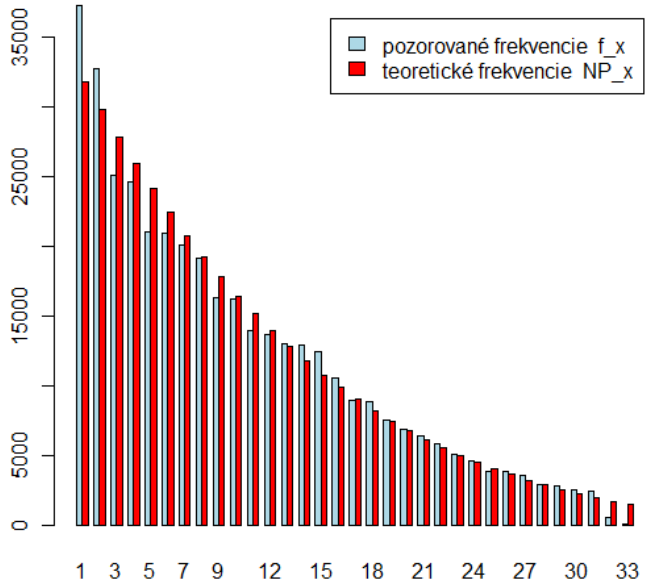
Obr. 3. **Ordov graf pre frekvencie grafém slovenského jazyka**

Ďalším testovacím kritériom je test dobrej zhody (bližšie opísaný v časti 1.2). P – hodnota testovacej štatistiky je uvedená dole v tabuľke (tab. 1). Keďže tento test zamietol nulovú hypotézu (1.11) (p - hodnota vyšla menšia ako 0.05, preto zamietame nulovú hypotézu), preto sa počíta koeficient *C* (uvedený dole v tabuľke (tab. 1)).

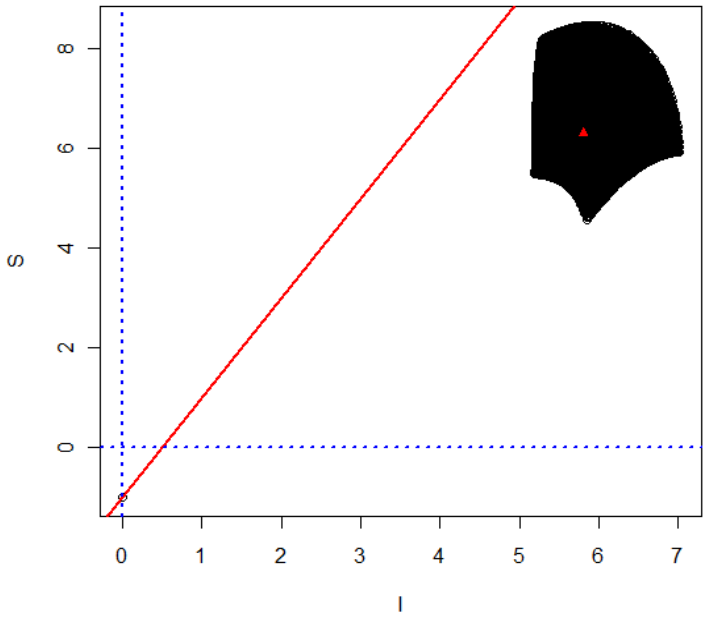
Analogický postup urobíme aj pre dostupné pozorované frekvencie z ukrajinského, slovinského a ruského jazyka, pričom *Ordov graf* uvedieme už len pre ukrajinské dáta.

Tab. 2. Pozorované frekvencie grafém v ukrajinskom jazyku

i	f_i	i	f_i	i	f_i
1	37267	***12***	13697	***23***	5074
2	32774	***13***	12959	***24***	4625
3	25080	***14***	12949	***25***	3876
4	24639	***15***	12398	***26***	3843
5	21053	***16***	10584	***27***	3565
6	20941	***17***	8944	***28***	2857
7	20075	***18***	8877	***29***	2790
8	19171	***19***	7487	***30***	2484
9	16296	***20***	6888	***31***	2407
10	16240	***21***	6406	***32***	506
11	13936	***22***	5850	***33***	78
n = 33			N = 386616		
p1 = 0.9370			C = 0.0135		
p2 = 0.9085			p_hodnota = 0.000		



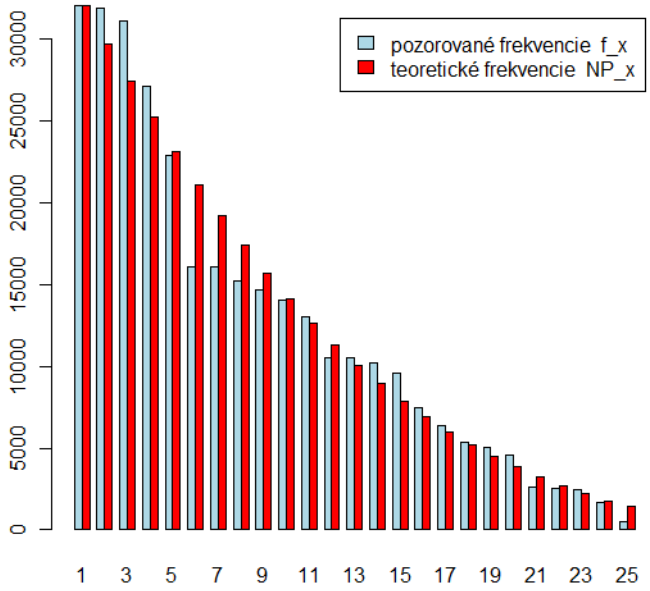
Obr. 4. **Porovnanie pozorovaných a teoretických frekvencii grafém ukrajinského jazyka**



Obr.5. **Ordov graf pre frekvencie grafém ukrajinského jazyka**

Tab. 3. Pozorované frekvencie grafém v slovinskom jazyku

i	f_i	i	f_i	i	f_i
1	32036	***10***	14043	***19***	5055
2	31891	***11***	13034	***20***	4608
3	31122	***12***	10517	***21***	2606
4	27150	***13***	10514	***22***	2554
5	22905	***14***	10216	***23***	2463
6	16088	***15***	9568	***24***	1675
7	16084	***16***	7446	***25***	497
8	15221	***17***	6413	***30***	2484
9	14668	***18***	5361	***31***	2407
n = 25			N = 313735		
p1 = 0.9317			C = 0.0145		
p2 = 0.8781			p_hodnota = 0.000		

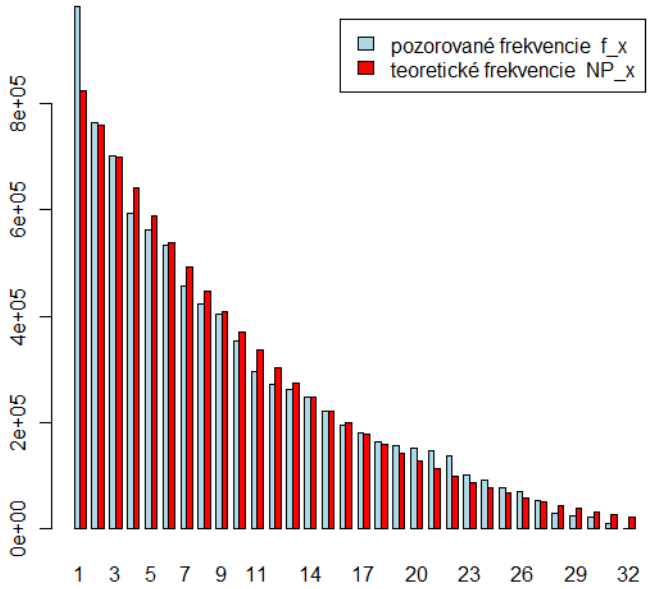


Obr. 6. **Porovnanie pozorovaných a teoretických frekvencii grafém slovinského jazyka**

Ak si pozrieme hodnotu koeficientu *C* v tabuľkách (tab. 1 - 4), tak hodnota tohto koeficientu vyšla vo všetkých prípadoch menšia ako 0.02. Teda sme sa dostali pod požadovanú hranicu, zhoda modelu s dátami o frekvenciách grafém je postačujúca.

Tab. 4. Pozorované frekvencie grafém v ruskom jazyku

i	f_i	i	f_i	i	f_i
1	982048	***12***	272216	***23***	101994
2	763584	***13***	262459	***24***	93156
3	701891	***14***	248196	***25***	77999
4	593949	***15***	222221	***26***	69870
5	563581	***16***	195629	***27***	54464
6	532783	***17***	181684	***28***	30584
7	456610	***18***	163449	***29***	24421
8	423657	***19***	156929	***30***	22314
9	403285	***20***	151944	***31***	9578
10	353818	***21***	146832	***32***	2257
11	295548	***22***	138459		
n = 32			N = 8697409		
p1 = 0.9242			C = 0.0152		
p2 = 0.8915			p_hodnota = 0.000		



Obr. 7. **Porovnanie pozorovaných a teoretických frekvencii grafém ruského jazyka**

3.5 Modifikácia prvej triedy

Pri modelovaní frekvencii usporiadaných grafém v niektorých jazykoch musíme modely modifikovať. Napríklad v nemčine alebo tamilčine sa jedna graféma v textoch vyskytuje výrazne častejšie ako ostatné grafémy. Teda v dátach máme na začiatku jednu frekvenciu, ktorá výrazne prevyšuje ostatné. Ak by sme na takéto dáta použili pôvodný modelu (3.1), tak zhoda modelu s dátami nie je postačujúca (napr. pre nemecké frekvencie v tabuľke (tab. 5) vyšiel koeficient C pre model (3.1) až 0.0362). Preto musíme modifikovať pôvodný model. Jedna z možnosti je použiť modifikáciu prvej triedy (modifikujeme pravdepodobnosť výskytu prvej grafémy), túto modifikáciu možno nájsť v [13].

$$\begin{aligned} Q_1 &= 1 - b(1 - P_1) \\ Q_x &= bP_x \qquad\qquad \text{pre} \quad x = 2, 3, \dots, n \quad , \end{aligned} \tag{3.10}$$

kde $0 < b < 1$ a P_x sú pravdepodobnosti z pôvodného modelu (3.1), teda aj Q_x tvorí pravdepodobnostné rozdelenie.

V modifikovanom modeli sa vyskytuje ďalší neznámy parameter b. Neznáme parametre sa budú opäť odhadovať metódou minimálneho χ^2, popísanou v kapitole 3.1, kde namiesto pravdepodobnosti P_i dosadíme pravdepodobnosti Q_i. Namiesto úlohy (3.7) riešime úlohu

$$\min_{p_1,p_2,b}\left\{\sum_{i=1}^{k}\frac{(f_i - NQ_i(p_1,p_2,b))^2}{NQ_i(p_1,p_2,b)}, 0 \le p_1 \le 1, 0 \le p_2 \le 1, 0 \le b \le 1\right\}.$$

Počiatočné hodnoty parametrov p_1^0 a p_2^0 zvolíme rovnako ako pri pôvodnom modeli (3.1) a počiatočnú hodnotu b^0 určíme z rovnice

$$NQ_1 = N\left(1 - b^0(1 - P_1)\right) = f_1,$$

odkiaľ

$$b^0 = \frac{1-\frac{f_1}{N}}{1-P_1}.$$

Ak sa vrátime k pôvodnému modelu (3.1), v ktorom $P_1 = c = \frac{f_1}{N}$, potom $b^0 = 1$.

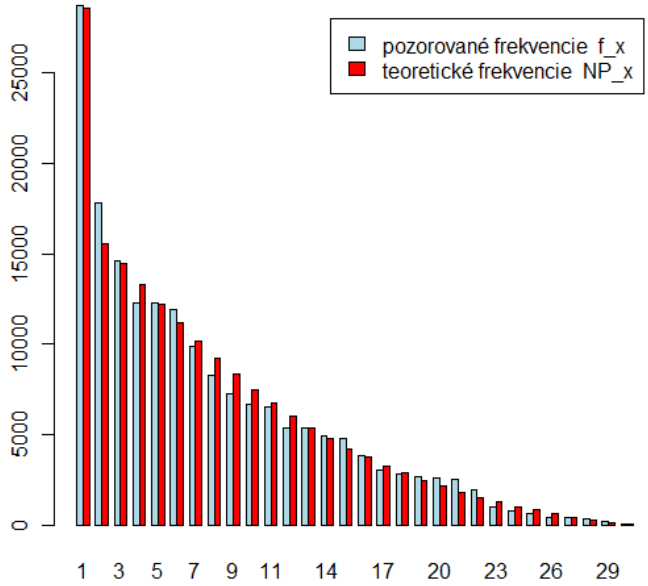
V tabuľke (tab. 5) sú zobrazené pozorované frekvencie z nemčiny. Je vidieť, že prvá frekvencia má oveľa vyššiu frekvenciu ako druhá.

Z obrázku (obr. 8) je vidieť, že modifikovaný model (3.10) správne zachytil prvú frekvenciu a aj ďalšie frekvencie. Podobne je to aj s tamilčinou, ktorej pozorované frekvencie sú zobrazené v tabuľke (tab. 6) a porovnanie pozorovaných frekvencii s modifikovaným modelom (3.10) je zobrazené na obrázku (obr. 9).

Tab. 5. Pozorované frekvencie grafém v nemeckom jazyku

i	f_i	i	f_i	i	f_i
1	28711	***11***	6559	***21***	2496
2	17814	***12***	5390	***22***	1948
3	14575	***13***	5386	***23***	974
4	12258	***14***	4923	***24***	785
5	12251	***15***	4808	***25***	633
6	11908	***16***	3825	***26***	446
7	9898	***17***	3028	***27***	388
8	8282	***18***	2804	***28***	312
9	7270	***19***	2686	***29***	212
10	6691	***20***	2626	***30***	35
n=30			N = 179922		
p1 = 0.9479			C = 0.0097		
p2 = 0.7917			p_hodnota = 0.0000		
b= 0.9332					

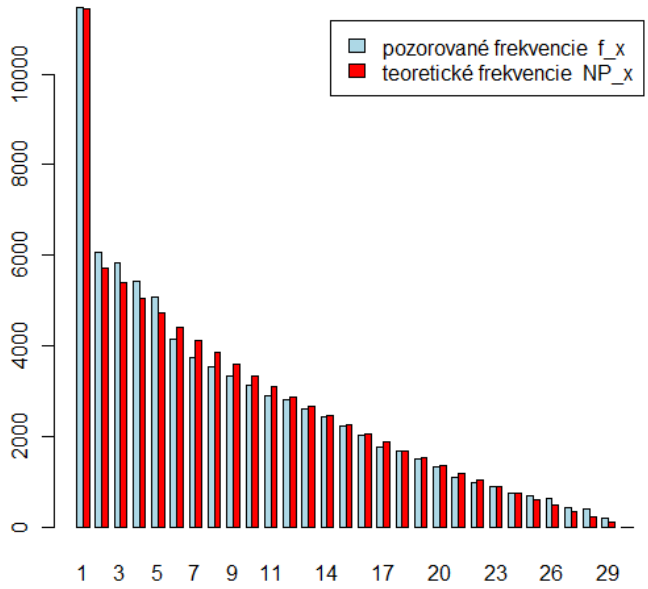
Hodnota koeficientu C uvedená v tabuľkách (tab. 5 - 6) je oveľa menšia ako 0.02, teda zhoda modifikovaného modelu (3.10) s dátami o frekvenciách z nemeckého a tamilského jazyka je „veľmi dobrá.“



Obr. 8. **Porovnanie pozorovaných a teoretických frekvencii grafém nemeckého jazyka**

Tab. 6. Pozorované frekvencie grafém v tamilskom jazyku

i	f_i	i	f_i	i	f_i
1	11462	***11***	2902	***21***	1087
2	6050	***12***	2807	***22***	970
3	5844	***13***	2599	***23***	899
4	5423	***14***	2440	***24***	755
5	5083	***15***	2214	***25***	674
6	4157	***16***	2019	***26***	623
7	3729	***17***	1763	***27***	420
8	3526	***18***	1669	***28***	382
9	3331	***19***	1510	***29***	201
10	3119	***20***	1329	***30***	0
n=30			N = 78987		
p1 = 0.9796			C = 0.0063		
p2 = 0.5964			p_hodnota = 0.0000		
b = 0.9310					



Obr. 9. **Porovnanie pozorovaných a teoretických frekvencii grafém tamilského jazyka**

Záver

Navrhnutý model bol testovaný na dátach z niekoľkých jazykov. Pre slovenčinu, ukrajinčinu, slovinčinu a ruštinu je miera zhody modelu s dátami postačujúca (vo všetkých prípadoch vyšiel koeficient C (1.12) menší ako 0.02). Pre iné jazyky (nemčina, tamilčina) je potrebne modifikovať pravdepodobnosť najčastejšie sa vyskytujúcej grafémy.

Práca prináša najmä teoretické výsledky, zároveň sa však môže stať základom pre aplikácie (nutným ďalším krokom je interpretácia parametrov).

Literatúra

[1] GRZYBEK, P. (2007). On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies. In: *Glottometrics*, 15; 82-91.

[2] GRZYBEK, P., KELIH, E. (2005). Towards a General Model of Grapheme Frequencies for Slavic Languages. In: Garabík, Radovan (Ed.), *Computer Treatment of Slavic and East European Languages*. Bratislava: Veda. s.73-87.

[3] GRZYBEK, P., KELIH, E., ALTMANN, G. (2006). Graphemhäufigkeiten im Slowakischen. (Teil II: Mit Digraphen). In: Kozmová, Ružena (Ed.), *Sprache und Sprachen im mitteleuropäischen Raum*. Trnava. (661-684).

[4] GRZYBEK, P., KELIH, E., ALTMANN, G. (2005). Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph. In: Altmann, Gabriel; Levickij, Viktor; Perebejnis, Valentina (Hrsg.), *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics*. Černivci: Ruta. (159-179).

[5] GRZYBEK, P., KELIH, E., ALTMANN, G. (2004). Häufigkeit russischer Grapheme. Teil II: Modelle von Häufigkeitsverteilungen. In: *Anzeiger für Slavische Philologie*, 32; 25-54.

[6] GRZYBEK, P., KELIH, E., STADLOBER, E. (2006). Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik. In: *Anzeiger für Slavische Philologie*, 34; 41-74.

[7] LAMOŠ, F., POTOCKÝ, R. (1998). *Pravdepodobnosť a matematická štatistika*. 2. vyd., Bratislava : Polygrafické stredisko UK.

[8] MAČUTEK, J. (2008). A generalization of geometric distribution and its application in quantitative linguistic. In: *Romanian Reports in Physics 60, 501-509.*

[9] ORD, J.K. (1967). On a system of discrete distributions, In: *Biometrika*, 54, 649-656.

[10] SIROMONEY, G. (1963). Entropy of Tamil prose, *Information and Control*, 6, 297–300, 1963.

[11] ŠTEFÁNIK, J., RUSKO, M., POVAŽANEC, D. (1999). The Frequency of Words,Graphemes, Phones and Other Elements in Slovak. In: *Jazykovedný časopis*, *50 ,B*ratislava, No. 2, s. 81 – 93.

[12] WIMMER, G. (2000). *Diskrétne jednorozmerné rozdelenia pravdepodobnosti*. Praha: Matfyzpress.

[13] WIMMER, G., WITKOVSKÝ, V., ALTMANN, G. (1999). Modification of probability distributions applied to word length research, In: *Journal of Quantitative Linguistics*, 6, 257–268.

Príloha

Zdrojový kód použitý v softvéry *R* na výpočet odhadov neznámych parametrov diskrétneho rozdelenia (3.1).

```
Px <- function(i,p1,p2,n)  {
  P <- c(n)
  c <- ((1/n)*( (p1^n + n)/(1-p1) - (1-p1^(n+1))/(1-p1)^2 + (1-p2^(n+1))/(1-p2)^2 -((n+1)*p2^n)/(1-p2) ))^(-1)
  return(c*((1-i/n)*p1^(i-1) + (i/n)*p2^(i-1) ))
}

chi_func <- function(x)   {
      p1 <- x[1]
      p2 <- x[2]
      s <- 0
      s_p <- 0
      f_p <- 0
      P <- c(k)
      c <- ((1/n)*( (p1^n + n)/(1-p1) - (1-p1^(n+1))/(1-p1)^2 + (1-p2^(n+1))/(1-p2)^2 -((n+1)*p2^n)/(1-p2) ))^(-
            1)
      for(i in 1:(k-1))     {
        P[i] <- c*((1-i/n)*p1^(i-1)+(i/n)*p2^(i-1))
        s <- s + (f[i]-N*P[i])^2/(N*P[i])
      }
      for(i in k:n)      {
        s_p <- s_p + c*((1-i/n)*p1^(i-1)+(i/n)*p2^(i-1))
        f_p <- f_p + f[i]
      }
      P[k] <- s_p
      s <- s + (f_p-N*s_p)^2/(N*s_p)
      return(s)
}

min_chi_method <- function(f) {
  f <<- t(f)
  n <<- length(f)
  N  <<- sum(f)
  k <<- n
  c_init <- f[1]/N
  p1_start <- 1 - N/sum((1:n)*f)
  p2_start <- p1_start
  je_splnene <- FALSE
  k_pom <- k
  while(je_splnene == FALSE)      {
      sol <- optim(c(p1_start,p2_start),chi_func,gr = NULL, method="L-BFGS-B",
                   lower = c(0.01,0.01), upper = c(0.99,0.99))
      p1_estim <- sol$par[1]
      p2_estim <- sol$par[2]
      N_P_chi <- c(k)
      N_P_chi <- N*Px(1:(k-1),sol$par[1],sol$par[2],n)
      N_P_chi[k] <- N*sum(Px(k:n,sol$par[1],sol$par[2],n))
      i <- 1
      while (N_P_chi[i] >= 5)   {
         i <- i+1
         if (i == (k+1)) break
      }
      k <- i-1
      if (k == k_pom) je_splnene <- TRUE
      k_pom <- k
  }
  solution <- c(p1_estim, p2_estim)
```

```
  N_P_teor <- c(n)
  N_P_teor <- N*Px(1:n,sol$par[1],sol$par[2],n)
  C <- chi_func(solution)/N
  p_value <- 1-pchisq(chi_func(solution), df = k - 3)
  return(solution,N_P_teor,C,p_value)
}

#---------------------------------------- Modifikacia prvej triedy ------------------------------------------------

Qx <- function(x,p1,p2,b,n)  {
  if(x == 1)  pom <- ( 1-b*(1-Px(1,p1,p2,n)))
  else  pom <-(b*Px(x,p1,p2,n))
  return(pom)
}

chi_func_mod <- function(x)   {
      p1 <- x[1]
      p2 <- x[2]
      b <- x[3]
      s <- 0
      s_k_trieda <- 0
      f_k_trieda <- 0
      Q <- c(k)
      P <- c(k)
      c <- ((1/n)*( (p1^n + n)/(1-p1) - (1-p1^(n+1))/(1-p1)^2 + (1-p2^(n+1))/(1-p2)^2 -((n+1)*p2^n)/(1-p2) ))^(-
            1)
      for(i in 1:(k-1))     {
        P[i] <- c*((1-i/n)*p1^(i-1)+(i/n)*p2^(i-1))
        if (i==1) Q[i] <- 1-b*(1-P[i])
          else Q[i] <- b*P[i]
        s <- s + (f[i]-N*Q[i])^2/(N*Q[i])
      }
      for(i in k:n)      {
        s_k_trieda <- s_k_trieda + b*c*((1-i/n)*p1^(i-1)+(i/n)*p2^(i-1))
        f_k_trieda <- f_k_trieda + f[i]
      }
      Q[k] <- s_k_trieda
      s <- s + (f_k_trieda-N*s_k_trieda)^2/(N*s_k_trieda)
      return(s)
}

min_chi_method_mod <- function(f) {
  f <<- t(f)
  n <<- length(f)
  N <<- sum(f)
  k <<- n
  c_init <- f[1]/N
  p1_start <- 1 - N/sum((1:n)*f)
  p2_start <- 1 - p1_start
  b_start <- 1
  je_splnene <- FALSE
  k_pom <- k
  while(je_splnene == FALSE)      {
      sol <- optim(c(p1_start,p2_start,b_start),chi_func_mod,method="L-BFGS-B",
                  lower = c(0.001,0.001,0.001), upper = c(0.999,0.999,0.999))
      p1_estim <- sol$par[1]
      p2_estim <- sol$par[2]
      b_estim <- sol$par[3]
      N_P_chi <- c(k)
      for( i in 1:k-1) {
        N_P_chi[i] <- N*Qx(i,sol$par[1],sol$par[2],sol$par[3],n)
      }
      suc <- 0
      for (i in k:n) {
```

```
      suc <- suc + N*(Qx(i,sol$par[1],sol$par[2],sol$par[3],n))
    }
    N_P_chi[k] <- suc
    i <- 1
    while (N_P_chi[i] >= 5)   {
       i <- i+1
      if (i == (k+1)) break
    }
    k <- i-1
    if (k == k_pom) je_splnene <- TRUE
    k_pom <- k
  }
  N_P_teor_mod <- c(n)
  for( i in 1:n) {
    N_P_teor_mod[i] <- N*Qx(i,sol$par[1],sol$par[2],sol$par[3],n)
  }
  solution <- c(p1_estim, p2_estim,b_estim)
  C <- chi_func_mod(solution)/N
  p_value <- 1-pchisq(C*N, df = k - 3)
  return(solution,N_P_teor_mod, C, p_value)
}

#------------------------------------------------ Ordov graf ------------------------------------------------------------

poc <- 80
p1 <- c(poc)
p1[1] <- output$solution[1]-0.04
for(i in 2:poc)  p1[i] <- p1[i-1] + 0.001

poc_1 <- 80
p2 <- c(poc_1)
p2[1] <- output$solution[2]-0.04
for(i in 2:poc)  p2[i] <- p2[i-1] + 0.001

#n <- c(20:80)
n <- length(f)
m <- 0
S <- 0
I <- 0

for(j in 1:length(n)){   #--- vypocet S a I ---
  for(k in 1:length(p1)){
    for(l in 1:length(p2)){
      m <- m+1
      i <- c(1:n[j])
      mi <- sum(i*Px(i,p1[k],p2[l],n[j]))
      mi_2 <- sum((i-mi)^2*Px(i,p1[k],p2[l],n[j]))
      mi_3 <- sum((i-mi)^3*Px(i,p1[k],p2[l],n[j]))
      S[m] <- mi_3/mi_2
      I[m] <- mi_2/mi

    }
  }
}
S[length(S)+1] <- -1
I[length(I)+1] <- 0

#---------------------------- Ordov graf pre pozorovane frekvencie-----------------------------------------

mi_poz <- sum(i*f)/sum(f)
mi_poz_2 <- sum(f*(i-mi_poz)^2)/sum(f)
mi_poz_3 <- sum(f*(i-mi_poz)^3)/sum(f)
S_poz <- mi_poz_3/mi_poz_2
I_poz <- mi_poz_2/mi_poz
```