UNIVERZITA KOMENSKÉHO V BRATISLAVA
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

MIERY DOBREJ ZHODY ZALOŽENÉ NA ŠTATISTICKÝCH VZDIALENOSTIACH

DIPLOMOVÁ PRÁCA

Bratislava 2013 Bc. Miloš Bella

UNIVERZITA KOMENSKÉHO V BRATISLAVA
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

MIERY DOBREJ ZHODY ZALOŽENÉ NA ŠTATISTICKÝCH VZDIALENOSTIACH

DIPLOMOVÁ PRÁCA

Študijný program: Ekonomická a finančná matematika
Študijný odbor: Aplikovaná matematika 1114
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Školiteľ: Mgr.Ján Mačutek, PhD.

Bratislava 2013 Bc. Miloš Bella

37393348

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Bc. Miloš Bella
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: Miery dobrej zhody založené na štatistických vzdialenostiach

Cieľ: Aplikácie niektorých štatistických vzdialeností (entropie, divergencie) na testovanie zhody medzi modelmi a dátami. Využitie náhodných čísel a simulovaných p-hodnôt.

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:	06.09.2012

prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

....

študent vedúci práce

Prehlásenie

Čestne prehlasujem, že som túto diplomovú prácu vypracoval samostatne s použitím citovaných zdrojov.

V Bratislava, apríl 2013

....

Bc. Miloš Bella

Poďakovanie

Ďakujem vedúcemu mojej práce Mgr. Jánovi Mačutkovi, PhD. za vedenie, ochotu a čas, ktorý mi venoval pri písaní práce.

Abstrakt

Bella Miloš: Miery dobrej zhody založené na štatistických vzdialenostiach: [Diplomová práca]. Univerzita Komenského v Bratislave. Fakulta matematiky, fyziky a informatiky; Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Mačutek, PhD. Bratislava 2013.

Motiváciou tejto práce sú problémy, ktoré sa vyskytujú pri použití chi-kvadrát testu dobrej zhody na testovanie dát v lingvistike. Tento test má dva základné nedostatky. V prípade veľkého počtu dát dáva tento test temer s istotou nulovú p-hodnotu. Druhým problémom tohto testu je požadovaná nezávislosť dát, čo je v lingvistike veľmi zriedkavé. Tento problém sa dá odstrániť použitím simulovaných p-hodnôt, ktoré nepožadujú nezávislosť dát.

Chi-kvadrát test dobrej zhody meria mieru podobnosti medzi modelom a dátami a patrí medzi štatistické vzdialenosti. Cieľom tejto práce je použitím simulovaných p-hodnôt na ligvistických dátach spočítať Hellingerovu vzdialenosť a totálnu variáciu a porovnať ich s chi-kvadrát vzdialenosťou.

Kľúčové slová: Chi-kvadrát dobrej zhody, štatistická vzdialenosť, diskrétne rozdelenie, generovanie náhodných čísel.

Abstract

Bella Miloš: Goodness-of-fit measures based on statistical distances: [Master thesis]. Comenius University in Bratislava. Faculty of mathematics, physics and informatics; Department of applied mathematics and statistics; supervisor: Mgr. Ján Mačutek, PhD. Bratislava 2013.

The motivation of this thesis are the problems of chi-square goodness-of-fit test when used on linguistics data. This test has two disadvantages. If the sample size is large the test rejects all null hypotheses. The stochastic independence of data, which is the requirement of this test, is problematic at linguistics. The solution is to evaluate the goodness-of-fit test by using simulated p-values, which does not require independence.

Chi-square goodness-of fit test belongs to statistical distances. The aim of this thesis is, by using simulated p-values, calculate Hellinger's distance, total variation and compare them to chi-square distance on linguistic's data.

Key words: Chi-square goodness-of-fit-test, statistical distance, discrete distribution, generation of random numbers.

Obsah

1 Úvod

χ^2 test dobrej zhody sa používa na meranie miery podobnosti medzi modelom a dátami. Ako bolo uvedené v článku [5], tento test má dve základné nevýhody. V prípade veľkého počtu dát dostaneme temer s istotou nulovú p-hodnotu. Test požaduje nezávislosť dát, čo je v prípade použitia na lingvistických dátach temer nesplniteľný predpoklad. Tento nedostatok možno obísť použitím simulovaných p-hodnôt.

V tejto práci pracujeme so štatistickými vzdialenosťami na dvoch typoch lingvistických dát. Pre tieto dáta bola v práci [2] vypočítaná χ^2 vzdialenosť. V prvom kroku počítame χ^2 vzdialenosť pre tieto dáta. Zhodu medzi modelom a dátami vyhodnocujeme použitím simulovaných p-hodnôt. Ďalej počítame Hellingerovu vzdialenosť a totálnu variáciu. Výstupom je porovnanie týchto troch vzdialeností a vyvodenie empirických záverov ohľadom potenciálnej náhrady χ^2 testu dobrej zhody inými štatistickými vzdialenosťami.

Pre naše postupy používame lingvistické dáta, pre ktoré bolo v práci [2] použitím softwaru Altmann Fitter návrhnuté predpokladané rozdelenie. Pre toto rozdelenie počítame v programe R optimalizovaný parameter (parametre) rozdelenia pre danú štatistickú vzdialenosť. Použitím tohto optimalizovaného parametra (parametrov) generujeme 2000 krát sadu náhodných čísel z predpokladaného rozdelenia. Pre každú vzdialenosť a text dostaneme hodnotu vzdialenosti a simulovanú p-hodnotu.

Táto práca je rozdelená na dva hlavné celky. V prvej časti sú uvádzané teoretické poznatky, ktoré je potrebné ovládať na pochopenie neskoršej aplikácie na dátach. Prvá časť je venovaná pojmom ako pravdepodobnostné rozdelenia, χ^2 test dobrej zhody a štatistické vzdialenosti. Druhá časť je zameraná na prácu s dátami. Použitím troch štatistických vzdialeností a pomocou simulovaných p-hodnôt na lingvistických dátach dostaneme empirické závery ohľadom použiteľnosti resp. nepoužiteľnosti týchto vzdialeností ako náhrady Pearsonovho χ^2 testu dobrej zhody.

2 Teoretická časť

V tejto kapitole uvedieme základné poznatky z teórie pravdepodobnosti a o rozdeleniach pravdepodobnosti. Ako zdroj nám slúžila kniha [10].

2.1 Kolmogorovova definícia pravdepodobnosti

Definícia 2.1. Nech Ω je ľubovoľná neprázdna množina. Neprázdny systém $\mathcal{A}$ podmnožín množiny Ω sa nazýva σ-algebra, ak platí

$$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A},$$

$$A_i \in \mathcal{A}, i = 1, 2, \ldots \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}.$$

Definícia 2.2 (Kolmogorov)**.** Nech Ω je neprázdna množina, nech $\mathcal{A}$ je σ-algebra náhodných javov definovaných na Ω. Pravdepodobnosťou sa nazýva reálna funkcia $P(A)$ definovaná na $\mathcal{A}$, ktorá pre $A \in \mathcal{A}, A_1, A_2, \ldots \in \mathcal{A}, A_i \bigcap A_j = \emptyset$ a pre všetky $i \neq j$, spĺňa

$$P(\Omega) = 1,$$

$$P(A) \geq 0,$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Najdôležitejšie vlastnosti takto zavedenej pravdepodobnosti uvedieme v nasledujúcej vete.

Veta 2.1. *Ak $A, B \in \mathcal{A}, A_1, \ldots, A_k \in \mathcal{A}, A_i \cap A_j = \emptyset$ pre všetky $i \neq j$, tak platí*

$$P(\emptyset) = 0,$$

$$0 \leq P(A) \leq 1,$$

$$P(A^c) = 1 - P(A)$$

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i),$$

$$A \subset B \Rightarrow P(A) \leq P(B),$$

$$A \subset B \Rightarrow P(B - A) = P(B) - P(A).$$

Trojicu $(\Omega, \mathcal{A}, P)$ nazývame pravdepodobnostný priestor.

2.2 Pravdepodobnostné rozdelenia

Definícia 2.3. Nech X je reálna funkcia definovaná na Ω, nech a, b, sú ľubovolné reálne čísla, resp. $-\infty, \infty$. Zaveďme užitočné označenie pre niektoré podmnožina množiny Ω:

$$[X < b] = \omega \in \Omega : X(\omega) < b,$$

$$[a < X < b] = \omega \in \Omega : a < X(\omega) < b.$$

Definícia 2.4. Majme pravdepodobnostný priestor $(\Omega, \mathcal{A}, P)$. Reálna funkcia X definovaná na Ω, pre ktorú platí

$$x \in \Re \Rightarrow [X < x] \in \mathcal{A}, \tag{1}$$

sa nazýva náhodná veličina.

Reálna funkcia X spĺňajúca (1) sa nazýva merateľná, prvky σ-algebry $\mathcal{B}$ sa nazývajú merateľné množiny. Pre každú množinu $B \in \mathcal{B}$ určuje náhodná veličina X pravdepodobnosť $P_X(B)$ náhodného javu B pomocou vzťahu

$$P_X(B) = P(\omega \in \Omega : X(\omega) \in B) = P(X^{-1}(B)).$$

Definícia 2.5. Množinová funkcia

$$P_X(B) = P(X^{-1}(B)), \qquad B \in \mathcal{B},$$

sa nazýva rozdelenie pravdepodobnosti náhodnej veličiny X.

Definícia 2.6. Nech X je náhodná veličina definovaná na pravdepodobnostnom priestore $(\Omega, \mathcal{A}, P)$. Reálna funkcia

$$F_X(x) = P[X < x]$$

sa nazýva distribučná funkcia náhodnej veličiny X.

Definícia 2.7 (Diskrétne rozdelenie pravdepodobnosti)**.** Nech existujú také $x_1, x_2, \ldots$, pre ktoré platí:

$$\sum_{i=1}^{\infty} P[X = x_i] = 1.$$

Zoznam hodnôt, ktoré nadobúda náhodná veličina s diskrétnym rozdelením, a zoznam pravdepodobností, s ktorými tieto hodnoty náhodná veličina nadobúda, udávajú diskrétne rozdelenie pravdepodobnosti.

Definícia 2.8 (Binomické rozdelenie)**.** Budeme hovoriť, že náhodná premenná X má binomické rozdelenie pravdepodobnosti, ak nadobúda hodnoty j s pravdepodobnosťami

$$p_j = P(X = j) = \binom{n}{j} p^j q^{n-j}, \qquad j = 0, 1, 2, \ldots, n$$

kde $p \in (0, 1)$ a $q = 1 - p$.

Definícia 2.9 (Poissonovo rozdelenie)**.** Hovoríme, že náhodná premenná X sleduje Poissonovo rozdelenie, ak nadobúda hodnoty j s pravdepodobnosťami

$$p_j = P(X = j) = \frac{\lambda^j e^{-\lambda}}{j!}, \qquad j = 0, 1, 2, \ldots,$$

kde $\lambda > 0$ je parameter.

2.3 Generovanie náhodných čísel z diskrétnych rozdelení

V programoch, ktoré boli vytvorené pre naše účely, bolo potrebné generovať náhodné čísla z diskrétnych rozdelení. Pre tento účel bola použitá metóda z knihy [3] uvedená v ďalších riadkoch.

Uvažujme diskrétnu náhodnú premennú Y s hodnotami y_k a s odpovedajúcimi pravdepodobnosťami p_k, $k = 0, 1, \ldots$. Všeobecná metóda vychádza z toho, že ak náhodná premenná X má rovnomerné rozdelenie na intervale $(0, 1)$, potom pre $0 \leq a < b \leq 1$ platí $P(a \leq X < b) = b - a$. To znamená, že

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

Náhodné číslo z daného rozdelenia získame teda tak, že generujeme náhodné číslo x z rovnomerného rozdelenia na intervale $(0, 1)$ rozdelenia a hľadáme index m, pri ktorom platí

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

2.4 Štatistické vzdialenosti

Jedným zo základných pojmov použitých v tejto práci sú štatistické vzdialenosti. V nasledujúcich riadkoch uvádzame poznatky z knihy [7].

Najčastejšie používanými štatistickými vzdialenosťami sú f-divergencie. V súvislosti s f-divergenciou budeme pod f rozumieť ľubovoľnú pevnú konvexnú funkciu s definičným oborom $(0, \infty)$ a s oblasťou hodnôt na rozšírenej reálnej priamke. Funkcia f je naviac striktne konvexná v bode $u = 1$. Ako vyplýva z dôsledku v prílohe, existuje práve jedno spojité rozšírenie $f(0), f(\infty)$ tak, že rozšírená funkcia je konvexná na $< 0, \infty >$ a $f(0) > -\infty$. Bez narušenia všeobecnosti môžeme teda predpokladať, že f je definovaná a konvexná na $< 0, \infty >$, striktne konvexná v $u = 1$ a $f(0) > -\infty$.

Veta 2.2. *Existuje limita*

$$f(*) = \lim_{u \to \infty} \frac{f(u)}{u} \in \Re,$$

pričom platí

$$-\infty < f(1) < f(0) + f(*). \tag{2}$$

Ďalej platí

$$\lim_{u \to 0+, v \to v_0} u f(\frac{v}{u}) = v_0 f(*), \tag{3}$$

$$\lim_{u \to 0+, v \to v_0} v f(\frac{u}{v}) = v_0 f(0), \tag{4}$$

pre každé $v_0 \in (0, \infty)$.

Dôkaz. Ak je funkcia f konvexná, je spojitá na $(0, \infty)$, a teda $f(1) \in \Re$, t.j. $f(1) > -\infty$. Ďalej z vety (4.1 v Dodatku) dostaneme (najskôr pri $(u', u, u'') = (0, 1, u)$ a potom pri $(u', u, u'') = (1, u, u')$) nerovnosti

$$\frac{f(1) - f(0)}{1 - 0} < \frac{f(u) - f(1)}{u - 1} < \frac{f(u') - f(1)}{u' - 1}$$

pre všetky $1 < u < u'$. Z monotónnosti prostrednej funkcie na intervale $u \in (1, \infty)$ vyplýva, že existuje

$$\lim_{u \to \infty} \frac{f(u) - f(1)}{u - 1} = \lim_{u \to \infty} \frac{f(u)}{u},$$

kde rovnosť vyplýva z konečnosti $f(1)$. Podľa ľavej nerovnosti platí aj (2). Keďže

$$\lim_{u \to 0+, v \to v_0} u f\left(\frac{v}{u}\right) = \lim_{u \to 0+, v \to v_0} v \frac{f(\frac{v}{u})}{\frac{v}{u}},$$

vzťahy (3) a (4) vyplývajú zo spojitosti f na $< 0, \infty >$ a z toho, že limity $f(*)$ a

$$f(0) = \lim_{u \to 0+} f(u)$$

existujú. $\square$

Dôsledok 2.1. *Ak je $f(u)$ konvexná na $(0,\infty)$, tak*

$$\bar{f}(u) = f(u) - f(1)$$

je dobre definovaná a konvexná na $(0,\infty)$, pričom $\bar{f}(1) = 0$. Ak je $f(u)$ striktne konvexná v $u = 1$, je aj $\bar{f}(u)$ striktne konvexná v $u = 1$, pričom $\bar{f}(0) + \bar{f}(*) > 0$.

Tento dôsledok nás oprávňuje predpokladať, že $f(1) = 0$ bez narušenia všeobecnosti definície f-divergencie.

Definícia 2.10. f-divergenciu hustôt p, q (resp. príslušných pravdepodobností P, Q) definujeme pre každú funkciu $f(u)$, konvexnú na $(0,\infty)$ a striktne konvexnú v $u = 1$, výrazom

$$D_f(P||Q) = \sum_{x \in X} q(x) f\left(\frac{p(x)}{q(x)}\right), \tag{5}$$

kde predpokladáme, že $f(1) = 0$ a kladieme

$$0f\left(\frac{0}{0}\right) = 0 \tag{6}$$

a

$$0f\left(\frac{p}{0}\right) = pf(*), \qquad \forall p \in (0,\infty). \tag{7}$$

Všimnime si, že podmienka v (7) vyplýva z definície limity $f(*)$ a z požiadavky spojitosti

$$0f\left(\frac{p}{0}\right) = \lim_{\epsilon \to 0+} \epsilon f\left(\frac{p}{\epsilon}\right).$$

Niektoré špeciálne prípady f-divergencie, prislúchajúce určitým konkrétnym konvexným funkciám $f(u)$, sa vo fyzike, matematickej analýze, pravdepodobnosti, štatistike i teórii informácie používajú už mnoho desaťročí. Nasledujúca veta má fundamentálny význam, lebo umožňuje chápať, v akom zmysle meria f-divergencia nepodobnosť pravdepodobnostných modelov. Poznamenajme, že za maximálne divergentné je prirodzené považovať také modely, v ktorých sú pravdepodobnosti P a Q ortogonálne. Znamená to, že existujú disjunktné podmnožiny $E, F \subset X$, pre ktoré platí $P(E) = 1$, $Q(F) = 1$.

Veta 2.3. *Pre f-divergenciu platí nerovnosť*

$$0 \leq D_f(P,Q) \leq f(0) + f(*),$$

pričom obidve rovnosti nemôžu nastať súčasne. Ľavá strana platí práve vtedy, keď $P = Q$, a pravá práve vtedy, keď P a Q sú ortogonálne.

Z tejto vety vyplýva, že modely (X, P) a (X, Q) sú navzájom podobné, ak ich f-divergencia $D_f(P||Q)$ je blízka číslu 0. Maximálna podobnosť je vlastne zhoda pravdepodobností P a Q na príslušnej σ-algebre podmnožín množiny X. Naopak modely sú navzájom nepodobné, ak sa $D_f(P||Q)$ blíži k svojej maximálnej možnej hodnote $f(0) + f(*)$. Maximálna nepodobnosť znamená, že v modeli (X, P) majú kladnú pravdepodobnosť iba tie javy, ktorých pravdepodobnosť je v modeli (X, Q) nulová, a naopak.

V nasledujúcej časti uvedieme prehľad najdôležitejších tried f-divergencií. Tie môžme rozdeliť na základe vytvárajúcich funkcií$f(u)$:

1. $f(u) = u \log u$
2. $f(u) = |u^\beta - 1|^{1/\beta}$
3. $f(u) = |u - 1|^\alpha$
4. $f(u) = sign(\alpha - 1)(u^\alpha - 1)$

1. Nech
$$f(u) = u \log(u),$$
Potom
$$I(P||Q) = \sum_{x \in X} q(x) \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}.$$
Táto divergencia sa zvykne označovať ako I-divergencia.

2. Nech
$$f(u) = \frac{1}{2}|u^\beta - 1|^{1/\beta}.$$
Ak $\beta \in (0, 1)$, platí
$$\begin{aligned} D_\beta(P||Q) &= \frac{1}{2} \sum_{x \in X} q(x) \left(\left(\frac{p(x)}{q(x)} \right)^\beta - 1 \right) = \\ &= \frac{1}{2} \sum_{x \in X} q(x) \left(\left| p(x)^\beta - q(x)^\beta \right|^{\frac{1}{2}} \right) = \\ &= \frac{1}{2} \sum_{x \in X} q(x) - 2\sqrt{p(x)q(x)} + p(x) = \end{aligned}$$

$$= 1 - \sum_{x \in X} \sqrt{p(x)q(x)}.$$

Táto divergencia sa označuje ako β-divergencia.

Špeciálnym prípadom β-divergencie je Hellingerova vzdialenosť, ktorú dostaneme, ak položíme $\beta = \frac{1}{2}$, teda pre

$$f(u) = \frac{1}{2}(1 - \sqrt{u})^2.$$

Dostaneme potom

$$\begin{aligned}
D_{\frac{1}{2}}(P||Q) &= \frac{1}{2} \sum_{x \in X} q(x) \left(1 - \sqrt{\frac{p(x)}{q(x)}}\right) = \\
&= \frac{1}{2} \sum_{x \in X} q(x) \left(1 - 2\sqrt{\frac{p(x)}{q(x)}} + \frac{p(x)}{q(x)}\right) = \\
&= \frac{1}{2} \sum_{x \in X} q(x) - 2\sqrt{p(x)q(x)} + p(x) = \\
&= 1 - \sum_{x \in X} \sqrt{p(x)q(x)}.
\end{aligned}$$

3. Nech

$$f(u) = |u - 1|^\alpha$$

Ak položíme $\alpha = 1$, dostaneme vzdialenosť, ktorá sa označuje ako totálna variácia. Je definovaná ako

$$\chi^1(P||Q) = V(P||Q) = \sum_{x \in X} q(x) \left| \frac{p(x)}{q(x)} - 1 \right| = \sum_{x \in X} |p(x) - q(x)|.$$

Ak položíme $\alpha \in (1, \infty)$ dostaneme χ^α-divergenciu

$$\chi^\alpha(P||Q) = \sum_{x \in X} q(x) \left| \frac{p(x)}{q(x)} - 1 \right|^\alpha = \sum_{x \in X} \frac{|p(x) - q(x)|^\alpha}{q(x)^{\alpha - 1}}.$$

χ^2-divergenciu

$$\chi^2(P||Q) = \sum_{x \in X} q(x) \left(\frac{p(x)}{q(x)} - 1 \right)^2 = \sum_{x \in X} \frac{(p(x) - q(x))^2}{q(x)}$$

získame dosadením $\alpha = 2$. Táto vzdialenosť sa používa pri χ^2 teste dobrej zhody.

4. Nech $f(u) = sign(\alpha - 1)(u^\alpha - 1)$. Ak $\alpha \in (0,1)$, platí

$$D^\alpha(P||Q) = \sum_{x\in X} -q(x)\left(\left(\frac{p(x)}{q(x)}\right)^\alpha - 1\right) = 1 - \sum_{x\in X} p(x)^\alpha q(x)^{1-\alpha}.$$

Pre $\alpha \in (1,\infty)$ máme

$$D^\alpha(P||Q) = \sum_{x\in X} q(x)\left(\left(\frac{p(x)}{q(x)}\right)^\alpha - 1\right) = \sum_{x\in X} \frac{p(x)^\alpha}{q(x)^{\alpha-1}} - 1.$$

Táto vzdialenosť sa označuje ako α-divergencia.

O význame I-divergencie svedčí okrem iného aj to, že v priebehu mnohých desaťročí bola táto divergencia objavovaná a skúmaná s väčšou alebo menšou mierou nezávislosti veľkým počtom autorov. Konkrétne mená autorov tu neuvádzame, niektoré možno nájsť v knihe [7].

$V(P||Q)$ je v teórii pravdepodobnosti používaná už dlho pod názvom totálna variácia, v matematickej štatistike je známa ako Kolmogorovova vzdialenosť.

V nasledujúcej vete bez dôkazu uvedieme známe vzťahy medzi divergenciami, ktoré sú dôležité pre pochopenie ich topologických vlastností.

Veta 2.4. *Pre všetky pravdepodobnosti P, Q na množine X platí:*

1. $D_1(P||Q) = \chi^1(P||Q) = V(P||Q)$,
2. $D_{\frac{1}{2}}(P||Q) = 2D^{\frac{1}{2}}(P||Q)$,
3. $\chi^2(P||Q) = D^2(P||Q)$,
4. $V(P||Q) < (\chi^\alpha(P||Q))^{\frac{1}{\alpha}} \quad \forall\, \alpha \geq 1$,
5. $\left[\left(1 - \frac{V(P||Q)}{2}\right)^\beta - \left(1 - \frac{V(P||Q)}{2}\right)^\beta\right]^{\frac{1}{\beta}} \leq D_\beta(P||Q) \leq V(P||Q)$ *pre* $0 < \beta \leq 1$,
6. $1 - \left(1 + \frac{V(P||Q)}{2}\right)^{max(\alpha,1-\alpha)} \left(1 - \frac{V(P||Q)}{2}\right)^{min(\alpha,1-\alpha)} \leq D^\alpha(P||Q) \leq \frac{V(P||Q)}{2}$ *pre* $0 < \alpha < 1$.

Ďalej platí, že $D^{\frac{1}{2}}(P||Q)$ a $\chi^1(P||Q) = V(P||Q)$ a taktiež $D_\beta(P||Q)^\beta$ pri ľubovoľnej β sú metriky na priestore všetkých pravdepodobností množiny X.

2.5 Pearsonov χ^2 test dobrej zhody

Nulová hypotéza Pearsonovho χ^2 testu dobrej zhody je, že pozorovania pochádzajú z konkrétneho teoretického rozdelenia pravdepodobnosti. Testovacia štatistika je definovaná ako

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - NP_i)^2}{NP_i}, \tag{8}$$

kde f_i je sledovaná empirická početnosť hodnôt v i-tej triede, P_i teoretická pravdepodobnosť hodnôt v i-tej triede, n je počet tried, do ktorých sú dáta rozdelené, a N je veľkosť pozorovanej vzorky (celkový počet pozorovaní). Ak platí nulová hypotéza, štatistika (8) asymptoticky sleduje χ^2 rozdelenie s $n - p - 1$ stupňami voľnosti (n je počet tried, p je počet odhadovaných parametrov). Očakávané početnosti by nemali byť príliš malé, Snedecor a Cochran [6] navrhovali, aby žiadna z početností nebola menšia ako 1. Triedy s očakávanými početnosťami menšími ako 1 majú byť zlúčené. Test (8) je matematicky korektný, ale podobne ako ostatné štatistické testy zamieta všetky nulové hypotézy, ak je vzorka príliš veľká. Toto je často prípad v lingvistike, kde vzorky o veľkosti desiatok tisíc a viac nie sú výnimkou. Pre takéto obrovské sady dát je tento test prakticky nepoužiteľný. Tento nedostatok do istej miery odstraňuje koeficient diskrepancie. Predpokladajúc, že χ^2 štatistika rastie lineárne s veľkosťou vzorky, ak sú teoretické a empirické relatívne početnosti fixné, koeficient diskrepancie (pozri [1]) je definovaný ako

$$C = \frac{\chi^2}{N}. \tag{9}$$

Za prijateľné sú podľa [5] považované hodnoty koeficientu $C \leq 0.02$, prípadne hodnoty $C \leq 0.05$. Test (8) je v lingvistike používaný na testovanie, ak je vzorka "primerane" veľká. Naopak koeficient diskrepancie je pre χ^2 test vhodný, ak je testovacia vzorka príliš veľká. Oveľa závažnejším problémom ako veľkosť vzorky je splnenie predpokladov testu (8). Jedným zo základných predpokladov Pearsonovho χ^2 testu dobrej zhody je, že pozorovania pochádzajú z náhodného výberu, čo je postupnosť nezávislých a rovnako rozdelených náhodných premenných. Predpokladanie nezávislosti by napríklad znamenalo, že ak jazyk používa grafému a, ľubovoľne dlhá postupnosť pozostávajúca iba z grafémy a by bola možná, jej výskyt by mal závisieť iba na dĺžke postupnosti a na očakávanej početnosti grafémy a v jazyku. Keďže a je jedna z najčastejšie používaných grafém, za predpokladu nezávislosti by sa slovo *aaaaa* malo vyskytovať v textoch častejšie ako ostatné zmysluplné slová, ktoré obsahujú menej používané grafémy.

3 Aplikácia štatistických vzdialeností na lingvistické dáta

3.1 Všeobecná schéma postupu - simulované p-hodnoty

Hlavnou náplňou tejto práce je výpočet troch druhov štatistických vzdialeností na vopred určené dáta. Na dátach, ktoré používame pre naše účely, bola v práci [2] spočítaná χ^2 vzdialenosť, čo je vlastne test dobrej zhody. Našim cieľom bolo okrem spomínaného χ^2 testu dobrej zhody spočítať Hellingerovu vzdialenosť a totálnu variáciu.

Pre naše účely sme v tejto diplomovej práci použili dva typy dát, na ktoré sme aplikovali naše postupy. V tomto odseku uvedieme všeobecnú schému postupu.

Východzím bodom je, že máme k dispozícii predpokladané rozdelenie dát. Potom postupujeme nasledovne:

1. Optimalizácia parametrov rozdelenia pre danú štatistickú vzdialenosť a pre dané dáta. Výsledkom je teda optimalizovaná hodnota parametra (parametrov), pre ktorú (ktoré) je daná vzdialenosť minimálna.

2. 2000 krát generujeme náhodné čísla z rozdelenia s optimalizovaným parametrom (parametrami), pričom počet náhodných čísel je zhodný s počtom dát v súbore. (Počet 2000 opakovaní bol zvolený preto, lebo program R bežne vykonáva 2000 opakovaní pri výpočte simulovaných p-hodnôt.)

3. Pre každý vygenerovaný súbor vypočítame vzdialenosť od predpokladaného rozdelenia s optimalizovaným parametrom.

4. Spočítame simulovanú p-hodnotu (teda počet tých vygenerovaných súborov, v ktorých je vzdialenosť väčšia ako v prípade optimalizovaného parametra, vydelený počtom všetkých generovaní).

3.2 Dáta od S.G. Čebanova

3.2.1 Čebanovove výsledky

V tejto podkapitole popíšeme postup, ktorý bol použitý v článku [2]. Tento postup uvádzame preto, lebo je v ňom závažná chyba urobená pri výpočte. Tieto lingvistické dáta a výpočty pochádzajú od ruského lekára Sergeja Grigorieviča Čebanova (1897-1966). Jeho záujem o lingvistiku sa zameriaval hlavne na vývoj jazyka. Čebanov považoval "distribúciu slov podľa počtu slabík" za jednu z najdôležitejších štatistických charakteristík štruktúry jazyka. Čebanov skúmal 127 rôznych jazykov a dialektov po dobu viac ako 20 rokov.

Čebanov hľadal všeobecný model pre rozdelenie slov podľa počtu slabík. Čebanovov počiatočný predpoklad bol špecifický vzťah medzi strednou hodnotou počtu slabík v texte $\bar{x}$ a relatívnymi frekvenciami p_i jednotlivých tried podľa počtu slabík. Poznajúc strednú hodnotu rozdelenia, Čebanov predpokladal, že Poissonovo rozdelenie je adekvátnym modelom pre jeho dáta.

Poissonovo rozdelenie je definované ako

$$P_x = \frac{e^{-a}a^x}{x!} \qquad x = 0, 1, 2 \dots \tag{10}$$

Keďže rovnica (10) je platná pre $x = 0, 1, 2, \dots$ a $a \geq 0$ a keďže sa v texte nenachádzajú žiadne nulovo-slabičné slová, tak je pre model vhodnejšie použiť posunuté Poissonovo rozdelenie

$$P_x = \frac{e^{-a}a^{x-1}}{(x-1)!} \qquad x = 1, 2, 3 \dots \tag{11}$$

V knihe [2] boli uvedené a použité 3 z Čebanovových dátových súborov, ktoré budú rovnako použité i pre naše účely. Jedná sa o nemecký text *Parzival*, starosaský text *Heliand* a tretí text je urývok z knihy od Leva N. Tolstoja *Vojna i mir*. Absolútne frekvencie (f_i) a zodpovedajúce relatívne frekvencie p_i sú uvedené v tabuľke. Počet slabík v slove označuje premenná i. Dáta predstavujeme v Tabuľke 1.

i	Parzival		Heliand		Vojna i mir	
	f_i	p_i	f_i	p_i	f_i	p_i
1	1823	0.628	1572	0.469	466	0.283
2	849	0.292	1229	0.367	541	0.328
3	194	0.067	452	0.135	391	0.237
4	37	0.013	83	0.025	172	0.104
5			14	0.004	64	0.039
6					15	0.009
$\sum$	2903		3350		1698	

Tabuľka 1: Relatívne a absolútne početnosti i-slabičných slov

Čebanov počítal χ^2 test dobrej zhody na daných dátach. Pre prvé dáta *Parzival*, s počtom tried $k = 4$ dostal $\chi^2 = 1.45$. Táto hodnota sa môže považovať za dobrý fit, nakoľko $p(\chi^2) = 0.48$ pre dva stupne voľnosti. Z dát *Heliand* Čebanov nedostal dobré výsledky. Hodnota $\chi^2 = 10.35$ je aj napriek horšiemu výsledku stále akceptovateľná na hladine významnosti $\alpha = 0.01$ ($p(\chi^2) = 0.016$ pre tri stupne voľnosti). Rovnako pre dáta *Vojna i mir* vyšli Čebanovovi dobré výsledky, $\chi^2 = 5.82$ ($p(\chi^2) = 0.213$ pre štyri stupne voľnosti).

3.2.2 Chyba u Čebanova

Dôležitou chybou u Čebanova pri počítaní χ^2 štatistiky bolo, že všetky teoretické pravdepodobnosti počítal podľa vzorca

$$P_x = \frac{e^{-a}a^x}{x!}, \quad x = 1, 2, \ldots, k, \tag{12}$$

kde k je maximálna dĺžka slova v analyzovanom texte. To nespĺňa základnú podmienku

$$\sum_{x=1}^{k} P_x = 1.$$

Tento nedostatok sa ľahko dá napraviť tým, že teoretické pravdepodobnosti počítame pomocou vzorca (12) pre $x = 1, 2, \cdots, k-1$ a

$$P_k = 1 - \sum_{x=1}^{k-1} P_x. \tag{13}$$

V našom postupe sme teda počítali pravdepodobnosti podľa vzora (12) doplneného podľa vzorca (13). Ako druhý krok sme hľadali optimalizovaný parameter $\bar{a}$. Tento postup sme robili v programe R. Všetky programy možno nájsť v prílohe. Po optimalizácii parametra $\hat{a}$ sme dostali nasledovné teoretické frekvencie a χ^2 štatistiky.

i	Parzival		Heliand		Vojna i mir	
	f_i	Np_i	f_i	Np_i	f_i	Np_i
1	1823	1821.58	1572	1611.48	466	440.78
2	849	848.93	1229	1179.29	541	581.56
3	194	197.82	452	431.51	391	383.65
4	37	34.67	83	105.26	172	168.72
5			14	22.46	64	55.65
6					15	18.64
$\sum$	2903		3350		1698	
χ^2	0.23		11.93		6.44	

Tabuľka 2: Početnosti i-slabičných slov a χ^2 štatistika

V Tabuľke 3 môžeme vidieť porovnanie našich χ^2 štatistík s prislúchajúcimi simulovanými p-hodnotami a výsledky od Čebanova.

	Parzival		Heliand		Vojna i Mir	
Čebanov	χ^2	p-hodnota	χ^2	p-hodnota	χ^2	p-hodnota
	1.45	0.48	10.35	0.016	5.82	0.213
Naše výsledky	χ^2	p-hodnota	χ^2	p-hodnota	χ^2	p-hodnota
	0.23	0.971	11.93	0.0185	6.44	0.2655

Tabuľka 3: Porovnanie výsledkov χ^2 štatistiky

Možno si všimnúť, že naše χ^2 štatistiky nadobúdajú iné hodnoty ako v prípade výsledkov od Čebanova. Tento rozdiel je spôsobený jeho iným (nesprávnym) postupom.

3.2.3 Iné štatistické vzdialenosti

V tejto práci venujeme pozornosť nielen χ^2 vzdialenosti (v ďalšom χ^2), ale aj ostatným štatistickým vzdialenostiam, a to Hellingerovej vzdialenosti (v ďalšom HV) a totálnej variácii (v ďalšom TV). Rovnako i pri týchto vzdialenostiach sme optimalizovali parameter $\hat{a}$ a následne sme spočítali dané vzdialenosti. Výsledky môžeme vidieť v nasledujúcej tabuľke.

	Parzival	Heliand	Vojna i mir
χ^2	0.231	11.929	6.439
$\hat{a}$	0.466	0.732	1.3194
P-value	0.971	0.019	0.267
HV	0.003	0.022	0.022
$\hat{a}$	0.466	0.727	1.318
P-value	0.976	0.011	0.247
TV	0.001	0.012	0.026
$\hat{a}$	0.465	0.757	1.333
P-value	0.999	0.318	0.176

Tabuľka 4: Štatistické vzdialenosti a ich simulované p-hodnoty

Možno si všimnúť, že v prvých dvoch textoch sme najlepšiu simulovanú p-hodnotu dostali v prípade totálnej variácie. Podobný jav sledujeme aj v prípade dát o dĺžke slov v dolnolužickej srbčine v ďalšej kapitole.

3.3 Dáta o dĺžke slov v dolnolužickej srbčine

Ako ďalšie dáta sme použili dĺžky slov v žurnalistických textoch v dolnolužickej srbčine. Dáta sme prebrali z práce [8]. Dolnolužickou srbčinou hovorí asi 20000 ľudí v nemeckom spolkovom štáte Brandenburg.

Dáta použité v tejto časti práce majú dĺžku 100-1000 slov. Štúdie boli vykonané na homogénnych textoch a ide o články z dolnosrbského časopisu Nowy Casnik. Pre každý analyzovaný text bol spočítaný počet slov spadajúcich do príslušnej skupiny podľa počtu slabík. Testovali sme zhodu dát s hyper-Poissonovým rozdelením. Hyper-Poissonovo rozdelenie (pozri [9]) je definované ako

$$P_x = \frac{a^{x-1}}{b^{(x-1)}{}_1F_1(1;b;a)}, \qquad x = 1, 2, \ldots \tag{14}$$

kde ${}_1F_1(1;b;a)$ je hypergeometrická funkcia s parametrami a a b.

Hypergeometrická funkcia je definovaná v [9] ako

$${}_1F_1(1;b;a) = 1 + \frac{b}{(a)1!} + \frac{b(b+1)}{a(a+1)2!} + \cdots$$

Rovnako i pri týchto dátach sme počítali s tromi štatistickými vzdialenosťami: χ^2 vzdialenosťou, Hellingerovou vzdialenosťou a totálnou variáciou. Pre každý z desiatich textov je vypočítaná optimalizovaná štatistika spolu s optimalizovanými parametrami daného rozdelenia a simulovaná p-hodnota. V nasledujúcej časti sú pre lepší prehľad uvedené iba výsledky. Podrobnejšie informácie spolu s konkrétnymi dátami sú uvedené v Prílohe.

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	10.695	1.155	1.167	0.029
TV	0.061	0.723	0.658	0.260
HV	0.902	1.012	1.054	0.069

Tabuľka 5: Text 1

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	8.061	1.396	1.618	0.210
TV	0.036	0.689	0.644	0.675
HV	0.062	0.991	1.064	0.455

Tabuľka 6: Text 2

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	1.601	1.851	2.449	0.904
TV	0.028	1.430	1.801	0.891
HV	0.035	1.820	2.424	0.896

Tabuľka 7: Text 3

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	6.602	1.985	2.005	0.170
TV	0.0458	2.740	3.132	0.656
HV	0.069	1.772	1.792	0.184

Tabuľka 8: Text 4

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	20.775	3.645	4.294	0.010
TV	0.112	4.667	6.647	0.050
HV	0.128	3.474	4.423	0.020

Tabuľka 9: Text 5

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	8.765	2.793	3.904	0.066
TV	0.053	3.796	5.867	0.523
HV	0.081	2.251	3.210	0.113

Tabuľka 10: Text 6

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	2.381	0.945	1.065	0.671
TV	0.027	1.156	1.391	0.848
HV	0.041	0.906	1.027	0.633

Tabuľka 11: Text 7

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	8.321	0.808	0.736	0.079
TV	0.053	0.514	0.424	0.469
HV	0.086	0.706	0.665	0.168

Tabuľka 12: Text 8

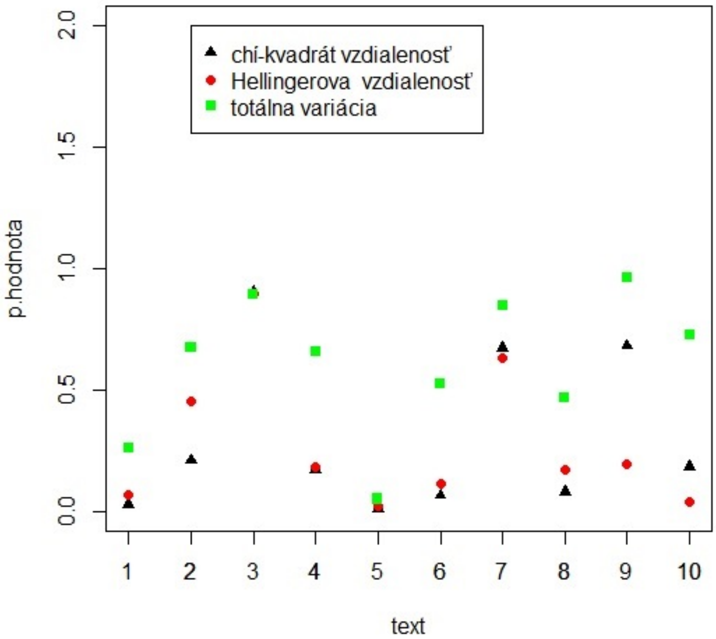
Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	3.101	1.109	1.216	0.680
TV	0.023	1.260	1.455	0.963
HV	0.085	0.814	0.845	0.195

Tabuľka 13: Text 9

Vzdialenosť	Štatistika	a	b	P-hodnota
χ^2	6.199	1.355	1.459	0.183
TV	0.044	1.098	1.197	0.728
HV	0.109	0.808	0.829	0.040

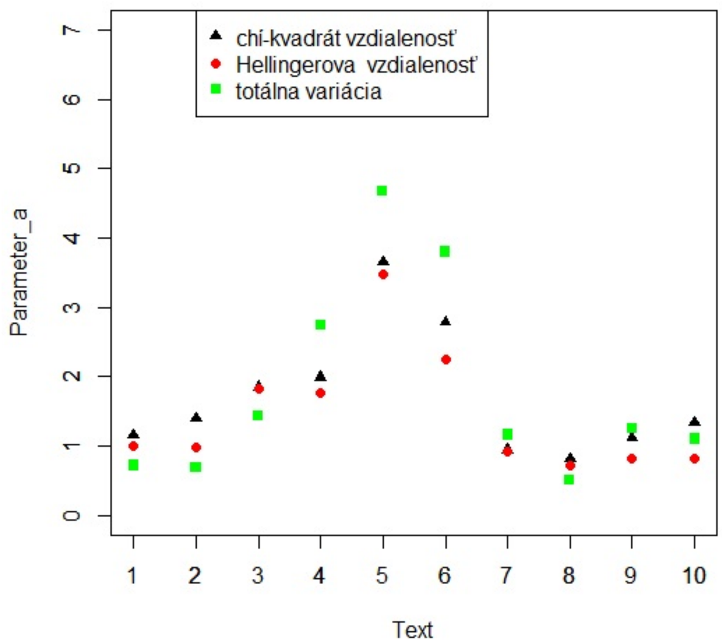
Tabuľka 14: Text 10

Pre ľahšiu orientáciu vo výsledkoch sme vyrobili grafy znázorňujúce porovnanie simulovaných p-hodnôt a optimalizovaných parametrov pre jednotlivé vzdialenosti.

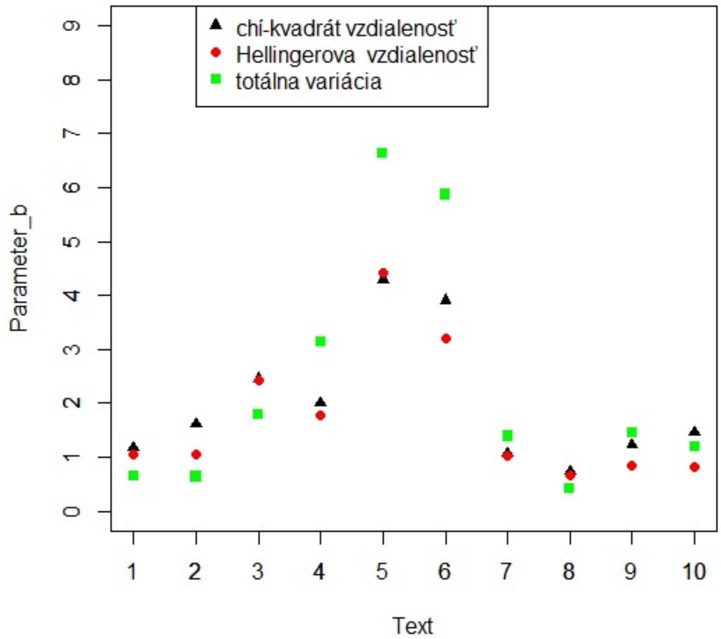


Obr. 1: Porovnanie simulovaných p-hodnôt

Z uvedených výsledkov a rovnako i z grafu je možné si všimnúť, že skoro vo všetkých desiatich textoch je simulovaná p-hodnota v prípade použitia χ^2 vzdialenosti a Hellingerovej vzdialenosti dosť podobná. Naopak v prípade totálnej variácie sú p-hodnoty oveľa vyššie. Podobný jav sme zaznamenali aj v dvoch z troch analyzovaných textoch od Čebanova [pozri Kapitolu 3.2]. Zdá sa, že Hellingerova vzdialenosť dáva podobné výsledky ako χ^2 test dobrej zhody a preto by Hellingerova vzdialenosť mohla byť použitá ako istá obmena práve spomenutého testu dobrej zhody. Čo sa skrýva za vyššími hodnotami p-hodnôt v prípade totálnej variácie nie je úplne zrejmé. Možno si všimnúť, že v prípade totálnej variácie nastáva optimalizácia parametrov a a b tak, že teoretické frekvencie dvoch hodnôt sa vždy rovnajú tým reálnym. Nie je teda zrejmé, či totálna variácia je nevhodná ako alternatíva testu dobrej zhody vďaka nízkej ochote zamietať výsledky, alebo skutočne dosahuje táto vzdialenosť lepšie optimalizované hodnoty.



Obr. 2: Porovnanie hodnôt optimalizovaného parametra a



Obr. 3: Porovnanie hodnôt optimalizovaného parametra b

Rovnaký empirický záver možno vyvodiť i v prípade optimalizovaných parametrov a a b. Optimalizované parametre v prípade χ^2 vzdialenosti dosahujú podobné hodnoty ako v prípade Hellingerovej vzdialenosti. Zdá sa, že parametre totálnej variácie nadobúdajú iné hodnoty. Samotné hodnoty týchto parametrov nám nehovoria nič z čisto matematického hľadiska, no z pohľadu lingvistiky môžu mať význam pri interpretácii jednotlivých štatistických vzdialeností.

4 Záver

Táto práca sa zaoberá porovnávaním χ^2 testu dobrej zhody, Hellingerovej vzdialenosti a totálnej variácie na lingvistických dátach.

Prvým hlavným výsledkom bolo nájdenie chyby pri počítaní v práci [2] na dátach od Čebanova. Pri počítaní χ^2 vzdialenosti sú dáta rozdelené do viacerých tried. Pre každú triedu je podľa predpokladaného rozdelenia vypočítaná očakávaná frekvencia p_i. V prípade použitia posunutého Poissonovho rozdelenia je potrebné pre poslednú triedu dát vypočítať očakávanú frekvenciu ako 1- $\sum(p_i)$. Chybou, ktorej sa dopustili v knihe [2], je, že aj posledná trieda bola počítaná podľa vzorca pre predpokladané rozdelenie, čo viedlo k problému, že súčet očakávaných pravdepodobností nie je rovný 1.

Druhým výstupom je vypočítanie týchto troch vzdialeností na dvoch typoch dát. V prvom prípade máme tri menšie sady dát. Keďže sa jedná o dáta z troch rôznych jazykov, nie je možné jednoznačne vyvodzovať výsledky o týchto vzdialenostiach. No i v tomto prípade je možné si všimnúť, že p-hodnoty pre totálnu variáciu nadobúdajú v dvoch z troch prípadov oveľa väčšie hodnoty ako pre ďalšie dve použité štatistiky.

V druhej sade dát, ktorá obsahuje dáta z 10 rôznych textov z dolnolužickej srbčiny, je možné si všimnúť, že temer v každom texte nadobúda totálna variácia vyššie simulované p-hodnoty ako Hellingerova vzdialenosť a χ^2 vzdialenosť. Naopak Hellingerova vzdialenosť dosahuje podobné hodnoty optimalizovaných parametrov a simulovaných p-hodnôt ako χ^2 vzdialenosť.

Literatúra

[1] Cressie, N., Read, T.R.C. (1984). Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society. Series B (Methodological), 46(3), 440-464.

[2] Grzybek, P. (2006). History and methodology of word length studies. In: Contributions to the Science of Text and Language. Word Length Studies and Related Issues (Grzybek, P., ed.), pp. 15-90. Dordrecht: Springer.

[3] Kalas J., Pekár J. (1991). Simulačné metódy. Bratislava: Polygrafické stredisko UK.

[4] Lamoš F.(1983). Základy teórie pravdepodobnosti. Bratislava: Polygrafické stredisko UK.

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

[6] Snedecor, G.W., Cochran, W.G. (1989). Statistical Methods (8th ed.). Ames (IA): Iowa State University Press.

[7] Vajda I. (1982). Teória informácie a štatistického rozhodovania. Bratislava: vydavateľstvo Alfa.

[8] Wilson, A.. (2006). Word-length distribution in present-day Lower Sorbian newspaper texts. In: Contributions to the Science of Text and Language. Word Length Studies and Related Issues (Grzybek, P., ed.), pp. 319-327. Dordrecht: Springer.

[9] Wimmer, G., Altmann, G. (1999). Thesaurus of univariate discrete probability distributions. Essen: Stamm.

[10] Zvára K., Štěpán J. (2001). Pravděpodobnosť a matematická statistika. Praha: Matfyzpress - Vydavatelství Matematicko-fyzikální Univerzity Karlovy.

Príloha

Veta 4.1 (I-interval)**.** *Ak $f : I(a,b) \to \Re$ je konvexná a $u' < u < u''$ sú ľubovoľné tri body z $I(a,b)$, tak*

$$\frac{f(u)-f(u')}{u-u'} \le \frac{f(u'')-f(u')}{u''-u'} \le \frac{f(u'')-f(u)}{u''-u}$$

pričom nerovnosti sú ostré, ak je f striktne konvexná v bode u.

Dôsledok 4.1. *Ak je $f : (a,b) \to R$ konvexná, tak existujú v $\Re$ limity*

$$f(a) = \lim_{u\to a^+} f(u), \qquad f(b) = \lim_{u\to b^-} f(u)$$

pričom takto (jednoznačne) rozšírená funkcia $f :< a, b >\to R*$ je konvexná. Rozšírená funkcia je striktne konvexná (v bode $u_0 \in (a,b)$) práve vtedy, keď f je striktne konvexná (v bode u_0).

Podrobné výsledky z dolnolužickej srbčiny

Nasledujúce tabuľky obsahujú výsledky z 10 textov,kde:

- x_i je počet slabík v slove
- f_i je sledovaná frekvencia i-slabičných slov
- NP_i je teoretická frekvencia i-slabičných slov
- a je optimalizovaný parameter pre danú štatistiku
- b je optimalizovaný parameter pre danú štatistiku
- χ^2 optimalizovaná chi-kvadrát štatistika
- VV optimalizovaná totálna variácia
- HV optimalizovaná Hellingerova vzdialenosť
- P_{value} p-hodnota pre danú vzdialenosť

Text 1			**Text 2**		
x_i	f_i	NP_i	x_i	f_i	NP_i
1	61	61.10	1	71	73.51
2	67	59.51	2	76	63.43
3	19	31.72	3	25	33.82
4	18	11.58	4	13	13.05
5	2	4.09	5	2	3.94
			6	1	0.98
			7	1	0.25
χ^2	10.69457		χ^2	8.06078	
a	1.15527		a	1.39572	
b	1.16686		b	1.61750	
P_{value}	0.0285		P_{value}	0.2095	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	61	63.29	1	71	73.40
2	67	60.77	2	76	68.64
3	19	29.95	3	25	33.03
4	18	9.93	4	13	10.70
5	2	3.07	5	2	2.61
			6	1	0.51
			7	1	0.10
HV	0.90243		HV	0.06242	
a	1.01238		a	0.99161	
b	1.05441		b	1.06371	
P_{value}	0.0685		P_{value}	0.4545	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	61	61.00	1	71	71.00
2	67	67.00	2	76	76.00
3	19	29.21	3	25	31.86
4	18	7.94	4	13	8.30
5	2	1.85	5	2	1.57
			6	1	0.23
			7	1	0.03
TV	0.06113		TV	0.03629	
a	0.72286		a	0.68909	
b	0.6581		b	0.64375	
P_{value}	0.2595		P_{value}	0.675	

Text 3			Text 4		
x_i	f_i	NP_i	x_i	f_i	NP_i
1	68	68.07	1	56	52.93
2	54	51.45	2	49	52.40
3	23	27.61	3	32	34.61
4	14	11.49	4	25	17.16
5	4	3.90	5	5	9.89
6	1	1.47			
χ^2	1.60087		χ^2	6.60175	
a	1.85100		a	1.98527	
b	2.44900		b	2.00530	
P_{value}	0.904		P_{value}	0.1695	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	68	68.64	1	56	54.53
2	54	51.53	2	49	53.93
3	23	27.39	3	32	34.22
4	14	11.27	4	25	15.99
5	4	3.78	5	5	8.33
6	1	1.40			
HV	0.03524		HV	0.06897	
a	1.82010		a	1.77167	
b	2.42433		b	1.79159	
P_{value}	0.8955		P_{value}	0.184	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	68	68.00	1	56	56.00
2	54	54.00	2	49	49.00
3	23	27.57	3	32	32.50
4	14	10.37	4	25	17.35
5	4	3.09	5	5	12.15
TV	0.02789316		TV	0.04578	
a	1.43047		a	2.74032	
b	1.80133		b	3.13179	
P_{value}	0.891		P_{value}	0.656	

Text 5			Text 6		
x_i	f_i	NP_i	x_i	f_i	NP_i
1	56	46.74	1	68	64.11
2	28	39.68	2	44	45.87
3	24	27.32	3	21	26.13
4	29	15.82	4	20	12.36
5	6	13.45	5	3	7.53
χ^2	20.77491		χ^2	8.76540	
a	3.64527		a	2.79295	
b	4.29435		b	3.90350	
P_{value}	0.01		P_{value}	0.066	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	56	51.63	1	68	67.20
2	28	40.56	2	44	47.12
3	24	25.99	3	24	25.19
4	29	14.06	4	20	10.88
5	6	10.77	5	3	5.62
HV	0.12760		HV	0.08066	
a	3.47404		a	2.25054	
b	4.42250		b	3.20969	
P_{value}	0.02		P_{value}	0.1125	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	56	56.00	1	68	68.00
2	28	39.32	2	44	44.00
3	24	24.00	3	21	24.32
4	29	12.95	4	20	11.74
5	6	10.72	5	3	7.94
TV	0.11221		TV	0.05296	
a	4.66746		a	3.79623	
b	6.64717		b	5.86690	
P_{value}	0.0495		P_{value}	0.5225	

Text 7			Text 8		
x_i	f_i	NP_i	x_i	f_i	NP_i
1	77	75.83	1	51	50.39
2	64	67.24	2	62	55.31
3	36	30.76	3	15	25.74
4	6	9.48	4	12	7.60
5	3	2.69	5	1	1.97
χ^2	2.38132		χ^2	8.32148	
a	0.94481		a	0.80789	
b	1.06548		b	0.73607	
P_{value}	0.671		P_{value}	0.079	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	77	76.64	1	51	52.71
2	64	67.62	2	62	56.79
3	36	30.22	3	15	23.87
4	6	9.05	4	12	6.24
5	3	2.47	5	1	1.38
HV	0.04149		HV	0.08600	
a	0.90599		a	0.70603	
b	1.02681		b	0.66541	
P_{value}	0.633		P_{value}	0.168	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	77	77.00	1	51	51.00
2	64	64.00	2	62	62.00
3	36	30.95	3	15	22.43
4	6	10.55	4	12	4.76
5	3	3.50	5	1	0.81
TV	0.02716		TV	0.05267	
a	1.15618		a	0.51493	
b	1.39103		b	0.42357	
P_{value}	0.8475		P_{value}	0.469	

Text 9			Text 10		
x_i	f_i	NP_i	x_i	f_i	NP_i
1	60	58.94	1	48	46.65
2	52	53.78	2	44	42.37
3	30	26.92	3	19	22.84
4	9	9.29	4	13	8.75
5	0	2.44	5	0	3.39
6	1	0.63			
χ^2	3.10105		χ^2	6.19936	
a	1.10921		a	1.32520	
b	1.21554		b	1.45910	
P_{value}	0.6795		P_{value}	0.1825	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	60	59.79	1	48	48.41
2	52	57.64	2	44	47.23
3	30	25.45	3	19	20.88
4	9	7.28	4	13	5.97
5	0	1.54	5	0	1.50
6	1	0.3			
HV	0.08450		HV	0.10922	
a	0.81439		a	0.80848	
b	0.84471		b	0.82867	
P_{value}	0.1945		P_{value}	0.0395	
x_i	f_i	NP_i	x_i	f_i	NP_i
1	60	60.00	1	48	48.00
2	52	52.00	2	44	44.00
3	30	26.71	3	19	21.98
4	9	9.75	4	13	7.54
5	0	2.76	5	0	2.48
6	1	0.79			
TV	0.02306		TV	0.04399	
a	1.26080		a	1.09764	
b	1.45476		b	1.19742	
P_{value}	0.9625		P_{value}	0.7275	

Programy v jazyku R

Nasledujúci program je kód z jazyka R pre prvú sadu dát od Čebanova.

```
Z<-c(1823,849,194,37)
y=length(Z)
w=sum(Z)
pp<-rep(0,y)
Np<-rep(0,y)
A<-rep(0,y)

#optimalizacia parametra a pre Chí kvadrát vzdialenosť
res<-function(a){
for(k in 1:y){
pp[k]=exp(-a)*a^(k-1)/factorial(k-1) #1-...
pp[y]<-1-sum(pp[1:3])
Np[k]=sum(Z)*pp[k]
A[k]=((Z[k]-Np[k]))^2/(Np[k])
P=sum(A)}
P}
Pmin<-optimize(res,c(0,1))
Pmin

#optimalizacia parametra a pre Totálnu Variáciu
res<-function(a)
{
for(k in 1:y){
pp[k]<- exp(-a)*a^(k-1)/factorial(k-1)
pp[y]<-1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=(abs((Z[k]-Np[k])))
P=sum(A)/(2*w)}
P}
Smin<-optimize(res,c(0,1))
Smin

#optimalizacia parametra a pre Hellingerovu vzdialenosť
res<-function(a)
{
for(k in 1:y){
pp[k]<-exp(-a)*a^(k-1)/factorial(k-1)
pp[y]<-1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=(sqrt(Z[k]/sum(Z))-sqrt(pp[k]))^2
P=(sum(A)^(1/2))/sqrt(2)}
P}
HellingerMin<-optimize(res,c(0,1))
HellingerMin
```

```
#Simulované p-hodnoty pre Chi kvadrát
ss=2000
R<-rep(0,ss)
for(e in 1:ss){
m=2903
a=Pmin$minimum
Y<-rep(0,m)
x<-rep(0,m)
Z<-rep(0,m)
while(m>0){
x[m]<-runif(1)
k=1
pp<-rep(0,100)
M=1
while(M>0){
pp[k]<-exp(-a)*a^(k-1)/factorial(k-1)
P=sum(pp)
M=x[m]-P
Y[m]=k
k=k+1}
Z[k-1]=Z[k-1]+1
m=m-1}
Y
Z[y]=length(Y[Y>y-1])
Z[1:y]
A<-rep(0,y)
pp<-rep(0,y)
for(k in 1:y){
pp[k]=exp(-a)*a^(k-1)/factorial(k-1)
pp[y]<-1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=((Z[k]-Np[k]))^2/(Np[k])
P=sum(A)}
R[e]=P
}
R
T=length(R[R>Pmin$objective])/ss
T

#Simulované p-hodnoty pre Totálnu Variáciu
U<-rep(0,ss)
for(e in 1:ss){
m=2903
a=Smin$minimum
Y<-rep(0,2903)
x<-rep(0,2903)
Z<-rep(0,2903)
while(m>0){
```

```
x[m]<-runif(1)
k=1
pp<-rep(0,100)
M=1
while(M>0){
pp[k]<-exp(-a)*a^(k-1)/factorial(k-1)
P=sum(pp)
M=x[m]-P
Y[m]=k
k=k+1}
Z[k-1]=Z[k-1]+1
m=m-1}
Y
Z[y]=length(Y[Y>y-1])
Z[1:y]
A<-rep(0,y)
pp<-rep(0,y)
for(k in 1:y){
pp[k]<- exp(-a)*a^(k-1)/factorial(k-1)
pp[y]<-1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=(abs((Z[k]-Np[k])))
P=sum(A)/(2*w)}
U[e]=P
}
U
V=length(U[U>Smin$objective])/ss

#Simulované p-hodnoty pre Hellingerovu vzdialenosť
Q<-rep(0,ss)
for(e in 1:ss){
m=2903
p=HellingerMin$minimum
Y<-rep(0,2903)
x<-rep(0,2903)
Z<-rep(0,2903)
while(m>0){
x[m]<-runif(1)
k=1
pp<-rep(0,2903)
M=1
while(M>0){
pp[k]<-exp(-a)*a^(k-1)/factorial(k-1)
P=sum(pp)
M=x[m]-P
Y[m]=k
k=k+1}
Z[k-1]=Z[k-1]+1
m=m-1}
```

```
Y
Z[y]=length(Y[Y>y-1])
Z[1:y]
A<-rep(0,y)
pp<-rep(0,y)
for(k in 1:y){
pp[k]<-exp(-a)*a^(k-1)/factorial(k-1)
pp[y]<-1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=(sqrt(Z[k]/sum(Z))-sqrt(pp[k]))^2
P=(sum(A)^(1/2))/sqrt(2)}
Q[e]=P
}
Q
HellingerSimPvalue=length(Q[Q>HellingerMin$objective])/ss

T
V
HellingerSimPvalue
```

Nasledujúci kód je pre 10 textov dolnolužickej srbčiny.

```
Z<-c(48,44,19,13,0)
y=length(Z)
w=sum(Z)
pp<-rep(0,y)
Np<-rep(0,y)
A<-rep(0,y)
B<-rep(0,y)

#optimalizacia parametrov a,b pre Chi kvadrát vzdialenosť
res<-function(X){
a<-X[1]
b<-X[2]
B[1]=1
B[2]=b
for(k in 2:y-1){B[k+1]=B[k]*(b+k-1)}
for(k in 1:y){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
pp[y]=1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=((Z[k]-Np[k]))^2/(Np[k])
P=sum(A)}
P}
Pmin<-optim(c(1.3,1.45),res)
Pmin

#optimalizacia parametrov a,b pre Totálnu Variáciu
res<-function(X){
```

```
a<-X[1]
b<-X[2]
B[1]=1
B[2]=b
for(k in 2:y-1){B[k+1]=B[k]*(b+k-1)}
for(k in 1:y){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
pp[y]=1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=(abs((Z[k]-Np[k])))
P=sum(A)/(2*w)}
P}
Smin<-optim(c(1.1,1.2),res)
Smin

#Optimalizácia parametrov pre Hellingerovu vzdialenosť
res<-function(X){
a<-X[1]
b<-X[2]
B[1]=1
B[2]=b
for(k in 2:y-1){B[k+1]=B[k]*(b+k-1)}
for(k in 1:y){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
pp[y]=1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=(sqrt(Z[k]/sum(Z))-sqrt(pp[k]))^2
P=(sum(A)^(1/2))/sqrt(2)}
P}
HellingerMin<-optim(c(0.8,0.8),res)
HellingerMin

#Simulované p-hodnoty pre Chi kvadrát vzdialenosť
ss=2000
R<-rep(0,ss)
for(e in 1:ss){
m=124
a=Pmin$par[1]
b=Pmin$par[2]
Y<-rep(0,m)
x<-rep(0,m)
Z<-rep(0,m)
while(m>0){
x[m]<-runif(1)
k=1
pp<-rep(0,y)
M=1
while(M>0){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
```

```
pp[5]<-1-sum(pp[1:4])
P=sum(pp[1:k])
M=x[m]-P
Y[m]=k
k=k+1}
Z[k-1]=Z[k-1]+1
m=m-1}
Y
Z[y]=length(Y[Y>y-1])
Z[1:y]
A<-rep(0,y)
pp<-rep(0,y)
B[1]=1
B[2]=b
for(k in 2:y-1){B[k+1]=B[k]*(b+k-1)}
for(k in 1:y){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
pp[y]=1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=((Z[k]-Np[k]))^2/(Np[k])  #chi kvadrat test
P=sum(A)}
R[e]=P
}
R
T=length(R[R>Pmin$value])/ss
T

#Simulované p-hodnoty pre Totálnu Variáciu
U<-rep(0,ss)
for(e in 1:ss){
m=124
a=Pmin$par[1]
b=Pmin$par[2]
Y<-rep(0,m)
x<-rep(0,m)
Z<-rep(0,m)
while(m>0){
x[m]<-runif(1)
k=1
pp<-rep(0,y)
M=1
while(M>0){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
pp[5]<-1-sum(pp[1:4])
P=sum(pp[1:k])
M=x[m]-P
Y[m]=k
k=k+1}
Z[k-1]=Z[k-1]+1
```

```
m=m-1}
Y
Z[y]=length(Y[Y>y-1])
Z[1:y]
A<-rep(0,y)
pp<-rep(0,y)
B[1]=1
B[2]=b
for(k in 2:y-1){B[k+1]=B[k]*(b+k-1)}
for(k in 1:y){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
pp[y]=1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=(abs((Z[k]-Np[k])))
P=sum(A)/(2*w)}
U[e]=P
}
U
V=length(U[U>Smin$value])/ss

#Simulované p-hodnoty pre Hellingerovu vzdialenosť
Q<-rep(0,ss)
for(e in 1:ss){
m=124
a=Pmin$par[1]
b=Pmin$par[2]
Y<-rep(0,m)
x<-rep(0,m)
Z<-rep(0,m)
while(m>0){
x[m]<-runif(1)
k=1
pp<-rep(0,y)
M=1
while(M>0){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
pp[5]<-1-sum(pp[1:4])
P=sum(pp[1:k])
M=x[m]-P
Y[m]=k
k=k+1}
Z[k-1]=Z[k-1]+1
m=m-1}
Y
Z[y]=length(Y[Y>y-1])
Z[1:y]
A<-rep(0,y)
pp<-rep(0,y)
B[1]=1
```

```
B[2]=b
for(k in 2:y-1){B[k+1]=B[k]*(b+k-1)}
for(k in 1:y){
pp[k]=a^(k-1)/((B[k])*genhypergeo(1,b,a))
pp[y]=1-sum(pp[1:y-1])
Np[k]=sum(Z)*pp[k]
A[k]=(sqrt(Z[k]/sum(Z))-sqrt(pp[k]))^2
P=(sum(A)^(1/2))/sqrt(2)}
Q[e]=P
}
Q
HellingerSimPvalue=length(Q[Q>HellingerMin$value])/ss

T
V
HellingerSimPvalue
```