O tom, čo robí štatistik s lingvistickými dátami a čo robia lingvistické dáta so štatistikom

Ján Mačutek
Department of Applied Mathematics and Statistics
Comenius University, Bratislava

22. januára 2018

Dve verzie dnešnej prednášky

- krátka verzia
- dlhšia verzia

Krátka verzia

O tom, čo robí štatistik s lingvistickými dátami a čo robia lingvistické dáta so štatistikom

- Čo robí štatistik s lingvistickými dátami?

Krátka verzia

O tom, čo robí štatistik s lingvistickými dátami a čo robia lingvistické dáta so štatistikom

- Čo robí štatistik s lingvistickými dátami?
  Modeluje ich.

Krátka verzia

O tom, čo robí štatistik s lingvistickými dátami a čo robia lingvistické dáta so štatistikom

- Čo robí štatistik s lingvistickými dátami?
  Modeluje ich.
- Čo robia lingvistické dáta so štatistikom?

Krátka verzia

O tom, čo robí štatistik s lingvistickými dátami a čo robia lingvistické dáta so štatistikom

- Čo robí štatistik s lingvistickými dátami?
  Modeluje ich.
- Čo robia lingvistické dáta so štatistikom?
  Modelujú ho.

Krátka verzia

- Čo robí štatistik s lingvistickými dátami?
  Modeluje ich (modelovať = vytvárať zjednodušenú reprezentáciu javu/procesu, ktorá slúži na jeho skúmanie a vysvetlenie).
- Čo robia lingvistické dáta so štatistikom?
  Modelujú ho (modelovať = formovať, tvarovať, v istom zmysle meniť).

Krátka verzia - záver

Dlhšia verzia - obsah

Kvantitatívna lingvistika - čo to je

- aplikácia matematických modelov a štatistických metód vo výskume jazyka a textu
- snaha vysvetliť a/alebo interpretovať parametre modelov
- ľahšie viditeľné vzťahy medzi jednotlivými vlastnosťami jazyka
- prehľad (spred 13 rokov...) v knihe *Quantitative Linguistics. An International Handbook* (Köhler, R., Altmann, G., Piotrowski, R.G., eds., 2005, de Gruyter)

Kvantitatívna lingvistika - ako to funguje

- budovanie vedeckej teórie jazyka
- vedecká teória - súbor vedeckých zákonov, ktoré sa dajú odvodiť z teórie (dedukcia)
- (veľmi) zjednodušene, teória je jeden (veľmi) všeobecný vzorec, zákony sú jeho špeciálnymi prípadmi
- teória a zákony majú platiť všeobecne
  - zákony pre všetky jazyky
  - teória pre všetky jazyky a pre pre všetky vlastnosti jazyka

Kvantitatívna lingvistika - ale ... toto je ROBUST!
Kde je náhodnosť, štatistika, ... ???

Kvantitatívna lingvistika - ale toto je ROBUST! Kde je náhodnosť, štatistika, ... ???

- zákony aj teória majú stochastický charakter, platia "približne" (s výnimkou existenčných výrokov)
- lingvistický exkurz - existujú jazykové univerzálie, ktoré nie sú stochastické?
- požiadavky na "dobré" jazykové zákony:
  - odvoditeľné z teórie
  - platné všeobecne
  - dostatočne (v mnohých textoch, jazykoch, ...) potvrdené (Mario Bunge - "corroboration") - treba odhadovať parametre, testovať zhodu modelu s dátami, ...

Kvantitatívna lingvistika - načo to je

Kvantitatívna lingvistika - načo to je

- generuje pekné vzorce a zaujímavé matematické a štatistické problémy

Kvantitatívna lingvistika - načo to je

- zákon (napr. dĺžka slov) má byť vyjadrený tým istým matematickým modelom vo všetkých jazykoch, textoch, u všetkých autorov, atď.
- rozličné miery vlastnosti (napr. rôzne preferencie pre dĺžky slov) by sa mali odzrkadliť v rôznych hodnotách parametrov modelu
- na základe hodnôt parametrov (z viacerých zákonov) sa dajú určovať autori, klasifikovať texty, budovať lingvistická typológia, atď.

Kvantitatívna lingvistika - načo to je

- zákon (napr. dĺžka slov) má byť vyjadrený tým istým matematickým modelom vo všetkých jazykoch, textoch, u všetkých autorov, atď.
- rozličné miery vlastnosti (napr. rôzne preferencie pre dĺžky slov) by sa mali odzrkadliť v rôznych hodnotách parametrov modelu
- na základe hodnôt parametrov (z viacerých zákonov) sa dajú určovať autori, klasifikovať texty, atď.
- generuje pekné vzorce a zaujímavé matematické a štatistické problémy

Lingvistická teória (takmer) všetkého

Wimmer, G., Altmann, G. (2005), Unified derivation of some linguistic laws. In: Quantitative Linguistics. An International Handbook

$$P_x = \left(1 + a_0 + \sum_{j \geq 1} \sum_{i=1}^{k_j} \frac{a_{ji}}{(x + b_{ji})^c}\right) P_{x-1}$$

Lingvistická teória (takmer) všetkého - nenamáhaj sa viac, ako je nutné?

$$P_x = \left(1 + a_0 + \sum_{j \geq 1} \sum_{i=1}^{k_j} \frac{a_{ji}}{(x + b_{ji})^c}\right) P_{x-1}$$

- základná idea - princíp najmenšieho úsilia (least effort principle, Zipf)
- "boj" medzi hovoriacim a počúvajúcim (speaker-hearer perspective)
  - S chce používať krátke slová aj za cenu, že rôzne pojmy označí tým istým slovom (lebo krátkych slov je málo) - unifikácia
  - H preferuje jednoznačnosť aj za cenu používania dlhých slov - diverzifikácia

Lingvistická teória (takmer) všetkého - nenamáhaj sa viac, ako je rozumné?

$$P_x = \left(1 + a_0 + \sum_{j \geq 1} \sum_{i=1}^{k_j} \frac{a_{ji}}{(x + b_{ji})^c}\right) P_{x-1}$$

- ak H bude používať príliš málo rôznych slov, jeho správa nebude zrozumiteľná a bude ju musieť niekoľkokrát zopakovať - to je proti least effort principle
- "boj" medzi H a S vyústi do stavu rovnováhy, kedy sú obaja ako-tak spokojní (rovnováha medzi "silami" unifikácie a diverzifikácie)
- všeobecná a zjednodušená interpretácia parametrov:
  - ak sú parametre a_{ji} malé a parametre b_{ji} veľké, tak P_x je výrazne menšie ako P_{x-1} - vyhráva unifikácia, teda viac sa presadil S
  - parametre a_{ji} - vplyv S, parametre b_{ji} - vplyv H

Menzerathov-Altmannov zákon

- čím väčší celok, tým menšie sú jeho časti

Menzerathov-Altmannov zákon

- čím väčší celok, tým menšie sú jeho časti
- hovorí o vzťahu "susedov" v hierarchii jazykových jednotiek
  - čím viac slabík obsahuje slovo, tým kratšia je priemerná dĺžka jeho slabík meraná počtom hlások (pôvodné znenie od Menzeratha z roku 1928)
  - čím dlhšia klauza (v slovách), tým kratšie slová (v slabikách)
  - čím dlhšia veta (v klauzách), tým kratšie klauzy (v slovách)

Menzerathov-Altmannov zákon

- čím väčší celok, tým menšie sú jeho časti
- hovorí o vzťahu "susedov" v hierarchii jazykových jednotiek
  - čím viac slabík obsahuje slovo, tým kratšia je priemerná dĺžka jeho slabík meraná počtom hlások (pôvodné znenie od Menzeratha z roku 1928)
  - čím dlhšia klauza (v slovách), tým kratšie slová (v slabikách)
  - čím dlhšia veta (v klauzách), tým kratšie klauzy (v slovách)
- pravdepodobne to má niečo spoločné s krátkodobou pamäťou aj s fyzickými obmedzeniami (napr. z času na čas sa musíme nadýchnuť)
- platí nielen v jazyku, ale aj i iných komunikačných systémoch (komunikácia zvierat, hudba, štruktúra DNA)

Menzerathov-Altmannov zákon

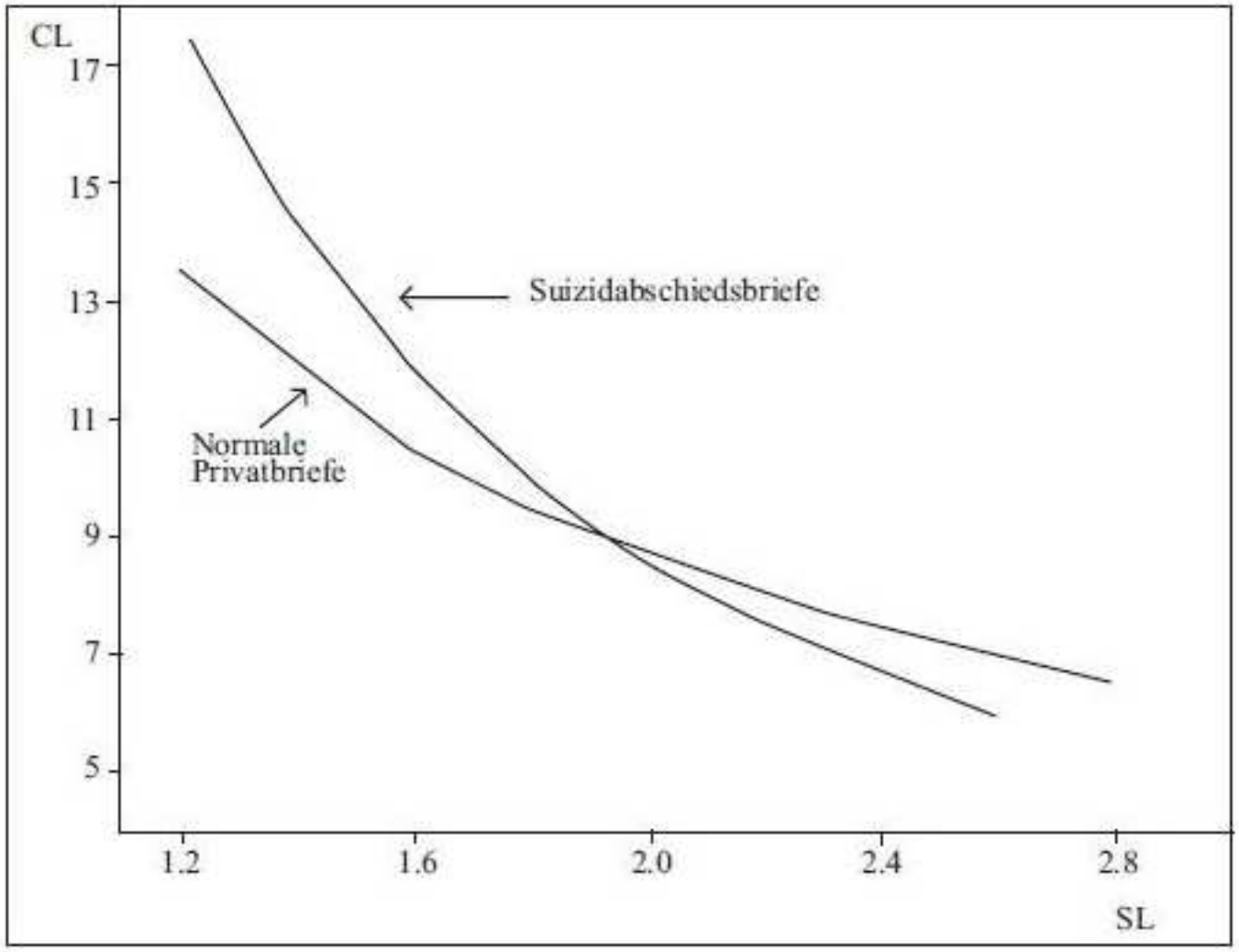
- čím väčší celok, tým menšie sú jeho časti
- hovorí o vzťahu "susedov" v hierarchii jazykových jednotiek
  - čím viac slabík obsahuje slovo, tým kratšia je priemerná dĺžka jeho slabík meraná počtom hlások (pôvodné znenie od Menzeratha z roku 1928)
  - čím dlhšia klauza (v slovách), tým kratšie slová (v slabikách)
  - čím dlhšia veta (v klauzách), tým kratšie klauzy (v slovách)
- pravdepodobne to má niečo spoločné s krátkodobou pamäťou aj s fyzickými obmedzeniami (napr. z času na čas sa musíme nadýchnuť)
- platí nielen v jazyku, ale aj v iných komunikačných systémoch (komunikácia zvierat, hudba, štruktúra DNA)
- matematické vyjadrenie je skoro vždy triviálne, $y = ax^b$

Menzerathov-Altmannov zákon

- čím väčší celok, tým menšie sú jeho časti
- hovorí o vzťahu "susedov" v hierarchii jazykových jednotiek
  - čím viac slabík obsahuje slovo, tým kratšia je priemerná dĺžka jeho slabík meraná počtom hlások (pôvodné znenie od Menzeratha z roku 1928)
  - čím dlhšia klauza (v slovách), tým kratšie slová (v slabikách)
  - čím dlhšia veta (v klauzách), tým kratšie klauzy (v slovách)
- pravdepodobne to má niečo spoločné s krátkodobou pamäťou aj s fyzickými obmedzeniami (napr. z času na čas sa musíme nadýchnuť)
- platí nielen v jazyku, ale aj v iných komunikačných systémoch (komunikácia zvierat, hudba, štruktúra DNA)
- matematické vyjadrenie je skoro vždy triviálne, $y = ax^b$
- fraktály???

Menzerathov-Altmannov zákon - nemecké listy

47. Das Menzerathsche Gesetz



Vzdialenosti medzi slovami rovnakej dĺžky

- dĺžka slova meraná počtom jeho slabík
- pozorujeme vzdialenosti medzi slovami dĺžky 1, 2, 3 a 4
- viac ako 4-slabičné slová sa vykytujú pomerne zriedkavo

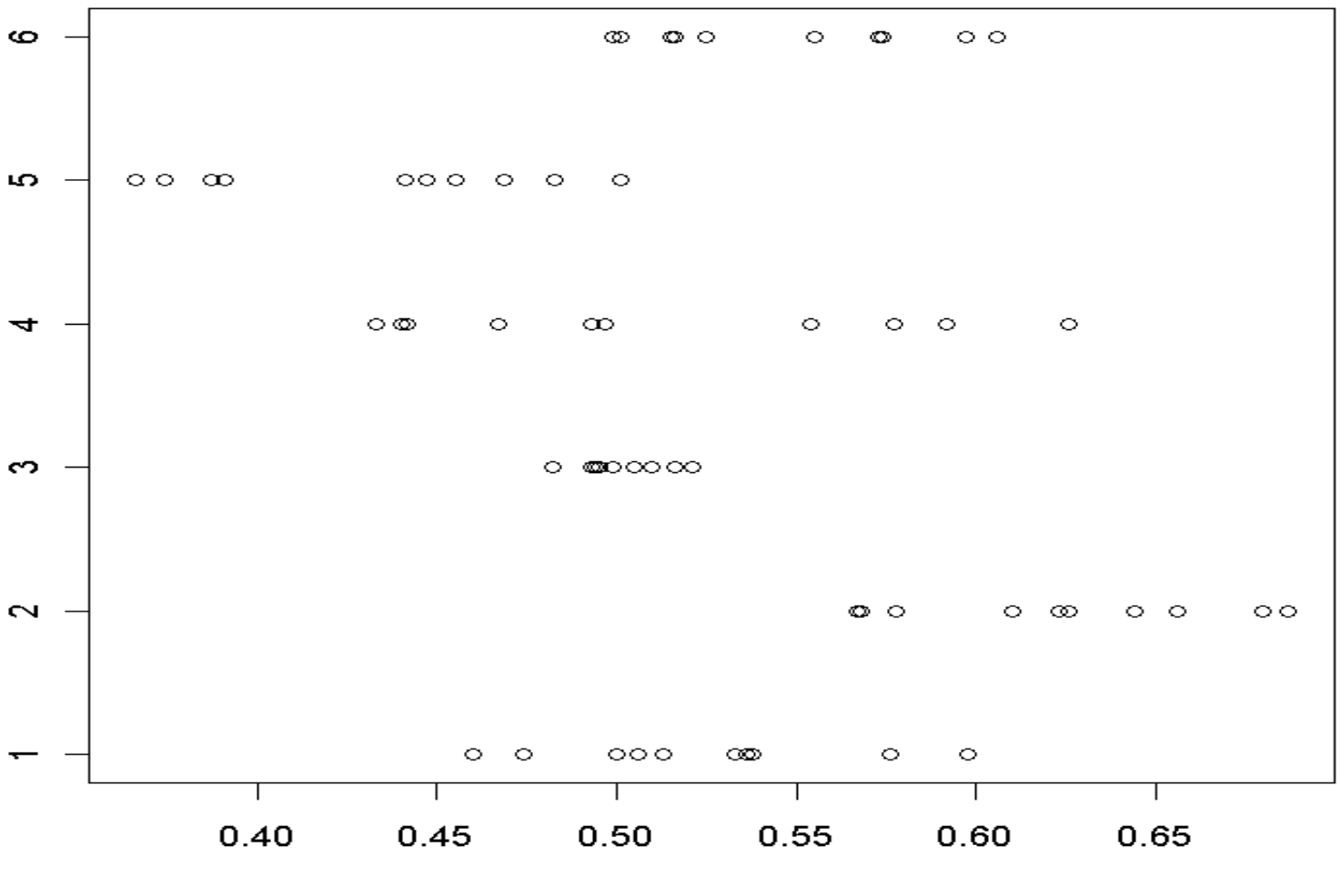
Vzdialenosti medzi slovami rovnakej dĺžky

Gross-Harris geometric distribution

$$
\begin{aligned}
P_0 &= 1 - q \\
P_x &= (1-q)(1-a)\left[a + (1-a)(1-q)\right]^{x-1}, \qquad x = 1, 2, \ldots
\end{aligned}
$$

- parameter a - len vylepšuje fit rozdelenia
- parameter q - zdá sa, že môže pomôcť pri automatickej klasifikácii textov
- model aplikovaný na malý korpus ukrajinských textov (blogy, vedecké články, umelecká próza, drámy, športové reportáže, kázne)
- pre každý text máme 4 hodnoty parametra p, na tie sme pustili klastrovú analýzu

Vzdialenosti medzi slovami dĺžky 1 - ukrajinské texty



Vzdialenosti medzi slovami dĺžky 1 - ukrajinské texty

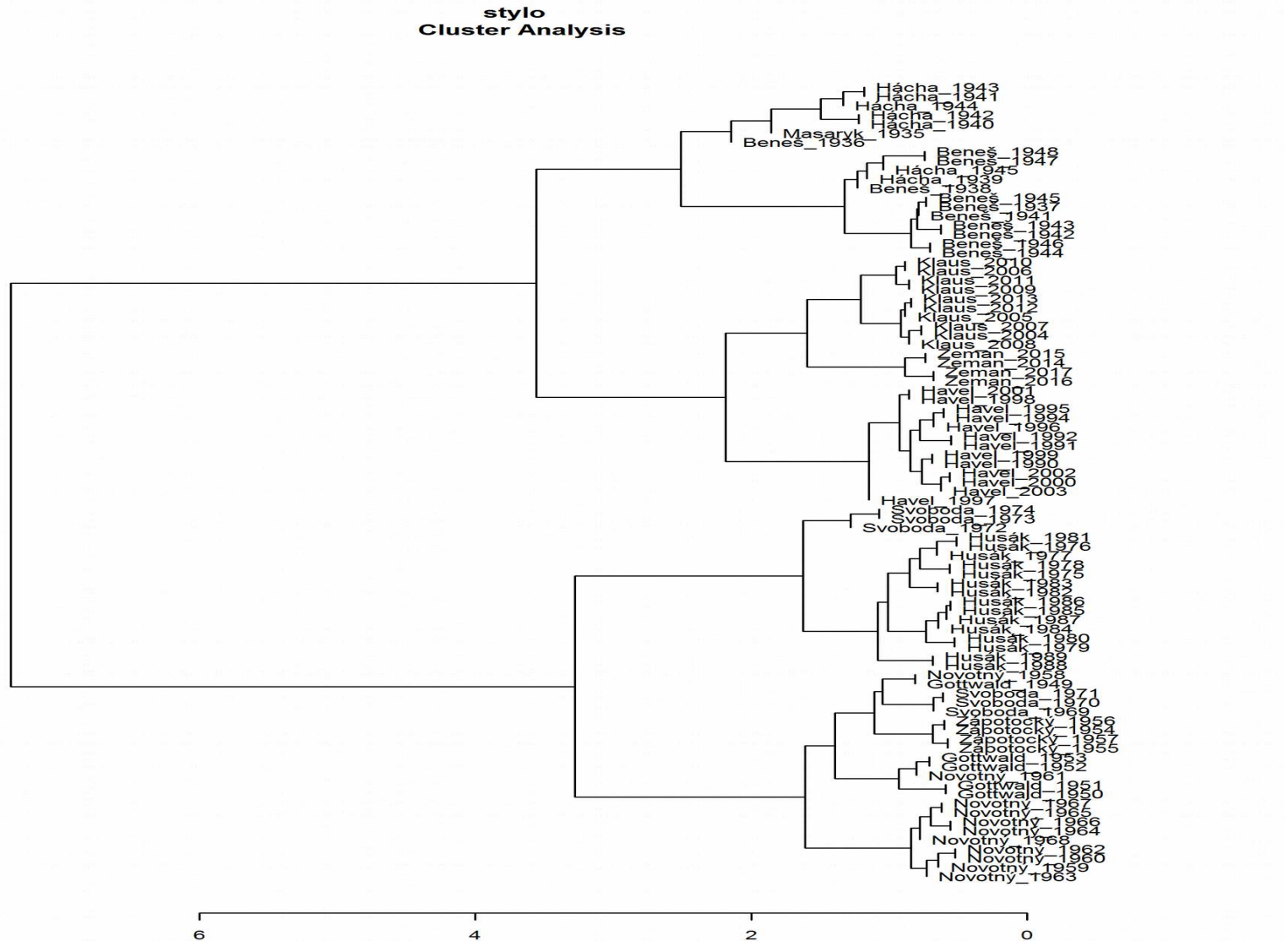
	cluster A	cluster B	cluster C	cluster D
blog	-	5/10	2/10	3/10
scientific text	10/10	-	-	-
prose	-	-	9/9	-
sermon	1/10	3/10	1/10	5/10
drama	-	-	1/10	9/10
sport reportage	-	8/10	-	2/10
cluster characteristic	scientific	“low style”, popular	“high style”, artistic	spoken word

Vianočné a novoročné príhovory prezidentov alebo ako spoznám prezidenta - komunistu

Vianočné a novoročné príhovory prezidentov alebo ako spoznám prezidenta - komunistu

- všetky vianočné a novoročné prejavy československých a českých prezidentov od r. 1935 do r. 2016
- prejavy v češtine (aj Husák), jazyk bol ponechaný tak, ako bol použitý - texty neboli prepisované do dnešnej češtiny
- z prejavov bol vytvorený jeden korpus
- zobrali sme 100 najfrekventovanejších slov s podmienkou, že sa musia vyskytovať aspoň v 60% textov
- klastrová analýza založená na vzdialenostiach medzi textami, ktoré závisia od frekvencií týchto slov

Vianočné a novoročné príhovory prezidentov alebo ako spoznám prezidenta - komunistu



Zipf-Mandelbrotovo rozdelenie

$$P_x = \frac{c}{(x+a)^b}, \; x = 1, 2, \ldots$$

- bežne sa používa ako model ore usporiadané frekvencie slov
- typické hodnoty parametrov a a b sú menšie ako 10
- pri pokuse modelovať týmto rozdelením frekvencie písmen hodnoty parametrov "vybuchli" (hodnoty rádovo aspoň v tisíckach), navyše veľmi nestabilné
- zároveň sú pomery a/b takmer konštantné
- podobné správanie bolo ohlásené pri modelovaní biodiverzity v botanike v roku 1991
- vysvetlenie - ak $a, b \to \infty$ a $a/b \to const$, tak Zipf-Mandelbrot konverguje ku geometrickému rozdeleniu

Varovanie

Kolegovia z Olomouca (Eva, Karel, Ondra, Kamila) môžu odísť, už to počuli.

Aký dobrý je χ^2 test dobrej zhody?

testovacia štatistika

$$\chi^2 = \sum_{j=1}^{k} \frac{(f_i - NP_i)^2}{NP_i},$$

kde

- N je rozsah náhodného výberu
- k je počet tried
- f_i je početnosť v i-tej triede
- P_i je pravdepodobnosť, že náhodne vybratý objekt je z i-tej triedy

rozdelenie testovacej štatistiky

- ak je rozdelenie z nulovej hypotézy plne špecifikované (teda ak poznáme všetky jeho parametre), testovacia štatistika má asymptoticky χ^2-rozdelenie s $k-1$ stupňami voľnosti
- ak musíme niektoré parametre odhadovať, testovacia štatistika má asymptoticky χ^2-rozdelenie s $k-p-1$ stupňami voľnosti, kde p je počet odhadovaných parametrov
- keďže poznáme len asymptotické rozdelenie, nechceme byť od neho priďaleko - v učebniciach sa uvádza podmienka $NP_i \geq 5 \ \forall i$
- ak táto podmienka nie je splnená, odporúča sa zlučovať triedy

ako sa zlučujú triedy v USA

- google...chi square test
- vybral som prvých 5 'rozumných' popisov χ^2 testu dobrej zhody z amerických univerzitných stránok

USA - krajina neobmedzených možností

USA - krajina neobmedzených možností

- University of California (campus Santa Cruz)
  minimálna hodnota NP_i má byť 4 alebo 5

USA - krajina neobmedzených možností

- University of California (campus Santa Cruz)
  minimálna hodnota NP_i má byť 4 alebo 5
- North Dakota State University
  nič

USA - krajina neobmedzených možností

- University of California (campus Santa Cruz)
  minimálna hodnota NP_i má byť 4 alebo 5
- North Dakota State University
  nič (botanici)

USA - krajina neobmedzených možností

- University of California (campus Santa Cruz)
  minimálna hodnota NP_i má byť 4 alebo 5
- North Dakota State University
  nič (botanici)
- Utah State University
  nič (psychológovia)

USA - krajina neobmedzených možností

- University of California (campus Santa Cruz)
  minimálna hodnota NP_i má byť 4 alebo 5
- North Dakota State University
  nič (botanici)
- Utah State University
  nič (psychológovia)
- University of California (campus Irvine)
  80% tried NP_i aspoň 5, vo všetkých ostatných aspoň 1

USA - krajina neobmedzených možností

- University of California (campus Santa Cruz)
  minimálna hodnota NP_i má byť 4 alebo 5
- North Dakota State University
  nič (botanici)
- Utah State University
  nič (psychológovia)
- University of California (campus Irvine)
  80% tried NP_i aspoň 5, vo všetkých ostatných aspoň 1
- University of Texas
  ...all would accept a minimum of 10,
  many would accept a minimum of 5,
  some would in certain circumstances accept a minimum of 1

ako zlučovať triedy - zhrnutie č. 1

Dámy a páni, robte si, čo chcete.

ako zlučovať triedy - zhrnutie č. 1 a varovanie

Dámy a páni, robte si, čo chcete. Ale...

ako zlučovať triedy - príklad

Aplikujme ‘texaské kritériá’ na umelo vytvorené dáta:

1	15
2	3
3	3
4	2
5	2
6	1
7	1
...	1
18	1
19	4
20	1
...	1
30	1

ako zlučovať triedy - príklad

- H_0: dáta pochádzajú z useknutého zeta rozdelenia $P_x = cx^{-a}$, $x = 1, 2, \ldots, R$
- a je parameter, c je normovacia konštanta, R je bod useknutia (u nás $R = 30$)

$NP_i \geq$	a	χ^2	df	p-hodnota
5	0.8707	10.35	4	0.0349
1	0.7946	14.16	21	0.8624

ako zlučovať triedy - zhrnutie č. 2

Dámy a páni, robte si, čo chcete, ale nezlučujte triedy.
Radšej používajte simulované p-hodnoty.

problém - výbery (veľmi) veľkých rozsahov

- ak je rozsah súboru veľmi veľký, nulovú hypotézu skoro iste zamietneme
- 'goodness-of-fit tests are often more a reflection on the size of the sample than on the adequacy of the model' (Browne, M.W., Cudeck, R., 1993, Alternative ways of assessing model fit, in Bollen, K.A., Long, J.S., eds., Testing structural equation models, Newbury Park, SAGE)

problém - výbery (veľmi) veľkých rozsahov

- aj modely, ktoré ‘prežijú’ testovanie, sú idealizáciou, ktorá niektoré aspekty reality nevyhnutne zanedbáva
- ‘essentially, all models are wrong, but some are useful; however, the approximate nature of the model must always be borne in mind’ (Box, G.E.P., Draper, N.R., 1987, Empirical Model Building and Response Surfaces, New York, Wiley)
- ‘remember that all models are wrong; the practical question is how wrong do they have to be to not be useful’ (ibid.)
- fixovaná p-hodnota je pre veľmi veľké výbery prakticky nepoužiteľná

problém - výbery (veľmi) veľkých rozsahov

- alternatívne prístupy upúšťajú od klasického testovania
- napr. v matematickom modelovaní v lingvistike sa často používa $C = \chi^2/N$
- $C \leq 0.02$ sa pokladá za veľmi dobrý fit

záver

- je matematicky korektný, poznáme asymptotické rozdelenie testovacej štatistiky
- 'najprirodzenejší' test pre diskrétne rozdelenia
- problém so zlučovaním tried (riešenie - simulované p-hodnoty)
- nepoužiteľný pre veľmi veľké výbery

D'akujem za pozornosť