Modellierung von Wortlängen durch die Singh-Poisson Verteilung

Gordana Đuraš[1] und Ernst Stadlober[2]

[1]Zentrum für Wirtschafts- und Innovationsforschung, Joanneum Research

[2]Institut für Statistik, Technische Universität Graz

22. Oktober 2010

Wort und Worthäufigkeiten

Text enthält n Wörter: $w_1, w_2, \ldots, w_n$

Wortlänge $l(w_j)$ ist Anzahl der Silben pro Wort, $j = 1, \ldots, n$

Addition aller $l(w_j) = i$ führt zu Häufigkeiten f_i, $i = 1, \ldots, k$

f_i ... absolute Häufigkeiten

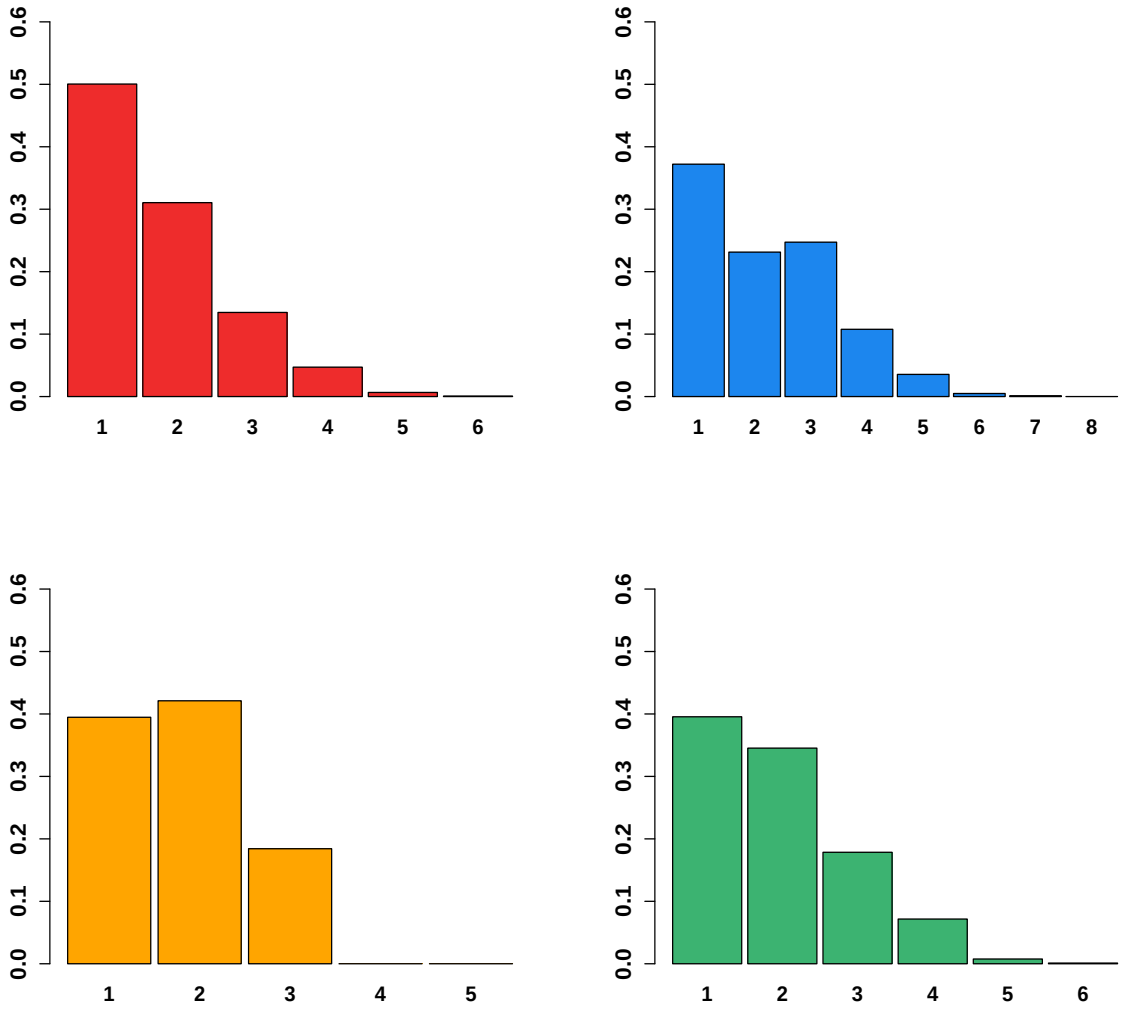
$p_i = f_i/n$ ... relative Häufigkeiten mit $\sum_{i=1}^{k} p_i = 1$

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

Zufallsvariable X = **Anzahl der Silben pro Wort**

Wertebereich $I = \{1, \ldots, k\}$

Einige Wortlängenverteilungen



Ziel: Finde ein "gutes Modell " für Häufigkeitsverteilung

Datenbasis

120 slowenische Texte (je 30 aus 4 Texttypen)

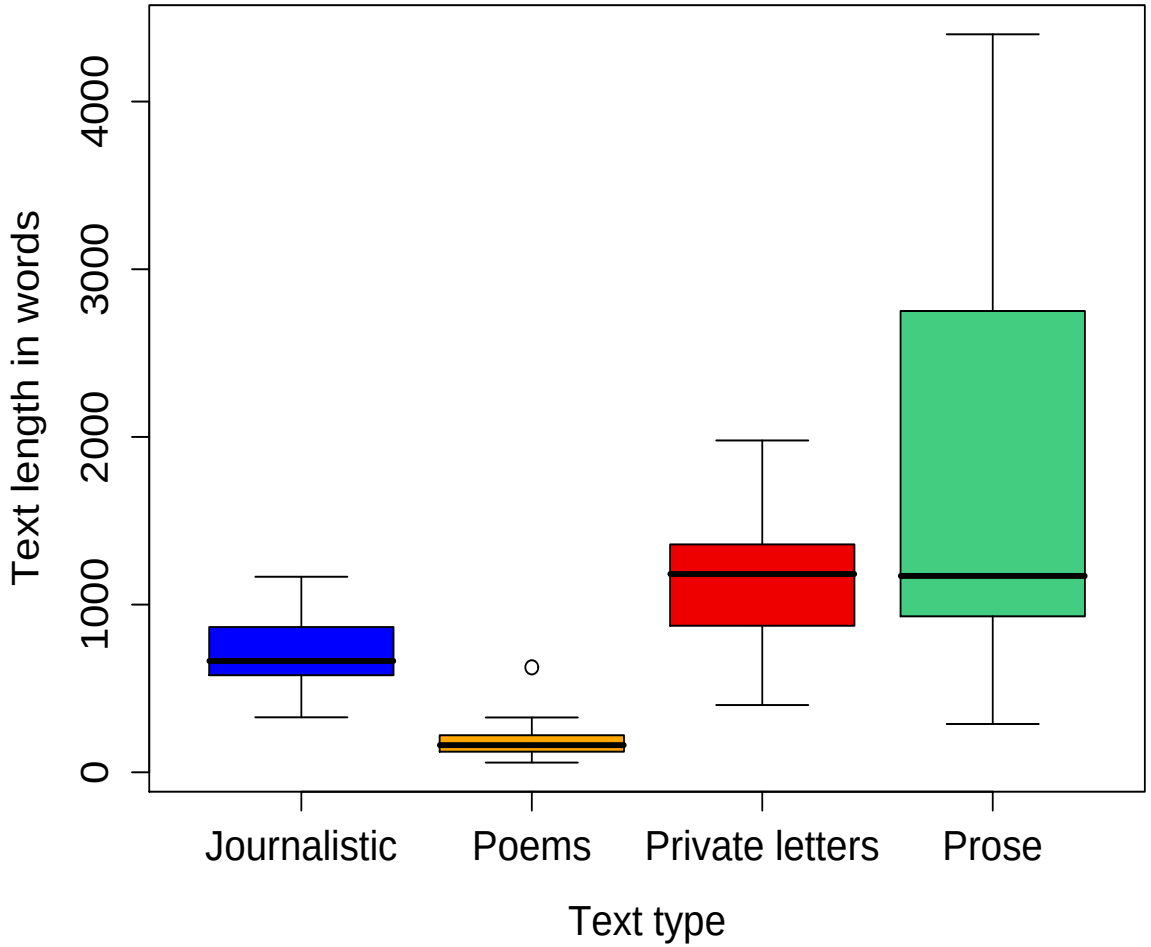
Texttyp	**# Texte**	$\bar{x}$		s^2		TL	
		min	max	min	max	min	max
Journalistisch	30	2.05	2.46	1.22	1.96	328	1166
Gedicht	30	1.48	1.90	0.37	0.84	58	626
Privatbrief	30	1.72	1.98	0.78	0.98	401	1979
Prosa	30	1.73	1.98	0.70	1.04	288	4401

Notation: $\bar{x}$ ... Empirischer Mittelwert

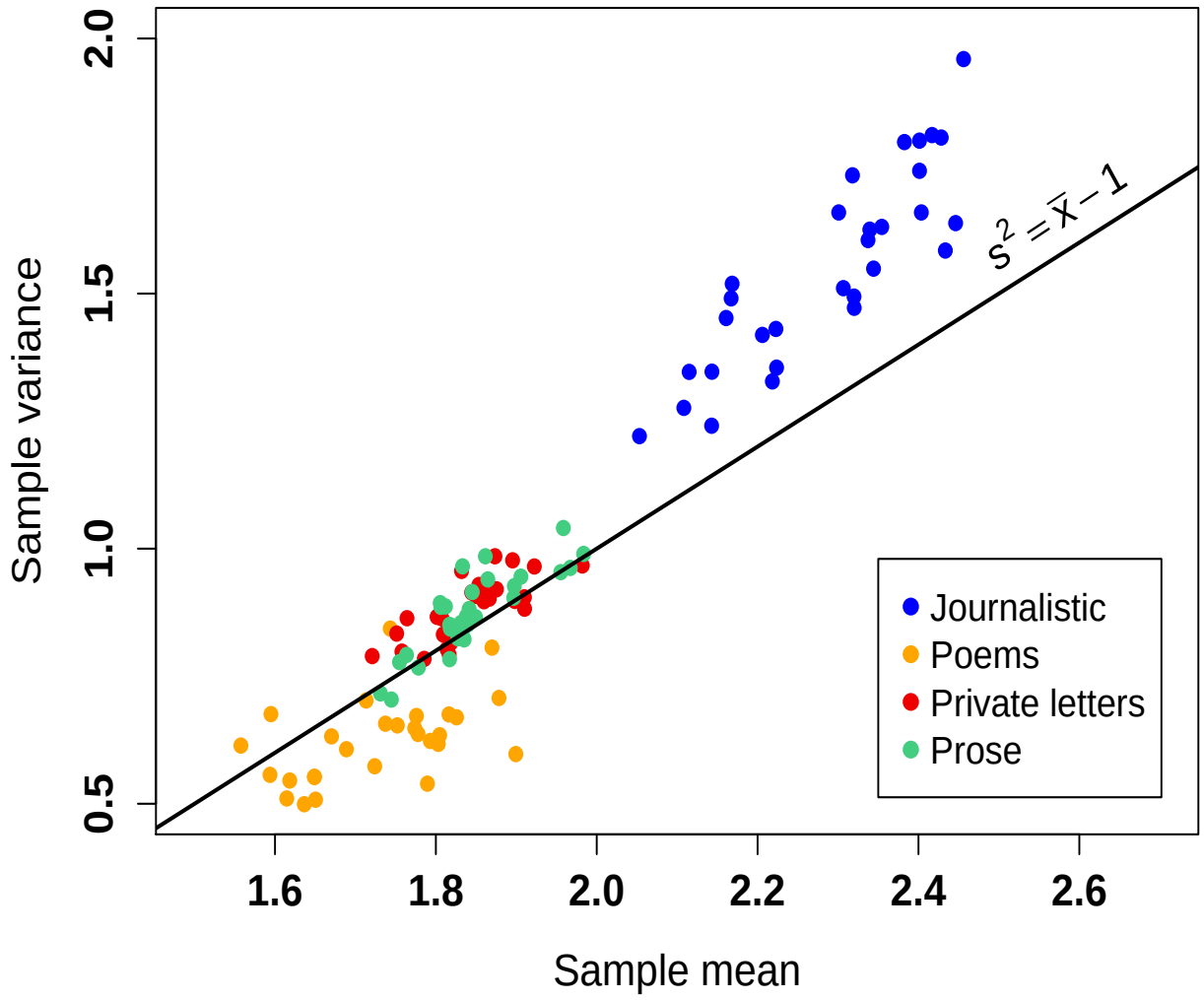
s^2 ... Empirische Varianz

TL ... Textlänge

Textlängen der 4 Texttypen



Mittelwert versus Varianz



Dispersionsindex

120 Texte: **Dispersionsindex** $\boldsymbol{d = \frac{s^2}{\bar{x} - 1}}$

Texttyp	d			Anzahl Texte		
	min	max	Mittel	$d < 1$	$d \approx 1$	$d > 1$
Journalistisch	1.086	1.346	1.202	0	0	30
Gedicht	0.605	1.136	0.854	27	0	3
Privatbrief	0.969	1.150	1.048	8	6	16
Prosa	0.946	1.159	1.036	8	13	9

Bekannte Modelle:

für $d > 1$ Negativ binomial(k, p)

für $d = 1$ Poisson(θ)

für $d < 1$ Dacey-Poisson(α, θ)

Ein Modell für gesamten Bereich von d?

Singh-Poisson Modell

Wahrscheinlichkeitsfunktion

$$P(X = x) = \begin{cases} 1 - \alpha + \alpha e^{-\theta}, & x = 1 \\ \alpha\theta^{x-1}e^{-\theta}/(x-1)!, & x = 2, 3, \ldots \end{cases}$$

Parameterbereich $\theta > 0$ und $0 < \alpha \leq \alpha_{\text{max}} = 1/(1 - e^{-\theta})$

Momente

$$\mathrm{E}(X) = 1 + \alpha\theta$$

$$\mathrm{var}(X) = \alpha\theta(1 + \theta - \alpha\theta)$$

$\alpha = 1 \quad \Longrightarrow \quad$ 1–verschobene Poisson(θ)

Rolle von Parameter α

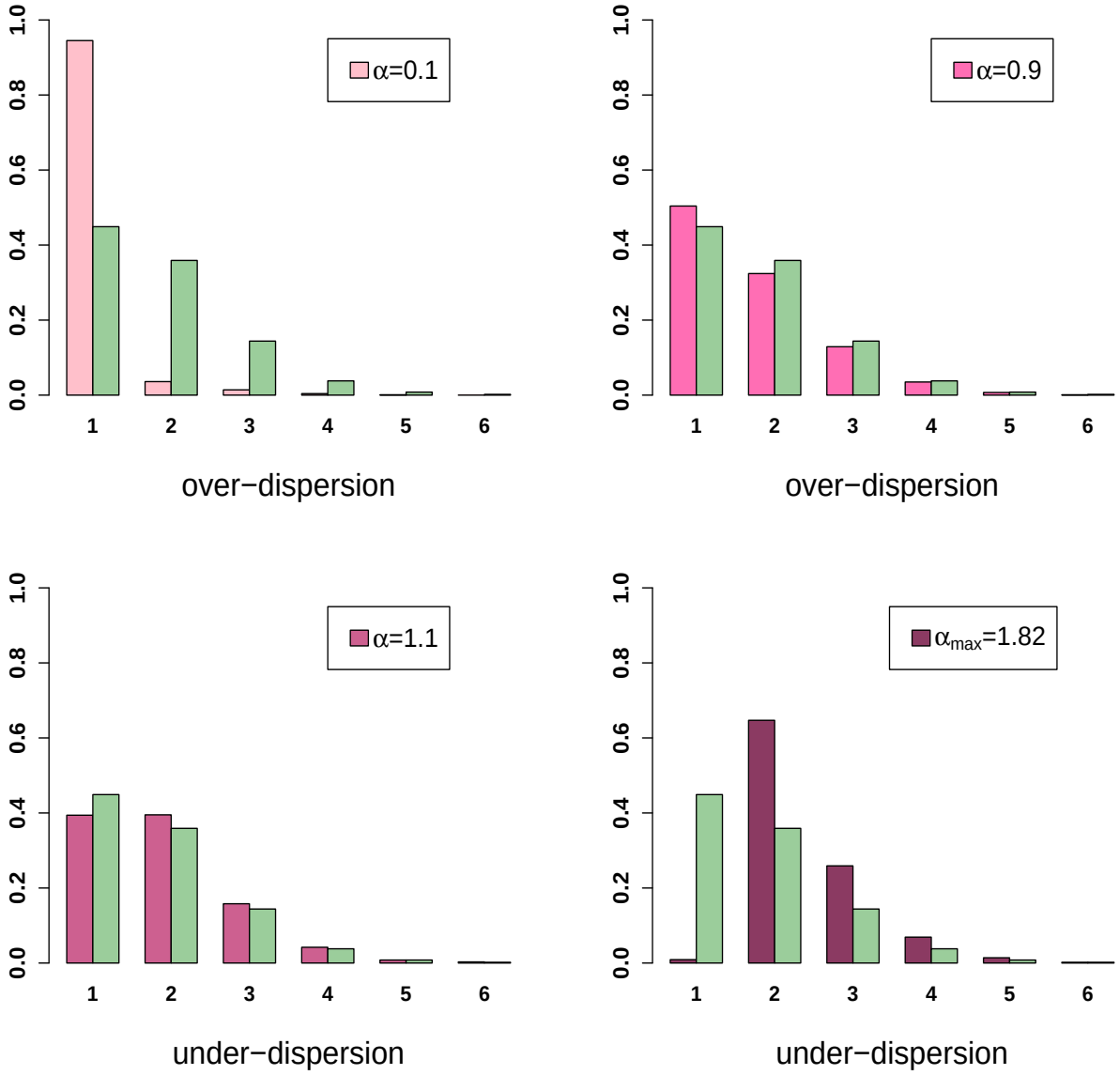
Dispersionsindex

$$\delta = \frac{\mathrm{var}(X)}{\mathrm{E}(X) - 1} = \frac{\alpha\theta(\theta + 1 - \alpha\theta)}{\alpha\theta} = 1 + \theta(1 - \alpha)$$

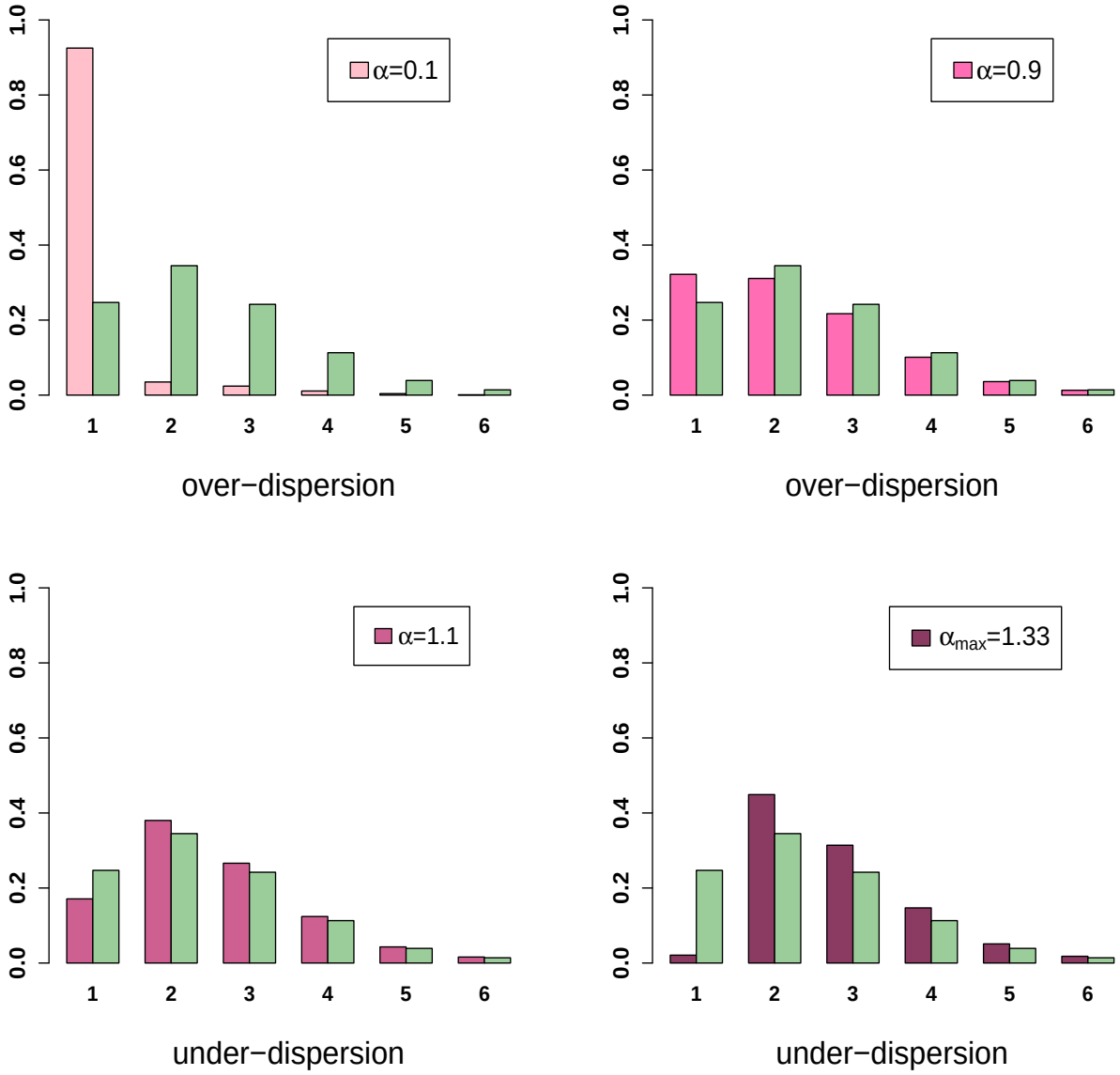
- $0 < \alpha < 1 \quad \Longrightarrow \quad \delta > 1 \quad \Rightarrow$ Über-Dispersion
- $\alpha = 1 \quad \Longrightarrow \quad \delta = 1 \quad \Rightarrow$ Gleiche Dispersion (Poisson(θ))
- $1 < \alpha \leq \alpha_{\text{max}} \quad \Longrightarrow \quad \delta < 1 \quad \Rightarrow$ Unter-Dispersion

$\Longrightarrow$ Parameter α steuert **Verteilungstyp**

Poisson(0.8) vs. Singh-Poisson(α,0.8)



Poisson(1.4) vs. Singh-Poisson(α,1.4)



Modellanpassung an Daten

Schätzung der Parameter α und θ ?

Schätzmethoden

- Momentenmethode
- Maximum Likelihood Methode
- Schätzung durch empirischen Mittelwert und erste Häufigkeitsklasse

  Brauchbar, falls:

  (1) in der Stichprobe $f_1 \gg f_2, f_3, \ldots, f_k$

  (2) Graph der empirischen Verteilung annähernd L-Form

Schätzung durch Momentenmethode

Relation zwischen Parametern und Momenten

$$\mu = 1 + \alpha\,\theta\,, \qquad \mu_{(2)} = 2\alpha\,\theta + \alpha\,\theta^2$$

Ersetze Parameter durch ihre empirischen Gegenstücke

$$\bar{x} = 1 + \hat{\alpha}\,\hat{\theta}\,, \qquad m_{(2)} = 2\hat{\alpha}\,\hat{\theta} + \hat{\alpha}\,\hat{\theta}^2$$

wobei

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} i f_i \qquad \ldots \text{ empirischer Mittelwert}$$

$$m_{(2)} = \frac{1}{n}\sum_{i=1}^{k} i(i-1) f_i \qquad \ldots \text{ zweites faktorielle Moment}$$

Schätzungen sind

$$\hat{\theta}_{\mathsf{MM}} = \frac{m_{(2)}}{\bar{x}-1} - 2 \qquad \text{and} \qquad \hat{\alpha}_{\mathsf{MM}} = \frac{\bar{x}-1}{\hat{\theta}_{\mathsf{MM}}}$$

Maximum Likelihood Schätzung

Likelihoodfunktion

$$L(\boldsymbol{\Theta}|f_1,\ldots,f_k) = \prod_{i=1}^{k} [P(X=i)]^{f_i}, \quad \text{where } \boldsymbol{\Theta} = (\theta_1,\ldots,\theta_m)$$

Log-likelihood

$$\log L(\boldsymbol{\Theta}|f_1,\ldots,f_k) = f_1 \log P(X=1) + \sum_{i=2}^{k} f_i \log P(X=i)$$

Unser Fall: $\boldsymbol{\Theta} = (\alpha,\theta)$

$$\log L(\alpha,\theta|f_i) = f_1 \log(1-\alpha+\alpha e^{-\theta}) + \sum_{i=2}^{k} f_i[\log\alpha + (i-1)\log\theta - \theta - \log(i-1)!]$$

Maximum Likelihood Schätzung

Scoregleichungen

$$\frac{\partial \log L(\alpha,\theta|f_i)}{\partial \alpha} = \frac{(e^{-\theta}-1)f_1}{1-\alpha+\alpha e^{-\theta}} + \frac{1}{\alpha}\sum_{i=2}^{k} f_i = 0$$

$$\frac{\partial \log L(\alpha,\theta|f_i)}{\partial \theta} = \frac{-\alpha e^{-\theta} f_1}{1-\alpha+\alpha e^{-\theta}} + \frac{1}{\theta}\sum_{i=2}^{k} f_i(i-1-\theta) = 0$$

Schätzer $\hat{\theta}_{\mathsf{ML}}$ Lösung der transzendenten Gleichung

$$\frac{\hat{\theta}(n-f_1)}{n(\bar{x}-1)} + e^{-\hat{\theta}} - 1 = 0$$

und Schätzer $\hat{\alpha}_{\mathsf{ML}}$ ist

$$\hat{\alpha}_{\mathsf{ML}} = \frac{n-f_1}{n(1-e^{-\hat{\theta}_{\mathsf{ML}}})}$$

Schätzung durch $\bar{x}$ und erste Häuf.-klasse

Gleichsetzen

$$\bar{x} = 1 + \hat{\alpha}\hat{\theta} \qquad \Longrightarrow \qquad \hat{\alpha}_{\mathsf{FF}} = \frac{\bar{x} - 1}{\hat{\theta}_{\mathsf{FF}}}$$

$$\frac{f_1}{n} = 1 - \hat{\alpha} + \hat{\alpha} e^{-\hat{\theta}} \qquad \overset{\hat{\theta}_{\mathsf{FF}}}{\Longrightarrow} \qquad \frac{\hat{\theta}(n - f_1)}{n(\bar{x} - 1)} = \mathbf{1} - \mathbf{e}^{-\hat{\theta}} \qquad \Longrightarrow \qquad \hat{\theta}_{\mathsf{ML}} = \hat{\theta}_{\mathsf{FF}}$$

Beziehung zwischen $\hat{\alpha}_{\mathsf{ML}}$ und $\hat{\alpha}_{\mathsf{FF}}$?

$$\hat{\alpha}_{\mathsf{ML}} = \frac{n - f_1}{n(\mathbf{1} - \mathbf{e}^{-\hat{\theta}_{\mathsf{ML}}})} = \frac{(\mathbf{n} - \mathbf{f_1})}{\mathbf{n}} \frac{\mathbf{n}(\bar{x} - 1)}{\hat{\theta}_{\mathsf{ML}}(\mathbf{n} - \mathbf{f_1})} = \frac{\bar{x} - 1}{\hat{\theta}_{\mathsf{ML}}} = \hat{\alpha}_{\mathsf{FF}}$$

$$\Longrightarrow (\hat{\alpha}_{\mathsf{ML}}, \hat{\theta}_{\mathsf{ML}}) \equiv (\hat{\alpha}_{\mathsf{FF}}, \hat{\theta}_{\mathsf{FF}})$$

Simulationsstudie

Eigenschaften des Singh-Poisson Modells für

- □ Über-Dispersion
- □ Gleiche Dispersion
- □ Unter-Dispersion

$B = 500$ Stichproben vom Umfang $n = 500$, $n = 1000$

Implementation in R mit numerischer Optimierung

Simulationsresultate

$n = 500$	α				θ			
	$\bar{\alpha}_{\text{MM}}$	se($\bar{\alpha}_{\text{MM}}$)	$\bar{\alpha}_{\text{ML}}$	se($\bar{\alpha}_{\text{ML}}$)	$\bar{\theta}_{\text{MM}}$	se($\bar{\theta}_{\text{MM}}$)	$\bar{\theta}_{\text{ML}}$	se($\bar{\theta}_{\text{ML}}$)
$d > 1$	0.824	0.041	0.822	0.034	1.579	0.098	1.581	0.083
$d \approx 1$	0.925	0.070	0.924	0.061	0.909	0.079	0.909	0.070
$d < 1$	1.155	0.103	1.150	0.094	0.627	0.066	0.629	0.062

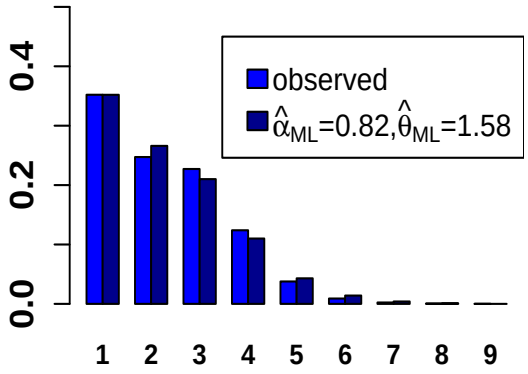
$n = 1000$	α				θ			
	$\bar{\alpha}_{\text{MM}}$	se($\bar{\alpha}_{\text{MM}}$)	$\bar{\alpha}_{\text{ML}}$	se($\bar{\alpha}_{\text{ML}}$)	$\bar{\theta}_{\text{MM}}$	se($\bar{\theta}_{\text{MM}}$)	$\bar{\theta}_{\text{ML}}$	se($\bar{\theta}_{\text{ML}}$)
$d > 1$	0.823	0.029	0.821	0.022	1.578	0.067	1.579	0.059
$d \approx 1$	0.923	0.048	0.921	0.042	0.910	0.057	0.911	0.051
$d < 1$	1.152	0.075	1.148	0.068	0.625	0.047	0.627	0.044

Theoretische Parameter (α, θ) in Simulationsstudie:
$d > 1$: $(0.82, 1.58)$, $d \approx 1$: $(0.92, 0.91)$, $d < 1$: $(1.14, 0.63)$

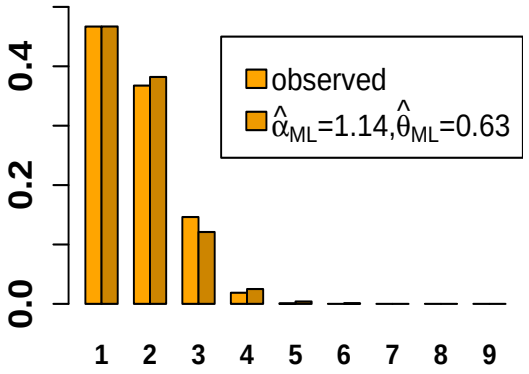
Momenten– und ML–Methode: Beide Parameter ohne Bias geschätzt

ML-Methode: kleinerer Standardfehler als Momentenmethode (80-90%)

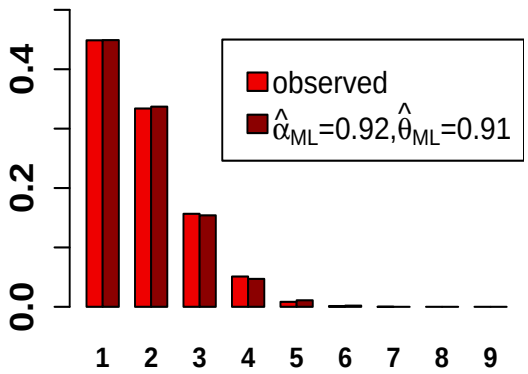
Aggregation nach Texttyp



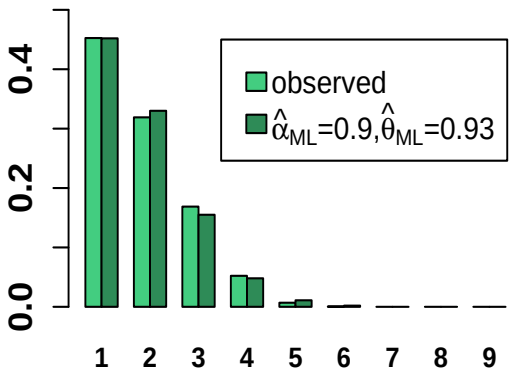
Journalistic



Poems

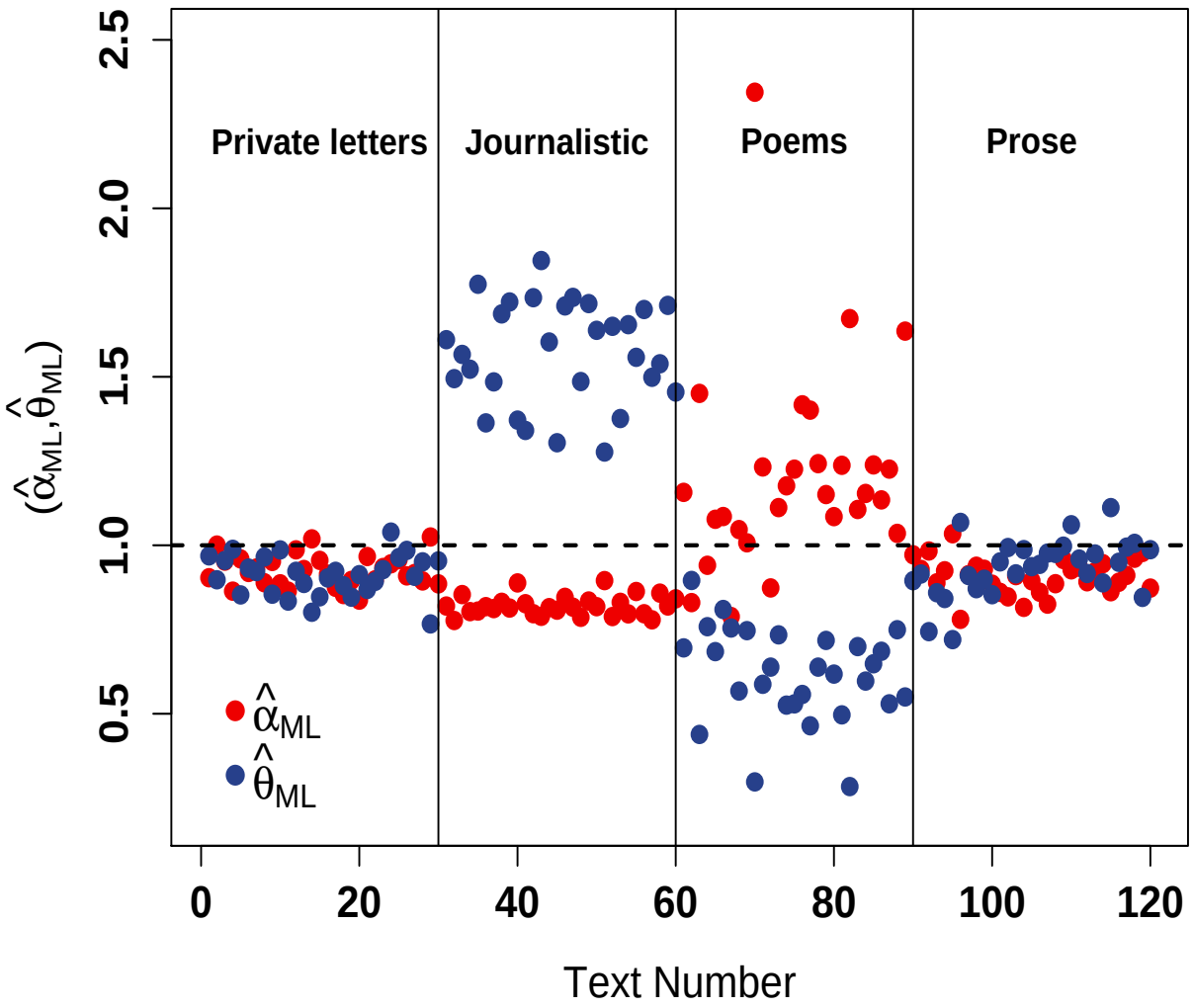


Private letters

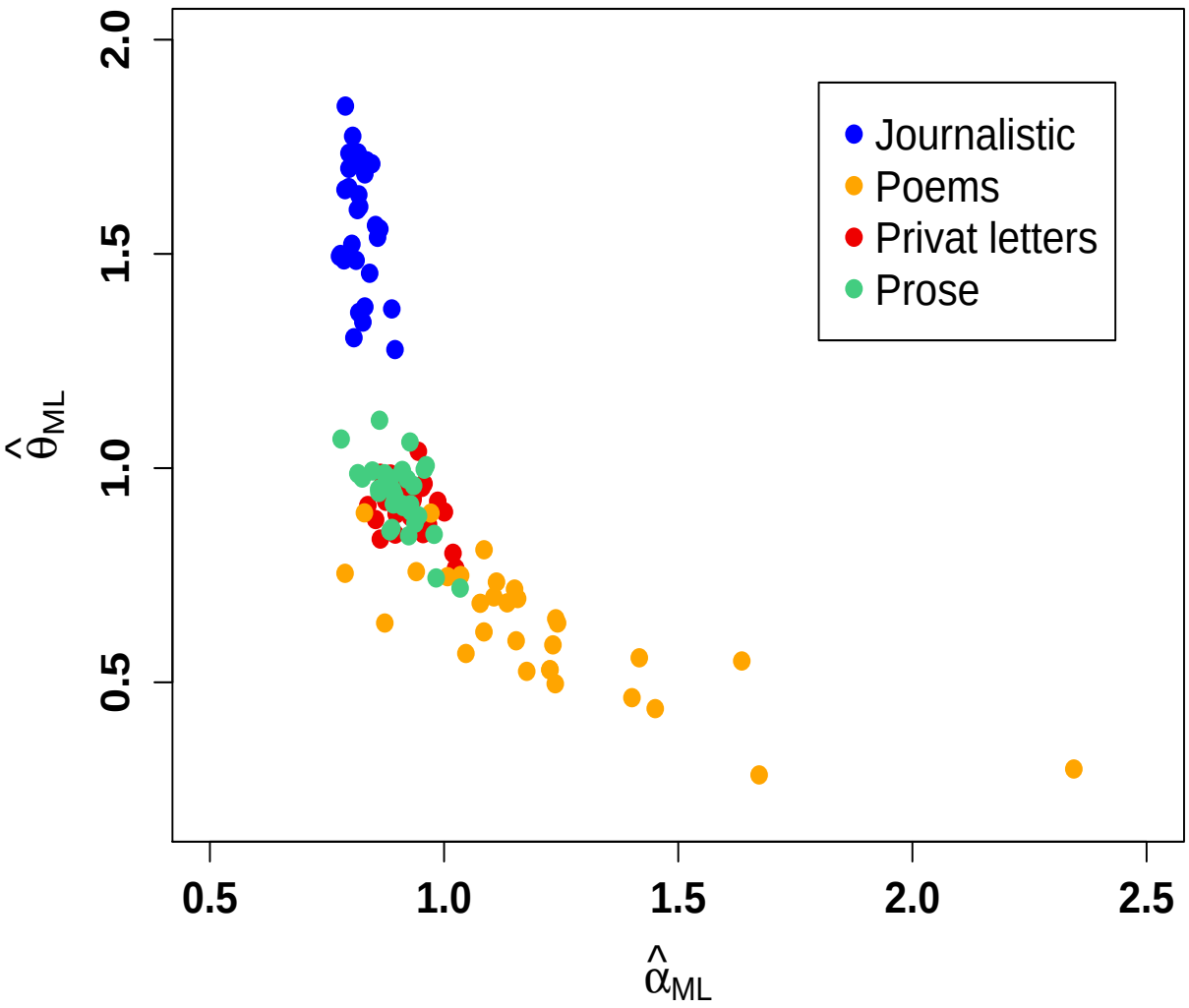


Prose

Anwendung auf slowenische Texte



Trennung von Texttypen durch (α, θ)



Güte der Anpassung

Teste Hypothese

$$H_0 : p_i = p_{0i}(\mathbf{\Theta}), \quad i = 1, 2, \ldots, k$$

mit

$$\mathbf{\Theta} = (\theta_1, \ldots, \theta_m) \qquad \ldots \quad \text{Parametervektor } (m < k-1)$$

$$p_0 = (p_{01}, p_{02}, \ldots, p_{0k}) \qquad \ldots \quad \text{Wahrsch.-vektor unter } H_0$$

Pearson's X^2 Teststatistik:

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - np_{0i}(\hat{\mathbf{\Theta}}))^2}{np_{0i}(\hat{\mathbf{\Theta}})} \sim \chi^2_{k-m-1}$$

In unserem Fall: $\mathbf{\Theta} = (\alpha, \theta)$

Güte der Anpassung

Nachteil des χ^2 Tests

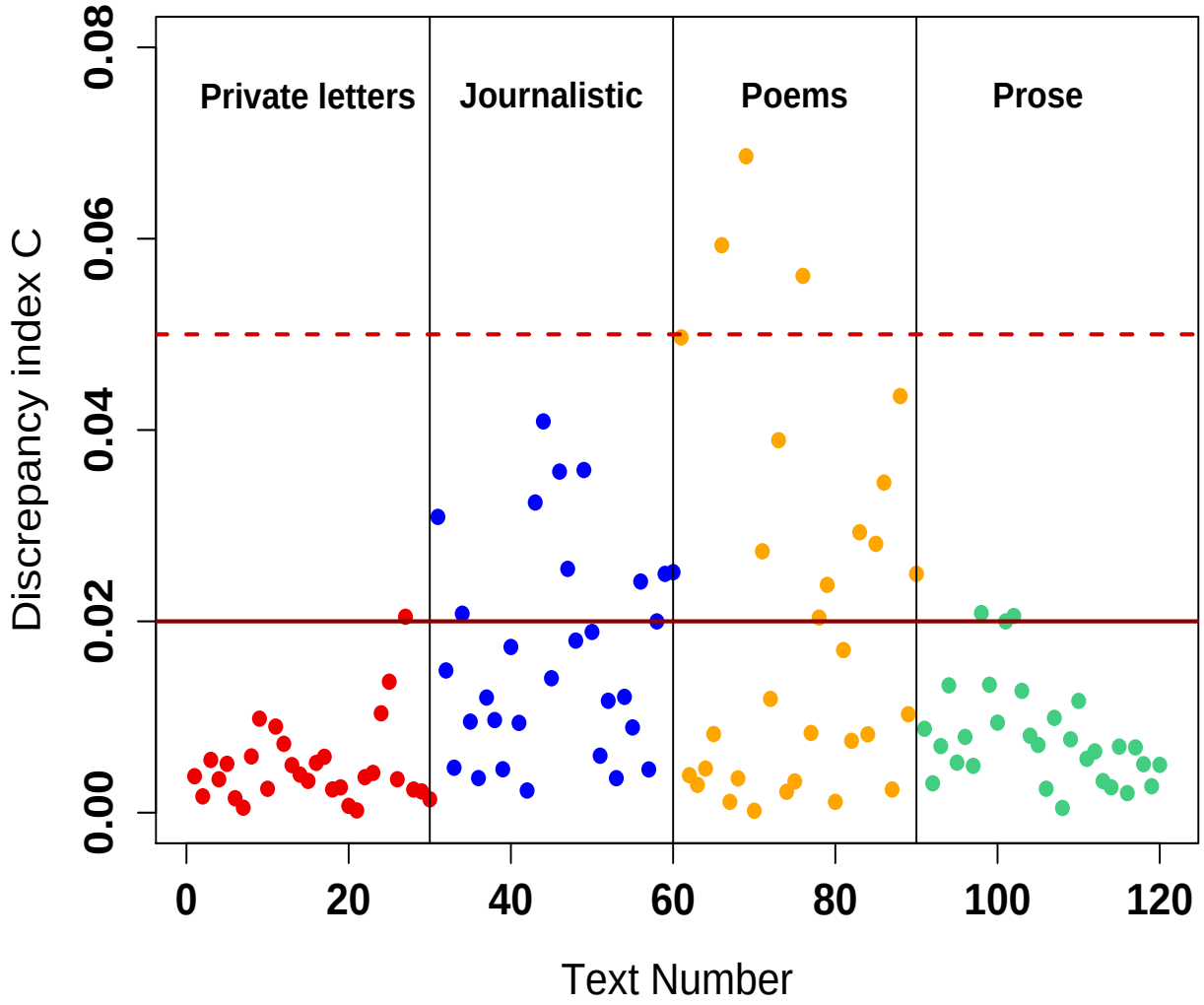
- direkter Vergleich nicht möglich wegen unterschiedlicher Stichprobengröße
- unterschiedliche Klassenzahl impliziert verschiedene Freiheitsgrade

$$\Longrightarrow \textbf{Diskrepanzindex } C = \frac{X^2}{n}$$

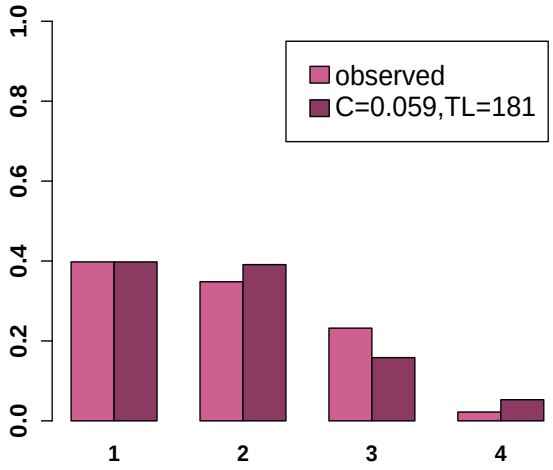
Empirische Faustregel

$C < 0.01$	...	sehr gute Anpassung
$C < 0.02$	...	gute Anpassung
$C < 0.05$	...	akzeptable Anpassung

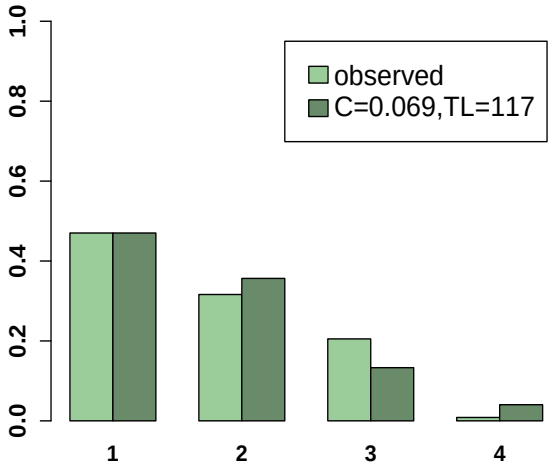
Anwendung auf slowenische Texte



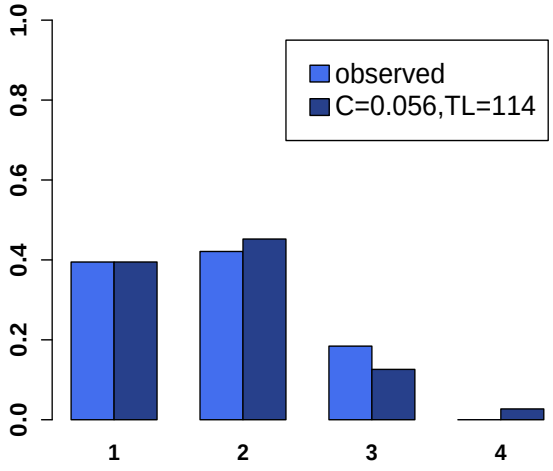
“Schlecht” angepasste Gregorčič Gedichte



Kesanje



Na sveti vecer



Pri zibelki

Zusammenfassung

- **Singh-Poisson** einfache, 2-parametrige Verallgemeinerung von Poisson
- **Ein Ansatz** für Fälle mit Unter-, gleicher und Über-Dispersion
- **Parameter** α identifiziert Poisson Über-/Unter-Dispersion
- **Schätzung** basierend auf ML $\Longleftrightarrow$ Mittelwert $+$ erste Häufigkeitsklasse
- **Gute Eigenschaften** von Parameterschätzungen für simulierte Daten
- **Brauchbare und stabile Schätzungen** für reale Daten
- **Texttypen quantifiziert** und **charakterisiert** durch Parameterbereiche von (α, θ)

Literatur

- Antić G., P. Grzybek, E. Stadlober (2006), *Mathematical aspects and modifications of Fucks´ Generalized Poisson distribution (GPD)*, in *Köhler R., G. Altmann and R.G. Piotrowski (Eds.): Quantitative Linguistics. International Handbook of Quantitative Lingustics, pp. 157–180*, Walter de Gruyter, Berlin.

- Antić G., E. Stadlober, P. Grzybek, E. Kelih (2006), *Word length and frequency distributions in different text genres*, in *Spiliopoulou, M. et al. (Eds.): From Data and Information Analysis to Knowledge Engineering, pp. 310–317*, Springer, Berlin.

- Đuraš, G., E. Stadlober (2010), *Modeling word length frequencies by the Singh-Poisson distribution*, in *Grzybek P. et al. (Eds.): Text and Language: Structures - Functions - Interrelations - Quantitative Perspectives, pp. 37–48*, Praesens Verlag, Wien.

Literatur

- Grzybek P., E. Stadlober, E. Kelih (2007), *The relationship of word length and sentence length: The inter-textual perspective*, in *Decker, R. and H.-J. Lenz (Eds.): Advances in Data Analysis, pp. 611–618*, Springer, Heidelberg.

- Grzybek P., E. Kelih, E. Stadlober (2008), *The relation between word length and sentence length: an intra–systemic perspective in the core data structure*, Glottometrics **16**, 111-121.

- Stadlober E., M. Djuzelic (2006), *Multivariate statistical methods in quantitative text analysis*, in *Grzybek P. (Ed.): Contributions to the Science of Text & Language: Word Length Studies and Related Issues. pp. 259–276*, Springer, Dordrecht.