

de wet van Zipf voor eigennamen

Gerrit Bloothoof
Uil-OTS Utrecht

de wet van Zipf

- vooral bekend door het tellen van de frequentie van woorden in een tekst
- er zijn een beperkt aantal woorden heel frequent, en een groot aantal woorden die we zelden tegenkomen
en daar zit een fraaie relatie tussen

2

op allerlei terreinen

- aantal aardbevingen met een bepaalde kracht
- aantal steden met een bepaalde grootte
- aantal hits op een website
- aantal verkochte boeken
- de frequentie van namen

3

de wet van Zipf is een machtsfunctie (power law)

$$n(f) = \text{constante} \cdot f^{-\alpha} = n(1) f^{-\alpha}$$

$n(f)$ = aantal namen met frequentie f

$$\log n(f) = \log n(1) - \alpha \log f$$

is rechte lijn op log-log vlak met helling $-\alpha$

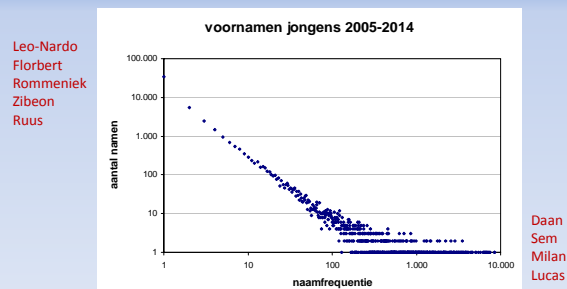
4

naamgegevens

- 20^e eeuw - 2014: gemeentelijke basisadministratie (voornamen van 23 miljoen personen)
- 2007: alle 16 miljoen achternamen
- 19^e eeuw: akten van geboorte, trouwen en overlijden (50 miljoen genoteerde voor- en achternamen)

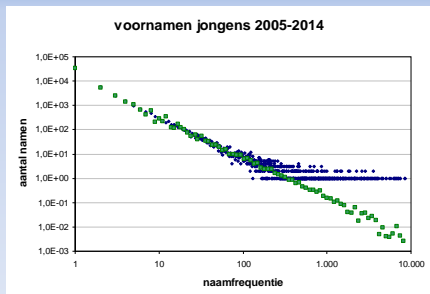
5

ruwe data



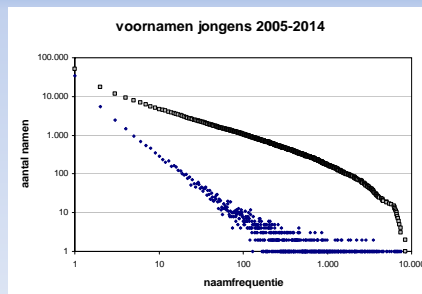
6

logaritmische bins



7

cumulatief



8

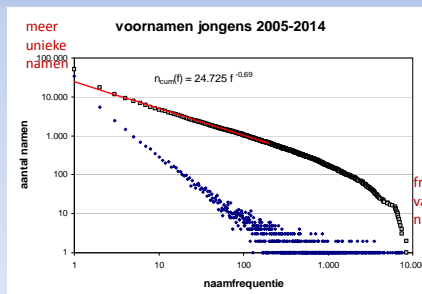
cumulatief

voor de cumulatieve verdeling geldt de machtsfunctie ook

$$n_{cum}(f) = n_{cum}(1) f^{-\alpha+1}$$

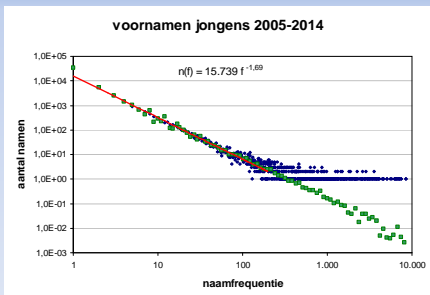
9

machtsfunctie $(\alpha = -1.69) + 1$



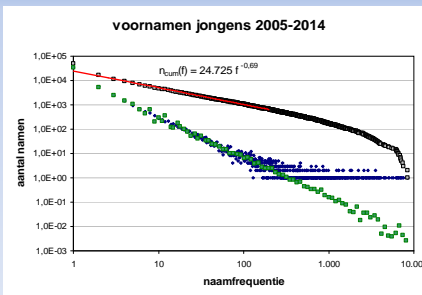
10

α gelijk (cumulatief en log bins)



11

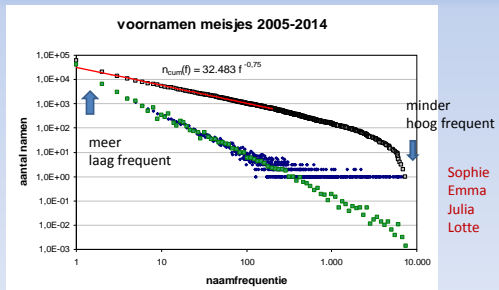
jongens $(\alpha = -1.69)$



12

meisjes ($\alpha = -1.75$)

Fobke
Lychee
Khristina
Joy-eux

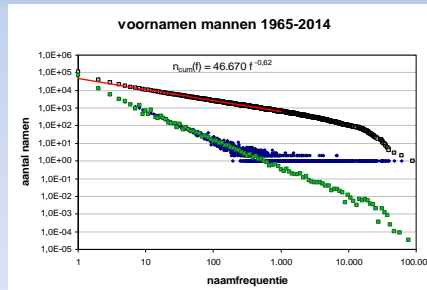


Sophie
Emma
Julia
Lotte

13

eind 20^{ste} eeuw (modenamen, $\alpha = -1.62$)

Tibbe-Jan
Meeuw
Disciple
Ritsjert

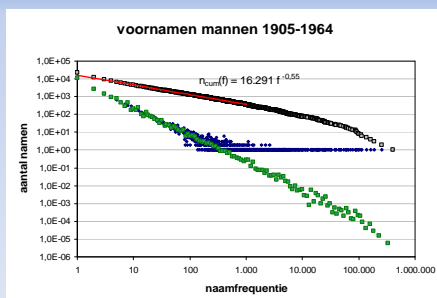


Johannes
Jan
Jeroen
Hendrik
Cornelis
Dennis

14

begin 20^{ste} eeuw (vernoemingsnamen, $\alpha = -1.55$)

Luca
Jesper
Jordi
Vince
Kelvin

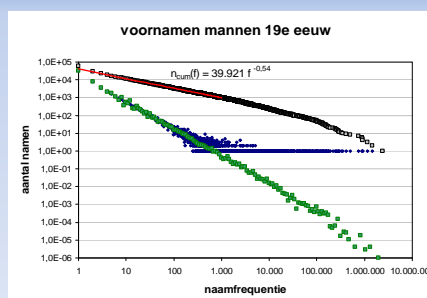


Johannes
Jan
Cornelis
Hendrik
Willem

15

19^e eeuw ($\alpha = -1.54$)

Benjemijn
Dirc
Jappie
Nopoleon
Zymon

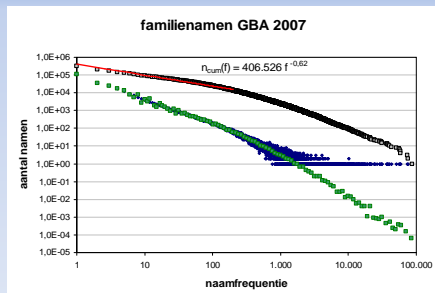


Jan
Johannes
Hendrik
Willem
Cornelis

16

achternamen ($\alpha = -1.62$)

Aaerde
Aegterberg
Ammeling
Zytveld
Zyp



de Jong
de Vries
Jansen
Janssen
van Dijk

17

verklaring?

Eerst: twee eigenschappen die helpen

(1) Zipf blijft behouden bij telling over verschillende groepen met eigen namenvoorraad (en gelijke α)

- Nederlands, Turks, Arabisch, ...
- jongens, meisjes

18

additiviteit

$$n_{\text{NL}}(f) = n_{\text{NL}}(1) f^{-\alpha}$$

$$n_{\text{Turks}}(f) = n_{\text{Turks}}(1) f^{-\alpha}$$

$$n_{\text{tot}}(f) = (n_{\text{NL}}(1) + n_{\text{Turks}}(1)) f^{-\alpha}$$

+

19

schaling

(2) Zipf blijft behouden bij telling over meerdere gelijksoortige groepen

voornamen met vergelijkbare distributie

- per aketype (19^e eeuwse geboorte-, huwelijks-, - overlijdensakte)
- per rol van persoon in akte (als boven)
- per periode van telling (per jaar, decade, 50 jaar)

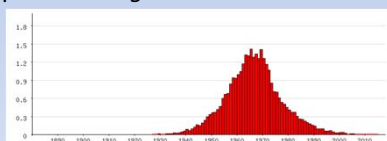
voor 2 gelijke distributies:

$$n_2(f) = n_1(f/2) = (n_1(1) 2^\alpha) f^{-\alpha}$$

20

mode voornamen >1950

populariteit *Ingrid*



- exponentiële groei (en daling) [logistische verdeling]
- gelijke snelheid van groei en daling

21

model

- jaarlijks worden 4000 nieuwe namen bedacht door ouders ($f = 1$)
- elk volgend jaar
 - fractie p van de namen houdt de aandacht en het aantal naamdragers groeit met **factor g**
 - tot er nog maar één naam over is, die dan maximaal populair is (en vervolgens ook daalt)
 - fractie $1-p$ van de namen verliest de aandacht en het aantal naamdragers neemt vervolgens jaarlijks af met **factor $1/g$**
 - tot de namen niet meer worden gegeven

22

model leidt tot machtsfunctie

$$n(f) = n_{\text{nieuw}} \cdot (1+p) f^{-\alpha}$$

$$p = g^{-\alpha}$$

bijvoorbeeld:

$$\alpha = 1.6$$

$$g = 1.33 \text{ (in 25 jaar groeien van 1 naar 1250 naamdragers)}$$

$$p = 0.82 \text{ (18\% van de nieuwe namen blijft uniek)}$$

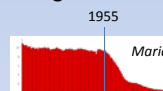
$$n(f) = 7200 f^{-1.6}$$

23

verstening

als Zipf gold, dan blijft dat zo bij 'verstening'

- voornamen: **vernoemen**
 - ouder modeverschijnsel?
- achternamen: **1811 vastgelegd** (daarna immigratie)
 - thema's volgden Zipf?
 - patroniemen (*Jansen, Janssen*)
 - beroepen (*Bakker, Visser, Smit*)
 - herkomst (*de Vries, van Dijk, van den Berg*)
 - eigenschappen (*de Jong*)



24

Zipf in combinatie

Voornamen:

Vernoeming	= Zipf (indien historisch Zipf)	
Modenamen	= Zipf	
Immigranten	= Zipf	
Bevolkingsgroei	blijft Zipf	
Combinatie	= Zipf	+

Achternamen:

1811	= Zipf (indien historisch Zipf)	
Immigranten	= Zipf	
Bevolkingsgroei	blijft Zipf	
Combinatie	= Zipf	+

25

conclusies

- Zipf geldt heel precies voor eigennamen
- weinig variatie in exponent
- additiviteit en schaling maken uitleg eenvoudiger
- model van modenamen kan Zipf verklaren
- maar dat lost niet alle problemen op

26