What Every Corpus Linguist Should Know About
Type-Token Distributions and Zipf’s Law

Tutorial Workshop #9, 22 July 2019

Stefan Evert
FAU Erlangen-Nürnberg

http://zipfr.r-forge.r-project.org/lrec2018.html
Licensed under CC-by-sa version 3.0
Outline

Introduction
- Motivation
- Notation & basic concepts
- Zipf’s law
- First steps (zipfR)

LNRE models
- Population & samples
- The mathematics of LNRE

Applications & examples
- Productivity & lexical diversity
- Practical LNRE modelling
- Bootstrapping experiments
- LNRE as Bayesian prior

Challenges
- Model inference
- Zipf’s law
- Non-randomness
- Significance testing
- Outlook
Outline

Introduction
Motivation
Notation & basic concepts
Zipf’s law
First steps (zipfR)

LNRE models
Population & samples
The mathematics of LNRE

Applications & examples
Productivity &
lexical diversity
Practical LNRE modelling
Bootstrapping experiments
LNRE as Bayesian prior

Challenges
Model inference
Zipf’s law
Non-randomness
Significance testing
Outlook
Some research questions

- How many words did Shakespeare know?
- What is the coverage of my treebank grammar on big data?
- How many typos are there on the Internet?
- Is -ness more productive than -ity in English?
- Are there differences in the productivity of nominal compounds between academic writing and novels?
- Does Dickens use a more complex vocabulary than Rowling?
- Can a decline in lexical complexity predict Alzheimer’s disease?
- How frequent is a hapax legomenon from the Brown corpus?
- What is appropriate smoothing for my n-gram model?
- Who wrote the Bixby letter, Lincoln or Hay?
- How many different species of ... are there? (Brainerd 1982)
Some research questions

- coverage estimates
- productivity
- lexical complexity & stylometry
- prior & posterior distribution
- unexpected applications
Type-token statistics

- These applications relate *token* and *type* counts
  - **tokens** = individual instances (occurrences)
  - **types** = distinct items
- Type-token statistics different from most statistical inference
  - not about probability of a specific event
  - but about diversity of events and their probability distribution
Type-token statistics

- These applications relate **token** and **type** counts
  - **tokens** = individual instances (occurrences)
  - **types** = distinct items
- Type-token statistics different from most statistical inference
  - not about probability of a specific event
  - but about diversity of events and their probability distribution
- Relatively little work in statistical science
- Nor a major research topic in computational linguistics
  - very specialized, usually plays ancillary role in NLP
- Corpus linguistics: TTR & simple productivity measures
  - often applied without any statistical inference
Zipf’s law (Zipf 1949)

A) Frequency distributions in natural language are highly skewed
Zipf’s law (Zipf 1949)

A) Frequency distributions in natural language are highly skewed

B) Curious relationship between rank & frequency

<table>
<thead>
<tr>
<th>word</th>
<th>r</th>
<th>f</th>
<th>r · f</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1.</td>
<td>142,776</td>
<td>142,776</td>
</tr>
<tr>
<td>and</td>
<td>2.</td>
<td>100,637</td>
<td>201,274</td>
</tr>
<tr>
<td>be</td>
<td>3.</td>
<td>94,181</td>
<td>282,543</td>
</tr>
<tr>
<td>of</td>
<td>4.</td>
<td>74,054</td>
<td>296,216</td>
</tr>
</tbody>
</table>

(Dickens)
Zipf’s law (Zipf 1949)

A) Frequency distributions in natural language are highly skewed

B) Curious relationship between rank & frequency

<table>
<thead>
<tr>
<th>word</th>
<th>r</th>
<th>f</th>
<th>r · f</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1.</td>
<td>142,776</td>
<td>142,776</td>
</tr>
<tr>
<td>and</td>
<td>2.</td>
<td>100,637</td>
<td>201,274</td>
</tr>
<tr>
<td>be</td>
<td>3.</td>
<td>94,181</td>
<td>282,543</td>
</tr>
<tr>
<td>of</td>
<td>4.</td>
<td>74,054</td>
<td>296,216</td>
</tr>
</tbody>
</table>

(Dickens)

C) Various explanations of Zipf’s law

- principle of least effort (Zipf 1949)
- optimal coding system, MDL (Mandelbrot 1953, 1962)
- random sequences (Miller 1957; Li 1992; Cao et al. 2017)
- Markov processes ➔ n-gram models (Rouault 1978)
Zipf’s law (Zipf 1949)

A) Frequency distributions in natural language are highly skewed

B) Curious relationship between rank & frequency

<table>
<thead>
<tr>
<th>word</th>
<th>r</th>
<th>f</th>
<th>r · f</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1.</td>
<td>142,776</td>
<td>142,776</td>
</tr>
<tr>
<td>and</td>
<td>2.</td>
<td>100,637</td>
<td>201,274</td>
</tr>
<tr>
<td>be</td>
<td>3.</td>
<td>94,181</td>
<td>282,543</td>
</tr>
<tr>
<td>of</td>
<td>4.</td>
<td>74,054</td>
<td>296,216</td>
</tr>
</tbody>
</table>

(Dickens)

C) Various explanations of Zipf’s law

- principle of least effort (Zipf 1949)
- optimal coding system, MDL (Mandelbrot 1953, 1962)
- random sequences (Miller 1957; Li 1992; Cao et al. 2017)
- Markov processes ➔ n-gram models (Rouault 1978)

D) Language evolution: birth-death-process (Simon 1955)

❉ not the main topic today!
Goals of this tutorial

▶ Introduce descriptive statistics, notation and terminology
▶ Explain mathematical foundations of LNRE models for statistical inference
▶ Practise application of models in R
▶ Discuss measures of productivity & lexical richness
▶ Address problems and advanced techniques
Outline

Introduction
  Motivation
  Notation & basic concepts
    Zipf’s law
    First steps (zipfR)
  LNRE models
    Population & samples
    The mathematics of LNRE

Applications & examples
  Productivity & lexical diversity
  Practical LNRE modelling
  Bootstrapping experiments
  LNRE as Bayesian prior

Challenges
  Model inference
  Zipf’s law
  Non-randomness
  Significance testing
  Outlook
Tokens & types

our sample: recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very

▶ $N = 15$: number of tokens = sample size
▶ $V = 7$: number of distinct types = vocabulary size
(recently, very, not, otherwise, much, merely, now)
## Tokens & types

our sample: `recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very`

- \( N = 15 \): number of tokens = sample size
- \( V = 7 \): number of distinct types = vocabulary size
  
  \( \) (`recently, very, not, otherwise, much, merely, now`)

### type-frequency list

<table>
<thead>
<tr>
<th>( w )</th>
<th>( f_w )</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>recently</code></td>
<td>1</td>
</tr>
<tr>
<td><code>very</code></td>
<td>5</td>
</tr>
<tr>
<td><code>not</code></td>
<td>3</td>
</tr>
<tr>
<td><code>otherwise</code></td>
<td>1</td>
</tr>
<tr>
<td><code>much</code></td>
<td>2</td>
</tr>
<tr>
<td><code>merely</code></td>
<td>2</td>
</tr>
<tr>
<td><code>now</code></td>
<td>1</td>
</tr>
</tbody>
</table>
Zipf ranking

our sample:  

\[ \text{recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very} \]

- \( N = 15 \): number of tokens = sample size
- \( V = 7 \): number of distinct types = vocabulary size
  \((\text{recently, very, not, otherwise, much, merely, now})\)

<table>
<thead>
<tr>
<th></th>
<th>( r )</th>
<th>( f_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{very}</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>\text{not}</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>\text{merely}</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>\text{much}</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>\text{now}</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>\text{otherwise}</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>\text{recently}</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>
Zipf ranking

our sample: *recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very*

▶ \(N = 15\): number of *tokens* = sample size
▶ \(V = 7\): number of distinct *types* = *vocabulary size*  
(\(recently, very, not, otherwise, much, merely, now\))

### Zipf ranking

<table>
<thead>
<tr>
<th>(w)</th>
<th>(r)</th>
<th>(f_r)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>very</em></td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td><em>not</em></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><em>merely</em></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td><em>much</em></td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td><em>now</em></td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td><em>otherwise</em></td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td><em>recently</em></td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>
## A realistic Zipf ranking: the Brown corpus

<table>
<thead>
<tr>
<th>rank range</th>
<th>randomly selected examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>7731 – 8271</td>
<td>schedules, polynomials, bleak</td>
</tr>
<tr>
<td>8272 – 8922</td>
<td>tolerance, shaved, hymn</td>
</tr>
<tr>
<td>8923 – 9703</td>
<td>decreased, abolish, irresistible</td>
</tr>
<tr>
<td>9704 – 10783</td>
<td>immunity, cruising, titan</td>
</tr>
<tr>
<td>10784 – 11985</td>
<td>geographic, lauro, portrayed</td>
</tr>
<tr>
<td>11986 – 13690</td>
<td>grigori, slashing, developer</td>
</tr>
<tr>
<td>13691 – 15991</td>
<td>sheath, gaulle, ellipsoids</td>
</tr>
<tr>
<td>15992 – 19627</td>
<td>mc, initials, abstracted</td>
</tr>
<tr>
<td>19628 – 26085</td>
<td>thar, slackening, deluxe</td>
</tr>
<tr>
<td>26086 – 45215</td>
<td>beck, encompasses, second-place</td>
</tr>
</tbody>
</table>

### Notation & basic concepts

#### A realistic Zipf ranking: the Brown corpus

<table>
<thead>
<tr>
<th>top frequencies</th>
<th>bottom frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>rank range</td>
</tr>
<tr>
<td>rank range</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>word</td>
</tr>
<tr>
<td>r</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>69836</td>
</tr>
<tr>
<td>2</td>
<td>36365</td>
</tr>
<tr>
<td>3</td>
<td>28826</td>
</tr>
<tr>
<td>4</td>
<td>26126</td>
</tr>
<tr>
<td>5</td>
<td>23157</td>
</tr>
<tr>
<td>6</td>
<td>21314</td>
</tr>
<tr>
<td>7</td>
<td>10777</td>
</tr>
<tr>
<td>8</td>
<td>10182</td>
</tr>
<tr>
<td>9</td>
<td>9968</td>
</tr>
<tr>
<td>10</td>
<td>9801</td>
</tr>
</tbody>
</table>
A realistic Zipf ranking: the Brown corpus
A realistic Zipf ranking: the Brown corpus
**Frequency spectrum**

- pool types with $f = 1$ (hapax legomena), types with $f = 2$ (dis legomena), …, $f = m$, …
- $V_1 = 3$: number of hapax legomena (*now, otherwise, recently*)
- $V_2 = 2$: number of dis legomena (*merely, much*)
- general definition: $V_m = |\{w \mid f_w = m\}|$

**Zipf ranking**

<table>
<thead>
<tr>
<th>$w$</th>
<th>$r$</th>
<th>$f_r$</th>
<th>frequency spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>very</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>not</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>merely</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>much</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>now</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>otherwise</td>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>recently</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

$V_m = |\{w \mid f_w = m\}|$
Frequency spectrum

- Pool types with $f = 1$ (hapax legomena), types with $f = 2$ (dis legomena), ... , $f = m$, ...
- $V_1 = 3$: number of hapax legomena (now, otherwise, recently)
- $V_2 = 2$: number of dis legomena (merely, much)
- General definition: $V_m = |\{w \mid f_w = m\}|$

Zipf ranking

<table>
<thead>
<tr>
<th>$w$</th>
<th>$r$</th>
<th>$f_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>very</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>not</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>merely</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>much</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>now</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>otherwise</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>recently</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

Frequency spectrum: adverbs

<table>
<thead>
<tr>
<th>$m$</th>
<th>$V_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Stefan Evert

T1: Zipf’s Law

22 July 2019 | CC-by-sa
A realistic frequency spectrum: the Brown corpus
Vocabulary growth curve

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

▶ $N = 1$, $V(N) = 1$, $V_1(N) = 1$
Vocabulary growth curve

our sample: *recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very*

- \( N = 1, \ V(N) = 1, \ V_1(N) = 1 \)
- \( N = 3, \ V(N) = 3, \ V_1(N) = 3 \)
Vocabulary growth curve

our sample: recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very

- $N = 1, \ V(N) = 1, \ V_1(N) = 1$
- $N = 3, \ V(N) = 3, \ V_1(N) = 3$
- $N = 7, \ V(N) = 5, \ V_1(N) = 4$
Vocabulary growth curve

our sample: recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very

- $N = 1$, $V(N) = 1$, $V_1(N) = 1$
- $N = 3$, $V(N) = 3$, $V_1(N) = 3$
- $N = 7$, $V(N) = 5$, $V_1(N) = 4$
- $N = 12$, $V(N) = 7$, $V_1(N) = 4$
Vocabulary growth curve

our sample: recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very

- $N = 1, \ V(N) = 1, \ V_1(N) = 1$
- $N = 3, \ V(N) = 3, \ V_1(N) = 3$
- $N = 7, \ V(N) = 5, \ V_1(N) = 4$
- $N = 12, \ V(N) = 7, \ V_1(N) = 4$
- $N = 15, \ V(N) = 7, \ V_1(N) = 3$
Vocabulary growth curve

our sample: *recently, very, not, otherwise, much, very, very, merely, not, now, very, much, merely, not, very*

- $N = 1$, $V(N) = 1$, $V_1(N) = 1$
- $N = 3$, $V(N) = 3$, $V_1(N) = 3$
- $N = 7$, $V(N) = 5$, $V_1(N) = 4$
- $N = 12$, $V(N) = 7$, $V_1(N) = 4$
- $N = 15$, $V(N) = 7$, $V_1(N) = 3$
A realistic vocabulary growth curve: the Brown corpus

![Vocabulary growth curve: Brown corpus](image)

Stefan Evert
Vocabulary growth in authorship attribution

- Authorship attribution by n-gram tracing applied to the case of the Bixby letter (Grieve et al. 2018)
- Word or character n-grams in disputed text are compared against large “training” corpora from candidate authors

**Gettysburg Address: Word 2-Grams**

The analysis was only run up to 4-word n-grams because from that point onward the Hay corpus contains none of the n-grams in the Gettysburg Address. The 3- and 4-word n-gram analyses also correctly attributed the Gettysburg Address to Lincoln: 18% of 3-grams for Lincoln vs. 14% of 3-grams for Hay and 2% of 4-grams for Lincoln vs. 0% of 4-grams for Hay. The 1-word n-gram analysis, however, incorrectly attributed the Gettysburg Address to Hay. Figure 3 presents the aggregated n-gram traces for all analyses. Notably, the 2-, 3- and 4-word n-gram analyses, which correctly attributed the document to Lincoln, appear to be far more definitive than the incorrect 1-word n-gram analysis.
Outline

Introduction
- Motivation
- Notation & basic concepts
- Zipf’s law
- First steps (zipfR)

LNRE models
- Population & samples
- The mathematics of LNRE

Applications & examples
- Productivity & lexical diversity
- Practical LNRE modelling
- Bootstrapping experiments
- LNRE as Bayesian prior

Challenges
- Model inference
- Zipf’s law
- Non-randomness
- Significance testing
- Outlook
Observing Zipf’s law
across languages and different linguistic units
Observing Zipf’s law

![Zipf ranking: Brown corpus](image)

**Zipf ranking: Brown corpus**

- **x-axis**: rank
- **y-axis**: frequency

The graph illustrates the frequency distribution of word ranks in the Brown corpus, showing that a small number of words are used very frequently, while the majority are used less often, following Zipf's law.
Observing Zipf’s law

Zipf ranking: Brown corpus

rank

frequency

1

10

100

1000

10000
Observing Zipf’s law

- Straight line in double-logarithmic space corresponds to **power law** for original variables
- This leads to Zipf’s (1949; 1965) famous law:

\[ f_r = \frac{C}{r^a} \]
Observing Zipf’s law

- Straight line in double-logarithmic space corresponds to **power law** for original variables
- This leads to Zipf’s (1949; 1965) famous law:
  \[ f_r = \frac{C}{r^a} \]
- If we take logarithm on both sides, we obtain:
  \[ \log f_r = \log C - a \cdot \log r \]
Observing Zipf’s law

- Straight line in double-logarithmic space corresponds to **power law** for original variables
- This leads to Zipf’s (1949; 1965) famous law:

\[ f_r = \frac{C}{r^a} \]

- If we take logarithm on both sides, we obtain:

\[ \log f_r = \log C - a \cdot \log r \]

Intuitive interpretation of \( a \) and \( \log C \):

- \( a \) is slope determining how fast log frequency decreases
- \( \log C \) is intercept, i.e. log frequency of most frequent word (\( r = 1 \Rightarrow \log r = 0 \))
Observing Zipf’s law

▶ Straight line in double-logarithmic space corresponds to **power law** for original variables

▶ This leads to Zipf’s (1949; 1965) famous law:

\[ f_r = \frac{C}{r^a} \]

▶ If we take logarithm on both sides, we obtain:

\[ \log f_r = \log C - a \cdot \log r \]

▶ Intuitive interpretation of \( a \) and \( C \):

▶ \( a \) is **slope** determining how fast log frequency decreases

▶ \( \log C \) is **intercept**, i.e. log frequency of most frequent word \((r = 1 \rightarrow \log r = 0)\)
Observing Zipf’s law

Least-squares fit = linear regression in log-space (Brown corpus)
Zipf-Mandelbrot law
Mandelbrot (1953, 1962)

- Mandelbrot’s extra parameter:

\[ f_r = \frac{C}{(r + b)^a} \]

- Zipf’s law is special case with \( b = 0 \)
Zipf-Mandelbrot law
Mandelbrot (1953, 1962)

▶ Mandelbrot’s extra parameter:

\[ f_r = \frac{C}{(r + b)^a} \]

▶ Zipf’s law is special case with \( b = 0 \)

▶ Assuming \( a = 1, \ C = 60,000, \ b = 1 \):
  ▶ For word with rank 1, Zipf’s law predicts frequency of 60,000; Mandelbrot’s variation predicts frequency of 30,000
  ▶ For word with rank 1,000, Zipf’s law predicts frequency of 60; Mandelbrot’s variation predicts frequency of 59.94
Zipf-Mandelbrot law
Mandelbrot (1953, 1962)

- Mandelbrot’s extra parameter:

\[ f_r = \frac{C}{(r + b)^a} \]

- Zipf’s law is special case with \( b = 0 \)
- Assuming \( a = 1, \ C = 60,000, \ b = 1: \)
  - For word with rank 1, Zipf’s law predicts frequency of 60,000; Mandelbrot’s variation predicts frequency of 30,000
  - For word with rank 1,000, Zipf’s law predicts frequency of 60; Mandelbrot’s variation predicts frequency of 59.94

- Zipf-Mandelbrot law forms basis of statistical LNRE models
  - ZM law derived mathematically as limiting distribution of vocabulary generated by a character-level Markov process
Zipf-Mandelbrot law

Non-linear least-squares fit (Brown corpus)
Outline

Introduction
Motivation
Notation & basic concepts
Zipf’s law
First steps (zipfR)

Applications & examples
Productivity & lexical diversity
Practical LNRE modelling
Bootstrapping experiments
LNRE as Bayesian prior

Challenges
Model inference
Zipf’s law
Non-randomness
Significance testing
Outlook
zipfR
Evert and Baroni (2007)

- [ ] http://zipfR.R-Forge.R-Project.org/
- [ ] Conveniently available from CRAN repository
- [ ] Package vignette = gentle tutorial introduction
First steps with zipfR

- Set up a folder for this course, and make sure it is your working directory in R (preferably as an RStudio project)
- Install the most recent version of the zipfR package
  - tutorial requires version 0.7 or newer
- Package, handouts, code samples & data sets available from http://zipfr.r-forge.r-project.org/lrec2018.html

```r
> library(zipfR)

> ?zipfR  # documentation entry point

> vignette("zipfr-tutorial")  # read the zipfR tutorial
```
Loading type-token data

▶ Most convenient input: sequence of tokens as text file in vertical format (“one token per line”)
  ⚠️ mapped to appropriate types: normalized word forms, word pairs, lemmatized, semantic class, n-gram of POS tags, ...  
  ⚠️ language data should always be in UTF-8 encoding!
  ⚠️ large files can be compressed (.gz, .bz2, .xz)
Loading type-token data

Most convenient input: sequence of tokens as text file in vertical format (“one token per line”)
- mapped to appropriate types: normalized word forms, word pairs, lemmatized, semantic class, n-gram of POS tags, ...
- language data should always be in UTF-8 encoding!
- large files can be compressed (.gz, .bz2, .xz)

Sample data: brown_adverbs.txt on tutorial homepage
- lowercased adverb tokens from Brown corpus (original order)
- download and save to your working directory
Loading type-token data

- Most convenient input: sequence of tokens as text file in vertical format ("one token per line")
  - mapped to appropriate types: normalized word forms, word pairs, lemmatized, semantic class, n-gram of POS tags, ...
  - language data should always be in UTF-8 encoding!
  - large files can be compressed (.gz, .bz2, .xz)

- Sample data: brown_adverbs.txt on tutorial homepage
  - lowercased adverb tokens from Brown corpus (original order)
  - download and save to your working directory

```r
> adv <- readLines("brown_adverbs.txt", encoding="UTF-8")
> head(adv, 30) # mathematically, a "vector" of tokens
> length(adv) # sample size = 52,037 tokens
```
Descriptive statistics: type-frequency list

```r
> adv.tfl <- vec2tfl(adv)
> adv.tfl
   k f type  
not 1 4859 not
n’t 2 2084 n’t
so  3 1464 so
only 4 1381 only
then 5 1374 then
now 6 1309 now
even 7 1134 even
as  8 1089 as
... 
N   V
52037 1907
```

```r
> N(adv.tfl)  # sample size
> V(adv.tfl)  # type count
```
Descriptive statistics: type-frequency list

- Visualize descriptive statistics with `plot` method

```r
> plot(adv.tfl)  # Zipf ranking
> plot(adv.tfl, log="xy")  # logarithmic scale recommended
```

![Type-Frequency List (Zipf ranking)](image)
Descriptive statistics: frequency spectrum

```r
> adv.spc <- tfl2spc(adv.tfl)  # or directly with vec2spc
> adv.spc
    m  Vm
 1 1 762
 2 2 260
 3 3 144
 4 4  99
 5 5  69
 6 6  50
 7 7  40
 8 8  34
  ...  ...
  N  V
52037 1907

> N(adv.spc)  # sample size
> V(adv.spc)  # type count
```
Descriptive statistics: frequency spectrum

> plot(adv.spc) # barplot of frequency spectrum
> ?plot.spc # see help page for further options
Descriptive statistics: vocabulary growth

- VGC lists vocabulary size $V(N)$ at different sample sizes $N$
- Optionally also spectrum elements $V_m(N)$ up to $m_{.\text{max}}$

```r
> adv.vgc <- vec2vgc(adv, m.max=2)
> plot(adv.vgc, add.m=1:2)  # plot all three VGCs
```

![Vocabulary Growth Graph](image)

$V(N)$, $V_1(N)$, $V_2(N)$
Further example data sets

- **?Brown** words from Brown corpus
- **?BrownSubsets** various subsets
- **?Dickens** words from novels by Charles Dickens
- **?ItaPref** Italian word-formation prefixes
- **?TigerNP** NP and PP patterns from German Tiger treebank
- **?Baayen2001** frequency spectra from Baayen (2001)
- **?EvertLuedeling2001** German word-formation affixes (manually corrected data from Evert and Lüdeling 2001)

**Practice:**

- Explore these data sets with descriptive statistics
- Try different plot options (from help pages ?plot.tfl, ?plot.spc, ?plot.vgc)
Outline

Introduction
  Motivation
  Notation & basic concepts
  Zipf’s law
  First steps (zipfR)

LNRE models
  Population & samples
  The mathematics of LNRE

Applications & examples
  Productivity & lexical diversity
  Practical LNRE modelling
  Bootstrapping experiments
  LNRE as Bayesian prior

Challenges
  Model inference
  Zipf’s law
  Non-randomness
  Significance testing
  Outlook
Why do we need statistics?

- Often want to compare samples of different sizes
  - extrapolation of VGC & productivity measures
Why do we need statistics?

▸ Often want to compare samples of different sizes
  ↳ extrapolation of VGC & productivity measures

▸ Interested in productivity of affix, vocabulary of author, . . .; not in a particular text or sample
  ↳ statistical inference from sample to population
  ↳ significance of differences in productivity
Why do we need statistics?

- Often want to compare samples of different sizes
  - extrapolation of VGC & productivity measures

- Interested in productivity of affix, vocabulary of author, . . . ; not in a particular text or sample
  - statistical inference from sample to population
  - significance of differences in productivity

- Discrete frequency counts are difficult to capture with generalizations such as Zipf’s law
  - Zipf’s law predicts many impossible types with $1 < f_r < 2$
  - population does not suffer from such quantization effects
This tutorial introduces the state-of-the-art LNRE approach proposed by Baayen (2001)

- LNRE = Large Number of Rare Events

LNRE uses various approximations and simplifications to obtain a tractable and elegant model

Of course, we could also estimate the precise discrete distributions using MCMC simulations, but . . .

1. LNRE model usually minor component of complex procedure
2. often applied to very large samples ($N > 1$ M tokens)
3. still better than naive least-squares regression on Zipf ranking
The LNRE population

- Population: set of $S$ types $w_i$ with occurrence probabilities $\pi_i$
- $S = \text{population diversity}$ can be finite or infinite ($S = \infty$)
- Not interested in specific types $\Rightarrow$ arrange by decreasing probability: $\pi_1 \geq \pi_2 \geq \pi_3 \geq \cdots$
  - Impossible to determine probabilities of all individual types
- Normalization: $\pi_1 + \pi_2 + \ldots + \pi_S = 1$
The LNRE population

- Population: set of $S$ types $w_i$ with occurrence probabilities $\pi_i$
- $S =$ population diversity can be finite or infinite ($S = \infty$)
- Not interested in specific types $\Rightarrow$ arrange by decreasing probability: $\pi_1 \geq \pi_2 \geq \pi_3 \geq \cdots$
  - impossible to determine probabilities of all individual types
- Normalization: $\pi_1 + \pi_2 + \ldots + \pi_S = 1$

- Need parametric statistical model to describe full population (esp. for $S = \infty$), i.e. a function $i \mapsto \pi_i$
  - type probabilities $\pi_i$ cannot be estimated reliably from a sample, but parameters of this function can
  - NB: population index $i \neq$ Zipf rank $r$
What should the population look like?
Zipf-Mandelbrot law as a population model

- Zipf-Mandelbrot law for type probabilities:

\[ \pi_i \coloneqq \frac{C}{(i + b)^a} \]

- Two free parameters: \( a > 1 \) and \( b \geq 0 \)
- Third parameter: \( S > 0 \) or \( S = \infty \)
- This is the Zipf-Mandelbrot population model (Evert 2004)
- \( ZM \) for Zipf-Mandelbrot model (\( S = \infty \))
- \( fZM \) for finite Zipf-Mandelbrot model
Zipf-Mandelbrot law as a population model

- Zipf-Mandelbrot law for type probabilities:
  \[ \pi_i := \frac{C}{(i + b)^a} \]

- Two free parameters: \( a > 1 \) and \( b \geq 0 \)
  - \( C \) is not a parameter but a normalization constant, needed to ensure that \( \sum_i \pi_i = 1 \)
Zipf-Mandelbrot law as a population model

- Zipf-Mandelbrot law for type probabilities:
  \[ \pi_i := \frac{C}{(i + b)^a} \]

- Two free parameters: \( a > 1 \) and \( b \geq 0 \)
  
  \( C \) is not a parameter but a normalization constant, needed to ensure that \( \sum_i \pi_i = 1 \)

- Third parameter: \( S > 0 \) or \( S = \infty \)
Zipf-Mandelbrot law as a population model

► Zipf-Mandelbrot law for type probabilities:

\[ \pi_i := \frac{C}{(i + b)^a} \]

► Two free parameters: \( a > 1 \) and \( b \geq 0 \)

\( C \) is not a parameter but a normalization constant, needed to ensure that \( \sum_i \pi_i = 1 \)

► Third parameter: \( S > 0 \) or \( S = \infty \)

► This is the **Zipf-Mandelbrot** population model (Evert 2004)
  
  ► **ZM** for Zipf-Mandelbrot model \((S = \infty)\)
  
  ► **fZM** for finite Zipf-Mandelbrot model
The parameters of the Zipf-Mandelbrot model

- $a = 1.2$, $b = 1.5$
- $a = 2$, $b = 10$
- $a = 2$, $b = 15$
- $a = 5$, $b = 40$
The parameters of the Zipf-Mandelbrot model

- $a=1.2$, $b=1.5$
- $a=2$, $b=10$
- $a=2$, $b=15$
- $a=5$, $b=40$
Sampling from a population model

Assume we believe that the population we are interested in can be described by a Zipf-Mandelbrot model:

Use computer simulation to generate random samples:

- Draw $N$ tokens from the population such that in each step, type $w_i$ has probability $\pi_i$ to be picked
- This allows us to make predictions for samples (= corpora) of arbitrary size $N$
Sampling from a population model

#1: 1 42 34 23 108 18 48 18 1 ...
Sampling from a population model

#1: 1 42 34 23 108 18 48 18 1 ...
  time order room school town course area course time ...

#2: 286 28 23 36 3 4 7 4 8 ...

#3: 2 11 105 21 11 17 17 1 16 ...

#4: 44 3 110 34 223 2 25 20 28 ...

#5: 24 81 54 11 8 61 1 31 35 ...

#6: 3 65 9 165 5 42 16 20 7 ...

#7: 10 21 11 60 164 54 18 16 203 ...

#8: 11 7 147 5 24 19 15 85 37 ...

...
Sampling from a population model

#1:  
1 42 34 23 108 18 48 18 1 ... 
time order room school town course area course time ... 

#2:  
286 28 23 36 3 4 7 4 8 ... 

Stefan Evert
# Sampling from a population model

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sequence</th>
<th>Time order</th>
<th>Room</th>
<th>School</th>
<th>Town</th>
<th>Course area</th>
<th>Course time</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1  42  34  23  108  18  48  18  1 ...</td>
<td>time order</td>
<td>room</td>
<td>school</td>
<td>town</td>
<td>course area</td>
<td>course time</td>
</tr>
<tr>
<td>#2</td>
<td>286  28  23  36  3  4  7  4  8 ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#3</td>
<td>2  11  105  21  11  17  17  1  16 ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Sampling from a population model

<table>
<thead>
<tr>
<th>#</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1 42 34 23 108 18 48 18 1 ...</td>
</tr>
<tr>
<td></td>
<td>time order room school town course area course time ...</td>
</tr>
<tr>
<td>#2</td>
<td>286 28 23 36 3 4 7 4 8 ...</td>
</tr>
<tr>
<td>#3</td>
<td>2 11 105 21 11 17 17 1 16 ...</td>
</tr>
<tr>
<td>#4</td>
<td>44 3 110 34 223 2 25 20 28 ...</td>
</tr>
<tr>
<td>#5</td>
<td>24 81 54 11 8 61 1 31 35 ...</td>
</tr>
<tr>
<td>#6</td>
<td>3 65 9 165 5 42 16 20 7 ...</td>
</tr>
<tr>
<td>#7</td>
<td>10 21 11 60 164 54 18 16 203 ...</td>
</tr>
<tr>
<td>#8</td>
<td>11 7 147 5 24 19 15 85 37 ...</td>
</tr>
</tbody>
</table>

...
### Samples: type frequency list & spectrum

<table>
<thead>
<tr>
<th>rank $r$</th>
<th>$f_r$</th>
<th>type $i$</th>
<th>$m$</th>
<th>$V_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37</td>
<td>6</td>
<td>1</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>1</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>3</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>31</td>
<td>7</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
<td>10</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>28</td>
<td>12</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
<td>2</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>24</td>
<td>4</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>16</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>23</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>22</td>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**sample #1**
## Samples: type frequency list & spectrum

<table>
<thead>
<tr>
<th>rank $r$</th>
<th>$f_r$</th>
<th>type $i$</th>
<th>$m$</th>
<th>$V_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>39</td>
<td>2</td>
<td>1</td>
<td>76</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>3</td>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>5</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>10</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>8</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>1</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>13</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>7</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>6</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>23</td>
<td>11</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>19</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

**sample #2**
Random variation in type-frequency lists

$r \leftrightarrow f_r$

$i \leftrightarrow f_i$
Random variation: frequency spectrum

Sample #1

\[ V_m \]

\[ m \]
Random variation: frequency spectrum

Sample #2

$v_m$

$m$
Random variation: frequency spectrum

Sample #3

V_m

m

0 20 40 60 80 100
Random variation: frequency spectrum

Sample #4

$V_m$

$m$
Random variation: vocabulary growth curve

Sample #1

$V(N)/V_1(N)$

$N$
Random variation: vocabulary growth curve

Sample #2

\[
\frac{V(N)}{V_1(N)}
\]

\(N\)
Random variation: vocabulary growth curve
Random variation: vocabulary growth curve
Expected values

▶ There is no reason why we should choose a particular sample to compare to the real data or make a prediction – each one is equally likely or unlikely

▶ Take the average over a large number of samples, called expected value or expectation in statistics

▶ Notation: $E[V(N)]$ and $E[V_m(N)]$
  ▶ indicates that we are referring to expected values for a sample of size $N$
  ▶ rather than to the specific values $V$ and $V_m$ observed in a particular sample or a real-world data set

▶ Expected values can be calculated efficiently without generating thousands of random samples
The expected frequency spectrum

Sample #1

![Chart showing frequency spectrum with two bars representing $V_m$ and $E[V_m]$.]
The expected frequency spectrum

Sample #2

- $V_m$
- $E[V_m]$
The expected frequency spectrum

Sample #3

- \( V_m \)
- \( E[V_m] \)
The expected frequency spectrum
The expected vocabulary growth curve

Sample #1

\[
E[V(N)]
\]

Sample #1

\[
E[V_1(N)]
\]
Prediction intervals for the expected VGC

“Confidence intervals” indicate predicted sampling distribution:

- for 95% of samples generated by the LNRE model, VGC will fall within the range delimited by the thin red lines
Parameter estimation by trial & error

\[ a = 1.5, \ b = 7.5 \]

\[ \begin{align*}
V_m/E[V_m] & \quad \text{observed} \\
& \quad \text{ZM model}
\end{align*} \]

\[ \begin{align*}
V(N)/E[V(N)] & \quad \text{observed} \\
& \quad \text{ZM model}
\end{align*} \]
Parameter estimation by trial & error

- LNRE models
- Population & samples

**Observed ZM model**

- $a = 1.3$
- $b = 7.5$

**Graphs**

- Bar graph: $V_m / E[V_m]$ vs. $m$
- Line graph: $V(N) / E[V(N)]$ vs. $N$

- Both graphs show the observed data and the ZM model predictions.

---

Stefan Evert
T1: Zipf’s Law
22 July 2019 | CC-by-sa
Parameter estimation by trial & error

\[ a = 1.3, \ b = 0.2 \]

\[ \frac{V_m}{E[V_m]} \]

\[ \frac{V(N)}{E[V(N)]} \]

\[ m \]

\[ N \]

\[ \text{observed} \]

\[ \text{ZM model} \]

Stefan Evert

T1: Zipf’s Law

22 July 2019 | CC-by-sa
Parameter estimation by trial & error

\[ a = 1.5, \ b = 7.5 \]

\[ V_m/E[V_m] \]

\[ V(N)/E[V(N)] \]

\[ \text{observed} \]  
\[ \text{ZM model} \]

- Stefan Evert
- T1: Zipf’s Law
- 22 July 2019 | CC-by-sa
Parameter estimation by trial & error

![Graph showing V_m/E[V_m] vs m with a = 1.7, b = 7.5]

![Graph showing V(N)/E[V(N)] vs N with a = 1.7, b = 7.5]

Observed ZM model

a = 1.7, b = 7.5

m

V_m/E[V_m]

N

V(N)/E[V(N)]
Parameter estimation by trial & error

$V_{m(E[V_m])}
\begin{align*}
    a &= 1.7, \quad b = 80 \\
\end{align*}$

$V_{(N(E[V_{(N)]})}$
\begin{align*}
    a &= 1.7, \quad b = 80 \\
\end{align*}$
Parameter estimation by trial & error

\[ a = 2, \ b = 550 \]

- Left graph: \( V_m/E[V_m] \) vs. \( m \)
- Right graph: \( V(N)/E[V(N)] \) vs. \( N \)

- Black line: observed
- Red line: ZM model
Automatic parameter estimation

By trial & error we found $a = 2.0$ and $b = 550$

Automatic estimation procedure: $a = 2.39$ and $b = 1968$
Outline

Introduction
- Motivation
- Notation & basic concepts
- Zipf’s law
- First steps (zipfR)

LNRE models
- Population & samples
- The mathematics of LNRE

Applications & examples
- Productivity & lexical diversity
- Practical LNRE modelling
- Bootstrapping experiments
- LNRE as Bayesian prior

Challenges
- Model inference
- Zipf’s law
- Non-randomness
- Significance testing
- Outlook
The sampling model

- Draw random sample of $N$ tokens from LNRE population
- Sufficient statistic: set of type frequencies $\{f_i\}$
  - because tokens of random sample have no ordering
- Joint **multinomial** distribution of $\{f_i\}$:

$$\Pr(\{f_i = k_i\} \mid N) = \frac{N!}{k_1! \cdots k_S!} \pi_1^{k_1} \cdots \pi_S^{k_S}$$
The sampling model

- Draw random sample of $N$ tokens from LNRE population
- Sufficient statistic: set of type frequencies $\{f_i\}$
  - because tokens of random sample have no ordering
- Joint **multinomial** distribution of $\{f_i\}$:

  $$
  \Pr(\{f_i = k_i\} \mid N) = \frac{N!}{k_1! \cdots k_S!} \pi_1^{k_1} \cdots \pi_S^{k_S}
  $$

- **Approximation:** do not condition on fixed sample size $N$
  - $N$ is now the average (expected) sample size
- Random variables $f_i$ have **independent Poisson** distributions:

  $$
  \Pr(f_i = k_i) = e^{-N\pi_i} \frac{(N\pi_i)^{k_i}}{k_i!}
  $$
Frequency spectrum

- Key problem: we cannot determine $f_i$ in observed sample
  - because we don’t know which type $w_i$ is
  - recall that population ranking $f_i \neq$ Zipf ranking $f_r$
- Use spectrum $\{ V_m \}$ and sample size $V$ as statistics
  - contains all information we have about observed sample
Frequency spectrum

- Key problem: we cannot determine $f_i$ in observed sample
  - because we don’t know which type $w_i$ is
  - recall that population ranking $f_i \neq$ Zipf ranking $f_r$
- Use spectrum $\{V_m\}$ and sample size $V$ as statistics
  - contains all information we have about observed sample
- Can be expressed in terms of indicator variables

$$l_{[f_i=m]} = \begin{cases} 
1 & f_i = m \\
0 & \text{otherwise} 
\end{cases}$$
Frequency spectrum

- Key problem: we cannot determine \( f_i \) in observed sample
  - because we don’t know which type \( w_i \) is
  - recall that population ranking \( f_i \neq \) Zipf ranking \( f_r \)
- Use spectrum \( \{ V_m \} \) and sample size \( V \) as statistics
  - contains all information we have about observed sample
- Can be expressed in terms of indicator variables

\[
I_{[f_i=m]} = \begin{cases} 
1 & f_i = m \\ 
0 & \text{otherwise}
\end{cases}
\]

\[
V_m = \sum_{i=1}^{S} I_{[f_i=m]}
\]
Frequency spectrum

- Key problem: we cannot determine \( f_i \) in observed sample
  - because we don’t know which type \( w_i \) is
  - recall that population ranking \( f_i \neq \) Zipf ranking \( f_r \)
- Use spectrum \( \{ V_m \} \) and sample size \( V \) as statistics
  - contains all information we have about observed sample
- Can be expressed in terms of indicator variables

\[
I[f_i=m] = \begin{cases} 1 & f_i = m \\ 0 & \text{otherwise} \end{cases}
\]

\[
V_m = \sum_{i=1}^{S} I[f_i=m]
\]

\[
V = \sum_{i=1}^{S} I[f_i>0] = \sum_{i=1}^{S} (1 - I[f_i=0])
\]
The expected spectrum

- It is easy to compute expected values for the frequency spectrum (and variances because the $f_i$ are independent)

$$E[l_{f_i=m}] = \Pr(f_i = m) = e^{-N\pi_i} \frac{(N\pi_i)^m}{m!}$$
The expected spectrum

It is easy to compute expected values for the frequency spectrum (and variances because the $f_i$ are independent)

\[
E[l_{f_i=m}] = Pr(f_i = m) = e^{-N \pi_i} \frac{(N \pi_i)^m}{m!}
\]

\[
E[V_m] = \sum_{i=1}^{S} E[l_{f_i=m}] = \sum_{i=1}^{S} e^{-N \pi_i} \frac{(N \pi_i)^m}{m!}
\]

\[\text{NB: } V_m \text{ and } V_m \text{ are not independent because they are derived from the same random variables}\]
The expected spectrum

- It is easy to compute expected values for the frequency spectrum (and variances because the $f_i$ are independent)

\[
E[l_{f_i=m}] = \Pr(f_i = m) = e^{-N\pi_i} \frac{(N\pi_i)^m}{m!}
\]

\[
E[V_m] = \sum_{i=1}^{S} E[l_{f_i=m}] = \sum_{i=1}^{S} e^{-N\pi_i} \frac{(N\pi_i)^m}{m!}
\]

\[
E[V] = \sum_{i=1}^{S} E[1 - l_{f_i=0}] = \sum_{i=1}^{S} (1 - e^{-N\pi_i})
\]

***NB:** $V_m$ and $V$ are not independent because they are derived from the same random variables.
The expected spectrum

It is easy to compute expected values for the frequency spectrum (and variances because the $f_i$ are independent)

$$E[l_{f_i=m}] = \Pr(f_i = m) = e^{-N\pi_i} \frac{(N\pi_i)^m}{m!}$$

$$E[V_m] = \sum_{i=1}^{S} E[l_{f_i=m}] = \sum_{i=1}^{S} e^{-N\pi_i} \frac{(N\pi_i)^m}{m!}$$

$$E[V] = \sum_{i=1}^{S} E[1 - l_{f_i=0}] = \sum_{i=1}^{S} (1 - e^{-N\pi_i})$$

**NB:** $V_m$ and $V$ are not independent because they are derived from the same random variables $f_i$
Sampling distribution of $V_m$ and $V$

- Joint sampling distribution of $\{V_m\}$ and $V$ is complicated
- **Approximation:** $V$ and $\{V_m\}$ asymptotically follow a **multivariate normal** distribution
  - motivated by the multivariate central limit theorem: sum of many independent variables $I_{[f_i=m]}$
- Usually limited to first spectrum elements, e.g. $V_1, \ldots, V_{15}$
  - approximation of discrete $V_m$ by continuous distribution suitable only if $E[V_m]$ is sufficiently large
Sampling distribution of $V_m$ and $V$

- Joint sampling distribution of $\{V_m\}$ and $V$ is complicated
- **Approximation:** $V$ and $\{V_m\}$ asymptotically follow a **multivariate normal** distribution
  - motivated by the multivariate central limit theorem: sum of many independent variables $l_{[f_i=m]}$
- Usually limited to first spectrum elements, e.g. $V_1, \ldots, V_{15}$
  - approximation of discrete $V_m$ by continuous distribution suitable only if $E[V_m]$ is sufficiently large
- Parameters of multivariate normal:
  $\mu = (E[V], E[V_1], E[V_2], \ldots)$ and $\Sigma =$ covariance matrix

$$
\Pr((V, V_1, \ldots, V_k) = v) \sim \frac{e^{-\frac{1}{2}(v-\mu)^T\Sigma^{-1}(v-\mu)}}{\sqrt{(2\pi)^{k+1} \det \Sigma}}
$$
Type density function

- Discrete sums of probabilities in $\mathbb{E}[V], \mathbb{E}[V_m], \ldots$ are inconvenient and computationally expensive

- **Approximation:** continuous type density function $g(\pi)$

\[
|\{w_i \mid a \leq \pi_i \leq b\}| = \int_a^b g(\pi) \, d\pi
\]

\[
\sum\{\pi_i \mid a \leq \pi_i \leq b\} = \int_a^b \pi g(\pi) \, d\pi
\]
Type density function

- Discrete sums of probabilities in $E[V]$, $E[V_m]$, ... are inconvenient and computationally expensive

- **Approximation**: continuous **type density function** $g(\pi)$

\[
|\{w_i \mid a \leq \pi_i \leq b\}| = \int_a^b g(\pi) \, d\pi
\]

\[
\sum\{\pi_i \mid a \leq \pi_i \leq b\} = \int_a^b \pi g(\pi) \, d\pi
\]

- Normalization constraint:

\[
\int_0^\infty \pi g(\pi) \, d\pi = 1
\]

- Good approximation for low-probability types, but probability mass of $w_1, w_2, ...$ “smeared out” over range
Type density function

Type density as continuous approximation

occurrence probability $\pi$

type density $g(\pi)$

Probability density vs. type probabilities

occurrence probability $\pi$

probability density $f(\pi)$

$\pi_1 = 0.376$
$\pi_2 = 0.215$
$\pi_3 = 0.130$
$\pi_4 = 0.082$
Type density function

Probability density vs. type probabilities

occurrence probability π

π₁ = 0.376
π₂ = 0.215
π₃ = 0.130
π₄ = 0.082
ZM and fZM as LNRE models

- Discrete Zipf-Mandelbrot population

\[ \pi_i := \frac{C}{(i + b)^a} \quad \text{for } i = 1, \ldots, S \]
ZM and fZM as LNRE models

- Discrete Zipf-Mandelbrot population

\[ \pi_i := \frac{C}{(i + b)^a} \quad \text{for } i = 1, \ldots, S \]

- Corresponding type density function (Evert 2004)

\[ g(\pi) = \begin{cases} 
C \cdot \pi^{-\alpha-1} & A \leq \pi \leq B \\
0 & \text{otherwise}
\end{cases} \]
ZM and fZM as LNRE models

- Discrete Zipf-Mandelbrot population

\[ \pi_i := \frac{C}{(i + b)^a} \quad \text{for } i = 1, \ldots, S \]

- Corresponding type density function (Evert 2004)

\[
g(\pi) = \begin{cases} 
C \cdot \pi^{-\alpha - 1} & A \leq \pi \leq B \\
0 & \text{otherwise}
\end{cases}
\]

with parameters

- \( \alpha = 1/a \ (0 < \alpha < 1) \)
- \( B = (1 - \alpha)/(b \cdot \alpha) \)
- \( 0 \leq A < B \) determines \( S \) (ZM with \( S = \infty \) for \( A = 0 \))
- \( C \) is a normalization factor, not a parameter
ZM and fZM as LNRE models
ZM and fZM as LNRE models

Type density of LNRE model

occurrence probability $\pi$

occurrence probability $\pi$

$g(\pi)$

$g(\pi)$

$fZM$
Expectations as integrals

- Expected values can now be expressed as integrals over $g(\pi)$

\[
E[V_m] = \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi) \, d\pi
\]

\[
E[V] = \int_0^\infty (1 - e^{-N\pi}) g(\pi) \, d\pi
\]
Expectations as integrals

- Expected values can now be expressed as integrals over $g(\pi)$

$$
E[V_m] = \int_{0}^{\infty} \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi) \, d\pi
$$

$$
E[V] = \int_{0}^{\infty} (1 - e^{-N\pi}) g(\pi) \, d\pi
$$

- Reduce to simple closed form for ZM with $b = 0$ ($\Rightarrow B = \infty$)

$$
E[V_m] = \frac{C}{m!} \cdot N^\alpha \cdot \Gamma(m - \alpha)
$$

$$
E[V] = C \cdot N^\alpha \cdot \frac{\Gamma(1 - \alpha)}{\alpha}
$$

- fZM and general ZM with incomplete Gamma function
Parameter estimation from training corpus

- For ZM, $\alpha = \frac{E[V_1]}{E[V]} \approx \frac{V_1}{V}$ can be estimated directly, but prone to overfitting.
- General parameter fitting by MLE:
  maximize likelihood of observed spectrum $\mathbf{v}$

$$\max_{\alpha, A, B} \Pr((V, V_1, \ldots, V_k) = \mathbf{v} | \alpha, A, B)$$
Parameter estimation from training corpus

- For ZM, \( \alpha = \frac{\mathbb{E}[V_1]}{\mathbb{E}[V]} \approx \frac{V_1}{V} \) can be estimated directly, but prone to overfitting.

- General parameter fitting by MLE:
  maximize likelihood of observed spectrum \( \mathbf{v} \)

  \[
  \max_{\alpha, A, B} \Pr((V, V_1, \ldots, V_k) = \mathbf{v} | \alpha, A, B)
  \]

- Multivariate normal approximation:

  \[
  \min_{\alpha, A, B} (\mathbf{v} - \mu)^T \Sigma^{-1} (\mathbf{v} - \mu)
  \]

- Minimization by gradient descent (BFGS, CG) or simplex search (Nelder-Mead)
Parameter estimation from training corpus

BNC (bare singular PPs)

α

Goodness-of-fit X2 (m = 10)

(0.65, −2.11)
Parameter estimation from training corpus

Brown Corpus (word forms)

Goodness-of-fit X2 (m = 5)
Goodness-of-fit
(Baayen 2001, Sec. 3.3)

- How well does the fitted model explain the observed data?
- For multivariate normal distribution:

\[ X^2 = (\mathbf{V} - \mu)^T \Sigma^{-1} (\mathbf{V} - \mu) \sim \chi^2_{k+1} \]

where \( \mathbf{V} = (V, V_1, \ldots, V_k) \)
Goodness-of-fit

(Baayen 2001, Sec. 3.3)

- How well does the fitted model explain the observed data?
- For multivariate normal distribution:

\[
X^2 = (V - \mu)^T \Sigma^{-1}(V - \mu) \sim \chi^2_{k+1}
\]

where \( V = (V, V_1, \ldots, V_k) \)

- Multivariate chi-squared test of **goodness-of-fit**
  - replace \( V \) by observed \( v \) \( \Rightarrow \) test statistic \( x^2 \)
  - must reduce df = \( k + 1 \) by number of estimated parameters

- NB: significant rejection of the LNRE model for \( p < .05 \)
Coffee break!
Outline

Introduction
  Motivation
  Notation & basic concepts
  Zipf’s law
  First steps (zipfR)

LNRE models
  Population & samples
  The mathematics of LNRE

Applications & examples
  Productivity & lexical diversity
  Practical LNRE modelling
  Bootstrapping experiments
  LNRE as Bayesian prior

Challenges
  Model inference
  Zipf’s law
  Non-randomness
  Significance testing
  Outlook
Measuring morphological productivity
example from Evert and Lüdeling (2001)

Vocabulary Growth Curves

\[ V(N) \]

- bar
- sam
- ös

Stefan Evert
Measuring morphological productivity
example from Evert and Lüdeling (2001)

\[ a = 1.45, \quad b = 34.59, \quad S = 20587 \]
Measuring morphological productivity
example from Evert and Lüdeling (2001)
Quantitative measures of productivity
(Tweedie and Baayen 1998; Baayen 2001)

- Baayen’s (1991) productivity index $\mathcal{P}$
  (slope of vocabulary growth curve)
  \[
  \mathcal{P} = \frac{V_1}{N}
  \]

- TTR = type-token ratio
  \[
  \text{TTR} = \frac{V}{N}
  \]

- Zipf-Mandelbrot slope
  \[
  a
  \]

- Herdan’s law (1964)
  \[
  C = \frac{\log V}{\log N}
  \]
Quantitative measures of productivity
(Tweedie and Baayen 1998; Baayen 2001)

- Baayen’s (1991) productivity index $P$
  (slope of vocabulary growth curve)
  \[ P = \frac{V_1}{N} \]

- TTR = type-token ratio
  \[ \text{TTR} = \frac{V}{N} \]

- Zipf-Mandelbrot slope
  \[ a \]

- Herdan’s law (1964)
  \[ C = \frac{\log V}{\log N} \]

- Yule (1944) / Simpson (1949)
  \[ K = 10000 \cdot \frac{\sum_m m^2 V_m - N}{N^2} \]

- Guiraud (1954)
  \[ R = \frac{V}{\sqrt{N}} \]

- Sichel (1975)
  \[ S = \frac{V_2}{V} \]

- Honoré (1979)
  \[ H = \frac{\log N}{1 - \frac{V_1}{V}} \]
## Productivity measures for bare singulars in the BNC

<table>
<thead>
<tr>
<th></th>
<th>spoken</th>
<th>written</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V$</td>
<td>2,039</td>
<td>12,876</td>
</tr>
<tr>
<td>$N$</td>
<td>6,766</td>
<td>85,750</td>
</tr>
<tr>
<td>$K$</td>
<td>86.84</td>
<td>28.57</td>
</tr>
<tr>
<td>$R$</td>
<td>24.79</td>
<td>43.97</td>
</tr>
<tr>
<td>$S$</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>$C$</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>$P$</td>
<td>0.21</td>
<td>0.08</td>
</tr>
<tr>
<td>TTR</td>
<td>0.301</td>
<td>0.150</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1.18</td>
<td>1.27</td>
</tr>
<tr>
<td>pop. $S$</td>
<td>15,958</td>
<td>36,874</td>
</tr>
</tbody>
</table>
# Productivity measures for bare singulars in the BNC

<table>
<thead>
<tr>
<th></th>
<th>spoken</th>
<th>written</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V$</td>
<td>2,039</td>
<td>12,876</td>
</tr>
<tr>
<td>$N$</td>
<td>6,766</td>
<td>85,750</td>
</tr>
<tr>
<td>$K$</td>
<td>86.84</td>
<td>28.57</td>
</tr>
<tr>
<td>$R$</td>
<td>24.79</td>
<td>43.97</td>
</tr>
<tr>
<td>$S$</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>$C$</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>$\mathcal{P}$</td>
<td>0.21</td>
<td>0.08</td>
</tr>
<tr>
<td>$\text{TTR}$</td>
<td>0.301</td>
<td>0.150</td>
</tr>
<tr>
<td>$a$</td>
<td>1.18</td>
<td>1.27</td>
</tr>
<tr>
<td>pop. $S$</td>
<td>15,958</td>
<td>36,874</td>
</tr>
</tbody>
</table>

![Vocabulary growth curves (BNC)](image)
Are these “lexical constants” really constant?
interactive demo
Outline

Introduction
Motivation
Notation & basic concepts
Zipf’s law
First steps (zipfR)

LNRE models
Population & samples
The mathematics of LNRE

Applications & examples
Productivity & lexical diversity
Practical LNRE modelling
Bootstrapping experiments
LNRE as Bayesian prior

Challenges
Model inference
Zipf’s law
Non-randomness
Significance testing
Outlook
Bootstrapping

- An empirical approach to sampling variation:
  - take many random samples from the same population
  - analyse distribution e.g. of productivity measures (mean, median, s.d., boxplot, histogram, ...)
  - alternatively, estimate LNRE model from each sample and analyse distribution of model parameters (➡️ later)
  - problem: how to obtain the additional samples?
Bootstrapping

▶ An empirical approach to sampling variation:
  ▶ take many random samples from the same population
  ▶ analyse distribution e.g. of productivity measures
    (mean, median, s.d., boxplot, histogram, ...)
  ▶ alternatively, estimate LNRE model from each sample and
    analyse distribution of model parameters (later)
  ▶ problem: how to obtain the additional samples?

▶ Bootstrapping (Efron 1979)
  ▶ resample from observed data with replacement
  ▶ this approach is not suitable for type-token distributions
    (resamples underestimate vocabulary size V!)
Bootstrapping

- An empirical approach to sampling variation:
  - take many random samples from the same population
  - analyse distribution e.g. of productivity measures (mean, median, s.d., boxplot, histogram, ...)
  - alternatively, estimate LNRE model from each sample and analyse distribution of model parameters (later)
  - problem: how to obtain the additional samples?

- Bootstrapping (Efron 1979)
  - resample from observed data with replacement
  - this approach is not suitable for type-token distributions (resamples underestimate vocabulary size $V$!)

- Parametric bootstrapping
  - use fitted LNRE model to generate samples, i.e. sample from the population described by the model
  - advantage: “correct” parameter values are known
Parametric bootstrapping with LNRE models

- Use simulation experiments to gain better understanding of quantitative measures
  - LNRE model = well-defined population

![Zipf–Mandelbrot spectrum](image)

- $E[V_m]$ can be computed directly in simple cases

- $a = 2$
- $a = 1.4$
- $a = 1.1$

Stefan Evert

T1: Zipf’s Law

22 July 2019 | CC-by-sa
Parametric bootstrapping with LNRE models

- Use simulation experiments to gain better understanding of quantitative measures
  - LNRE model = well-defined population
- Parametric bootstrapping based on LNRE population
  - dependence on sample size
  - controlled manipulation of confounding factors
  - empirical sampling distribution → variability
- $E[P]$ etc. can be computed directly in simple cases
Experiment: sample size
Experiment: frequent lexicalized types
Outline

Introduction
Motivation
Notation & basic concepts
Zipf’s law
First steps (zipfR)

LNRE models
Population & samples
The mathematics of LNRE

Applications & examples
Productivity &
lexical diversity
Practical LNRE modelling
Bootstrapping experiments
LNRE as Bayesian prior

Challenges
Model inference
Zipf’s law
Non-randomness
Significance testing
Outlook
Posterior distribution

Posterior distribution \((m = 1)\) for ZM model with \(\alpha = 0.4\)

- posterior
- log-adjusted

Good-Turing
MLE
95% confidence
99.9% confidence

Posterior distribution \(Pr(\pi|m)\)

expected frequency \((N\pi)\)

\(10^{-8}\) \(10^{-5}\) \(10^{-2}\) \(10^{1}\)

0.0 0.2 0.4 0.6 0.8 1.0

Stefan Evert
T1: Zipf’s Law
22 July 2019 | CC-by-sa

85 / 117
Posterior distribution

Posterior distribution \((m = 1)\) for ZM model with \(\alpha = 0.9\)

- **posterior**
- **log–adjusted**

Expected frequency \((N\pi)\):
- \(10^{-8}\)
- \(10^{-5}\)
- \(10^{-2}\)
- \(10^{1}\)

\[\text{posterior distribution } Pr(\pi|m) \text{ for } ZM \text{ model with } \alpha = 0.9\]

Good–Turing posterior log–adjusted

95% confidence

99.9% confidence

Stefan Evert
Posterior distribution

Posterior distribution $(m = 2)$ for ZM model with $\alpha = 0.9$

posterior distribution $Pr(\pi | m)$

expected frequency $(N\pi)$

- posterior
- log–adjusted

MLE
95% confidence
99.9% confidence
MAP
Good–Turing
posterior
log–adjusted

Stefan Evert

T1: Zipf’s Law
22 July 2019 | CC-by-sa
Outline

Introduction
  Motivation
  Notation & basic concepts
  Zipf’s law
  First steps (zipfR)

LNRE models
  Population & samples
  The mathematics of LNRE

Applications & examples
  Productivity & lexical diversity
  Practical LNRE modelling
  Bootstrapping experiments
  LNRE as Bayesian prior

Challenges
  Model inference
  Zipf’s law
  Non-randomness
  Significance testing
  Outlook
How reliable are the fitted models?

Three potential issues:
How reliable are the fitted models?

Three potential issues:

1. Model assumptions $\neq$ population (e.g. distribution does not follow a Zipf-Mandelbrot law) $\Rightarrow$ model cannot be adequate, regardless of parameter settings
How reliable are the fitted models?

Three potential issues:

1. Model assumptions ≠ population (e.g. distribution does not follow a Zipf-Mandelbrot law)
   - model cannot be adequate, regardless of parameter settings

2. Parameter estimation unsuccessful (i.e. suboptimal goodness-of-fit to training data)
   - optimization algorithm trapped in local minimum
   - can result in highly inaccurate model
How reliable are the fitted models?

Three potential issues:

1. Model assumptions ≠ population (e.g. distribution does not follow a Zipf-Mandelbrot law)
   - model cannot be adequate, regardless of parameter settings

2. Parameter estimation unsuccessful (i.e. suboptimal goodness-of-fit to training data)
   - optimization algorithm trapped in local minimum
   - can result in highly inaccurate model

3. Uncertainty due to sampling variation (i.e. training data differ from population distribution)
   - model fitted to training data, may not reflect true population
   - another training sample would have led to different parameters
   - especially critical for small samples ($N < 10,000$)
How reliable are the fitted models?

Three potential issues:

1. Model assumptions ≠ population (e.g. distribution does not follow a Zipf-Mandelbrot law)
   - model cannot be adequate, regardless of parameter settings

2. Parameter estimation unsuccessful (i.e. suboptimal goodness-of-fit to training data)
   - optimization algorithm trapped in local minimum
   - can result in highly inaccurate model

3. Uncertainty due to sampling variation (i.e. training data differ from population distribution)
   - model fitted to training data, may not reflect true population
   - another training sample would have led to different parameters
   - especially critical for small samples ($N < 10,000$)
Bootstrapping
parametric bootstrapping with 100 replicates

**Zipfian slope** \( a = 1/\alpha \)
Bootstrapping
parametric bootstrapping with 100 replicates

Offset $b = \frac{(1 - \alpha)}{(B \cdot \alpha)}$
Bootstrapping
parametric bootstrapping with 100 replicates

\textbf{fZM probability cutoff} \( A = \pi S \)
Bootstrapping
parametric bootstrapping with 100 replicates

**Goodness-of-fit statistic** $X^2$ (model not plausible for $X^2 > 11$)
Bootstrapping

c parametric bootstrapping with 100 replicates

Population diversity $S$
Bootstrapping
parametric bootstrapping with 100 replicates

Population diversity $S$
Sample size matters!

Brown corpus is too small for reliable LNRE parameter estimation (bare singulars)
How reliable are the fitted models?

Three potential issues:

1. **Model assumptions ≠ population**  
   (e.g. distribution does not follow a Zipf-Mandelbrot law)  
   - model cannot be adequate, regardless of parameter settings

2. **Parameter estimation unsuccessful**  
   (i.e. suboptimal goodness-of-fit to training data)  
   - optimization algorithm trapped in local minimum  
   - can result in highly inaccurate model

3. **Uncertainty due to sampling variation**  
   (i.e. training data differ from population distribution)  
   - model fitted to training data, may not reflect true population  
   - another training sample would have led to different parameters  
   - especially critical for small samples ($N < 10,000$)
Outline

Introduction
  Motivation
  Notation & basic concepts
  Zipf’s law
  First steps (zipfR)

LNRE models
  Population & samples
  The mathematics of LNRE

Applications & examples
  Productivity & lexical diversity
  Practical LNRE modelling
  Bootstrapping experiments
  LNRE as Bayesian prior

Challenges
  Model inference
  Zipf’s law
  Non-randomness
  Significance testing
  Outlook

Stefan Evert
T1: Zipf’s Law
22 July 2019 | CC-by-sa
How well does Zipf’s law hold?
How well does Zipf’s law hold?

- Z-M law seems to fit the first few thousand ranks very well, but then slope of empirical ranking becomes much steeper
  - similar patterns have been found in many different data sets
How well does Zipf’s law hold?

- Z-M law seems to fit the first few thousand ranks very well, but then slope of empirical ranking becomes much steeper
  - similar patterns have been found in many different data sets

- Various modifications and extensions have been suggested (Sichel 1971; Kornai 1999; Montemurro 2001)
  - mathematics of corresponding LNRE models are often much more complex and numerically challenging
  - may not have closed form for $E[V]$, $E[V_m]$, or for the cumulative type distribution $G(\rho) = \int_{\rho}^{\infty} g(\pi) \, d\pi$
How well does Zipf’s law hold?

- Z-M law seems to fit the first few thousand ranks very well, but then slope of empirical ranking becomes much steeper
  - similar patterns have been found in many different data sets

- Various modifications and extensions have been suggested (Sichel 1971; Kornai 1999; Montemurro 2001)
  - mathematics of corresponding LNRE models are often much more complex and numerically challenging
  - may not have closed form for $E[V]$, $E[V_m]$, or for the cumulative type distribution $G(\rho) = \int_\rho^\infty g(\pi) \, d\pi$

- E.g. Generalized Inverse Gauss-Poisson (GIGP; Sichel 1971)

$$g(\pi) = \frac{(2/bc)^{\gamma+1}}{K_{\gamma+1}(b)} \cdot \pi^{\gamma-1} \cdot e^{-\frac{\pi}{c} - \frac{b^2c}{4\pi}}$$
The GIGP model (Sichel 1971)

Type density of LNRE model

occurrence probability \( \pi \)

type density \( g(\pi) \)

- Type density \( f_{ZM} \)
The GIGP model (Sichel 1971)

Type density of LNRE model

Type density $g(\pi)$

occurrence probability $\pi$

$fZM$

$GIGP$

Stefan Evert

T1: Zipf’s Law

22 July 2019 | CC-by-sa
Outline

Introduction
  Motivation
  Notation & basic concepts
  Zipf’s law
  First steps (zipfR)

LNRE models
  Population & samples
  The mathematics of LNRE

Applications & examples
  Productivity &
  lexical diversity
  Practical LNRE modelling
  Bootstrapping experiments
  LNRE as Bayesian prior

Challenges
  Model inference
  Zipf’s law
  Non-randomness
  Significance testing
  Outlook
How accurate is LNRE-based extrapolation?  
(Baroni and Evert 2005)
How accurate is LNRE-based extrapolation?
(Baroni and Evert 2005)

Suffix –lich (25%)
How accurate is LNRE-based extrapolation?
(Baroni and Evert 2005)
How accurate is LNRE-based extrapolation?

(Baroni and Evert 2005)
Reasons for poor extrapolation quality

- Major problem: non-randomness of corpus data
  - LNRE modelling assumes that corpus is random sample
Reasons for poor extrapolation quality

- Major problem: **non-randomness** of corpus data
  - LNRE modelling assumes that corpus is random sample

- Cause 1: **repetition** within texts
  - most corpora use entire text as unit of sampling
  - also referred to as “term clustering” or “burstiness”
  - well-known in computational linguistics (Church 2000)
Reasons for poor extrapolation quality

- Major problem: **non-randomness** of corpus data
  - LNRE modelling assumes that corpus is random sample

- Cause 1: **repetition** within texts
  - most corpora use entire text as unit of sampling
  - also referred to as “term clustering” or “burstiness”
  - well-known in computational linguistics (Church 2000)

- Cause 2: **non-homogeneous** corpus
  - cannot extrapolate from spoken BNC to written BNC
  - similar for different genres and domains
  - also within single text, e.g. beginning/end of novel
The ECHO correction
(Baroni and Evert 2007)

▶ Empirical study: quality of extrapolation $N_0 \rightarrow 4N_0$ starting from random samples of corpus texts

![Graph showing the relative error of E[V] vs. V (DEWAC) for ZM, IZM, and GIGP. The graph plots relative error (%) against ZM, IZM, and GIGP.]

![Graph showing the goodness-of-fit vs. accuracy for V (3N_0). The graph plots $\chi^2$ against $\sqrt{\text{MSE}}$ (%) with a line and a scatter plot.]

Goodness–of–fit vs. accuracy for V (3N_0)

$\chi^2$

$\sqrt{\text{MSE}}$ (%)
The ECHO correction
(Baroni and Evert 2007)

- Empirical study: quality of extrapolation $N_0 \rightarrow 4N_0$ starting from random samples of corpus texts
The ECHO correction
(Baroni and Evert 2007)

▶ Assumption: repetition of type within short span is not a new lexical access or spontaneous formation

A fine example. A very fine example. Only the finest examples. The examples are fine. . . .

The cat sat on the mat. Another very fine cat sat down on the mat. Two mats are fine. . . .
The ECHO correction
(Baroni and Evert 2007)

- Assumption: repetition of type within short span is not a new lexical access or spontaneous formation

A fine example. A very fine example. Only the finest examples. The examples are fine. ...

The cat sat on the mat. Another very fine cat sat down on the mat. Two mats are fine. ...
The ECHO correction
(Baroni and Evert 2007)

- Assumption: repetition of type within short span is not a new lexical access or spontaneous formation
- Replace every repetition within span by special type ECHO
  - \( N, V \) and \( V_1 \) are not affected \( \rightarrow \) same VGC and \( P \)
  - ECHO correction as pre-processing step \( \rightarrow \) no modifications to LNRE models or other analysis software needed

A fine example. ECHO very ECHO ECHO. Only the ECHO ECHO. ECHO ECHO are ECHO. . .

The cat sat on ECHO mat. Another very fine ECHO ECHO down ECHO ECHO ECHO ECHO. Two ECHO are ECHO. . .
The ECHO correction
(Baroni and Evert 2007)

- Assumption: repetition of type within short span is not a new lexical access or spontaneous formation
- Replace every repetition within span by special type ECHO
  - $N, V$ and $V_1$ are not affected $\Rightarrow$ same VGC and $P$
  - ECHO correction as pre-processing step $\Rightarrow$ no modifications to LNRE models or other analysis software needed
- What is an appropriate span size?
  Repetition within textual unit ($\Rightarrow$ document frequencies)

A fine example. **ECHO** very **ECHO ECHO**. Only the **ECHO ECHO**. **ECHO ECHO** are **ECHO**. . .

The cat sat on **ECHO** mat. Another very fine **ECHO ECHO** down **ECHO ECHO ECHO ECHO**. Two **ECHO** are **ECHO**. . .
The ECHO correction
(Baroni and Evert 2007)

- ECHO correction: replace every repetition within same text by special type ECHO (= document frequencies)
The ECHO correction
(Baroni and Evert 2007)

- **ECHO correction**: replace every repetition within same text by special type ECHO (= document frequencies)

![Graph showing goodness-of-fit vs. accuracy for V (3N0)]

- Goodness-of-fit vs. accuracy for V (3N0):
- X²: chi-square statistic
- rMSE: root mean square error

- Data points:
  - ○ standard echo model
  - △ partition-adjusted
Outline

**Introduction**
- Motivation
- Notation & basic concepts
- Zipf’s law
- First steps (zipfR)

**LNRE models**
- Population & samples
- The mathematics of LNRE

**Applications & examples**
- Productivity & lexical diversity
- Practical LNRE modelling
- Bootstrapping experiments
- LNRE as Bayesian prior

**Challenges**
- Model inference
- Zipf’s law
- Non-randomness

**Significance testing**

Outlook
Case study: Iris Murdoch & early symptoms of AD
(Evert et al. 2017)

- Renowned British author (1919–1999)
- Published a total of 26 novels, mostly well received by critics
- Murdoch experienced unexpected difficulties composing her last novel, received “without enthusiasm” (Garrard et al. 2005)
- Diagnosis of Alzheimer’s disease shortly after publication

http://news.bbc.co.uk/2/hi/health/4058605.stm

**Murdoch novel reveals Alzheimer’s**

The last novel by the author Iris Murdoch reveals the first signs of Alzheimer’s disease, experts say.

A team from University College London say their examination of works from throughout Dame Iris’s career could be used to help diagnose others.

They found the structure and grammar of her novels was relatively unchanged, but her language was noticeably simpler in her last novel, 'Jackson’s Dilemma'.

The study is published online by the journal Brain.

http://news.bbc.co.uk/2/hi/health/4058605.stm
Case study: Iris Murdoch & early symptoms of AD  
(Evert et al. 2017)

- Renowned British author (1919–1999)
- Published a total of 26 novels, mostly well received by critics
- Murdoch experienced unexpected difficulties composing her last novel, received “without enthusiasm” (Garrard et al. 2005)
- Diagnosis of Alzheimer’s disease shortly after publication

Conflicting results:
- Decline of lexical diversity in last novel  
  (Garrard et al. 2005; Pakhomov et al. 2011)
- No clear effects found  
  (Le et al. 2011)
Case study: Iris Murdoch & early symptoms of AD
(Evert et al. 2017)

▶ Corpus data
  ▶ 19 out of 26 novels written by Iris Murdoch
  ▶ including 9 last novels, spanning a period of almost 20 years
  ▶ acquired as e-books (no errors due to OCR)

▶ Pre-processing and annotation
  ▶ Stanford CoreNLP (Manning et al. 2014) for tokenization, sentence splitting, POS tagging, and syntactic parsing
  ▶ exclude dialogue based on typographic quotation marks (following Garrard et al. 2005; Pakhomov et al. 2011)

▶ The challenge
  ◄ assess significance of differences in productivity for single texts
  ◄ might explain conflicting results in prior work
Measures of vocabulary diversity = productivity
(Evert et al. 2017)

Yule’s $\kappa$
Measures of vocabulary diversity = productivity

(Evert et al. 2017)

Yule’s $\kappa$

Honoré $H$
Measures of vocabulary diversity = productivity

(Evert et al. 2017)

**Type count / TTR**

**Honoré H**
Cross-validation for productivity measures
(Evert et al. 2017)

As a first step:

▶ Partition each novel into folds of 10,000 consecutive tokens

⇒ $k \geq 6$ folds for each novel (leftover tokens discarded)
Cross-validation for productivity measures
(Evert et al. 2017)

As a first step:
- Partition each novel into folds of 10,000 consecutive tokens
- \( k \geq 6 \) folds for each novel (leftover tokens discarded)

Then:
- Evaluate complexity measure of interest on each fold

\[ y_1, \ldots, y_k \]
Cross-validation for productivity measures

(Evert et al. 2017)

As a first step:
► Partition each novel into folds of 10,000 consecutive tokens
   $k \geq 6$ folds for each novel (leftover tokens discarded)

Then:
► Evaluate complexity measure of interest on each fold
   $y_1, \ldots, y_k$

► Compute macro-average as overall measure for the entire text
   $\bar{y} = \frac{y_1 + \cdots + y_k}{k}$

► Instead of value $x$ obtained by evaluating measure on full text
Cross-validation for productivity measures

(Evert et al. 2017)

Significance testing procedure:

- Standard deviation $\sigma$ of individual folds estimated from data

$$\sigma^2 \approx s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2$$
Cross-validation for productivity measures
(Evert et al. 2017)

Significance testing procedure:

- Standard deviation $\sigma$ of individual folds estimated from data

$$
\sigma^2 \approx s^2 = \frac{1}{k - 1} \sum_{i=1}^{k} (y_i - \bar{y})^2
$$

- Standard deviation of macro average can be computed as

$$
\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{k}} \approx \frac{s}{\sqrt{k}}
$$
Cross-validation for productivity measures

(Evert et al. 2017)

Significance testing procedure:

- Standard deviation $\sigma$ of individual folds estimated from data

$$\sigma^2 \approx s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2$$

- Standard deviation of macro average can be computed as

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{k}} \approx \frac{s}{\sqrt{k}}$$

- Asymptotic 95% confidence intervals are then given by

$$\bar{y} \pm 1.96 \cdot \sigma_{\bar{y}}$$
Cross‐validation for productivity measures
(Evert et al. 2017)

Significance testing procedure:

- Standard deviation $\sigma$ of individual folds estimated from data

$$\sigma^2 \approx s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2$$

- Standard deviation of macro average can be computed as

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{k}} \approx \frac{s}{\sqrt{k}}$$

- Asymptotic 95% confidence intervals are then given by

$$\bar{y} \pm 1.96 \cdot \sigma_{\bar{y}}$$

- Comparison of samples with Student’s $t$-test, based on pooled cross-validation folds (feasible even for $n_1 = 1$)
Productivity measures with confidence intervals

(Evert et al. 2017)

- Type count / TTR
- Honoré H

Significance testing

Significance test vs. first 17 novels

$t = -6.1, \text{df}=5.52, p = .0012^*$
Productivity measures with confidence intervals

(Evert et al. 2017)

t = −6.1, df = 5.52, p = .0012**
Productivity measures with confidence intervals
(Evert et al. 2017)

significance test vs. first 17 novels
\( t = -6.1, \text{ df}=5.52, \ p = .0012^{**} \)
Cross-validated measures depend on fold size!

**Frequency spectrum at 100k tokens**

- $E[V_m]$ vs. $m$
- $E[V_m]$ at different $a$ values ($a = 3.0$, $a = 2.0$, $a = 1.5$)

**Type-token ratio (TTR)**

- $E[y] = E[y_i]$ vs. fold size
- $E[y] = E[y_i]$ at different $a$ values ($a = 3.0$, $a = 2.0$, $a = 1.5$)

**Yule's K**

- $E[y] = E[y_i]$ vs. fold size
- $E[y] = E[y_i]$ at different $a$ values ($a = 3.0$, $a = 2.0$, $a = 1.5$)

**Honoré H**

- $E[y] = E[y_i]$ vs. fold size
- $E[y] = E[y_i]$ at different $a$ values ($a = 3.0$, $a = 2.0$, $a = 1.5$)
Cross-validated measures depend on fold size!

Frequency spectrum at 100k tokens

Yule’s K

Honoré H

Type–token ratio (TTR)
Cross-validated measures depend on fold size!

Frequency spectrum at 100k tokens

Type–token ratio (TTR)

Yule's K

Honoré H

Stefan Evert
Outline

Introduction
  Motivation
  Notation & basic concepts
  Zipf’s law
  First steps (zipfR)

LNRE models
  Population & samples
  The mathematics of LNRE

Applications & examples
  Productivity & lexical diversity
  Practical LNRE modelling
  Bootstrapping experiments
  LNRE as Bayesian prior

Challenges
  Model inference
  Zipf’s law
  Non-randomness
  Significance testing

Outlook
Research programme for LNRE models

- Improve efficiency & numerical accuracy of implementation
  - numerical integrals instead of differences of Gamma functions
  - better parameter estimation (gradient, aggregated spectrum)
- Analyze accuracy of LNRE approximations
  - comprehensive simulation experiments, esp. for small samples
- Specify more flexible LNRE population models
  - my favourite: piecewise Zipfian type density functions
  - Baayen (2001): mixture distributions (different parameters)
- Develop hypothesis tests & confidence intervals
  - key challenge: goodness-of-fit vs. confidence region
  - prediction intervals for model-based extrapolation
- Simulation experiments for productivity measures
  - Can we find a quantitative measure that is robust against confounding factors and corresponds to intuitive notions of productivity & lexical diversity?
Research programme for LNRE models

- Is non-randomness a problem?
  - not for morphological productivity ➔ ECHO correction
  - tricky to include explicitly in LNRE approach

- Do we need LNRE models for practical applications?
  - better productivity measures + empirical sampling variation
  - based on cross-validation approach (Evert et al. 2017)

- How important is semantics & context?
  - Does it make sense to measure productivity and lexical diversity purely in terms of type-token distributions?
  - e.g. register variation for morphological productivity
  - e.g. semantic preferences in productive slots of construction
  - type-token ratio ≠ complexity of author’s vocabulary
Thank you!
References I

Proceedings of the 29th Annual Meeting of the Association for Computational
Linguistics, pages 271–278.

Publishers, Dordrecht.

Baroni, Marco and Evert, Stefan (2005). Testing the extrapolation quality of word
frequency models. In P. Danielsson and M. Wagenmakers (eds.), Proceedings of
Corpus Linguistics 2005, volume 1, no. 1 of Proceedings from the Corpus
Linguistics Conference Series, Birmingham, UK. ISSN 1747-9398.

Baroni, Marco and Evert, Stefan (2007). Words and echoes: Assessing and mitigating
the non-randomness problem in word frequency distribution modeling. In
Proceedings of the 45th Annual Meeting of the Association for Computational
Linguistics, pages 904–911, Prague, Czech Republic.

Brainerd, Barron (1982). On the relation between the type-token and species-area

Cao, Yong; Xiong, Fei; Zhao, Youjie; Sun, Yongke; Yue, Xiaoguang; He, Xin; Wang,
Lichao (2017). Pow law in random symbolic sequences. Digital Scholarship in the
Humanities, 32(4), 733–738.


References III


