Aalto University
School of Science

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

DAMMOC
DATA MINING TOOLS
for changing modalities of communication

Are you talking Bernoulli to me? Comparing methods of assessing word frequencies

Jefrey Lijffijt *

Joint work with Panagiotis Papapetrou *, Tanja Säily **,
Kai Puolamäki *, Terttu Nevalainen **, and Heikki Mannila *

* Department of Information and Computer Science, Aalto University
** Department of Modern Languages, University of Helsinki

Problem setting

- Given two corpora *S* and *T*
- Find all words that are *significantly* more frequent in *S* than in *T*, or vice versa

Word	Freq in *S*	Freq in *T*
mr	1637	2084
Total	567.138	656.714

- Is this statistically significant*?*

Motivation

- Find differences between groups
  - Speaker groups of different ages
    - S = 20−30 , T = 40−50
  - Genres
    - S = newspaper, T = magazines
  - Author gender
    - S = male, T = female
  - Time periods
    - S = 1600−1639, T = 1640−1681

Data

- Text = sequence of words
- Corpus of Early English Correspondence (CEEC)
  - Version with normalized spelling
- Compare periods 1600-1639 and 1640-1681
  - 1.2+ M words
  - 3000+ letters
- Preprocessing: remove punctuation from all words

Problem setting

- Input:
  - Two corpora: *S* and *T*
  - A significance threshold: α $(0 < \alpha < 1)$
- Word *q* is *significant* at level α if and only if $p \leq \alpha$
- *p* gives the probability for *S* and *T* having equal means
  - Two-tailed tests: test direction separately

Log-likelihood ratio test (Dunning 1993)

- **Assume all words are independent**
  - Bag-of-words model
- Significance test using 2x2 table

Word	Freq in *S*	Freq in *T*
mr	1637	2084
Total	567.138	656.714

- $Bin(k,n,fr) = \binom{n}{k} fr^k (1-fr)^{n-k}$
- $\lambda = \frac{Bin(k_S,n_S,fr_{S+T}) \cdot Bin(k_T,n_T,fr_{S+T})}{Bin(k_S,n_S,fr_S) \cdot Bin(k_T,n_T,fr_T)}$
- $-2\log\lambda \sim \chi^2 \rightarrow p = 0.004$

Bag-of-words model (log-likelihood ratio test, χ^2-test, Fisher's exact test, binomial test)

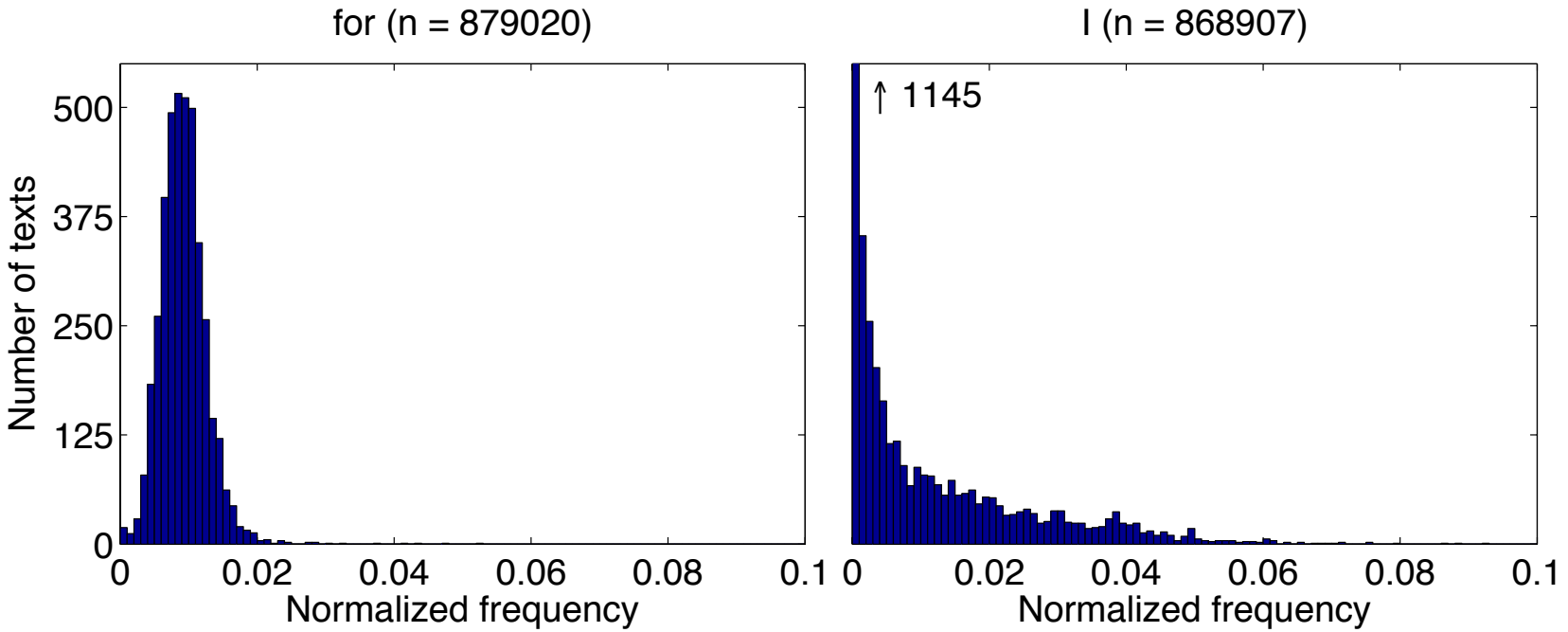
- **Assume all words are independent**
- *However:* texts have structure!

- Why the bag-of-words model then?
  - Mathematically simple
  - Computationally efficient

- Core questions:
  - Can we provide more realistic models?
  - Does it matter?

Many words are bursty

Data: British National Corpus, 4049 texts



- Frequency distribution differs per word
  - Depends on frequency and word ‘type’

Proposed method 1: Inter-arrival times (Lijffijt et al. 2011)

- Count space between consecutive occurrences (of **and**)

Finnair believes that it will be able to resume its scheduled service to **and** from New York on Monday, after two days of cancellations caused by hurricane Irene. All three airports serving New York City have been closed because of the hurricane **and** Finnair was forced to cancel flights on Saturday **and** Sunday. The airline is not certain when its scheduled service can be resumed, but the assumption is that Monday afternoon's flight from Helsinki will depart. Some Finnair passengers whose final destination is not New York have been rerouted **and** some have delayed travel plans. The company has also offered ticket holders a refund. *YLE*

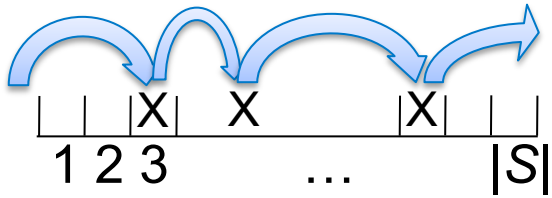
- $IAT_{and} = \{29, 9, 39, 29\}$

- Hypothesis: this captures the behavior pattern of words
  - Altmann et al. 2009: Predict word class based on IAT

Proposed method 1: Inter-arrival times (Lijffijt et al. 2011)

- Count space between consecutive occurrences
  → IAT distribution

- Resampling of *S* and *T*
  1. Pick random first index
  2. Sample random inter-arrival time
  3. Repeat 2. until size of *S* exceeded



- Produce random corpora: $S_1, \ldots, S_N$ and $T_1, \ldots, T_N$
- $$p_1 = \frac{\sum_{i=1}^{N} I(freq(q,S_i) \leq freq(q,T_i))}{N} \qquad p_2 = \frac{1 + N \cdot 2 \cdot \min(p_1, 1 - p_1)}{1 + N}$$

Proposed method 2: Bootstrapping (Lijffijt et al. 2011)

- Resampling based on word frequency distribution
  - Number of texts equal to number of texts in *S*

- Produce random corpora: $S_1, \ldots, S_N$ and $T_1, \ldots, T_N$

- $$p_1 = \frac{\sum_{i=1}^{N} I(freq(q,S_i) \leq freq(q,T_i))}{N}$$

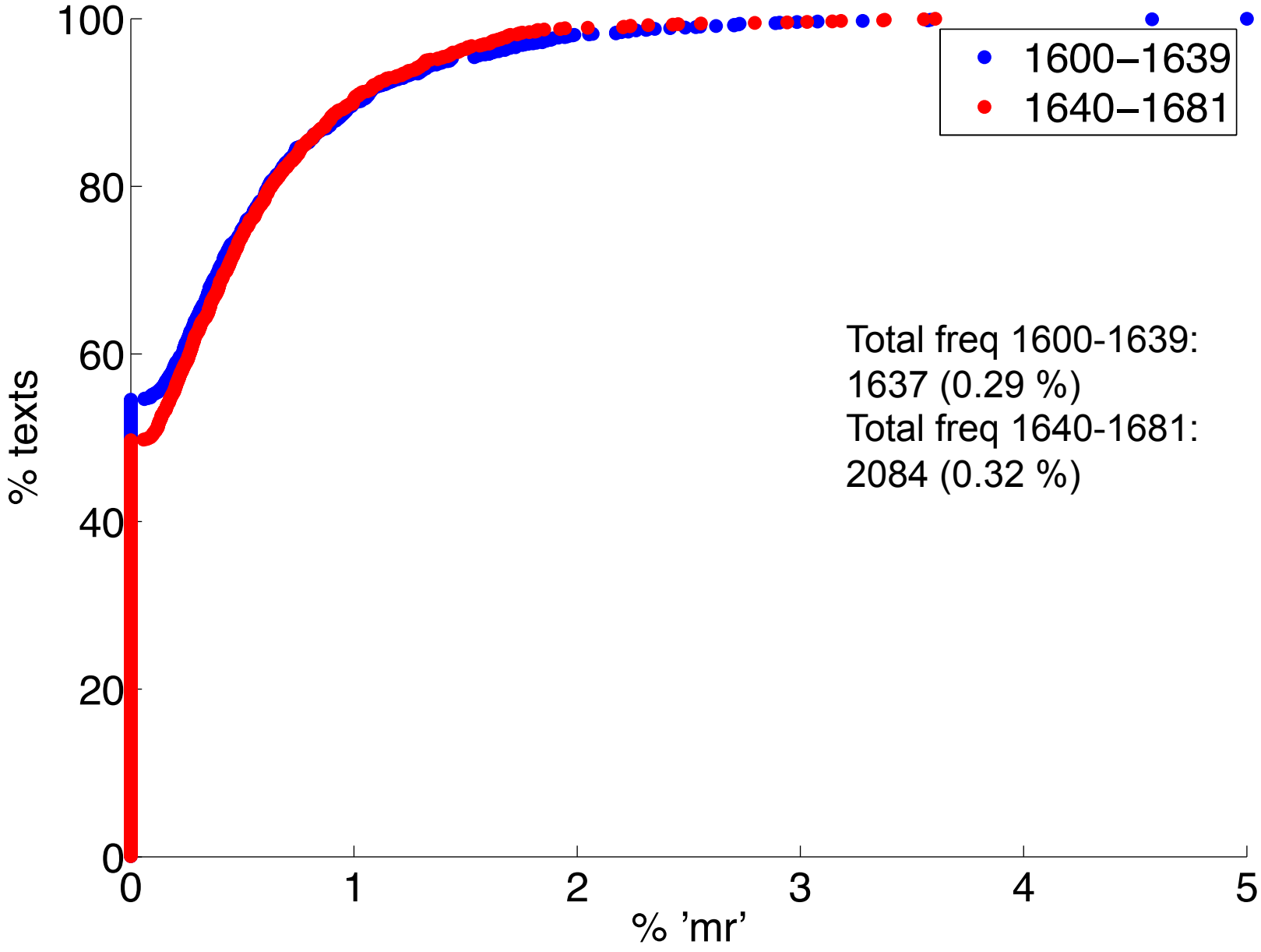
- $$p_2 = \frac{1 + N \cdot 2 \cdot \min(p_1, 1 - p_1)}{1 + N}$$

Comparison for *mr*

Word	Freq in *S*	Freq in *T*
mr	1637	2084
Total	567.138	656.714

- $p_{\chi 2}$ = 0.0043
- $p_{\text{log-likelihood}}$ = 0.0040
- p_{IAT} = 0.0747
- $p_{\text{bootstrap}}$ = 0.1043

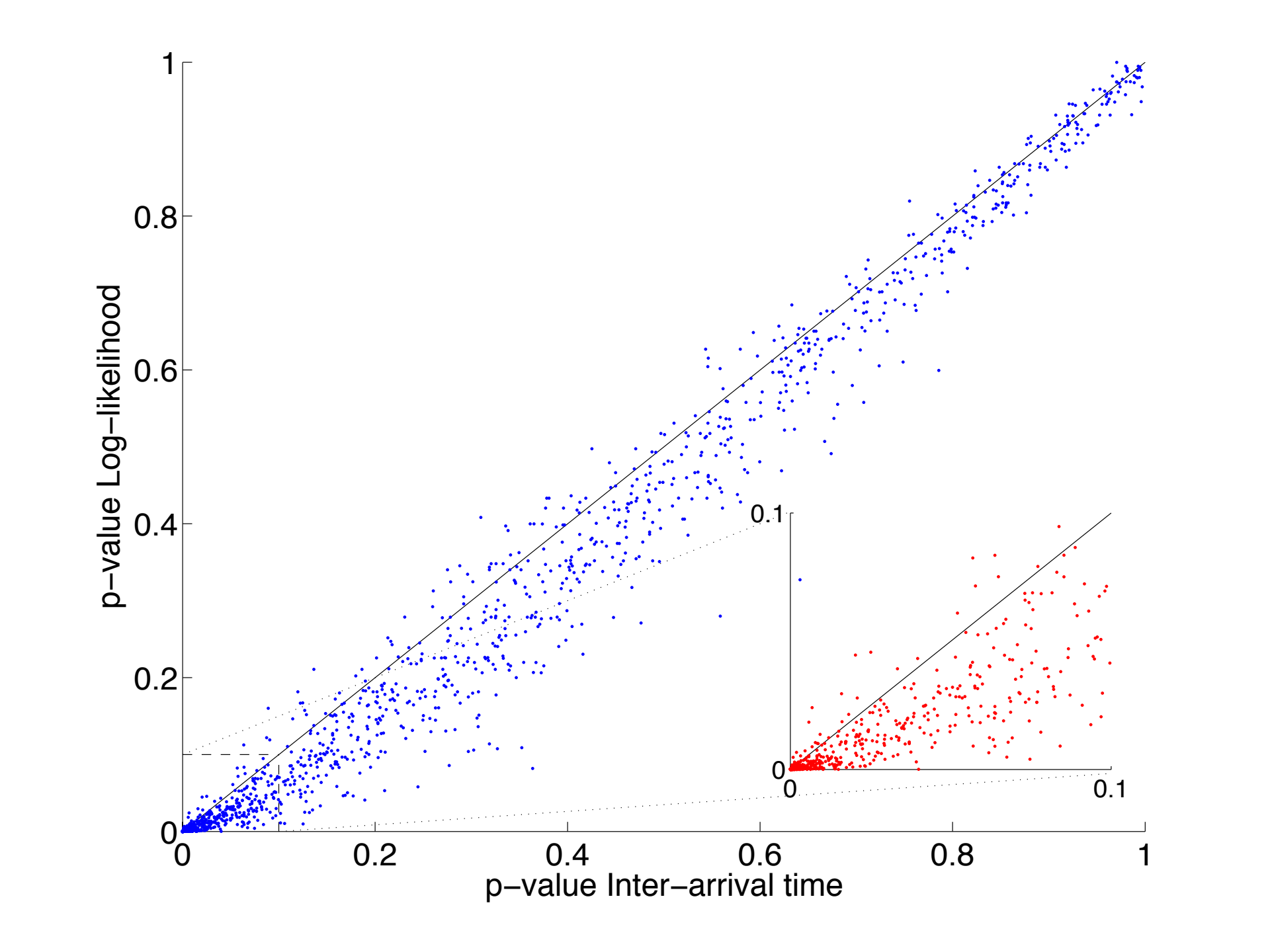
- Maybe the difference is not so significant!

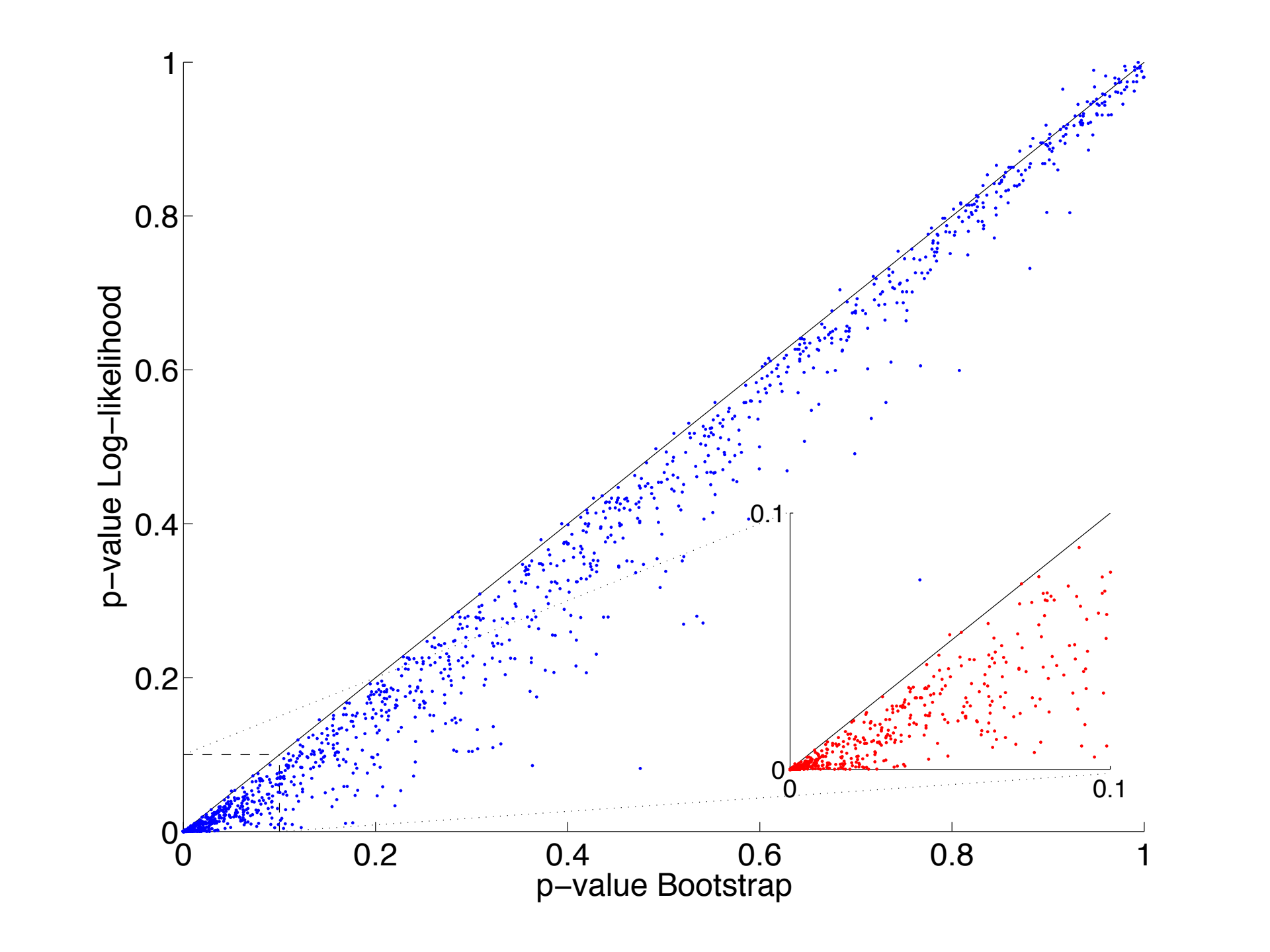

100
80
60
40
20
0
% texts
0
1
2
3
4
5
% 'mr'
1600–1639
1640–1681
Total freq 1600-1639:
1637 (0.29 %)
Total freq 1640-1681:
2084 (0.32 %)

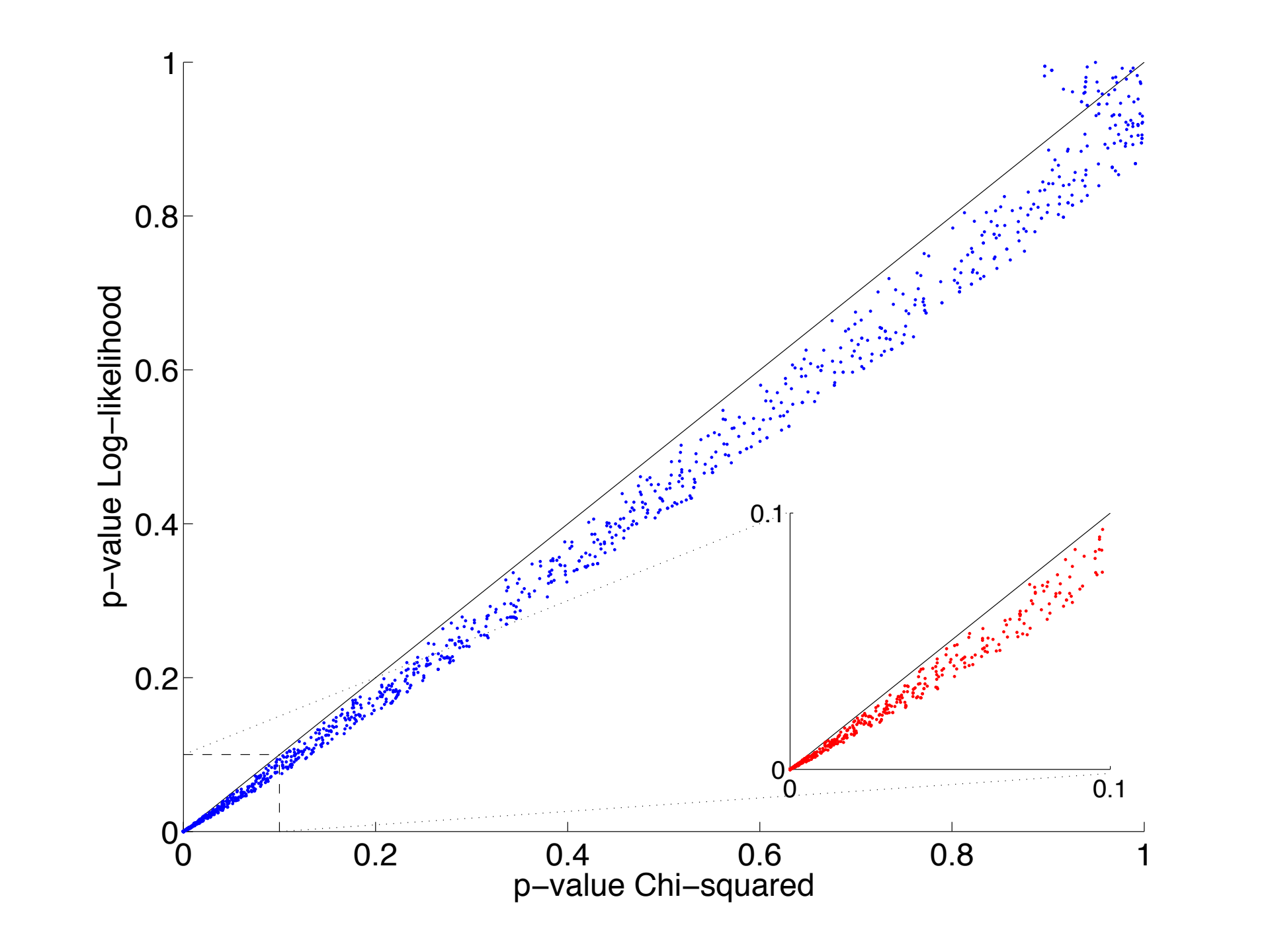
Experiments

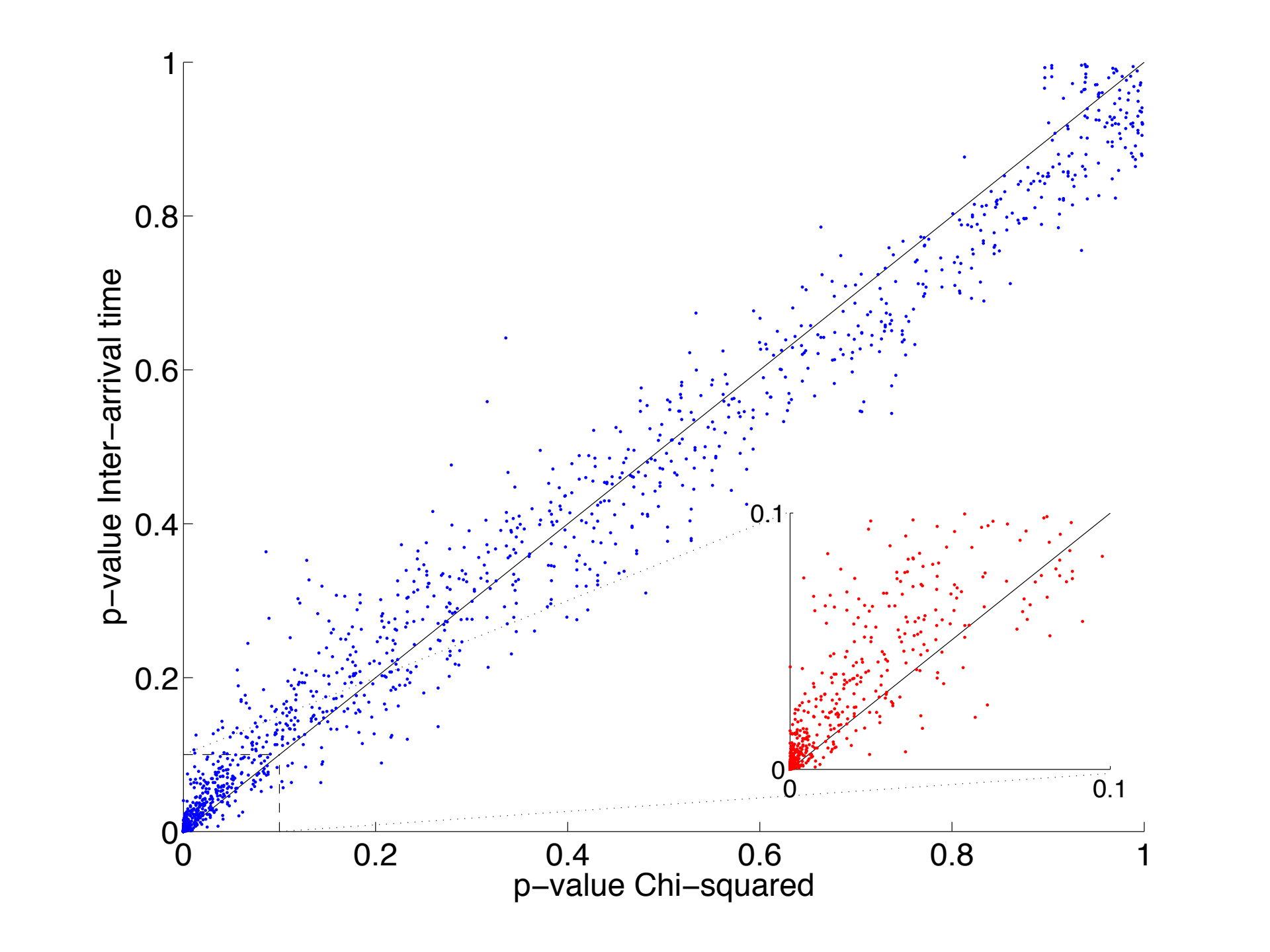
- Compare four methods
  - χ^2-test
  - Log-likelihood ratio test
  - Inter-arrival time test
  - Bootstrap test

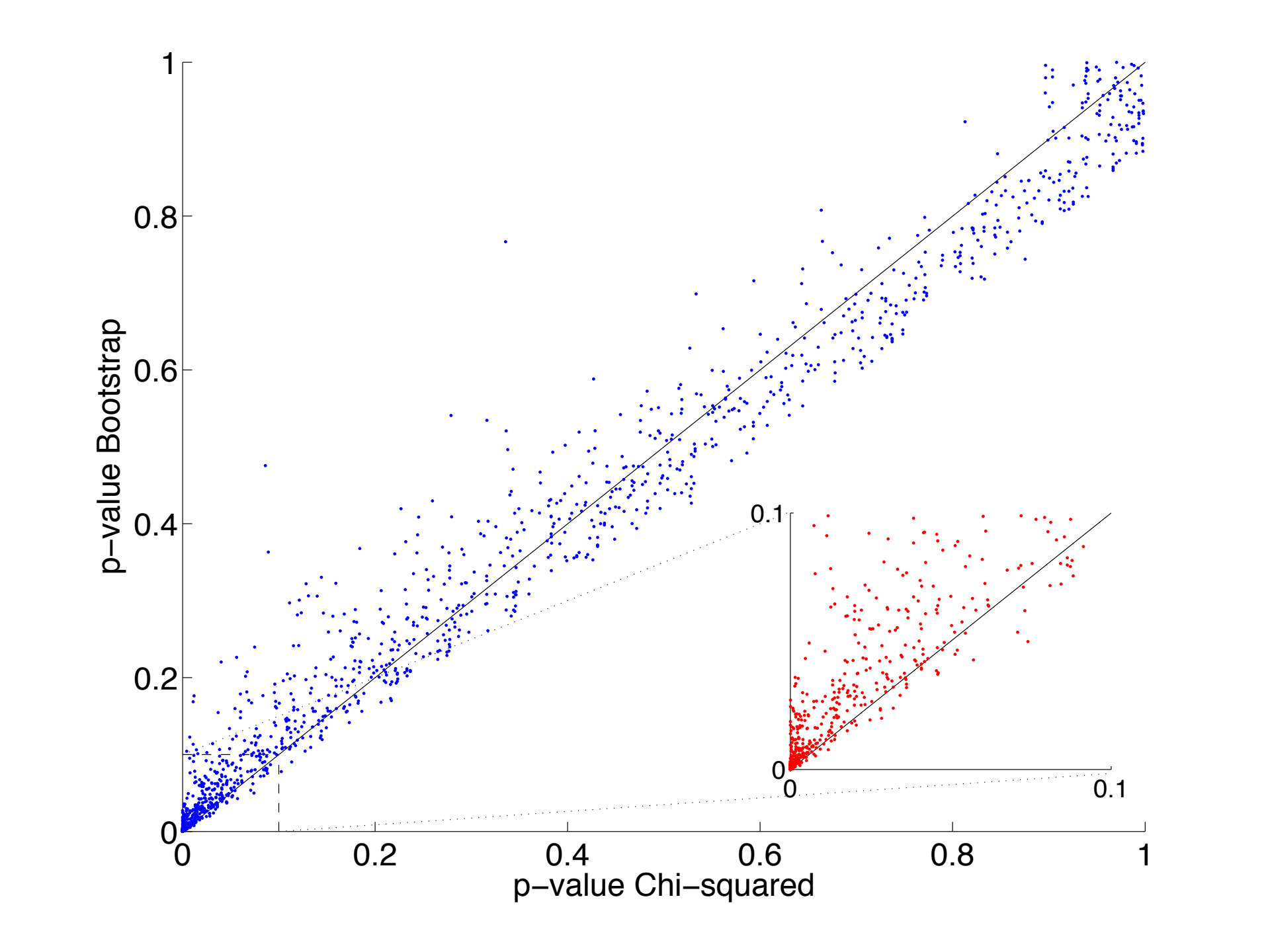
- Compute *p-values* for all words
  - Hypothesis: mean frequency in 1600-1639 and 1640-1681 are equal

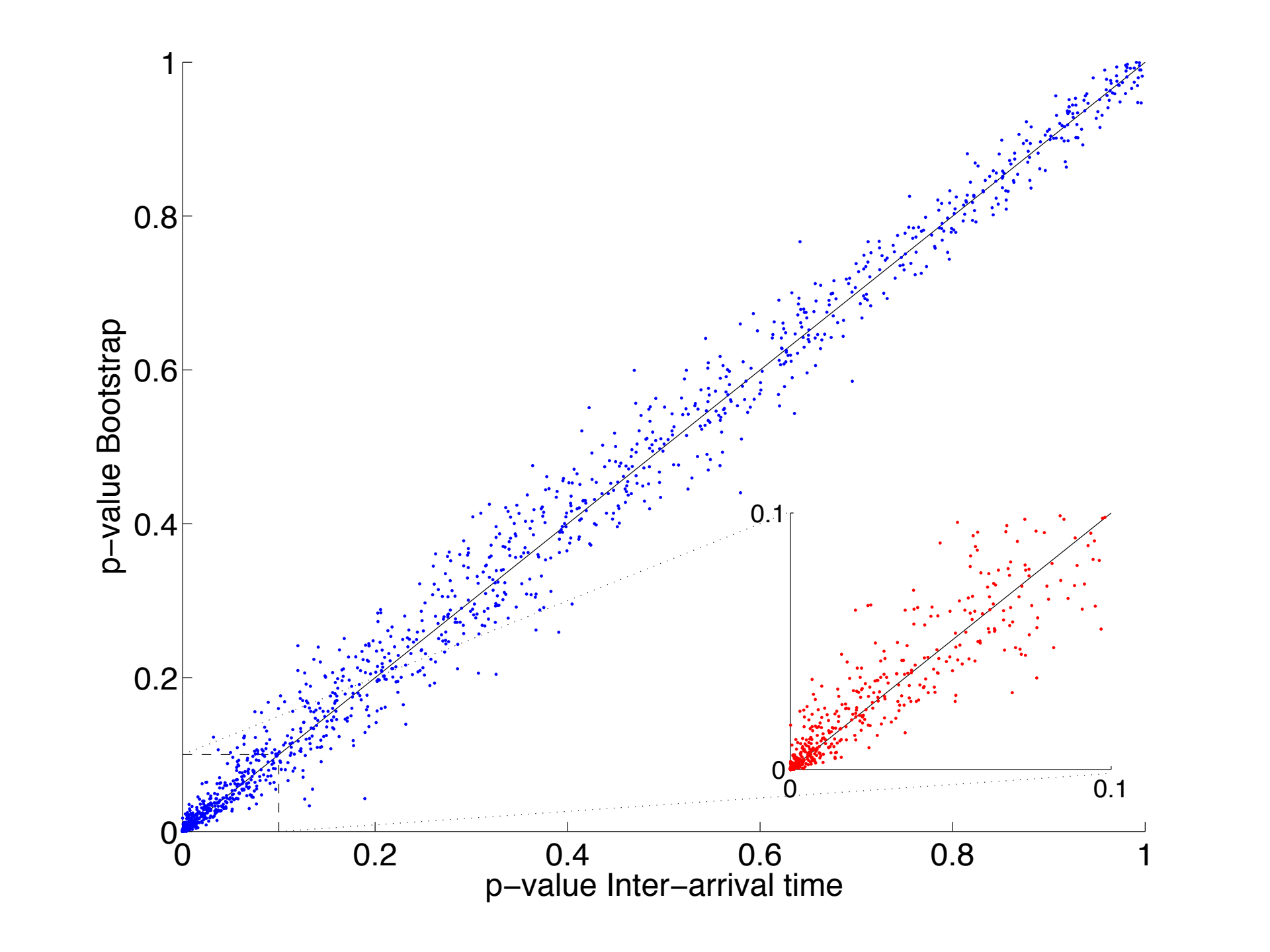

p−value Log−likelihood
p−value Inter−arrival time
0
0.2
0.4
0.6
0.8
1
0.1


p−value Log−likelihood
p−value Bootstrap
0
0.2
0.4
0.6
0.8
1
0
0.1


p−value Log−likelihood
p−value Chi−squared
0
0.2
0.4
0.6
0.8
1
0.1


p−value Inter−arrival time
p−value Chi−squared
0
0.2
0.4
0.6
0.8
1
0.1


p−value Bootstrap
p−value Chi−squared
0
0.2
0.4
0.6
0.8
1
0.1


p−value Bootstrap
p−value Inter−arrival time
0
0.2
0.4
0.6
0.8
1
0.1

Examples of words with > 10-fold difference

Word	% 1600-1639	% 1640-1681	p LL	p Boot
him	0.540	0.507	.011	.17
we	0.280	0.305	.011	.18
mr	0.289	0.317	.0040	.10
horse	0.019	0.026	.0091	.091
prince	0.027	0.020	.0065	.076
goods	0.013	0.019	.0091	.099
patent	0.008	0.004	.0051	.12
pound	0.019	0.013	.0090	.11
merchant	0.007	0.004	.010	.11
li	0.007	0.003	.0048	.095

Conclusion

- Bag-of-words model poorly represents frequency distributions
- New methods: inter-arrival times and bootstrap method (Lijffijt et al. 2011)
  - Take into account *burstiness* (or dispersion) of words
  - More conservative p-values
- Not covered in presentation:
  - Correction for multiple hypotheses
- Future work:
  - More in-depth analysis of differences between methods
  - Dependency between words

References

- **Altmann, E.G., Pierrehumbert, J.B., Motter, A.E.** (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distribution of words, *PLoS ONE*, **4**(11):e7678.
- **Dunning, T.** (1993). Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, **19**:61-74.
- **Lijffijt, J., Papapetrou, P., Puolamäki, K., Mannila, H.** (2011). Analyzing Word Frequencies in Large Text Corpora using Inter-arrival Times and Bootstrapping. In *ECML PKDD 2011, Part II*, pp. 341–357.

- http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/standardized.html
- http://users.ics.tkk.fi/lijffijt/
- http://tauchi.cs.uta.fi/virg/projects.html

Most frequent significant words

Word	% 1600-1639	% 1640-1681	p LL	p Boot
my	1.75	1.45	< .0001	.0001
that	1.48	1.72	< .0001	.0001
your	1.38	1.12	< .0001	.0001
it	1.16	1.33	< .0001	.0001
is	0.92	1.04	< .0001	.0001
and	3.35	2.95	< .0001	.0001
with	0.89	0.79	< .0001	.0001
but	0.78	0.95	< .0001	.0001
in	1.50	1.74	< .0001	.0001

- 292 words significant at $\alpha = 0.05$ in all four methods