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? Significance testing and burstiness of words in text corpora

Jefrey Lijffijt
Doctoral Student and Researcher
Department of Information and Computer Science, Aalto University

Helsinki Institute for Information Technology (HIIT)
Finnish Centre of Excellence for Algorithmic Data Analysis Research (ALGODAN)
Finnish Doctoral Programme in Computational Sciences (FICS)

My research group

- **Panagiotis Papapetrou** (post-doc)
- **Kai Puolamäki** (acting group leader)
- **Heikki Mannila** (vice president)

  Department of Information and Computer Science

  Aalto University

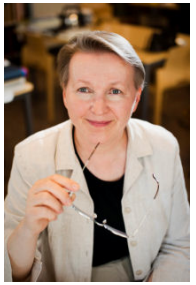


Outside collaborators

- **Tanja Säily** (PhD student)
- **Terttu Nevalainen** (academy professor)

  Department of Modern Languages

  University of Helsinki

Background research group

- Main topic: (algorithmic aspects of) data mining
- Pattern mining
  - binary matrices and sequences
- Significance testing of data mining results
  - Patterns, classification, clustering etc.
  - Permutation testing and constrained randomization
    - Swap randomization for binary matrices (Gionis et al. 2007)
    - Swap randomization for real-valued matrices (Ojala et al. 2008, Ojala 2010)

Outline

- Background and motivation
- Statistical tests for comparing corpora
- Comparing statistical tests for comparing corpora
- Directions for further research

Background and motivation

Initial motivation: repetitions of words

- General concept: **priming**
  - People repeat themselves and each other
  - Evidence for priming of words, phrases, syntax
  - Whole field of study
  - Open problems
    - Unclear at instance level
    - Unknown what ‘constructs’ can be primed
- Working concept: **persistence** = *unexpected* repetition
  - Interpretation free
- Main question: what is *(un-) expected?*

What is *(un-) expected?*

- Too general question, lots of factors in text/speech
  - Genre/register
  - Speaker/writer
    - Age
    - Gender
    - Background
  - Topic
  - Audience
  - …
- Huge natural variation

What is *(un-) expected?*

- Simpler questions, addressed today

1. Modeling expected behavior of words using inter-arrival times

2. How to take into account natural variation when comparing corpora

Definitions

- Corpus is a set of documents $S = \{D_1, \ldots, D_N\}$
- Document is a sequence of words $D_i = (D_{i,1}, \ldots, D_{i,|D_i|})$
- Frequency defined as $fr(w_j, D_i) = \sum_{j=1}^{|D_i|} I\left(w_i = D_{i,j}\right)$

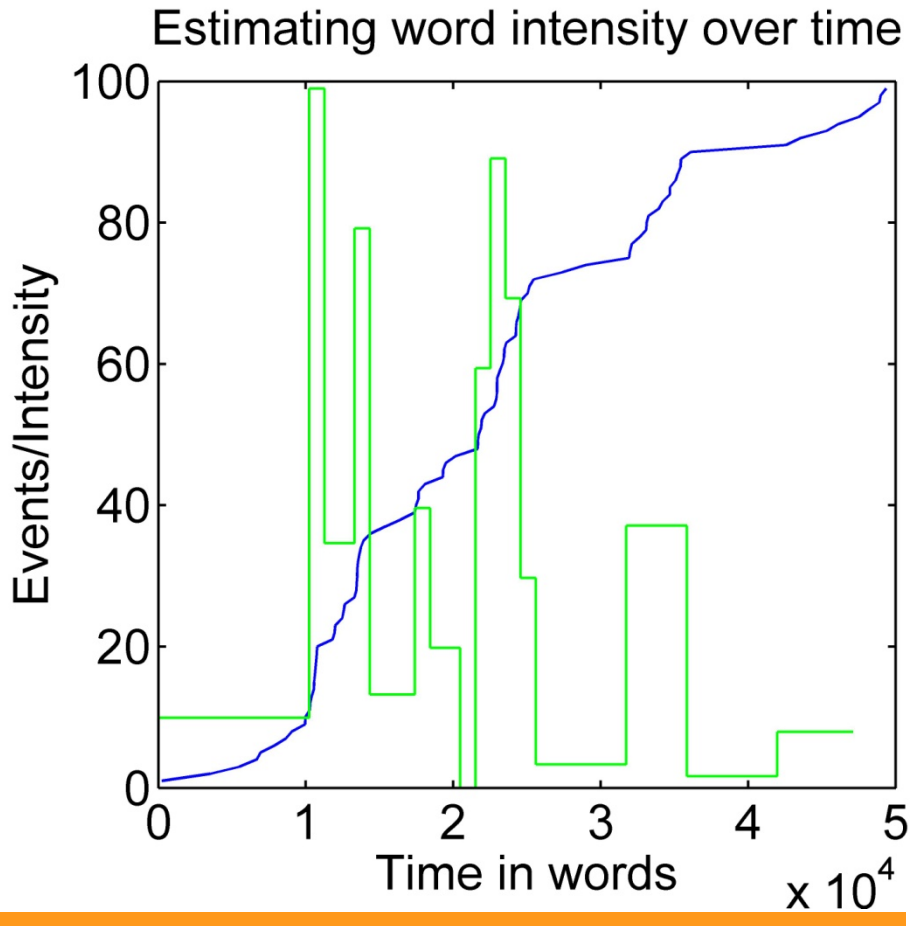
  $fr(w_j, S) = \sum_{i=1}^{|S|} fr(w_j, D_i)$
- Size defined as $size(S) = \sum_{i=1}^{N} |D_i|$

Definitions

- **Bag-of-words** assumption: words are generated from a multinomial distribution with fixed parameters
  - $p_1, \ldots, p_n$, such that $\sum_{i=1}^{n} p_i = 1$

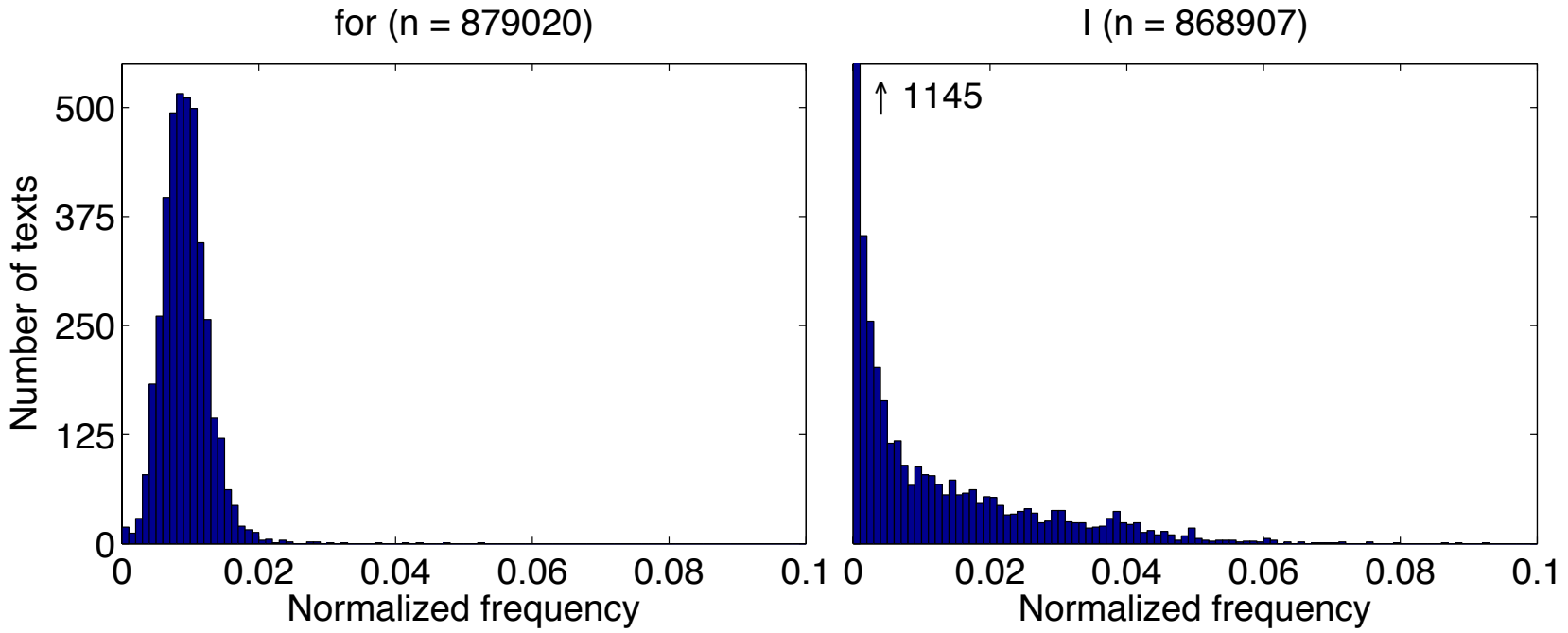
Example: text segmentation based on piecewise constant parts

- Blue line is incremented when we encounter an instance
- Green line is piecewise constant segmentation
- *Is the Poisson distribution a good model for this data?*



What is *(un-) expected?*

- Frequency distribution differs greatly per word
  - Depends on frequency and burstiness/dispersion



Data: British National Corpus, 4049 texts

Modeling expected behavior using inter-arrival times

- 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

Modeling expected behavior using inter-arrival times

- Denote the occurrence positions $q_i^1,\ldots,q_i^n = 14,43,52,91$
- j[th] inter-arrival time $a_{i,j} = q_i^{j+1} - q_i^j$, for $j = 1,\ldots,n-1$
- n[th] inter-arrival time $a_{i,n} = q_i^1 + size(S) - q_i^n$
- $size(S) = 106$
- IAT_{and} = {29, 9, 39, 29}

Modeling expected behavior using inter-arrival times (Altmann et al. 2009)

- Fit Weibull (stretched exponential) distribution to IAT

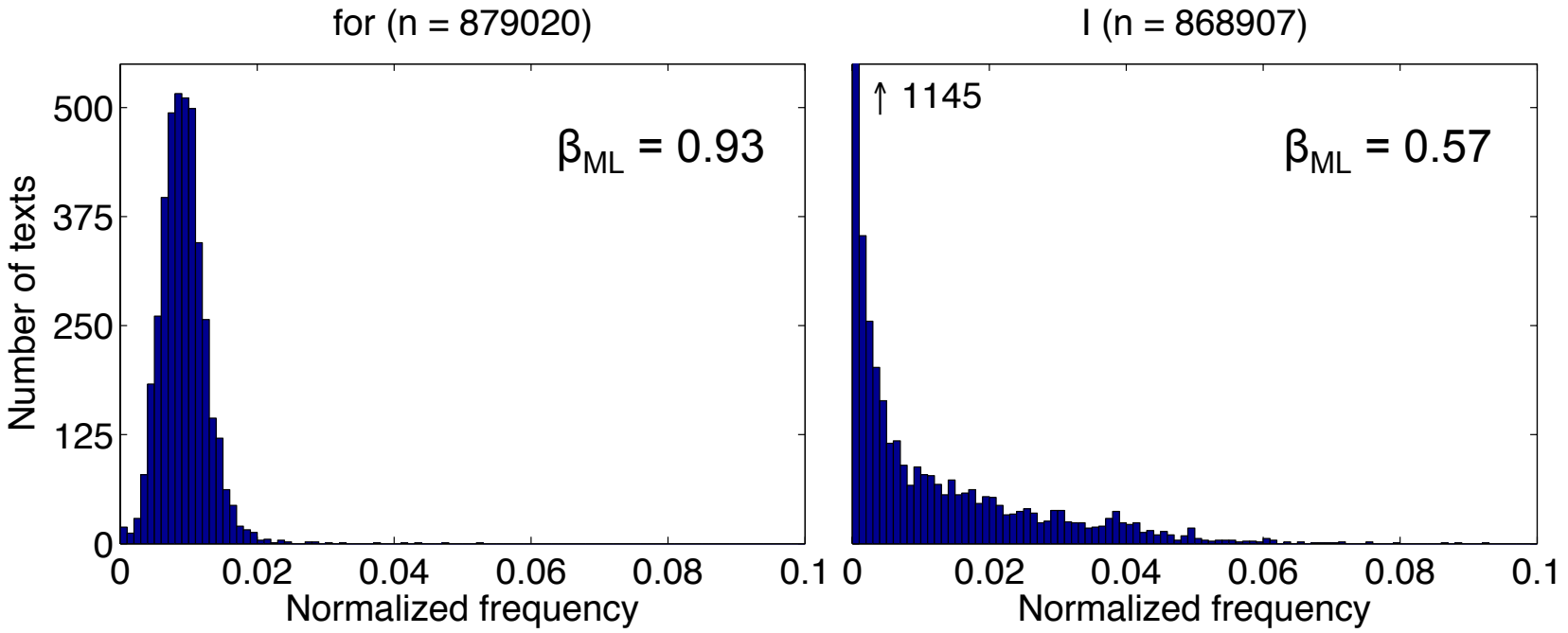
$$f(x) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^{\beta}}$$

  - $\alpha > 0$ is the scale parameter
  - $\beta > 0$ is the shape parameter
    - $\beta = 1 \rightarrow$ exponential distribution

- Predict word class based on β

What is *(un-) expected?*

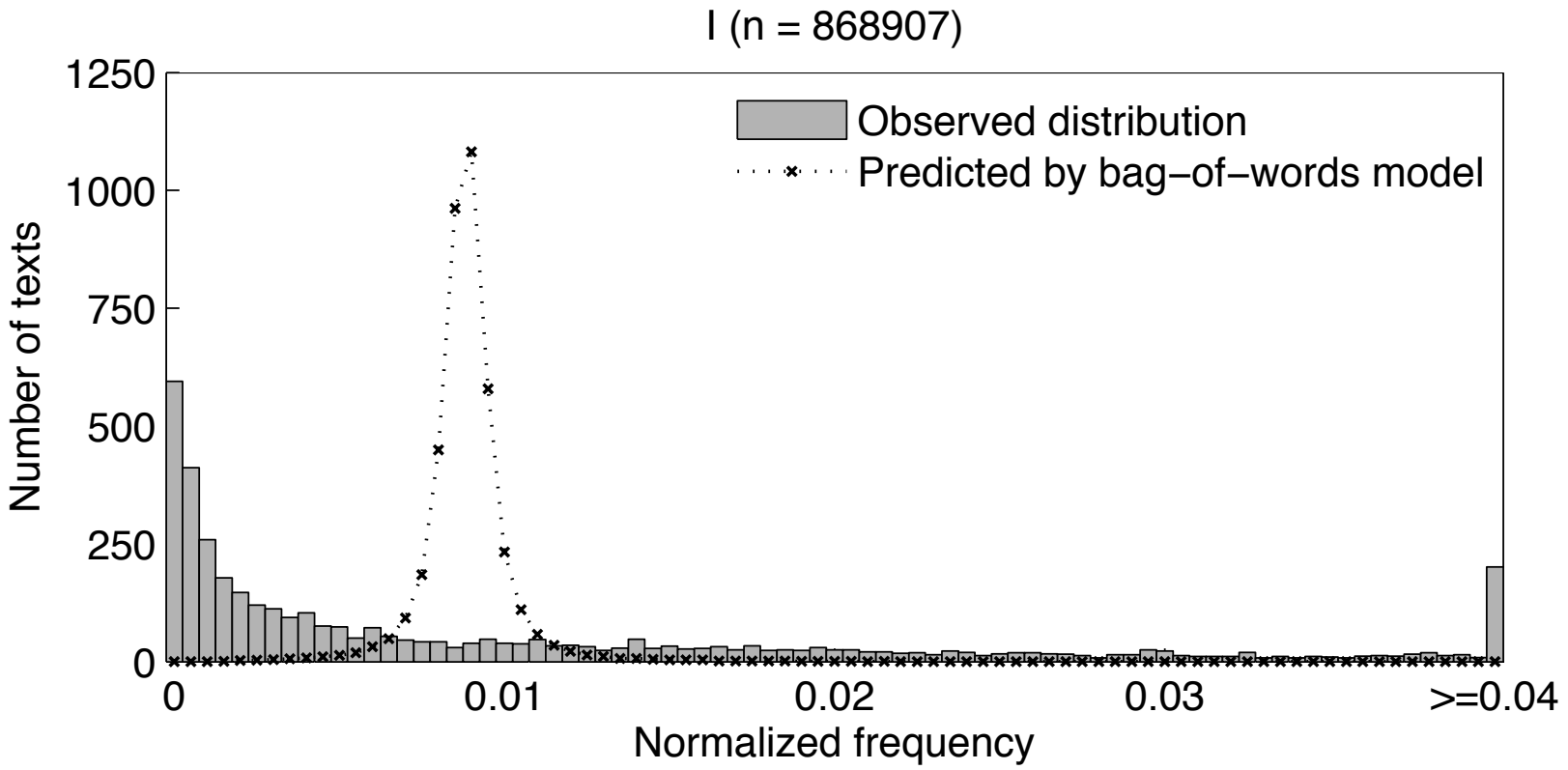
- Frequency distribution differs greatly per word
  - Depends on frequency and burstiness/dispersion



Data: British National Corpus, 4049 texts

What is *(un-) expected?*

- Bag-of-words prediction for *I* is quite poor



All models are wrong, but some are useful (G.E.P. Box, several articles)

- Linguists have done fine with statistical tests based on bag-of-words assumption
  - Paper on *log-likelihood test* for comparing corpora (Dunning 1993) has 1585 citations

- Information retrieval is based on burstiness of semantically rich words
  - Term frequency-inverse document frequency

$$idf(w_i) = \log \frac{|S|}{|\{D \in S : w_i \in D\}|}$$

Statistical tests for comparing corpora

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*
time	13,072	16,112
Total	7,196,688	8,365,458

- Is this difference 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

Definitions revisited

- Corpus is a set of documents $S = \{D_1, \ldots, D_N\}$
- Document is a sequence of words $D_i = (D_{i,1}, \ldots, D_{i,|D_i|})$
- Frequency defined as $fr(w_j, D_i) = \sum_{j=1}^{|D_i|} I\left(w_i = D_{i,j}\right)$

  $fr(w_j, S) = \sum_{i=1}^{|S|} fr(w_j, D_i)$
- Size defined as $size(D_i) = |D_i|$

  $size(S) = \sum_{i=1}^{N} |D_i|$

Definitions revisited

- Normalized frequency $\phi_{D_i}(w_j) = \frac{fr(w_j, D_i)}{size(D_i)}$

  and $\overline{\phi}_S(w_j) = \frac{\sum_{i=1}^{|S|} \phi_{D_i}(w_j)}{|S|}$ or $\phi_S(w_j) = \frac{fr(w_j, S)}{size(S)}$

- **Bag-of-words** assumption: words are generated from a multinomial distribution with fixed parameters
  - $p_1, \ldots, p_n$, such that $\sum_{i=1}^{n} p_i = 1$

- We are interested in only one parameter $p_i = \phi_S(w_i)$

Data (1/2)

- British National Corpus (BNC), XML edition (2007)
  - General language corpus
- Restrict to *fiction prose* genre
- Compare male *(S)* against female authors *(T)*
  - 7.2 M versus 8.4 M words
  - 203 versus 206 texts (stories and book parts)
- Preprocessing: remove punctuation, ignore multi-word tags, include titles etc.

Data (2/2)

- Corpus of Early English Correspondence (CEEC)
  - Contains personal letters
  - 1410 – 1681
  - Version with normalized spelling
- Compare periods 1600-1639 and 1640-1681
  - 1.2+ M words
  - 3000+ letters
- Preprocessing: remove punctuation from all words, include titles etc.

Formal 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 the normalized frequency of word *q* being equal in *S* and *T*
  - H_0: $\phi(q,S) = \phi(q,T)$
  - H_1: $\phi(q,S) \neq \phi(q,T)$
  - Two-tailed tests: test direction separately

Pearson's χ^2-test

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

Word	Freq in *S*	Freq in *T*
time	13,072	16,112
Total	7,196,688	8,365,458

- $X^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$
- With Yates' correction $X^2 = \sum_{i=1}^{2} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$
- $X^2 \sim \chi_1^2 \rightarrow p \approx 0.00000066$

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*
time	13,072	16,112
Total	7,196,688	8,365,458

- $Bin(k,n,p_i) = \binom{n}{k} {p_i}^k (1-p_i)^{n-k}$
- $\lambda = \frac{Bin(k_S,n_S,p_i^{S+T}) \cdot Bin(k_T,n_T,p_i^{S+T})}{Bin(k_S,n_S,p_i^{S}) \cdot Bin(k_T,n_T,p_i^{T})}$
- $-2\log\lambda \sim \chi_1^2 \rightarrow p \approx 0.00000061$

Welch's T-test

- Based on frequency distribution of *q* over documents
- Allows for unequal variances in the two sets
- $t = \frac{\bar{\phi}_S - \bar{\phi}_T}{\sqrt{\frac{s_S^2}{|S|} + \frac{s_T^2}{|T|}}}$, where *s* is the sample variance
- *t* ~ t-distribution with *v* degrees of freedom

Wilcoxon rank-sum test (Mann-Whitney U test)

- Based on frequency distribution of *q* over documents

- Order all texts by frequency
- R_1 and R_2 are the smaller and larger rank sums
- $$U = R_1 - \frac{N_1(N_1+1)}{2}$$
- For small U, consult table, for large $U \sim \mathrm{N}(\mu_U, \sigma_U)$

$$\mu_U = \frac{N_1 N_2}{2},\ \sigma_U = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}$$

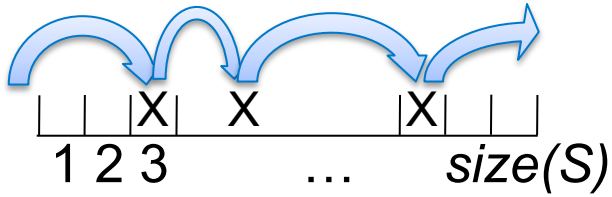
Inter-arrival time test (Lijffijt et al. 2011)

- Produce random corpora: $S'_1, \ldots, S'_R$ and $T'_1, \ldots, T'_R$
- Use two-tailed version of empirical *p*-value* (North et al. 2002)
- $$p_1 = \frac{\sum_{i=1}^{R} H(fr(q,S_i') \leq fr(q,T_i'))}{R}$$

  where $$H(x \leq y) = \begin{cases} 1 & if\ x < y \\ 0.5 & if\ x = y \\ 0 & if\ x > y \end{cases}$$
- $$p_2 = \frac{1 + R \cdot 2 \cdot \min(p_1, 1 - p_1)}{1 + R}$$

Inter-arrival time test (Lijffijt et al. 2011)

- Count space between consecutive occurrences
  → IAT distribution



- Resampling of *S* and *T*
  1. Pick random first index from $g(x)$
  2. Sample random inter-arrival time from $f(x)$
  3. Repeat 2. until size of *S* exceeded

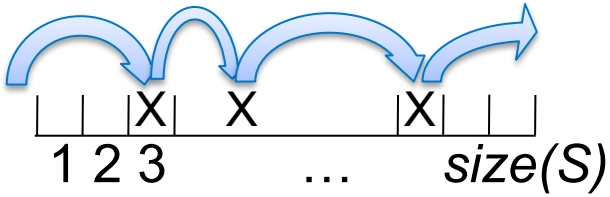
- $f(x) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^{\beta}}$

$$G(x) = 1 - \frac{\Gamma\left(1+\frac{1}{\beta}, \left(\frac{x}{\alpha}\right)^{\beta}\right)}{\Gamma\left(1+\frac{1}{\beta}\right)}$$

- $g(x) = C \cdot x \cdot f(x)$ s.t. $\sum_x C \cdot x \cdot f(x) = 1$

Inter-arrival time test (Lijffijt et al. 2011)

- Count space between consecutive occurrences
  → IAT distribution



- Resampling of *S* and *T*
  1. Pick random first index from $g(x)$
  2. Sample random inter-arrival time from $f(x)$
  3. Repeat 2. until size of *S* exceeded

- Alternatively, $f(x)$ is the empirical IAT-distribution
  – Lijffijt et al. (2011) suggests Weibull is often not a good fit

- $g(x) = C \cdot x \cdot f(x)$ s.t. $\sum_x C \cdot x \cdot f(x) = 1$

Bootstrap test (Lijffijt et al. 2011)

- Resampling based on word frequency distribution
  - Sample |*S*| texts (with replacement) from *S* or *T*
- Produce random corpora: $S'_1, \ldots, S'_R$ and $T'_1, \ldots, T'_R$
- Use two-tailed version of empirical *p*-value
- $$p_1 = \frac{\sum_{i=1}^{R} H(\phi_{S_i'}(q) \leq \phi_{T_i'}(q))}{R}$$
- $$p_2 = \frac{1 + R \cdot 2 \cdot \min(p_1, 1 - p_1)}{1 + R}$$

Comparison for *time* ($\beta = 0.88$)

Word	Freq in *S*	Freq in *T*
time	13,072	16,112
Total	7,196,688	8,365,458

- $p_{\chi 2} = 0.00000066$
- $p_{\text{log-likelihood}} = 0.00000061$
- $p_{\text{t-test}} = 0.048$
- $p_{\text{rank-sum}} = 0.028$
- $p_{\text{IAT}} = 0.00030$
- $p_{\text{bootstrap}} = 0.021$

Example: frequency thresholds (Lijffijt et al. 2011)

- $\alpha = 0.01$ in a text of 2000 words

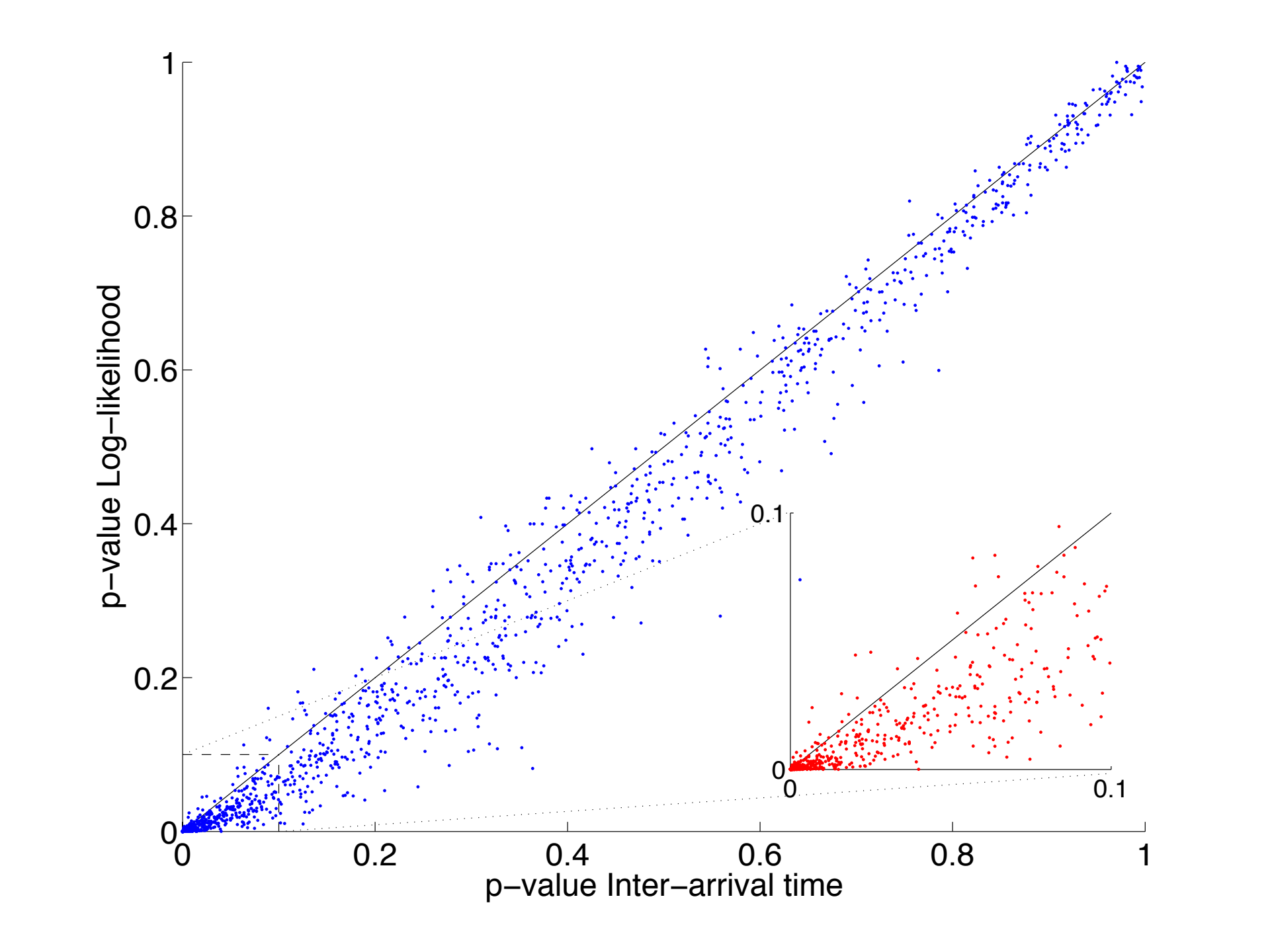
Word	Freq in BNC ($x10^6$)	Weibull β	Binomial	Weibull Inter-arrival	Bootstrap
a	2.2	1.01	61	61	72
for	0.9	0.93	29	30	37
I	0.9	0.57	29	48	110

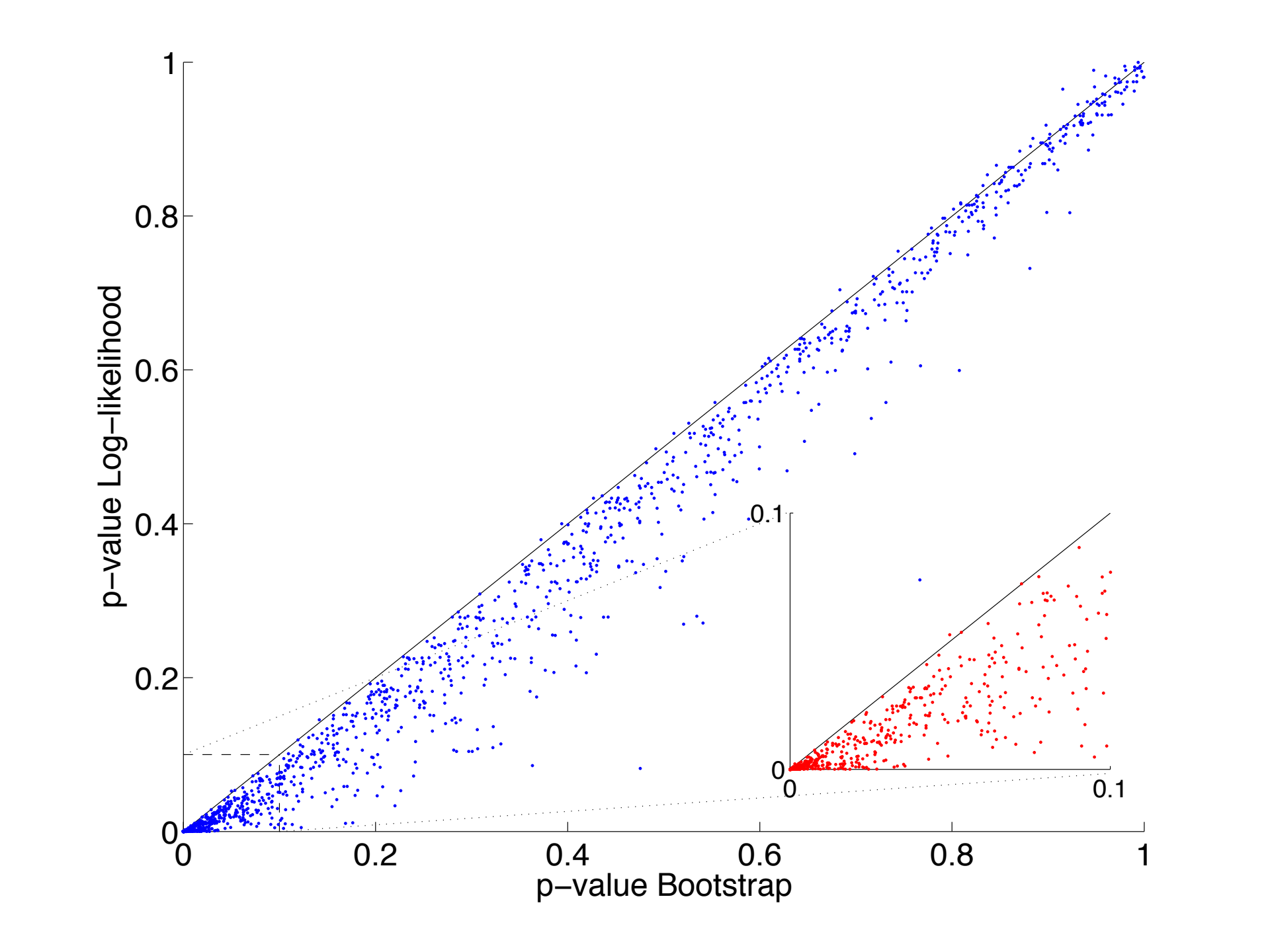
- Smaller β gives larger differences

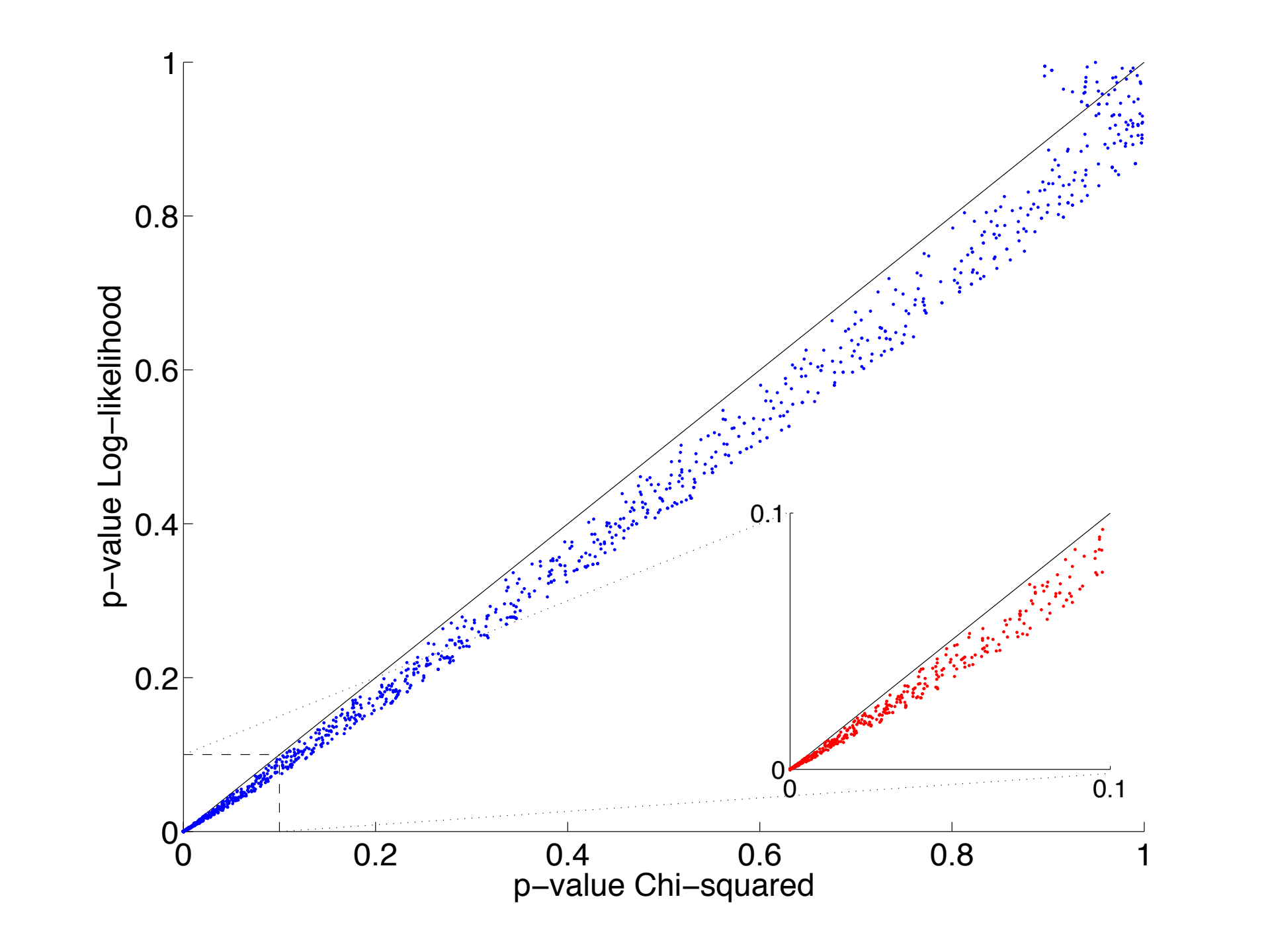
Experiments (Säily et al., forthcoming)

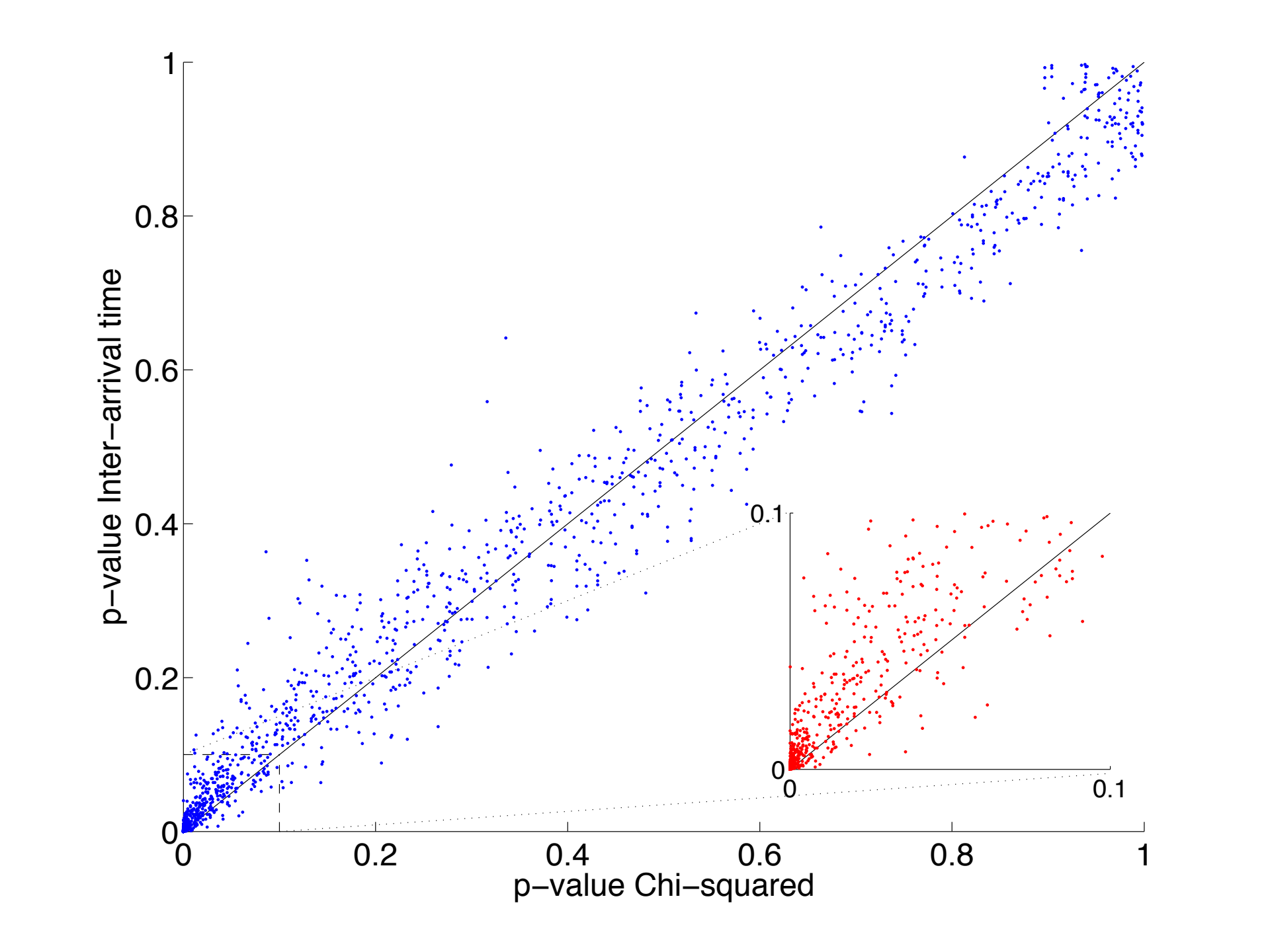
- Compare four methods using CEEC
  - χ^2-test
  - Log-likelihood ratio test
  - Inter-arrival time test
  - Bootstrap test

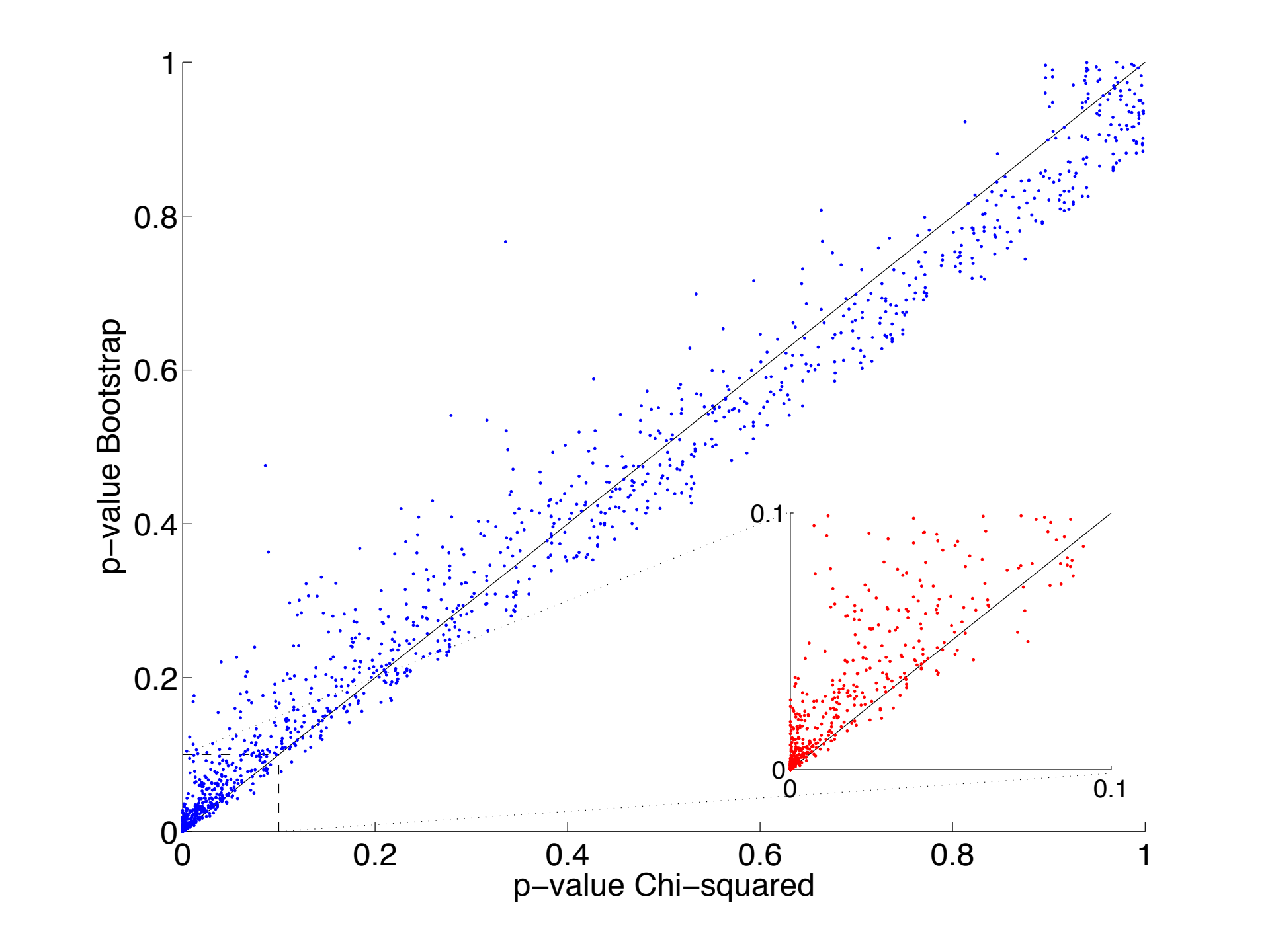
- Compute *p-values* for all words
  - H_0: normalized frequency in 1600-1639 and 1640-1681 are equal
  - H_1: normalized frequency in 1600-1639 and 1640-1681 are not equal

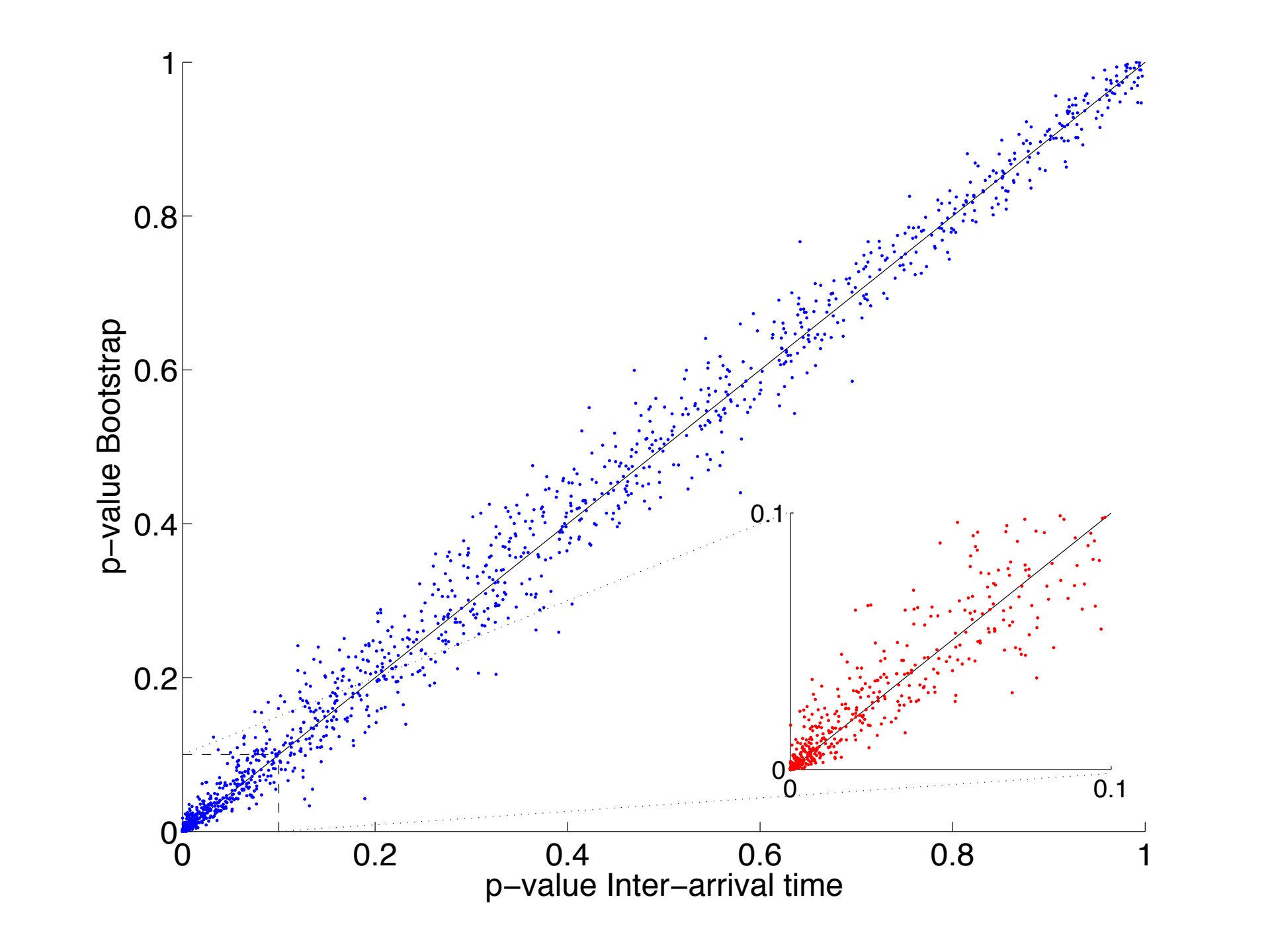

p−value Log−likelihood
p−value Inter−arrival time
0
0.2
0.4
0.6
0.8
1
0.1


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

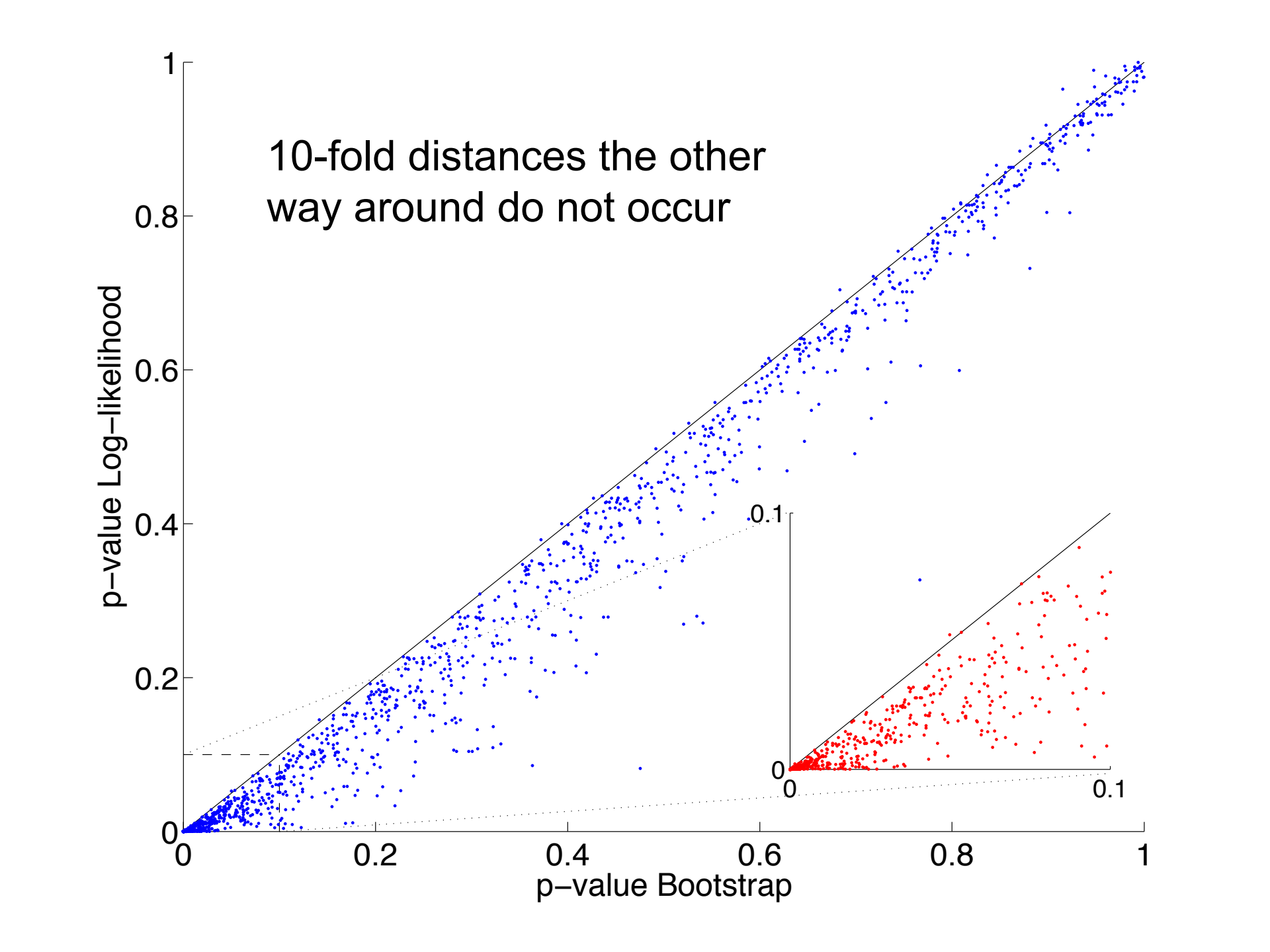

p−value Inter−arrival time
p−value Chi−squared
0
0.2
0.4
0.6
0.8
1
0.1


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


10-fold distances the other
way around do not occur
p−value Log−likelihood
p−value Bootstrap
0
0.2
0.4
0.6
0.8
1
0.1

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*

Conclusion part 1

- Bag-of-words model poorly represents frequency distributions (especially of *bursty* words)
- New models: inter-arrival times and bootstrap test
  - More conservative p-values
  - Weibull β predicts difference between models
- Not covered so far:
  - Correction for multiple hypotheses
- **How do we really know which test is *better*?**

Comparing statistical tests for comparing corpora

Uniformity of p-values

- By definition p-values should be uniform
  - $\Pr(p \leq x) = x$
- Test is *conservative* iff p-value are too high
  - $\Pr(p \leq x) < x$
- Test is *anti-conservative* iff p-values are too low
  - $\Pr(p \leq x) > x$

Testing uniformity of p-values

- Experiment using random splitting of data
  - BNC fiction-prose subcorpus (2000 word samples from each text)
- Do for each word with frequency ≥ 50
  - Assign half of the texts to *S* and other half to *T*
  - Compute p-value for each of the six methods
  - Repeat 500 times
- Data is generated under the null hypothesis
- Use Kolmogorov-Smirnov test to compute p-value for uniformity of p-values for each word for each test
  - Total of $3{,}302 \cdot 6 = 19{,}812$ p-values

Testing uniformity of p-values

- Use Kolmogorov-Smirnov test to compute p-value for uniformity of p-values for each word for each test
  - Total of $3{,}302 \cdot 6 = 19{,}812$ p-values
- Use Bonferroni correction for multiple hypothesis
  - We really do not want any false-positives
- $\alpha = 0.01$

DPnorm: normalized measure of dispersion (Gries 2008, Lijffijt & Gries *to appear*)

- Difference between expected and observed frequencies

- o_i = relative number of occurrences in D_i
- e_i = relative size of D_i

- $DP = \frac{\sum_{i=1}^{N} |o_i - e_i|}{2}$

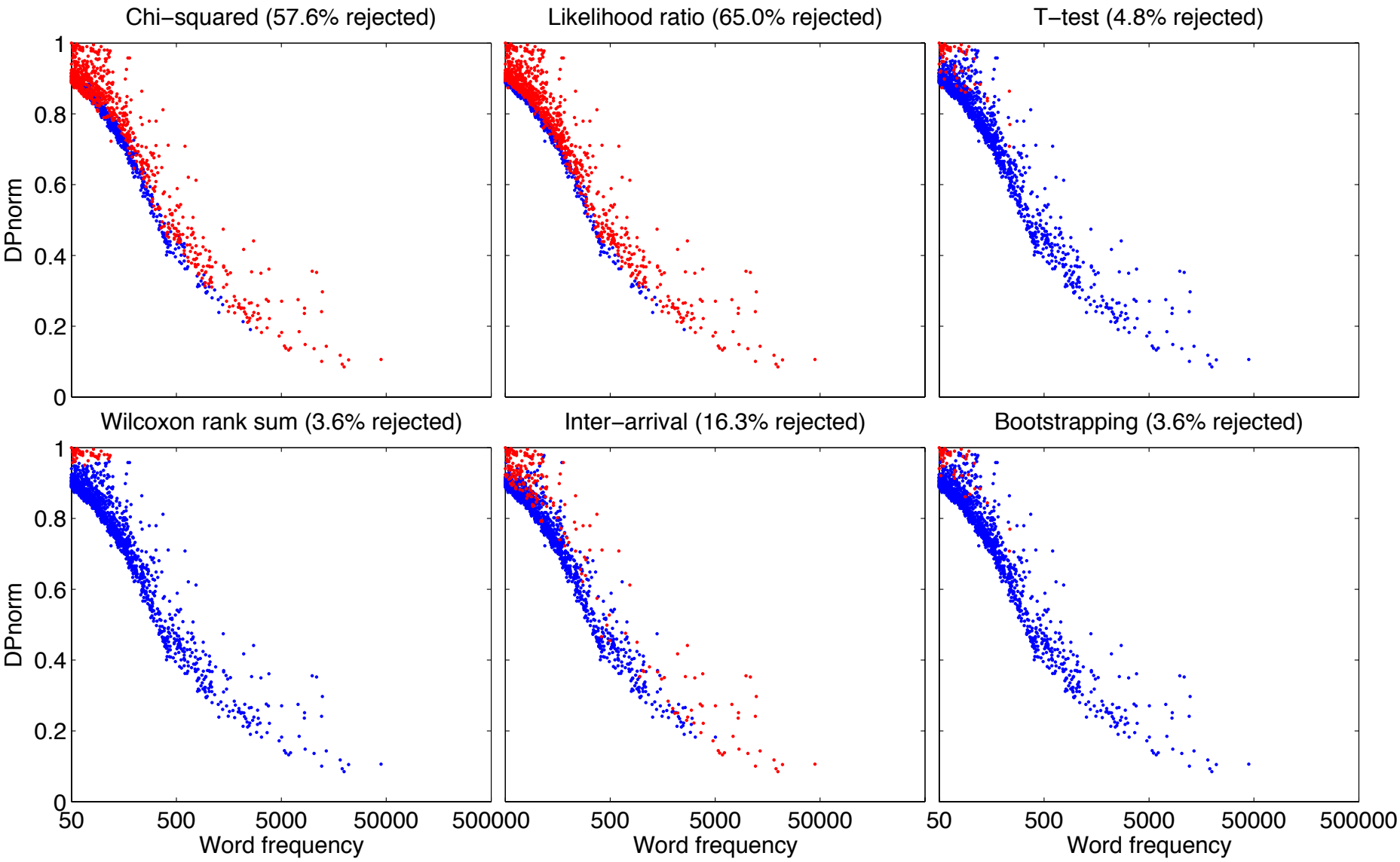
- $DP_{norm} = \frac{DP}{1 - \min_i(e_i)}$

e_i	0.4	0.4	0.2
o_i	0.0	0.0	1.0
$\lvert o_i - e_i \rvert$	0.4	0.4	0.8

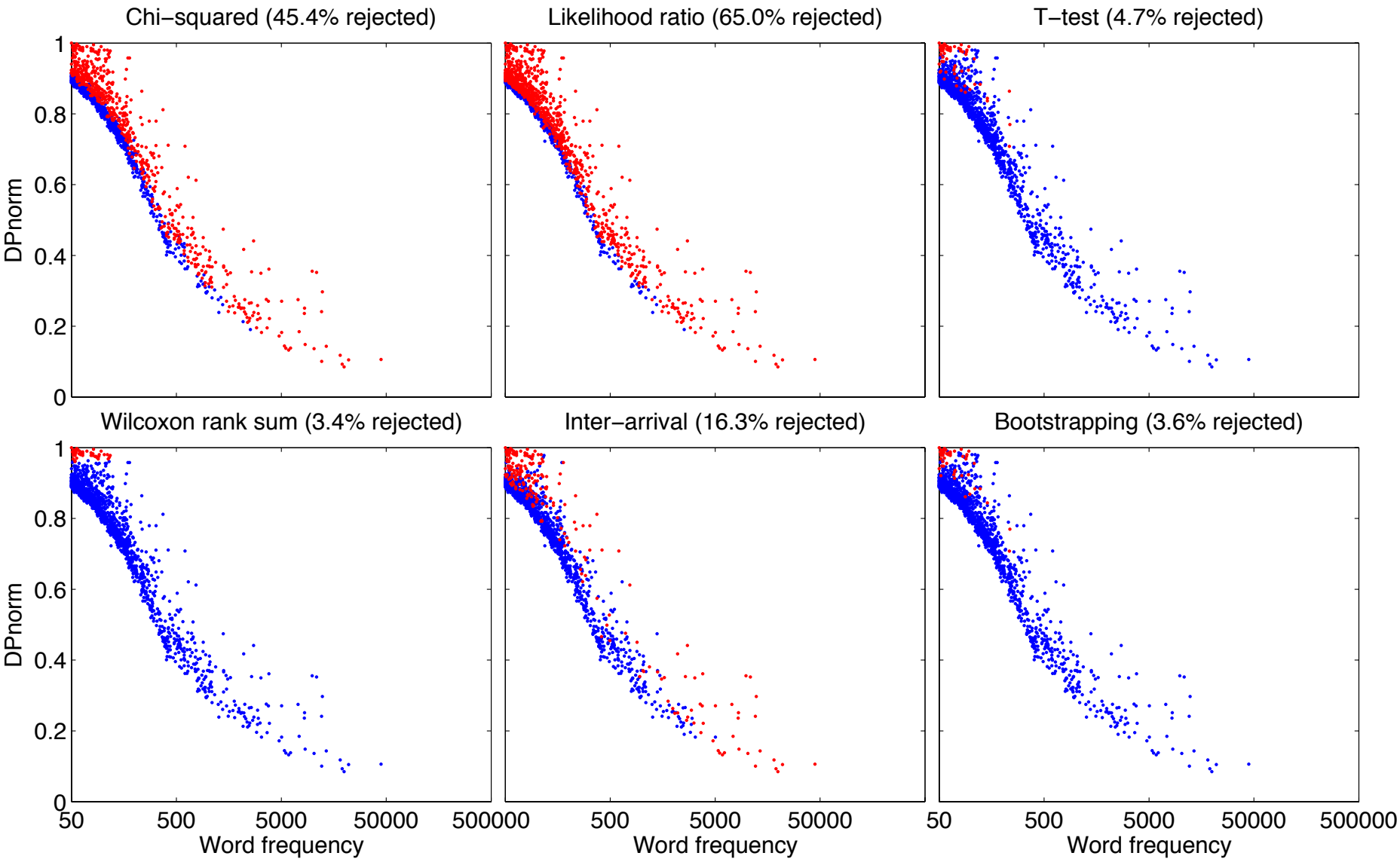
$$DP = 1.6/2 = 0.8$$

$$DP_{norm} = 0.8/(1 - 0.2) = 1$$

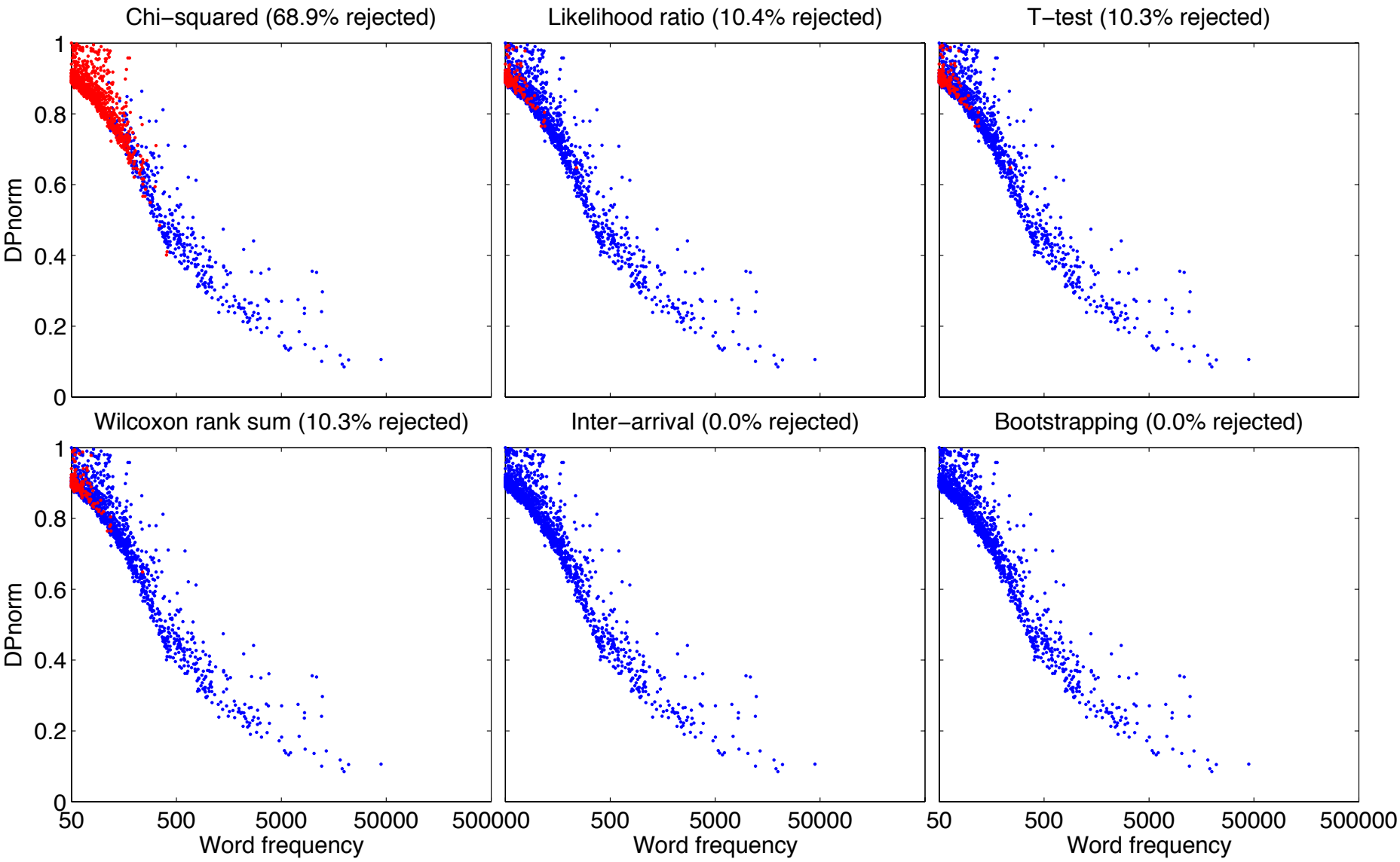
Uniformity under random text assignment



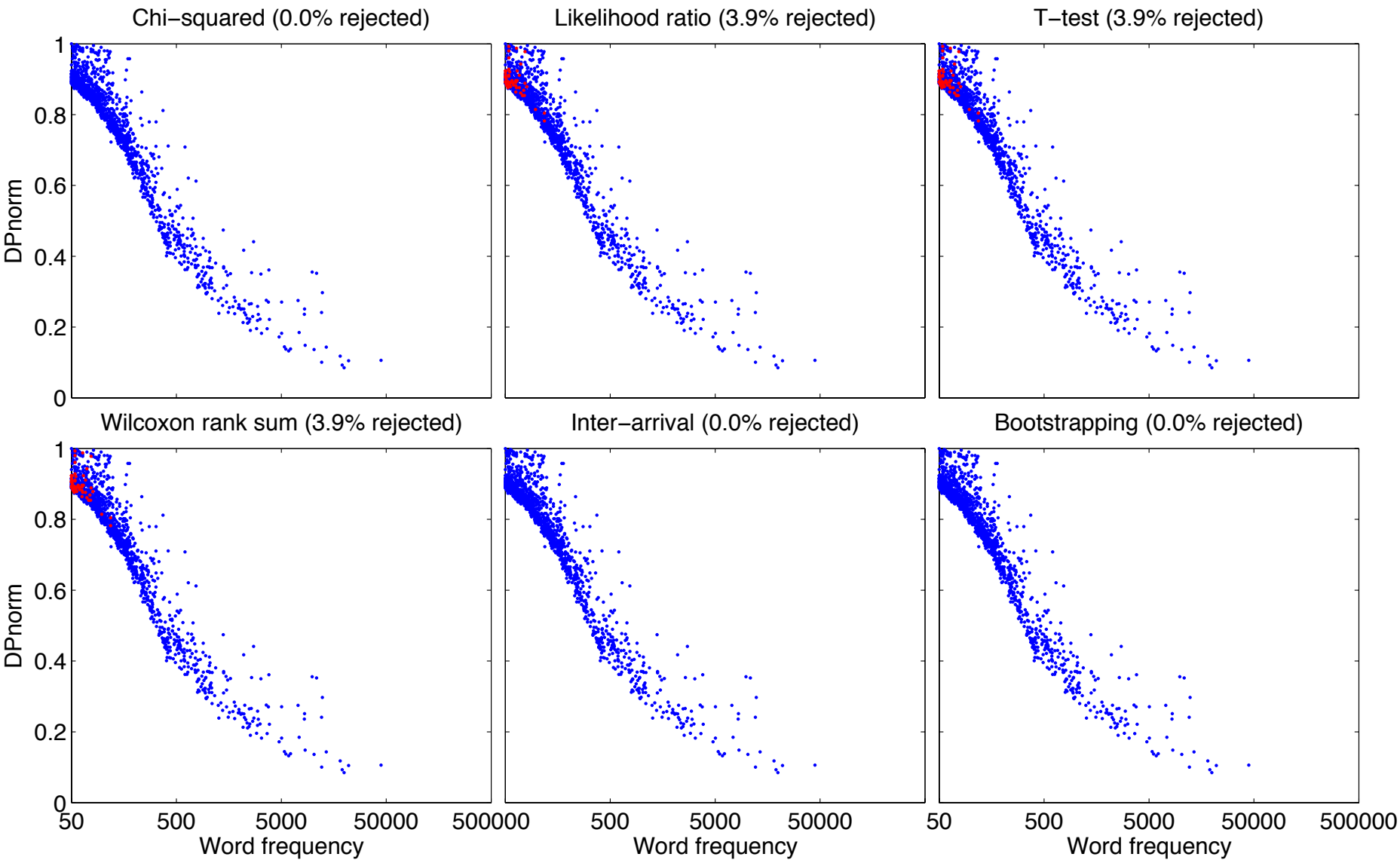
Anti-conservativeness under random text assignment



Uniformity under random word assignment

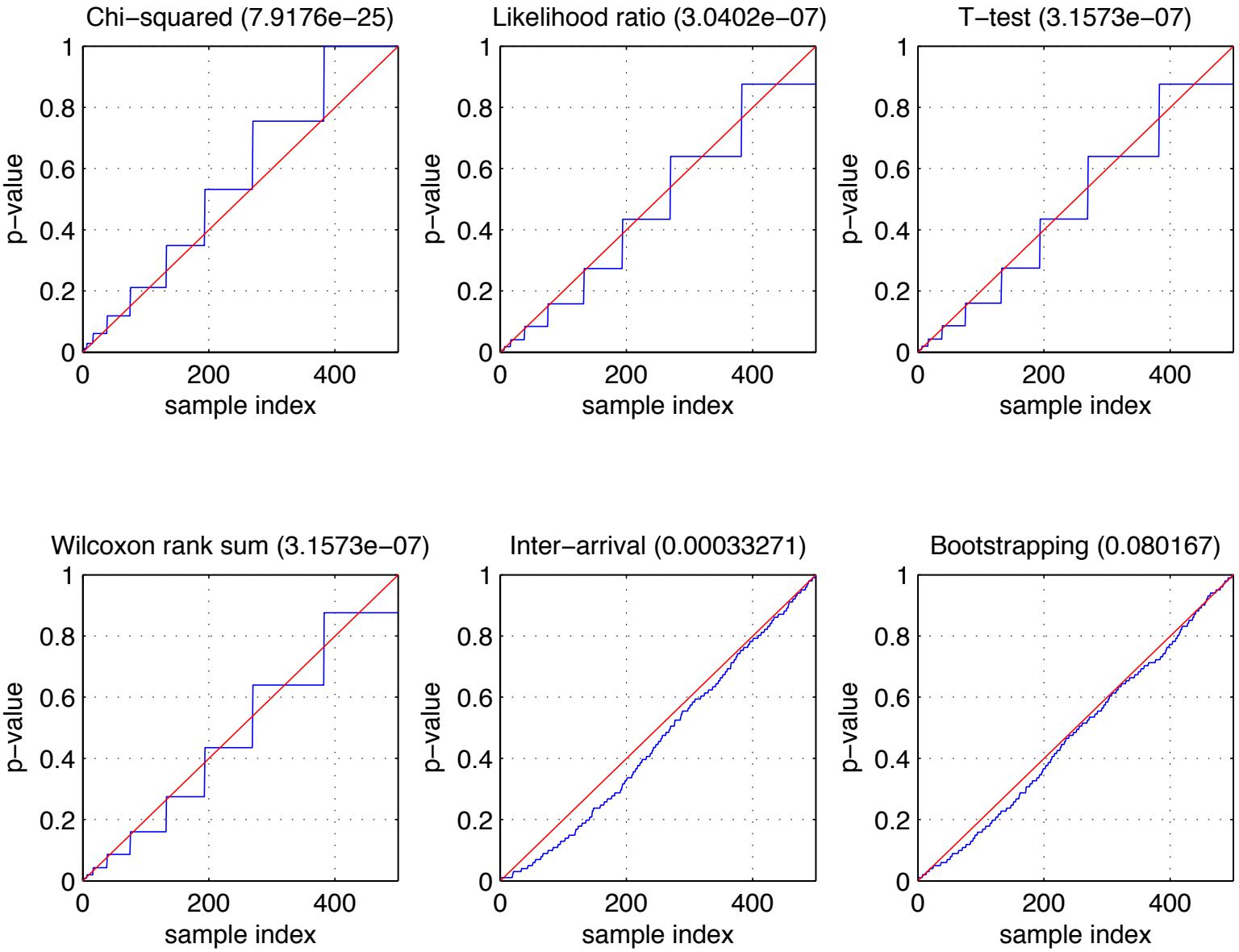


Anti-conservativeness under random word assignment



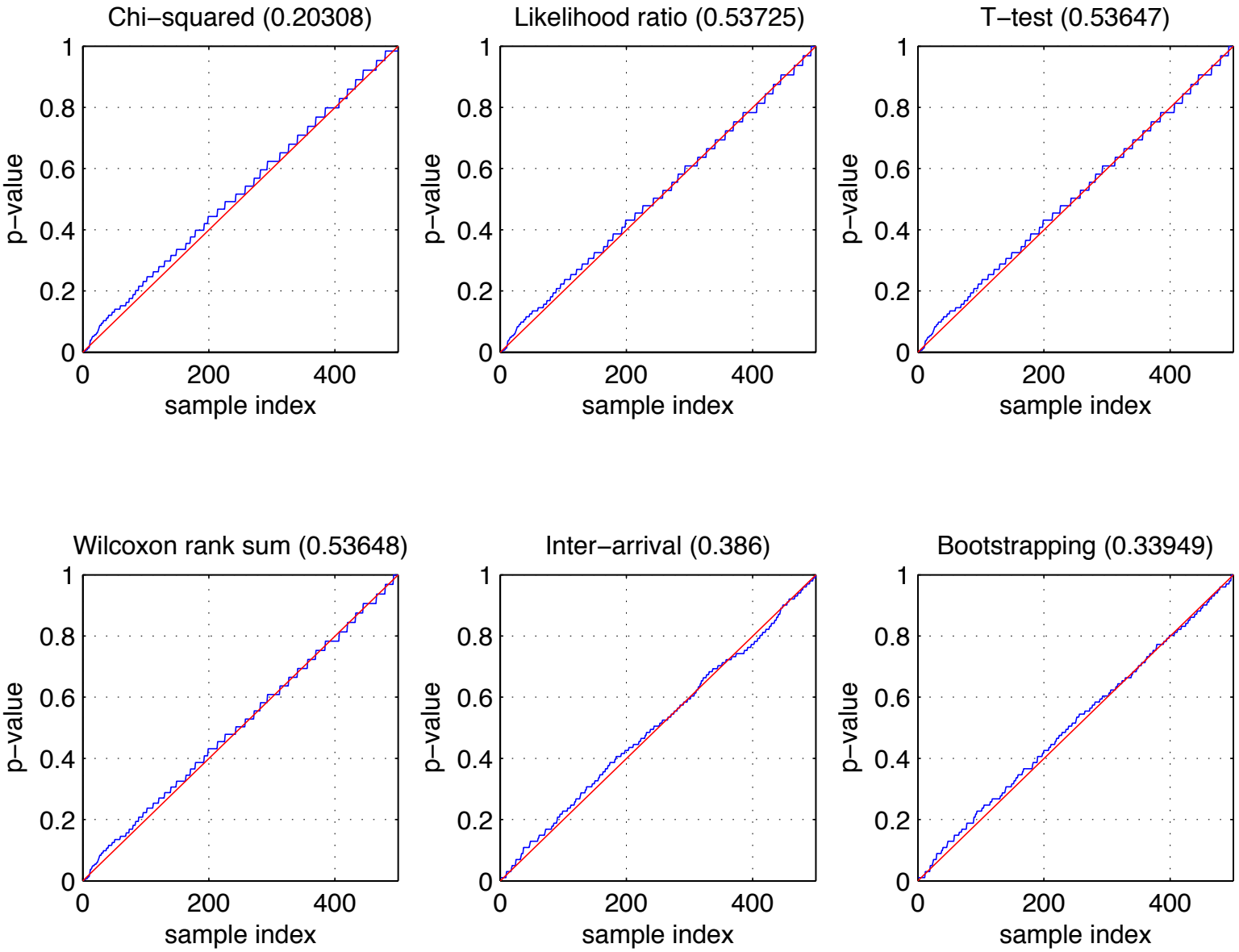
Discrete test problem

trip (41)



Discrete test problem

would (2590)



Conclusion part 2

- Bag-of-words based test should not be used by linguists
- Bootstrap test and Wilcoxon give best results
- T-test gives similar results
  - T-test is faster to compute
- Open problems
  - A fair test for all tests?
  - Why is resampling *S* and *T* required in bootstrap test?

Further research

Further research: Burstiness and inter-arrival times

- Segmentation of text
  - Find change-points in sequence of inter-arrival times
- Clustering of words
  - We know burstiness is related to parts-of-speech
  - Can we group words based on inter-arrival times?
- Stochastic model for inter-arrival times
  - Current best (Weibull) does not fit that well
  - How to take into account correlations in inter-arrival time test

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.
- **North, B. V., Curtis, D. and Sham, P. C.** (2002). A note on the calculation of empirical p-values from Monte Carlo procedures. *The American Journal of Human Genetics,* **71**(2): 439-441.
- **Gries, S. Th.** (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* **13**(4): 403-437.
- **Lijffijt, J. and Gries, S. Th.** (*to appear*). Correction to 'Dispersions and adjusted frequencies in corpora'. *International Journal of Corpus Linguistics.*

- Slides available soon at http://users.ics.tkk.fi/lijffijt