

Adapting random-instance sampling variance estimates and Binomial models for random-text sampling

1. Introduction

χ

corpus linguistics

- *invalid*
- *very different tests*
- *degree*
- *degree of certainty*
the best estimate of the observation

2. Previous research

χ

•

•

χ

•

infer

•

•

•

specific

3. Adjusting the Binomial model

non-empty texts p i t' n_i t' n

$$\text{standard deviation } S \equiv \sqrt{P - P n}$$

$$\text{variance } S \equiv P - P n$$

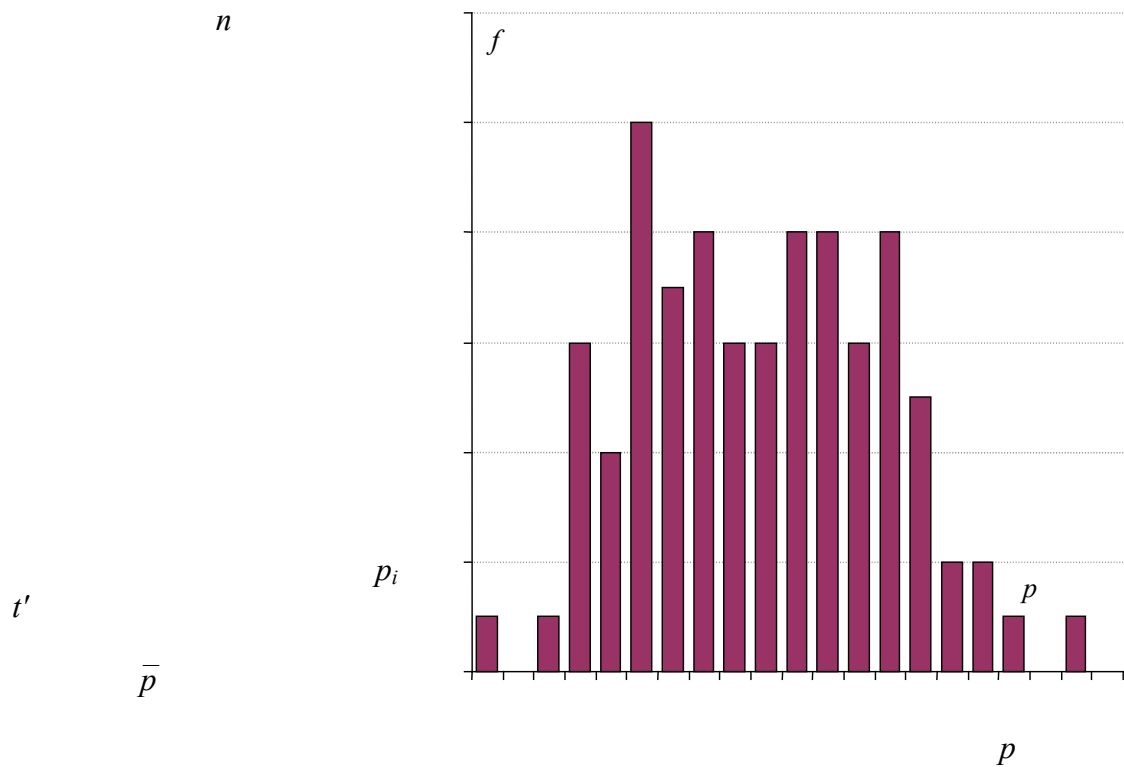
$$P \quad w^- \quad p \quad P \quad z_\alpha \quad S \quad n$$

$$\text{Wilson score interval } w^- \quad w^+ \equiv \left(p + \frac{z_\alpha}{n} \pm z_\alpha \sqrt{\frac{p - p}{n} + \frac{z_\alpha}{n}} \right) / \left(+ \frac{z_\alpha}{n} \right)$$

z_α

χ

α



subsample mean $\bar{p} = \frac{\sum p_i}{t'}$

predicted

predicted between-subsample variance $S = \frac{\bar{p} - \bar{p}}{t'}$

actual

unbiased estimate of the population variance

observed between-subsample variance $s = \frac{\sum p_i - \bar{p}}{t' - 1}$

t'

ratio of variances F

n

$$F = \frac{s^2}{S^2} = \frac{p - p}{n} \frac{n}{n'}$$

$F = \frac{s^2}{S^2}$
adjusted sample size $n' = n - F$

$$n' n = S^2 s$$

$$n - t'$$

t'

adjusted sample size $n' = n - t' F$

$$\frac{n - t' n'}{n' n} = t' F$$

n

n

finite population correction

et al.

$$v^2 = n N \frac{N}{n}$$

$$v^2 = N n N$$

n

Adapting variance for random-text sampling

CL	CL(inter)	Words	$p(\text{inter})$

4. Example 1: interrogative clause probability, direct conversations

observed probability p p $\frac{f}{f}$ p p

n f
 standard deviation s
 Wilson interval w w

measures of uncertainty an underestimate?

to what extent are these

text

p

\bar{p}

Adapting variance for random-text sampling

$$\bar{p} = \frac{\sum p_i}{t'}$$

$$S = \sqrt{\frac{\bar{p} - \bar{p}}{t'}}$$

$$s = \sqrt{\frac{\sum p_i - \bar{p}}{t'}}$$

s \bar{p}

ratio F S s
 number of cases n'
 standard deviation s
 95% Wilson interval w w

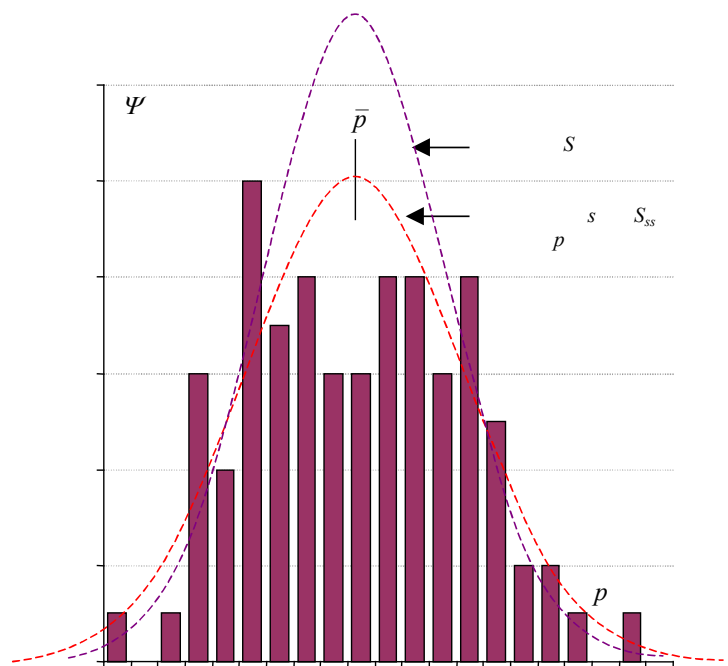
corp.ling.stats

n

s

$f p$

Z
 s
 $e \sum_{p=} Z \bar{p} s p - f p$
 $Z \bar{p} s p$
 $e s s \approx p$
 $f p$



s

cluster-adjustment ratio F

number of cases n'

standard deviation s

95% Wilson interval w w

p

5. Example 2: Clauses per word, direct conversations

exposure

choice

*the number of
clauses per word*

observed probability p p f f

number of cases n f

standard deviation s

95% Wilson interval w w

n

distribution mean \bar{p}

Adapting variance for random-text sampling

predicted standard deviation S
observed standard deviation s
cluster-adjustment ratio $F_{ss} = S^2 / s^2$

smaller

greater

s

p p
every word in the corpus could be the first word in a clause
 p

p

p

p

s

n

F

number of cases n'
standard deviation s
95% Wilson interval w w

6. Uneven-size subsamples

p

\bar{p}

n_i

\bar{p} p

$$s \sum p x_i p_i - p$$

$$p x_i \frac{n_i}{n}$$

P

$$s \sqrt{\sum n C r P^r - P^{n-r} r n - P} \equiv \sqrt{P - P n}$$

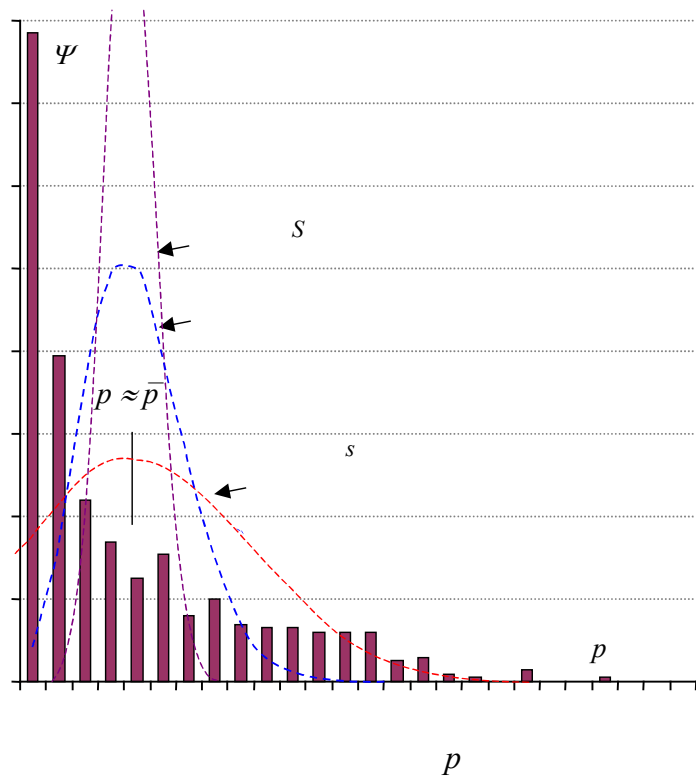
$$t' \quad t' -$$

$$s \frac{t'}{t' -} \sum p x_i p_i - p$$

$$t' \quad n_i$$

S_{SS}

7. Example 3: Interrogative clause probability, all ICE-GB data



observed probability p f f
 number of cases n f
 standard deviation s
 95% Wilson interval w w

et al

n

$r n P$

$r n P$

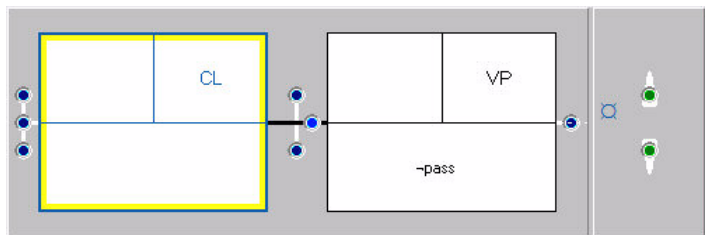
S
 s
 ratio F S s
 p
 n

p p
 number of cases n'
 standard deviation s
 95% Wilson interval w w

n
 p
 p
 n

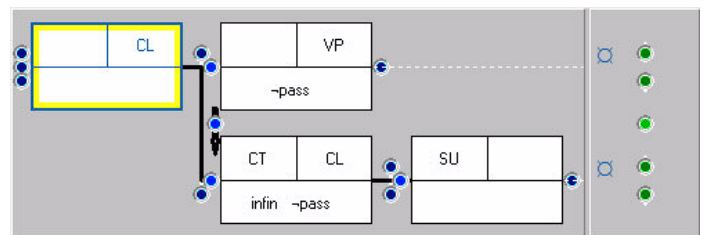
8. Example 4: Rate of transitive complement addition

Fuzzy Tree Fragments et al



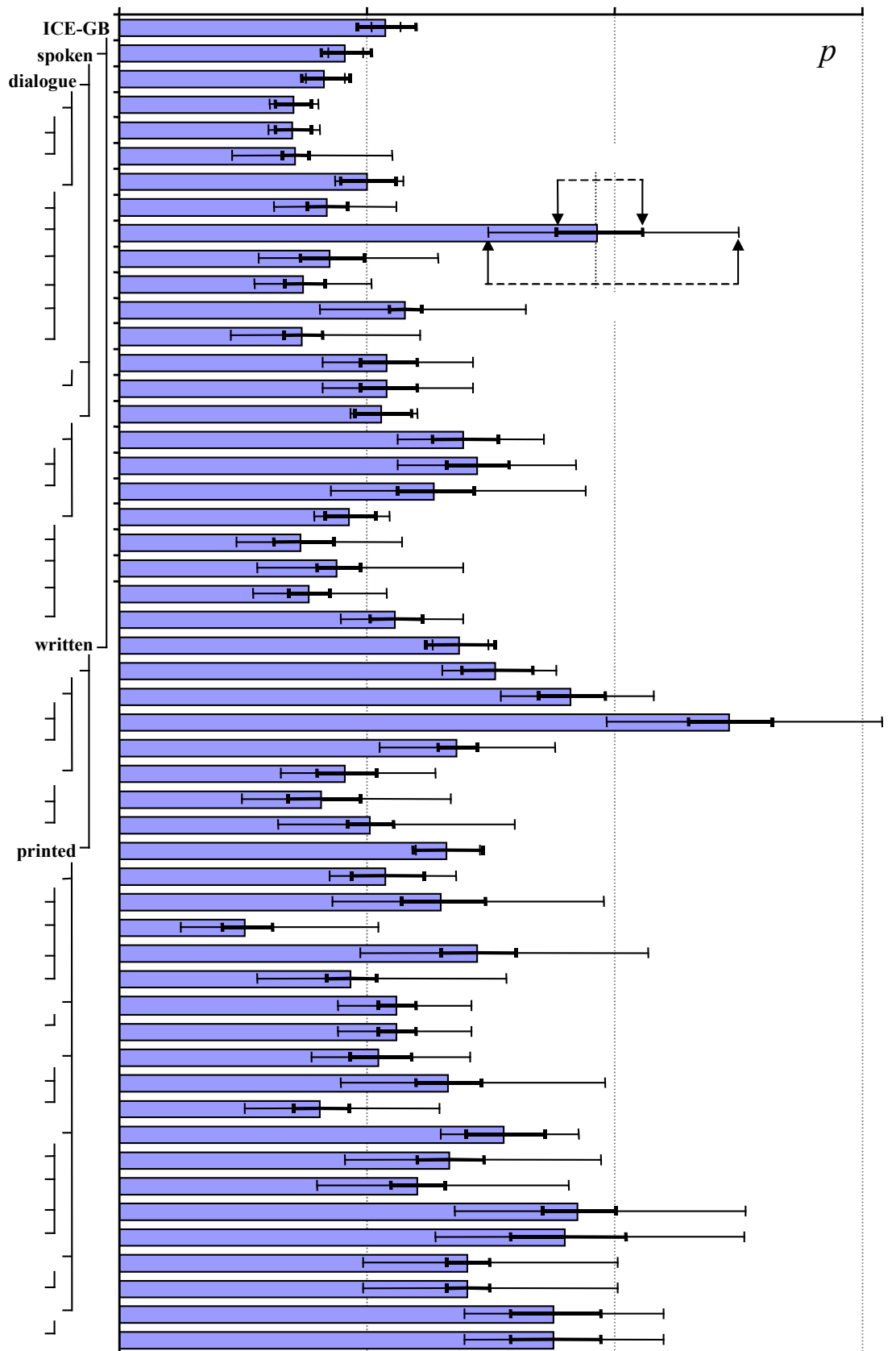
p

p



p

Adapting variance for random-text sampling



interviews p

Adapting variance for random-text sampling

broadcast
 p

t' p n w^- w^+ F w^- w^+

**spoken
dialogue**



written



printed



F_{ss}

p

p

p

p

Adapting variance for random-text sampling

F p *spoken dialogue written* *printed*
F

9. Conclusions

n

p

p

p

p

References

International Journal of Corpus Linguistics
Handbook of Parametric and Nonparametric Statistical Procedures

Approaches to Social

Research

Adapting variance for random-text sampling

Mathematical Statistics with

Applications

*Final Report to EPSRC: Next Generation Tools for Linguistic
Research in Grammatical Treebanks*

corp.ling.stats

corp.ling.stats

corp.ling.stats

corp.ling.stats

*Journal of Quantitative Linguistics
Statistics in Corpus Linguistics Research*