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

Adapting variance for random-text sampling

3. Adjusting the Binomial model

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

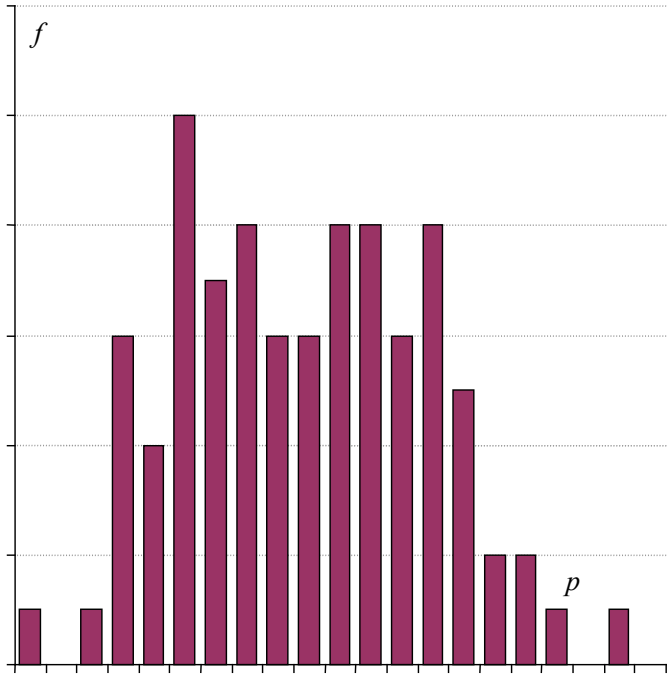
standard deviation $S \equiv \sqrt{P \quad - P \quad n}$
variance $S \equiv P \quad - P \quad n$

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

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

z_α α
χ

n



p_i

t'

$\bar{p}$

p

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

predicted

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

actual

unbiased estimate of the population variance

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

t'

ratio of variances F

n

p

$n \quad s \equiv p - p \quad n$

$n' \qquad n \quad F$

$F = S \quad s$

adjusted sample size $n' \quad n \quad F$

$n' \, n \qquad S \quad s$

t'

$n - t'$

adjusted sample size $n' \quad n \quad t' \quad F \quad t'$

$n \quad t' \; n' \qquad t' \quad F$

$n' \quad n$

n

n

---

finite population correction

et al. n

$\nu^2 \quad n\,N \quad N \qquad \nu^2 \quad N \quad n \quad N$

n

	CL	CL(inter)	Words	*p*(inter)

4. Example 1: interrogative clause probability, direct conversations

p *p*

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

n f
standard deviation s
Wilson interval w w

to what extent are these measures of uncertainty an underestimate?

text

p

$\bar{p}$

Adapting variance for random-text sampling

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

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

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

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

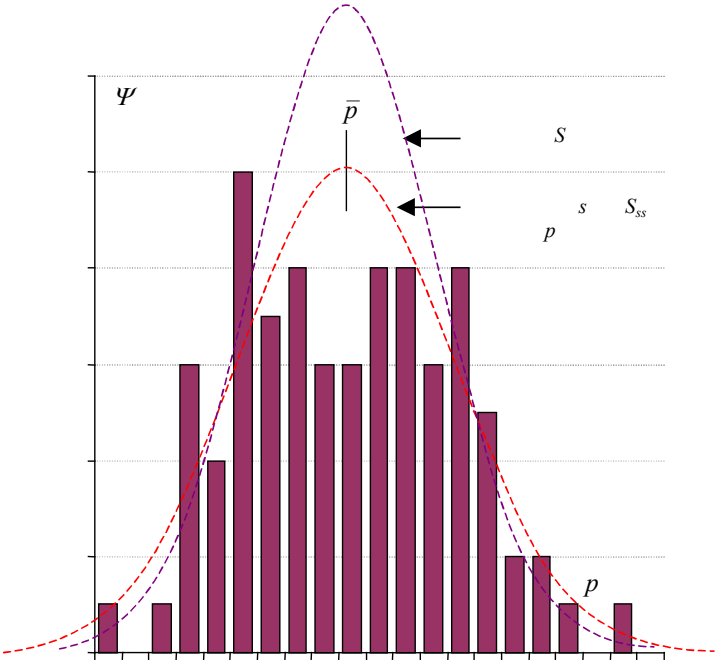
s

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

$Z\,\bar{p} \;\; s \;\; p$

$\bar{p}$ s p

e $s \approx$



$f\,p$

Adapting variance for random-text sampling

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 s*

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

$\bar{p}$
p

n_i

$\bar{p}$ *p*

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

$p\ x_i \quad n_i \quad n$
i

P

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

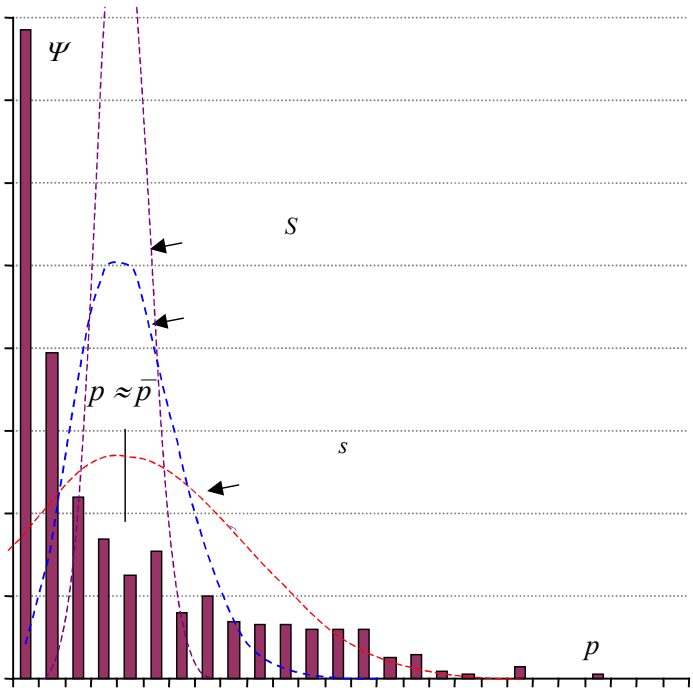
$t' \qquad t' -$

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

t' n_i

s_{ss}

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



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

---

et al

n $r\ n$ P r nP

p

S
s
ratio F S s

p
n

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

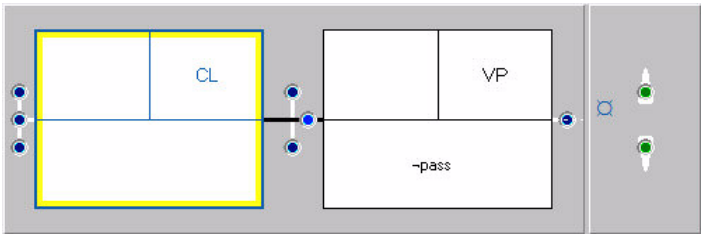
n
n

p

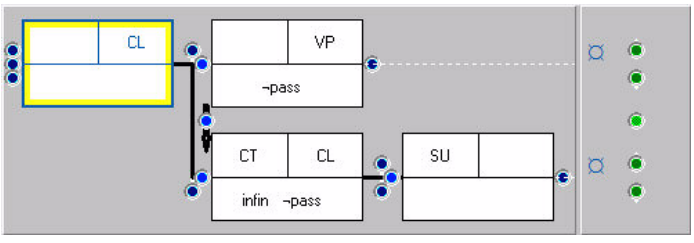
p

8. Example 4: Rate of transitive complement addition

Fuzzy Tree Fragments
et al

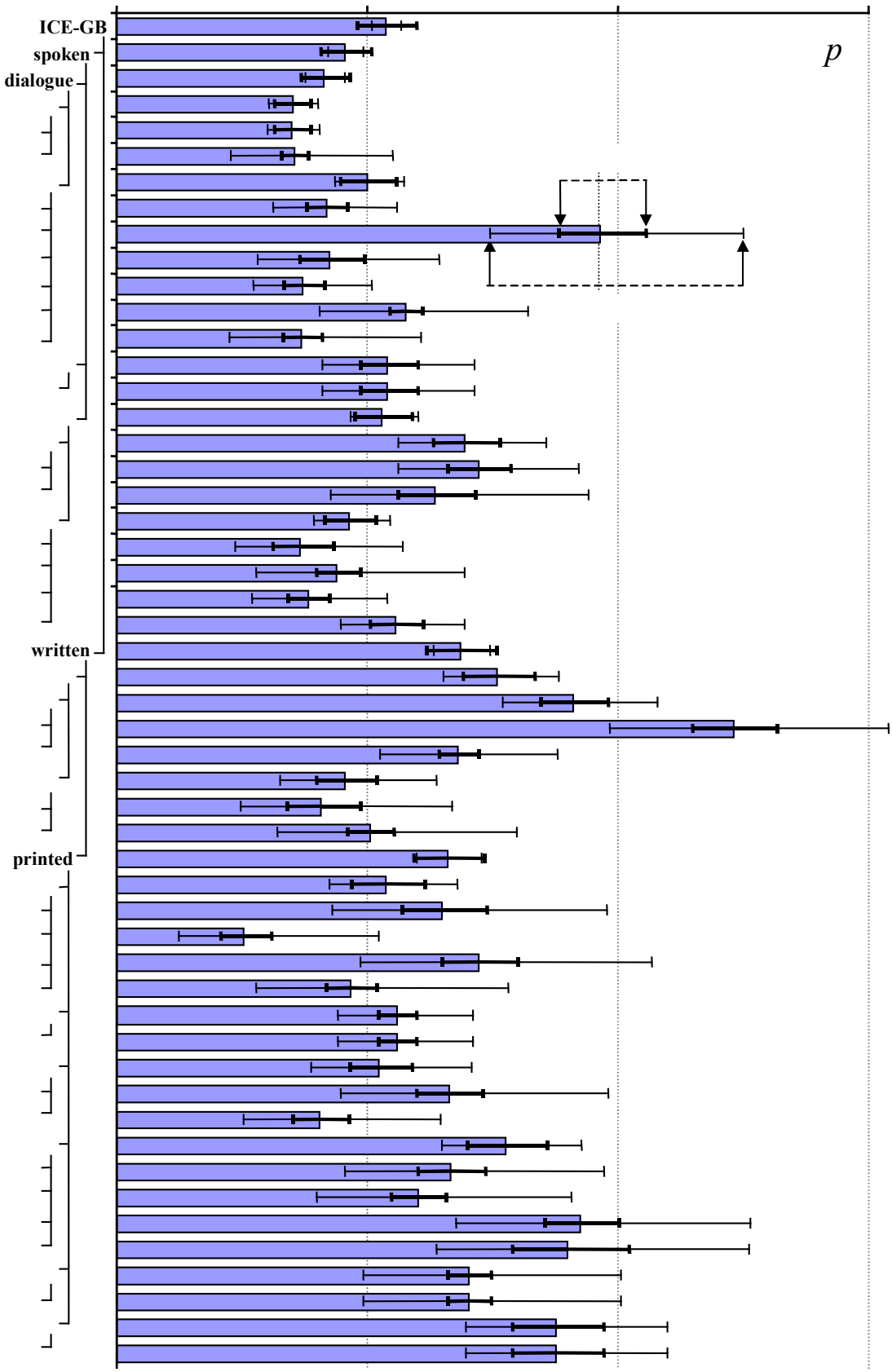


p



p

p



broadcast

interviews p *p*

Adapting variance for random-text sampling

	t'	p	n	w^-	w^+	F	w^-	w^+
spoken dialogue								
written								
printed								

F_{ss}

p p

p

p

Adapting variance for random-text sampling

p

F *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