Statistical Testing in NLP (II)

CS 690N, Spring 2018

Advanced Natural Language Processing

http://people.cs.umass.edu/~brenocon/anlp2018/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

Statistical variability in NLP

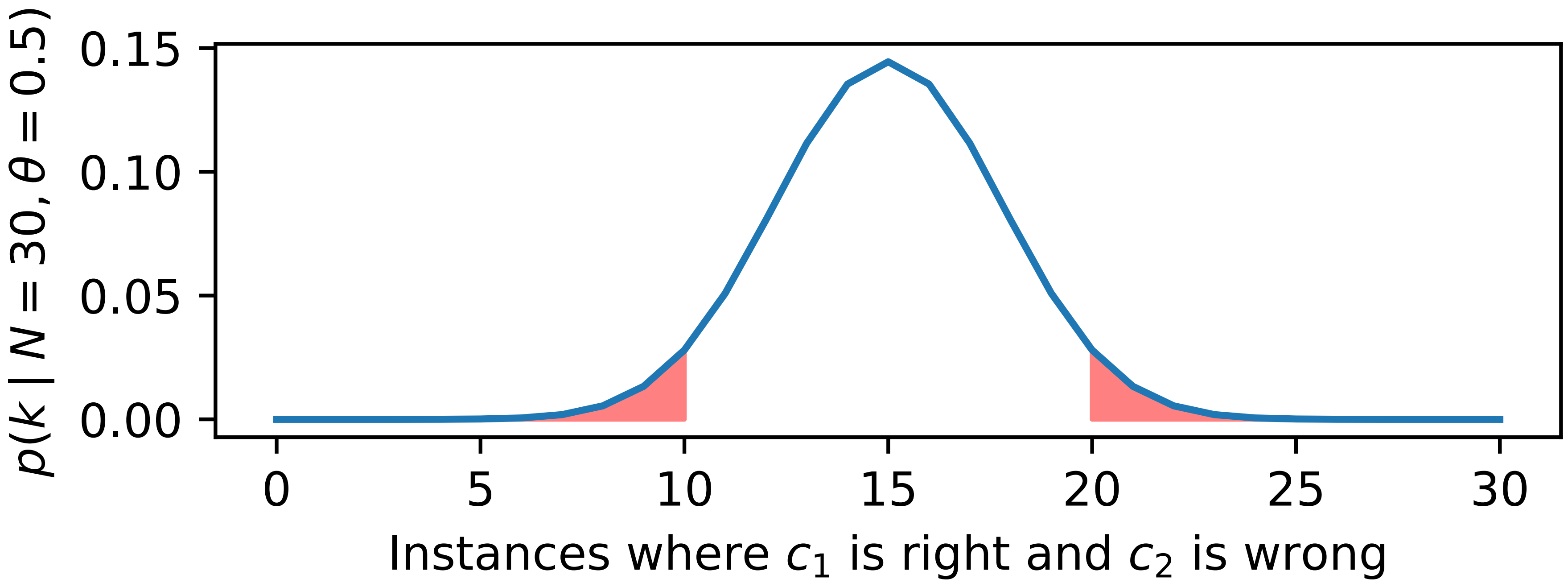
- How to trust experiment results, given many ***sources of variability***?
  - How was the text data sampled?
  - How were the annotations sampled?
    - How variably do the human annotators behave?
  - How variable are the computational algorithms?
- Today: ***Variability due to small sample size***

Text data variability

- Mathematically, the easiest case to analyze:
  What if we resampled the tokens/sentences/documents from a similar population as our current data sample?
- Assume units are sampled i.i.d.; then apply your favorite statistical significance/confidence interval testing technique
  - T-tests, binomial tests, ...
  - Bootstrapping
  - Paired tests
- For
  - 1. Null hypothesis testing
  - 2. Confidence intervals

Null hypothesis test

- Must define a null hypothesis you wish to ~disprove
- pvalue = Probability of a result as least as extreme, if the null hypothesis was active
- Example: paired testing of classifiers with exact binomial test (R: *binom.test*)



$$p_{\text{Binom}}(k; N, \theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

Statistical tests

- Closed-form tests
  - t-tests, exact binomial test, chi-square tests....
- Bootstrapping

- All methods can give both p-values and confidence intervals

Bootstrapping

- Bootstrapped CI methods
  - Percentile
  - Standard error-based normal approx, etc.
- Theoretical guarantees (under various regularity conditions... for a slightly different CI method...):

$$\mathbb{P}(\theta \in C_n) = 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right)$$

- How many samples? 10,000-100,000
  (governs monte carlo error; can always make nearly 0)
- Paired bootstrap
- Bootstrapped p-values

- (stopped here 2/27)

Berg-Kirkpatrick et al. 2012

- Paired bootstrap test
    - (Subtle, debatable bug?)
- Stat. sig results may not transfer domains
- Researcher effects? Or is paired testing working correctly?

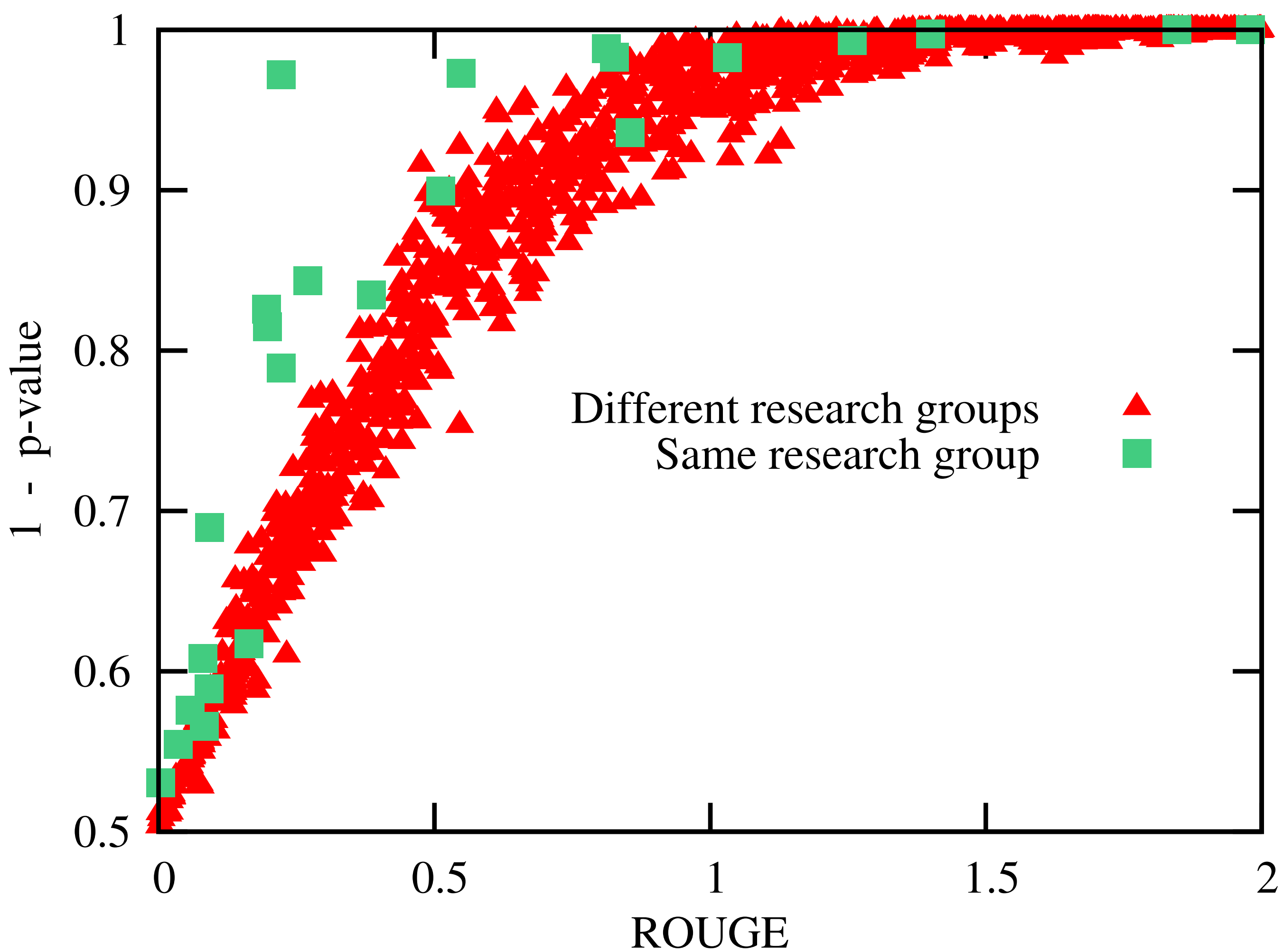


Figure 2: **TAC 2008 Summarization:** Confidence vs. ROUGE improvement on TAC 2008 test set for comparisons between all pairs of the 58 participating systems at TAC 2008.

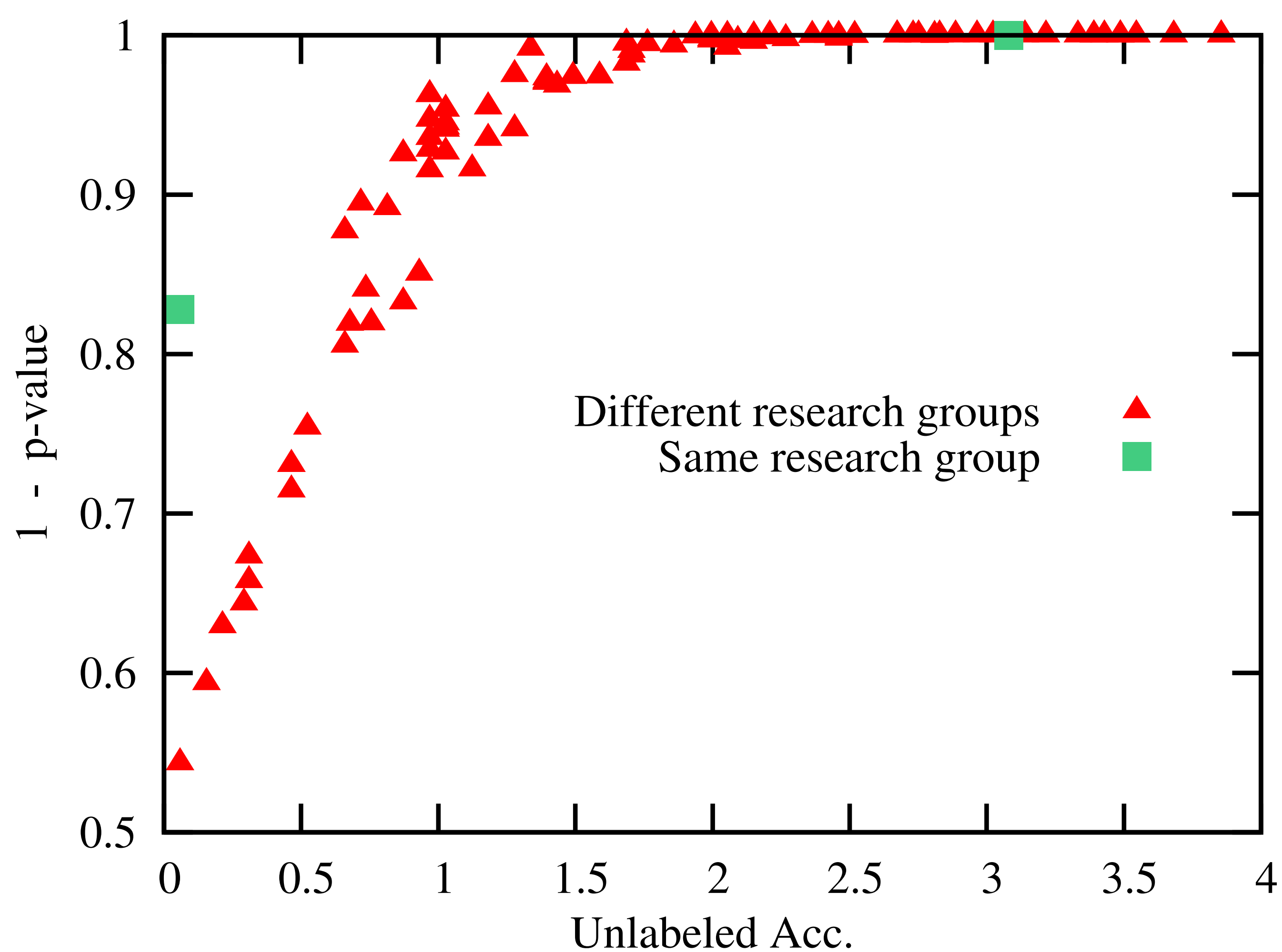


Figure 3: **CoNLL 2007 Dependency parsing:** Confidence vs. unlabeled dependency accuracy improvement on the Chinese CoNLL 2007 test set for comparisons between all pairs of the 21 participating systems in CoNLL 2007 shared task. Com-

Sec. 23 p-value	% Sys. A > Sys. B		
	Sec. 22	Sec. 24	Brown
0.00125 - 0.0025	97%	95%	73%
0.0025 - 0.005	92%	92%	60%
0.005 - 0.01	92%	85%	56%
0.01 - 0.02	88%	92%	54%
0.02 - 0.04	87%	78%	51%
0.04 - 0.08	83%	74%	48%

Table 1: **Empirical calibration:** p-value on section 23 of the WSJ corpus vs. fraction of comparisons where system A beats system B on section 22, section 24, and the Brown corpus. Note that system pairs are ordered so that A always outperforms B on section 23.

- Statistical significance != practical significance
- CI width, statistical power, data size
- Many other confounds we don't have models for, but know can be very significant
    - Researcher bias
    - File-drawer bias
    - Generalization (e.g. across domains)
    - Tuning on test sets
    - Reusing test set over multiple papers