

Statistical uncertainty in NLP

CS 685, Fall 2025

Advanced Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs685_f25/

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 variable are the computational algorithms?
 - How were the annotations sampled?
 - How was the text data sampled?
 - How representative is the text sample, compared to the greater population of possible texts?

Text data variability

- Do results generalize to
 - new domains?
 - new authors?
 - new documents?
 - new sentences?
- (Typically things get worse if anything changes)
- Also of interest: even if only care about text similar to our current one, did we “get lucky” in our selection of sentences/documents/etc?

Text data variability

- A simpler setting: variability due to a **small sample size**
 - What if we resampled the tokens/sentences/documents from a similar population as our current data sample?
- Rest of today: focus on **classifier accuracy evaluation.**
 - Is the result you see real, or due to chance?

Null hypothesis tests

- Core idea: compare your observed result to what you'd observe, to what you'd expect if results were "random" in some way
 - formally, the null hypothesis

- Example #1: are your predictions better than chance?



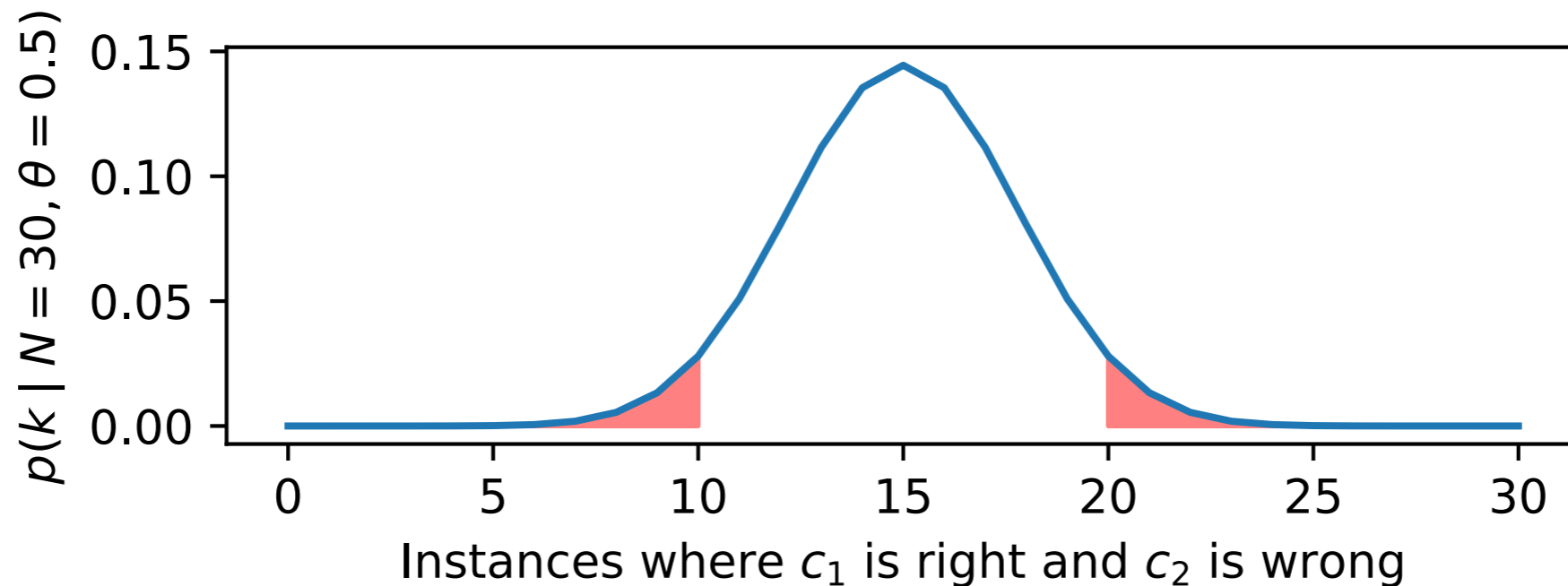
- Example #2: is the diff. in two classifiers' accuracies better than chance?

Null hypothesis tests

- Must define a null hypothesis you wish to disprove
 - H_0 = the "null hypothesis". Observations were generated in an uninteresting way, "due to chance"
- p-value: probability you could see a result at least as extreme as what you have, if H_0 was true
 - $pval = P(T(obs) > T(gendata) \mid gendata \sim H_0)$
- If you can't beat the null hypothesis, take your results with a grain of salt!

Null hypothesis test

- pvalue = Probability of a result as least as extreme, if the null hypothesis was active
- Example: paired testing of classifiers. Two equivalent methods:
 - 1. Randomized simulation
 - 2. Exact binomial test (R: *binom.test*)



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

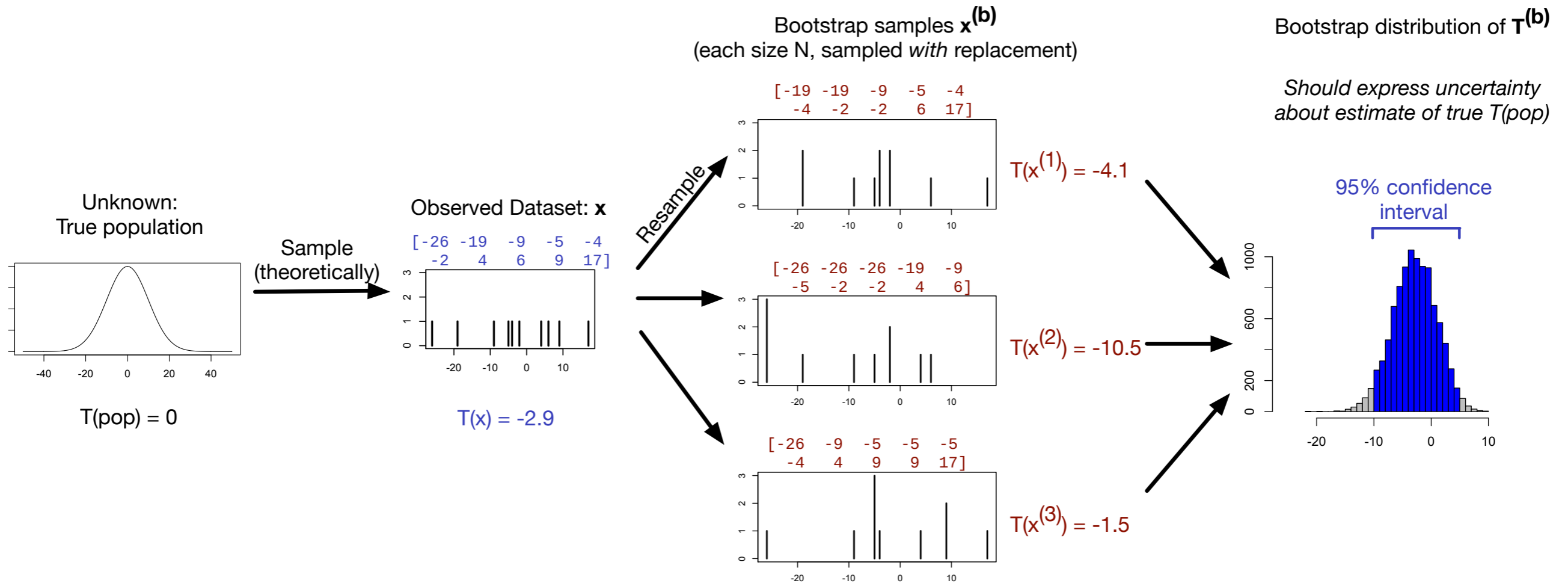
8

Statistical tests

- Two types of information
 - p-values \Leftrightarrow null hypothesis test
 - confidence intervals
- Simulation-based testing
 - 1. Randomized null hypothesis simulation
 - 2. Bootstrapped confidence intervals
- Closed-form tests
 - t-tests, exact binomial test, chi-square tests....

Bootstrapping

- Excellent, flexible method to infer **confidence intervals**



Paired testing

- Bootstrap sampling implicitly does a "paired test"

- Statistical significance testing may be necessary, but never sufficient, for a meaningful result!
 - Statistical significance *vs.*
 - Substantive significance

- Statistical significance \neq 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