Statistics in NLP: a complicated relation

Giovanni Cassani

6th June - CLiPS

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp
(almost) No statistical tests, significance values, confidence intervals, effect sizes and alike in NLP.

Instead, claims of superiority on a given benchmark when some accuracy score is higher than that of a competing model.

Why is this a problem? How can it be countered?
Overview

- Søgaard & al, 2014
  - Are samples representative?
  - Are covariates explaining significance?
  - Are there biases in metrics?
- Multiple comparisons
  - Parameter selection
- Interim conclusions
The paper tackles four main problems: sample biases, overlooked covariates, metric biases, and significance levels.

They evaluate on simulated data, PoS tagging, dependency parsing, and Named Entity Recognition, using several datasets.

Main research question: **how likely are results in NLP to be false?**
Overview

- Søgaard & al, 2014
  - Are samples representative?
  - Are covariates explaining significance?
  - Are there biases in metrics?
- Multiple comparisons
  - Parameter selection
- Interim conclusions
The WSJ fallacy

There is a risk to keep evaluating models on the same data, without caring about whether they are representative or not → **overfitting**

**Cross-validation** is not enough: the WSJ has different topics, but it’s all news articles from a same source!
In general, small samples make it harder to find true effects (low N → large SE).

However, if the data are not representative, adding more can make significant effects disappear!

Small samples → higher chance of false negatives (reject a true effect) and of false positives (accept a false effect). This happens when small samples are artificially coherent → lower variance then it should (low σ → low SE)!
Overview

- Søgaard & al, 2014
  - Are samples representative?
  - Are covariates explaining significance?
  - Are there biases in metrics?
- Multiple comparisons
  - Parameter selection
- Interim conclusions
By simple probability theory, if we subset data along a factor, and test differences over all factor levels, we have higher chances of discovering an effect.

Is it real? Does it make sense?
Søgaard & al, 2014
- Are samples representative?
- Are covariates explaining significance?
- Are there biases in metrics?

Multiple comparisons
- Parameter selection

Interim conclusions
By the same reasoning, if we can evaluate systems on multiple metrics, it becomes easier to find significant differences.

What does it mean to significantly outperform another system on only one of three metrics?
Overview

- Søgaard & al, 2014
  - Are samples representative?
  - Are covariates explaining significance?
  - Are there biases in metrics?
- Multiple comparisons
  - Parameter selection
- Interim conclusions
Significance is about reliability

$P$ values don’t tell about how strong an effect is, but about its **reliability**: is the difference I see in my samples real? If I take different samples, will I still see it?

When we compare so many **independent events**, the likelihood of detecting **an** effect increases (and if events are dependent, even more).
In NLP systems, parameter spaces are huge, and evaluation options are often countless (which sample, which metric, which systems, and so on).

There are methods to correct for multiple comparisons (Bonferroni, Tukey’s, Dunnett’s, …) but when comparisons become too many, they are not interesting: design issues!
Overview

- Søgaard & al, 2014
  - Are samples representative?
  - Are covariates explaining significance?
  - Are there biases in metrics?
- Multiple comparisons
  - Parameter selection
- Interim conclusions
When we need to select across a large parameter space, it’s advisable to have two evaluation procedures, intrinsic and extrinsic.

Use the first to select the best parametrization(s) and the second to compare results. The choice of evaluation become critical and should be theoretically, not practically, motivated.*

*when running many evaluations, use appropriate statistical methods to check which are the best parameters (sign tests, approximate randomization, …)
Overview

- Søgaard & al, 2014
  - Are samples representative?
  - Are covariates explaining significance?
  - Are there biases in metrics?
- Multiple comparisons
  - Parameter selection
- Interim conclusions
Søgaard and colleagues suggest to make $\alpha$ more stringent and avoid the issues with multiple comparisons, sample biases, metrics, and alike to avoid false positives. They suggest to run more comparisons and test on more samples, more metrics, more parameters. However, the big questions remain unanswered: what are we testing and why?
Variation is your friend

NLP systems are usually deterministic: given a choice of parameters and a dataset they output the same results.

However, **statistics only works when there’s variation**: for more reliable statistical inference in NLP tasks, identify theoretically or practically interesting sources of variation and run your system exploiting that variation. Test afterwards, to look at the effects **in the light of your hypothesis**.
Statistics can make your data tell pretty much whatever you want. That doesn’t make them good statistics though.

Too rarely in NLP initial assumptions and theoretically relevant research questions are spelled out. Too often, only the tip of the experimental iceberg is reported.

**Statistical significance is not practical significance, and statistics depend on theory.**
Thank you!
Questions?
What does significance depend on?

**Group means** →
the more far apart, the lower the *p* value

\[ SE \left[ \frac{\sigma}{\sqrt{N}} \right] \rightarrow \]
the lower it is, the lower the *p* value

*SE* is low when

\[ \sigma \text{ is low} \]
\[ N \text{ is high} \]
Appendix

What does significance depend on?

• Distance = 2
• Sigma = 15
• N = 30

• Distance = 6
• Sigma = 10
• N = 50

• Distance = 10
• Sigma = 5
• N = 100