Technion
Israel Institute of
Technology

The Hitchhiker’s Guide to Testing Statistical Significance in NLP

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart

ACL 2018

https://github.com/rtmdrr/testSignificanceNLP

I want to be…

state of the art

- – my new algorithm
- – current SOTA algorithm
- Data -
- Evaluation measure

- Apply algorithm on
- Apply algorithm on
- Test if

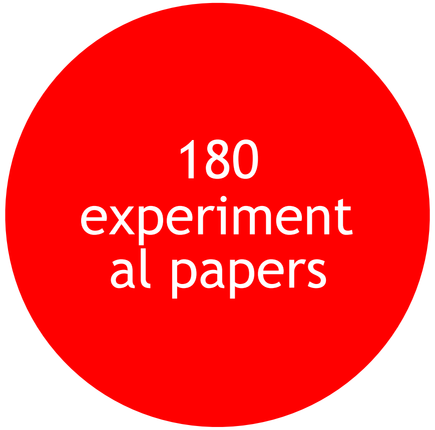
This is not enough!

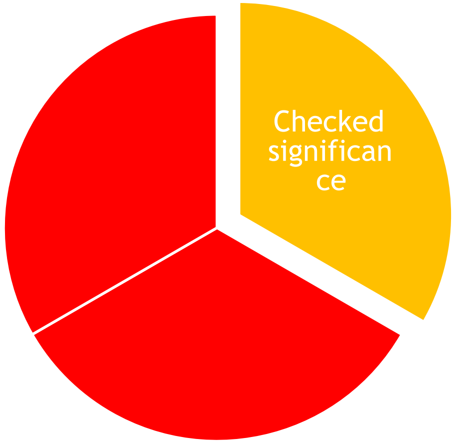
- The difference between the performance of algorithm and could be coincidental!
- We need to make sure that the probability of making a false claim is very small.
- We can do so by…

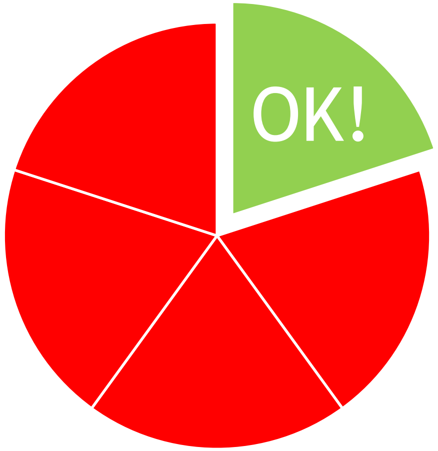
Testing Statistical Significance!

NLP & Hypothesis Testing – Survey ACL 2017

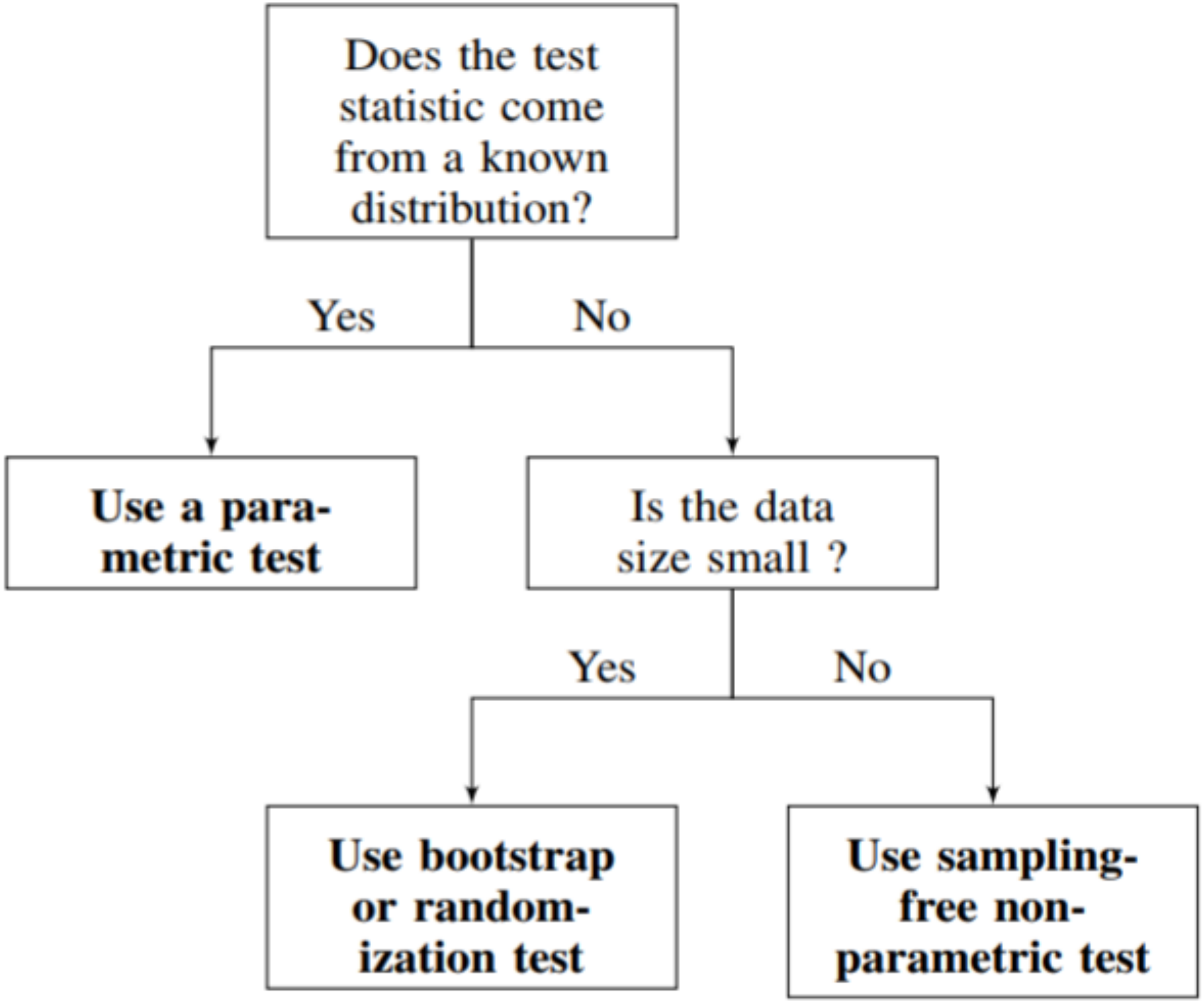
- 180 experimental long papers
- 63 checked statistical significance
- Only 42 mentioned the name of the statistical test
- **Only 36 used the correct statistical test - of all papers!**







Simple Guide



Does the test statistic come from a known distribution?

Yes

No

Use a parametric test

Is the data size small ?

Yes

No

Use bootstrap or randomization test

Use sampling-free non-parametric test

Statistical Significance Hypothesis Testing

- Let: .

Statistical Significance Hypothesis Testing

- The smaller the p-value is, the higher the indication that the null hypothesis, , does not hold.

- We reject the null hypothesis if

Statistical Significance Hypothesis Testing

- **Type I error** – rejecting the null hypothesis when it is true

- **Type II error** –not rejecting the null hypothesis when the alternative is true

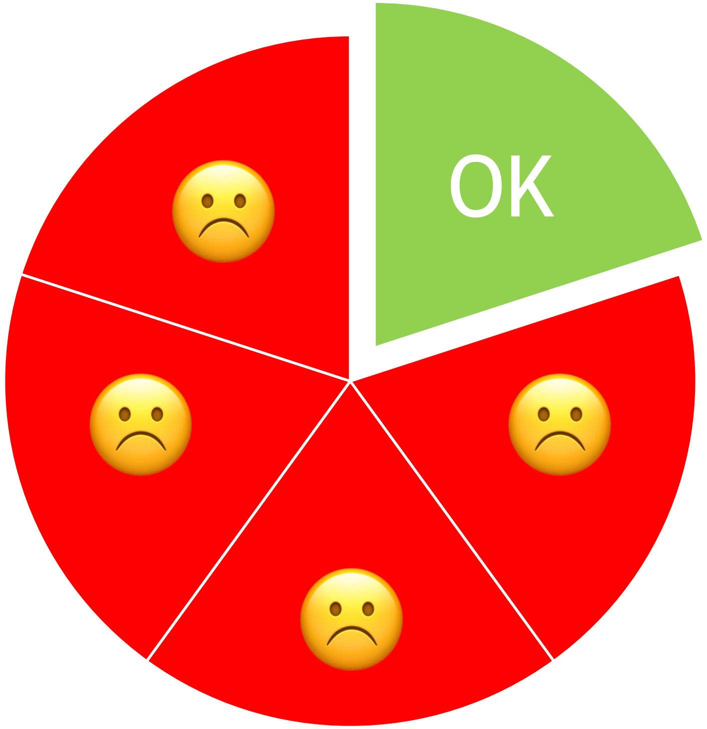
- **Significance level** – probability of making type I error ()

- **Significance Power** – probability of **not** making type II error

So…

Let’s all test for statistical significance!

Why not?



NLP & Hypothesis Testing - Problems

- Both algorithms are applied on the **same data**.
- What is the distribution of ?
- Data samples are not independent.

Paired Statistical Tests

- Both algorithms are applied on the **same data** – dependent

- Paired sample: sample selected from the first population is related to the corresponding sample from the second population

- **Solution:** apply paired-version of statistical test
    - Paired t-test, Wilcoxon signed-rank test, paired bootstrap…

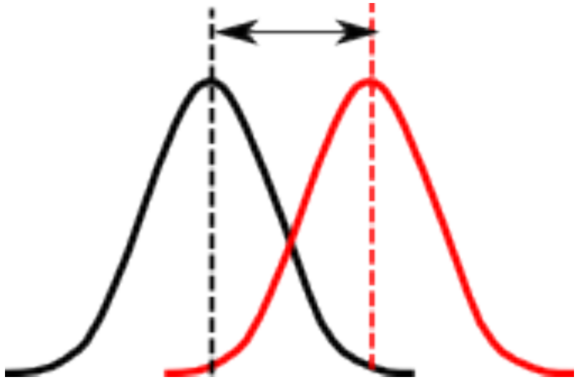
NLP & Hypothesis Testing - Problems

☑ Both algorithms are applied on the **same data**.

? What is the distribution of ?

? Data samples are not independent.

Parametric Tests



- First case: the distribution of is Normal

- Parametric tests make assumptions about the test statistic distribution, particularly - normal distribution.

- When the parametric test meets assumptions it has high statistical power
  - Linear regression analyses
  - **T-tests** and analyses of variance on the difference of means
  - Normal curve Z-tests of the differences of means and proportions

Parametric Tests – Check for Normality

- **Shapiro-Wilk:** tests if a sample comes from a normally distributed population

```
scipy.stats.shapiro([a-b for a, b in zip(res_A, res_B)])
```

- **Anderson-Darling:** tests if a sample is drawn from a given distribution

```
scipy.stats.anderson([a-b for a, b in zip(res_A, res_B)], 'norm')
```

- **Kolmogorov-Smirnov:** goodness of fit test. Samples are standardized and compared with a standard normal distribution.

```
scipy.stats.kstest([a-b for a, b in zip(res_A, res_B)], 'norm')
```

Non-Parametric Tests

- Second case: the distribution of is unknown\not normal
- Non parametric tests do not assume anything about the test statistic distribution
- Two types – *sampling-free* and *sampling-based* tests

Sampling-Free Non-Parametric Tests

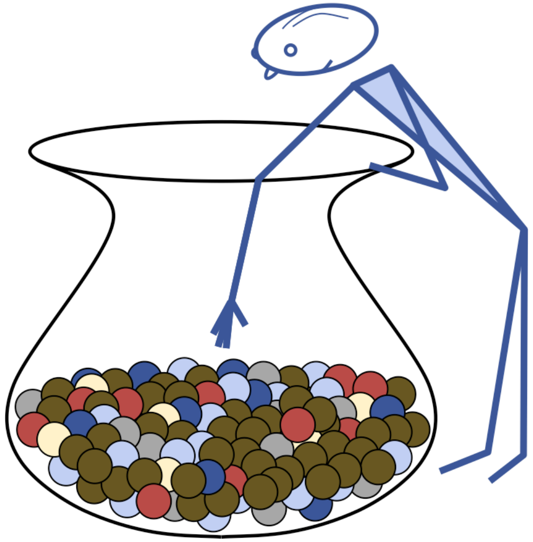
Binomial\
Multinomial

Not Normal

rouge - red bleu - bue

Sampling-Based Non-Parametric Tests

- Permutation tests: resamples drawn at random from the original data. **Without replacements**.
  - Paired design – consider all possible choices of signs to attach to each difference.



- Bootstrap: resamples drawn at random from the original data. **With replacements**.
  - Paired design – sample with repetitions from the set of all differences.

NLP & Hypothesis Testing - Problems

☑ Both algorithms are applied on the **same data**.

☑ What is the distribution of ?

? Data samples are not independent.

NLP Data and I.I.D Assumption

- Many NLP datasets have dependent samples

- All statistical test assume independency => all tests are invalid, impact hard to quantify

- **Solution:** come up with statistical tests that allow dependencies

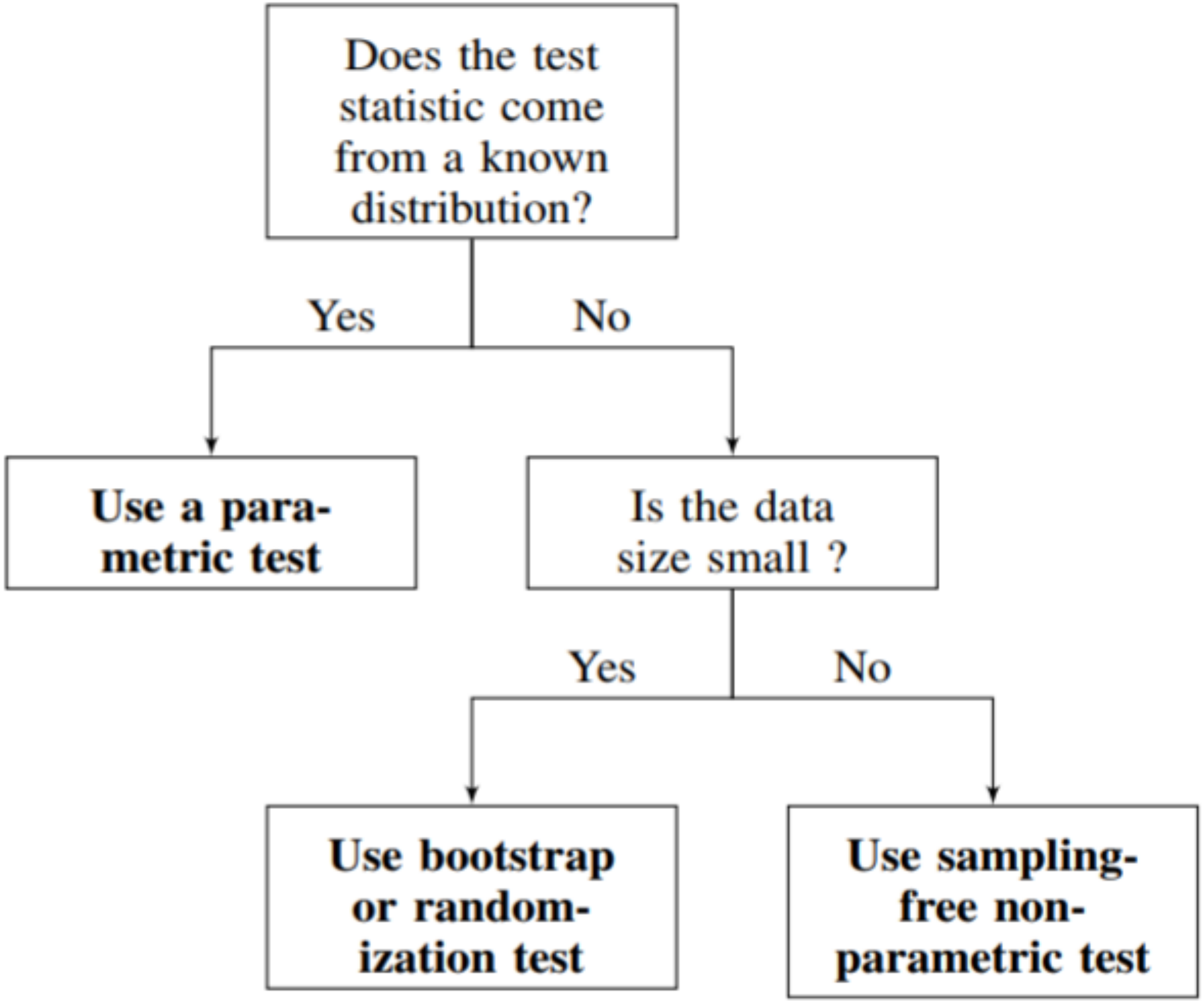
NLP & Hypothesis Testing

☑ Both algorithms are applied on the **same data**.

☑ What is the distribution of ?

❓ Data samples are not independent.

Simple Guide



Does the test statistic come from a known distribution?

Yes

No

Use a parametric test

Is the data size small ?

Yes

No

Use bootstrap or randomization test

Use sampling-free non-parametric test

Thank You for Listening
Questions?

https://github.com/rtmdrr/testSignificanceNLP