Statistical Significance Testing
In Theory and In Practice

Ben Carterette
University of Delaware

http://ir.cis.udel.edu/ICTIR13tutorial
Hypotheses and Experiments

• Hypothesis:
  – Using an SVM for classification will give better accuracy than using Naïve Bayes
  – A “Symbol-Refined Tree Substitution Grammar” will give better parsing results than a simple TSG
  – Expanding a short keyword query with synonyms will improve search engine effectiveness

• Experiment:
  – Build a baseline system
  – “Improve” it based on your hypothesis
  – Test both systems on one or more datasets
Experimental Results

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 (small)</th>
<th>F1 (full)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFG</td>
<td>61.9</td>
<td>63.6</td>
</tr>
<tr>
<td>TSG</td>
<td>77.1</td>
<td>85.0</td>
</tr>
<tr>
<td>SR-TSG ($P^{sr-tsg}$)</td>
<td>73.0</td>
<td>86.4</td>
</tr>
<tr>
<td>SR-TSG ($P^{sr-tsg}, P^{sr-cfg}$)</td>
<td>79.4</td>
<td>89.7</td>
</tr>
<tr>
<td>SR-TSG ($P^{sr-tsg}, P^{sr-cfg}, P^{ru-cfg}$)</td>
<td><strong>81.7</strong></td>
<td><strong>91.1</strong></td>
</tr>
</tbody>
</table>

Table 1: Comparison of parsing accuracy with the small and full training sets.

from Shindo et al., *Bayesian Symbol-Refined Tree Substitution Grammars for Syntactic Parsing*, ACL 2012
So What?

• “Do these results support my hypothesis?"

• “Are these results meaningful?”

• “Is it possible that my results are due to chance?”

→ statistical significance testing!
Part 1

TESTING STATISTICAL SIGNIFICANCE
Using R

• R is a software environment for statistical computing

• Includes built-in implementations of many common tests
  – Also has its own programming language for implementing your own

• Download from [http://r-project.org](http://r-project.org)
  – Download TREC-8 evaluation data from [http://ir.cis.udel.edu/ICTIR13tutorial/trec8.RData](http://ir.cis.udel.edu/ICTIR13tutorial/trec8.RData)
Commonly-Used Tests

• Parametric:
  – Student’s t-test
  – ANOVA

• Non-parametric:
  – Wilcoxon signed rank test
  – Sign test/binomial test

• Distribution-free:
  – Randomization test
  – Bootstrap test
## Student’s t-test

\[
\hat{\mu} = B - A = 0.214
\]

\[
\hat{\sigma}_{B-A} = 0.291
\]

\[
t = \frac{\hat{\mu}}{\hat{\sigma}_{B-A} \sqrt{n}} = 2.33
\]

<table>
<thead>
<tr>
<th>Example</th>
<th>A</th>
<th>B</th>
<th>B-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
<td>.35</td>
<td>+.10</td>
</tr>
<tr>
<td>2</td>
<td>.43</td>
<td>.84</td>
<td>+.41</td>
</tr>
<tr>
<td>3</td>
<td>.39</td>
<td>.15</td>
<td>-.24</td>
</tr>
<tr>
<td>4</td>
<td>.75</td>
<td>.75</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>.43</td>
<td>.68</td>
<td>+.25</td>
</tr>
<tr>
<td>6</td>
<td>.15</td>
<td>.85</td>
<td>+.70</td>
</tr>
<tr>
<td>7</td>
<td>.20</td>
<td>.80</td>
<td>+.60</td>
</tr>
<tr>
<td>8</td>
<td>.52</td>
<td>.50</td>
<td>-.02</td>
</tr>
<tr>
<td>9</td>
<td>.49</td>
<td>.58</td>
<td>+.09</td>
</tr>
<tr>
<td>10</td>
<td>.50</td>
<td>.75</td>
<td>+.25</td>
</tr>
</tbody>
</table>
Student’s t-test

\[ \hat{\mu} = B - A = 0.214 \]

\[ \hat{\sigma}_{B-A} = 0.291 \]

\[ t = \frac{\hat{\mu}}{\hat{\sigma}_{B-A}} \sqrt{n} = 2.33 \]

\[ p-value = 0.02 \]
## Wilcoxon Signed-Rank Test

<table>
<thead>
<tr>
<th>Example</th>
<th>A</th>
<th>B</th>
<th>B-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
<td>.35</td>
<td>+.10</td>
</tr>
<tr>
<td>2</td>
<td>.43</td>
<td>.84</td>
<td>+.41</td>
</tr>
<tr>
<td>3</td>
<td>.39</td>
<td>.15</td>
<td>-.24</td>
</tr>
<tr>
<td>4</td>
<td>.75</td>
<td>.75</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>.43</td>
<td>.68</td>
<td>+.25</td>
</tr>
<tr>
<td>6</td>
<td>.15</td>
<td>.85</td>
<td>+.70</td>
</tr>
<tr>
<td>7</td>
<td>.20</td>
<td>.80</td>
<td>+.60</td>
</tr>
<tr>
<td>8</td>
<td>.52</td>
<td>.50</td>
<td>-.02</td>
</tr>
<tr>
<td>9</td>
<td>.49</td>
<td>.58</td>
<td>+.09</td>
</tr>
<tr>
<td>10</td>
<td>.50</td>
<td>.75</td>
<td>+.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>B-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-.02</td>
</tr>
<tr>
<td>2</td>
<td>+.09</td>
</tr>
<tr>
<td>3</td>
<td>+.10</td>
</tr>
<tr>
<td>4</td>
<td>-.24</td>
</tr>
<tr>
<td>5.5</td>
<td>+.25</td>
</tr>
<tr>
<td>5.5</td>
<td>+.25</td>
</tr>
<tr>
<td>7</td>
<td>+.41</td>
</tr>
<tr>
<td>8</td>
<td>+.60</td>
</tr>
<tr>
<td>9</td>
<td>+.70</td>
</tr>
</tbody>
</table>

\[ W = \left| 40 - 5 \right| = 35 \]
Wilcoxon Signed-Rank Test

\[ W = |40 - 5| = 35 \]

\[ p-value = 0.03 \]
## Sign Test

<table>
<thead>
<tr>
<th>Example</th>
<th>A</th>
<th>B</th>
<th>B-A</th>
<th>B &gt; A?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
<td>.35</td>
<td>+.10</td>
<td>+1</td>
</tr>
<tr>
<td>2</td>
<td>.43</td>
<td>.84</td>
<td>+.41</td>
<td>+1</td>
</tr>
<tr>
<td>3</td>
<td>.39</td>
<td>.15</td>
<td>-.24</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>.75</td>
<td>.75</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>.43</td>
<td>.68</td>
<td>+.25</td>
<td>+1</td>
</tr>
<tr>
<td>6</td>
<td>.15</td>
<td>.85</td>
<td>+.70</td>
<td>+1</td>
</tr>
<tr>
<td>7</td>
<td>.20</td>
<td>.80</td>
<td>+.60</td>
<td>+1</td>
</tr>
<tr>
<td>8</td>
<td>.52</td>
<td>.50</td>
<td>-.02</td>
<td>-1</td>
</tr>
<tr>
<td>9</td>
<td>.49</td>
<td>.58</td>
<td>+.09</td>
<td>+1</td>
</tr>
<tr>
<td>10</td>
<td>.50</td>
<td>.75</td>
<td>+.25</td>
<td>+1</td>
</tr>
</tbody>
</table>

\[ S = 7 \]

\[ p(7 \mid 10 \, \text{trials, } \frac{1}{2} \, \text{probability}) = 0.05 \]
Randomization Test

<table>
<thead>
<tr>
<th>Example</th>
<th>A</th>
<th>B</th>
<th>B-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
<td>.35</td>
<td>+.10</td>
</tr>
<tr>
<td>2</td>
<td>.84</td>
<td>.43</td>
<td>-.41</td>
</tr>
<tr>
<td>3</td>
<td>.39</td>
<td>.15</td>
<td>-.24</td>
</tr>
<tr>
<td>4</td>
<td>.75</td>
<td>.75</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>.68</td>
<td>.43</td>
<td>-.25</td>
</tr>
<tr>
<td>6</td>
<td>.15</td>
<td>.85</td>
<td>+.70</td>
</tr>
<tr>
<td>7</td>
<td>.80</td>
<td>.20</td>
<td>-.60</td>
</tr>
<tr>
<td>8</td>
<td>.50</td>
<td>.52</td>
<td>+.02</td>
</tr>
<tr>
<td>9</td>
<td>.58</td>
<td>.49</td>
<td>0.09</td>
</tr>
<tr>
<td>10</td>
<td>.75</td>
<td>.50</td>
<td>-.25</td>
</tr>
</tbody>
</table>

\[ \hat{\mu}_0 = \overline{B} - A = 0.214 \]
\[ \hat{\mu}_1 = -0.008 \]
\[ \hat{\mu}_2 = -0.093 \]
Randomization Test

\[ \hat{\mu}_0 = \bar{B} - A = 0.214 \]

\[ p - value = 0.02 \]
# Bootstrap Test

<table>
<thead>
<tr>
<th>Example</th>
<th>A</th>
<th>B</th>
<th>B-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
<td>.35</td>
<td>+.10</td>
</tr>
<tr>
<td>2</td>
<td>.43</td>
<td>.84</td>
<td>+.41</td>
</tr>
<tr>
<td>3</td>
<td>.39</td>
<td>.15</td>
<td>-.24</td>
</tr>
<tr>
<td>4</td>
<td>.75</td>
<td>.75</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>.43</td>
<td>.68</td>
<td>+.25</td>
</tr>
<tr>
<td>6</td>
<td>.15</td>
<td>.85</td>
<td>+.70</td>
</tr>
<tr>
<td>7</td>
<td>.20</td>
<td>.80</td>
<td>+.60</td>
</tr>
<tr>
<td>8</td>
<td>.52</td>
<td>.50</td>
<td>-.02</td>
</tr>
<tr>
<td>9</td>
<td>.49</td>
<td>.58</td>
<td>+.09</td>
</tr>
<tr>
<td>10</td>
<td>.50</td>
<td>.75</td>
<td>+.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>s1</th>
<th>s2</th>
<th>s3</th>
</tr>
</thead>
<tbody>
<tr>
<td>-24</td>
<td>+.25</td>
<td>-.24</td>
<td></td>
</tr>
<tr>
<td>+.41</td>
<td>+.10</td>
<td>+.60</td>
<td></td>
</tr>
<tr>
<td>-.02</td>
<td>+.25</td>
<td>-.70</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>+.60</td>
<td>+.25</td>
<td></td>
</tr>
<tr>
<td>+.25</td>
<td>+.70</td>
<td>+.70</td>
<td></td>
</tr>
<tr>
<td>+.10</td>
<td>-.02</td>
<td>+.41</td>
<td></td>
</tr>
<tr>
<td>+.25</td>
<td>+.10</td>
<td>-.02</td>
<td></td>
</tr>
<tr>
<td>+.10</td>
<td>+.25</td>
<td>-.24</td>
<td></td>
</tr>
<tr>
<td>+.25</td>
<td>0</td>
<td>+.70</td>
<td></td>
</tr>
<tr>
<td>+.10</td>
<td>-.02</td>
<td>+.25</td>
<td></td>
</tr>
</tbody>
</table>
Bootstrap Distribution

$p-value = 0.005$
ANOVA

- Compare variance due to system to variance due to topic

<table>
<thead>
<tr>
<th>Example</th>
<th>A</th>
<th>B</th>
<th>B-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.25</td>
<td>0.35</td>
<td>+0.10</td>
</tr>
<tr>
<td>2</td>
<td>0.43</td>
<td>0.84</td>
<td>+0.41</td>
</tr>
<tr>
<td>3</td>
<td>0.39</td>
<td>0.15</td>
<td>-0.24</td>
</tr>
<tr>
<td>4</td>
<td>0.75</td>
<td>0.75</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.43</td>
<td>0.68</td>
<td>+0.25</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
<td>0.85</td>
<td>+0.70</td>
</tr>
<tr>
<td>7</td>
<td>0.20</td>
<td>0.80</td>
<td>+0.60</td>
</tr>
<tr>
<td>8</td>
<td>0.52</td>
<td>0.50</td>
<td>-0.02</td>
</tr>
<tr>
<td>9</td>
<td>0.49</td>
<td>0.58</td>
<td>+0.09</td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.75</td>
<td>+0.25</td>
</tr>
</tbody>
</table>

\[ \hat{\sigma}^2 = MSE = 0.042 \]

\[ \hat{\sigma}_S^2 = MST = 0.229 \]

\[ F = \frac{MST}{MSE} = 5.41 \]
Summary

• These are 6 of the most common tests seen in IR experimentation
  – Many others in the literature:
    • Chi-squared
    • Proportion test
    • ANCOVA/MANOVA/MANCOVA

• All have in common:
  – The use of some probability distribution, computation of a p-value from that distribution
Part 2

FUNDAMENTALS OF SIGNIFICANCE TESTING
Testing Paradigms

Ronald Fisher

Harold Jeffreys

Jerzy Neyman

Egon Pearson
What Are Tests Really Telling Us?

• Formal set-up:
  - $H_0: \mu = 0$
  - $H_1: \mu \neq 0$

• The null hypothesis is a model
  - We are looking to prove the model false

• The p-value is the probability that you would have found the same results if $H_0$ were true
  - If that probability is low, conclude $H_0$ is false
What Are Tests Really Telling Us?

• Fisher: p-value is the likelihood of the data under $H_0$
  – The p-value is a conclusion about this particular experiment only
  – Nothing more, nothing less

• Neyman-Pearson: $p < 0.05$ means we can reject $H_0$ as being unlikely to be true
  – p-values lead to inference about the population
  – The p-value itself is not interesting; the inference is
  – Note that we do not accept that $H_1$ is true!

• Jeffreys: posterior probability of $H_0$ being true can be compared to posterior probability of other models
Terms and Definitions

• Single-sample vs two-sample tests
  – A single-sample test is generally based on applying one or more “treatments” (search algorithms) to a single sample of “subjects” (queries/topics)
  – In a two-sample test, different samples are used for each treatment

• Paired vs unpaired
  – Paired tests are a special case of single-sample tests: subtract evaluation results for each example to obtain the measurements to summarize
  – Unpaired tests can be single-sample too
Test Statistics and Distributions

• Test statistic
  – A summary of the data, usually designed to have specific distribution guarantees (asymptotically)

• Parametric vs non-parametric
  – If the test statistic distribution has any free parameters, the test is said to be “parametric”

• Confidence interval
Sizes and Values

• Sample size
  – The number of subjects/examples in the experiment
  – Assumed to be sampled i.i.d. from a much larger population

• Effect size
  – A measure of the difference between two “treatments” or algorithms in the population
  – Independent of sample size
  – $H_0$: no effect

• p-value
  – The likelihood of observing the results assuming $H_0$ is true

• Critical value
  – The minimum effect size necessary to obtain $p < \alpha$ with a given sample size
  – $\alpha$ usually = 0.05
Variance

• Total variance
  – The sum of the square differences between measurements and the overall mean

• Within-group variance
  – Variance due to instances in the sample
  – Paired tests subtract this variance out

• Between-group variance
  – Variance due to the treatments/systems
Accuracy and Power

- **Accuracy**
  - The probability of getting $p \geq \alpha$ when $H_0$ is actually true
  - Probability of correctly not rejecting $H_0$
  - Proportional to false positive rate

- **Power**
  - The probability of getting $p < \alpha$ when the null hypothesis is actually false
  - The probability of correctly rejecting $H_0$
  - True positive rate

- **Most tests are defined to have a false positive rate of $\alpha$ when $H_0$ is true**
  - Achieving a certain power level involves estimating effect size and sample size

<table>
<thead>
<tr>
<th></th>
<th>$H_0$</th>
<th>true</th>
<th>false</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not</td>
<td>accuracy</td>
<td>Type I</td>
</tr>
<tr>
<td>rejected</td>
<td></td>
<td></td>
<td>error</td>
</tr>
<tr>
<td>rejected</td>
<td></td>
<td>Type II</td>
<td>power</td>
</tr>
</tbody>
</table>
The Linear Model

• Statistical tests are classifiers
  – Like classifiers, they are based on an underlying model
  – Unlike classifiers, we cannot evaluate them directly

• The t-test is based on the linear regression model
  \[ y_i = \beta_0 + \beta_1 x + \epsilon_i \]
All models are wrong, but some are useful.

George E. P. Box
Myths and Misconceptions

• Significance tests lend rigor to our experimentation
  – Without them, the usual differences of < 5% would be difficult to interpret

• But they are widely misunderstood
  – p-values can be incorrectly interpreted
  – p-values can be easily manipulated (even unintentionally)

• They are fundamentally no more rigorous than any AI approach to classification
  – Though they may have a much deeper theoretical basis
Myth: $H_0$ is a Realistic Model

• The first and biggest misconception: the null hypothesis is sometimes true
  — That is, there is a chance that there really is no effect

• In AI, the null hypothesis is almost never true
  — Really only when the experimenter made a mistake

• The only question is how big of a sample size will it take to reject it
  — There is always some sample big enough to reject it
Myth: Rejecting $H_0$ Means it is False

• First, $H_0$ is always false

• But even if it were true, we could still reject it for many reasons:
  – something about our sample
  – violations of test model assumptions
  – failure to model important sources of variance
  – unintentional overfitting

• Rejecting $H_0$ should not be taken to mean our system is definitely better
Myth: Test Assumptions Are Important

• Consider the t-test based on the linear model

• Assumptions:
  – y is unbounded
  – linearity and additivity
  – homoscedasticity
  – normality of errors
  – (note: normality of data is not an assumption)

• All of these are false!
  – But we can evaluate how much their falseness affects accuracy and power
Myth: Test Assumptions Are Important

• OK, so t-test assumptions are false. Why not use a different test?

• Every test is based on some model, and every model is false
  – Even so-called “assumption-free” tests like Fisher’s exact test or the bootstrap actually do involve assumptions

• The tradeoff is between simplicity and power
  – Fewer assumptions → less power → fewer significant results

• t-test is popular because it is powerful, robust to violations of its assumptions, and computationally easy
Myth: p-Values Have Intrinsic Meaning

- p < 0.05 is often taken as a “gold standard” of proof

- Two things to keep in mind:
  - The p-value comes out of a model; “all models are wrong”
  - 0.05 is an arbitrary value that was probably first used as an example

- Any meaning given to a p-value is extrinsic
  - Usually granted by a community of scientists
Myth: p-Values Have Intrinsic Meaning

• The real gold standard is whether it helps users

• Any IR evaluation based on the Cranfield paradigm cannot directly answer that
Myth: Lower p-Values are Better

• If a p-value of 0.04 is better than a p-value of 0.06, then a p-value of 0.02 is even better, right?

• A p-value can be lower for three reasons:
  – The effect size is bigger (good)
  – The sample size is bigger (bad)
  – Modeling effects, including random effects

• There’s no way to know which of these is the reason
Myth: Lower p-Values are Better

• \( p\)-value = \( P(\text{data} \mid H_0, \text{test model}, \text{inputs}) \)

• Any change to the underlying model results in a different probability distribution
  — That includes changes to the systems being tested

• \( p\)-values should not be compared directly
  — Fisher and Neyman/Pearson would have agreed on this!
Myth: Running Many Tests is OK

• AI experimentation often happens like this:
  – Implement system, compare to baseline, run test
  – Not significant?
    • Re-implement system, compare to baseline, run test
  – Significant?
    • Start writing a paper

• How many tests does it take to get to the endpoint?
Sequential Testing

• Suppose (hypothetically) that the null hypothesis is actually true

• The probability of concluding it is false after one test is $\alpha$ (normally 0.05)
  – The probability of concluding it is false after two tests is $0.05 + 0.95 \times 0.05 = 0.0975$
  – After three tests, $0.05 + 0.95 \times 0.05 + 0.95 \times 0.95 \times 0.05 = 0.143$
  – After 14 tests, $\sim 0.5$
  – After 27 tests, $\sim 0.75$
  – After 90 tests, $\sim 0.99$
Multiple Testing

• Suppose three different people have the same null hypothesis
  – If each of them does one experiment, probability that there will be one false positive is 0.143
  – If each of them does three experiments, probability goes to ~0.4

• Result: very high probability that any given published result is false!
  – “Why Most Published Research Findings Are False”, Ioannidis, PLoS Medicine, 2005
Multiple Testing

![Graph showing the probability of at least one false significant result versus the number of experiments with H₀ true for different significance levels (α = 0.10, α = 0.05, α = 0.01).]
Correcting for Multiple Comparisons

• We should adjust our p-values up for the fact that we have made multiple comparisons

• Many different approaches:
  – Bonferroni correction
  – Tukey’s Honest Significant Differences
  – Multivariate t test
Tukey’s HSD

• Omnibus hypothesis:
  – $H_0$: $S_1 = S_2 = S_3 = \ldots = S_n$
  – ANOVA fits a linear model to all data; rejects null if there is any difference between any pair of systems

• The maximum difference is the one most likely to be the cause of rejection

• Tukey: compute a distribution of maximum difference, base all p-values on that
Tukey’s HSD
Effect on TREC-8 Evaluation
Families of Experiments

• p-values should be adjusted based on “families” of experiments
  – All experiments testing the same hypothesis

• How do we define a family of experiments?
  – Suppose we are testing hypotheses about clustering for IR
    • H: augmenting LM retrieval with clusters improves ad hoc retrieval
    • H: augmenting BM25 retrieval with clusters improves ad hoc retrieval
    • H: augmenting any ranking function with clusters improves ad hoc retrieval
    • H: clusters are good for retrieval
What is a “Family”?

• In TREC data, families could be:
  – All pairs of submitted systems
  – Pairs of systems submitted by each participating group in the context of the full set of systems
  – Pairs of systems submitted by each participating group in the context of just that group’s systems

• p-values can be corrected based on each family type
  – Which results in different adjustments for each

• The third is the least “honest”, yet is really the only thing you can do on your own
Summary

• **Significance tests are just models**
  – When we use them “out of the box”, we fail to model many sources of variance in IR
    • Variance in relevance, in user behavior, in interactions between system components, …
  – The things we do model are probably being modeled wrong
    • In particular additivity of system and topic effects
  – More correct models could change our conclusions about systems

• We know that modeling multiple testing changes our conclusions dramatically
  – Most other concerns are extremely minor in comparison
  – But we don’t really truly know how to adjust for multiple comparisons
    • It depends very much on what other researchers are thinking and doing

• The one thing I want you to take away from this talk:
  – Never trust a p-value, even one you computed yourself
Part 3

SIGNIFICANCE TESTING IN IR RESEARCH
What Does it Mean?

• You can *always* find significance
  – With the right sample, the right sample size, the right test, enough iterations of testing
  – “Fishing expeditions”

• Significance is only a rough proxy for “interestingness”
  – A heuristic
Searching for Interesting Results

• How do we use significance tests in research?
  – Conference program committees/journal editors use them as a guide for determining what to publish
    • Publication determines research directions that people follow
      – Published systems implemented as baselines
      – Essentially as a heuristic in a search for the best algorithms

• They can easily be used as a substitute for human judgment
  – Like most AI, they should be used as an aide to human judgment
  – There isn’t one right way to do it
    • No Free Lunch Theorem applies to significance testing
Searching for Interesting Results

• What if significance was granted more conservatively? e.g. by:
  – Correcting for multiple comparisons
  – Using tests that make fewer assumptions
  – Using a lower value of alpha (0.01 for instance)

• Is a more conservative heuristic always better?
All hypotheses

Statistically significant results

Published results

Interesting results

The State of Research Today

InteresKng results
The State of Research When “Significance” is Granted More Conservatively

Fewer publications overall ... which means fewer uninteresting publications ... but also that fewer truly interesting results can be published
Thought Experiment

• Suppose statistical significance is a necessary and sufficient condition for publication
  – Consequences:
    • Many published papers are not interesting
    • Some interesting results are not published
    • Most uninteresting results are not published

  – Published uninteresting papers ->
    • time wasted reading, re-implementing

  – Unpublished interesting results ->
    • time wasted each time results are re-discovered

  – Unpublished uninteresting papers ->
    • time wasted each time experiment is tried and fails
Example: NLP for IR

• NLP generally doesn’t work for IR
  – Maybe in some domains, for some tasks, but in general not

• But almost every IR grad student has had some idea for using NLP to improve IR
  – Result: a handful of published papers from a very large number of experiments, mostly due to randomness
    • e.g. multiple testing
  – … which gives hope to the next generation of students (who don’t know about the very low success rate)
  – … which results in a lot of wasted time as they re-do experiments already done by every previous generation
Example: NLP for IR

• Would we have been better off had that handful of papers never been published?

• Would we have been better off if all those negative results *had* been published?

• Or are we better off with grad students having done the work to gain some intuition about why it doesn’t work?
Takeaways

• Always do significance tests
  – But don’t worry too much about which ones to do
  – The t-test is always a good option
  – Correcting for multiple testing is probably not necessary

• Don’t just report p-values or * to indicate significance
  – Always report estimated effect sizes and confidence intervals

• Always take results of tests with a grain of salt
  – Especially when the effect size is low
  – Build your intuition and use it

• Never say “barely significant” or “just missed being significant”