Not All Claims are Created Equal: Choosing the Right Statistical Approach to Assess Hypotheses

Erfan Sadeqi-Azer
Indiana University (now at Google)

Daniel Khashabi
Allen Institute for AI

Ashish Sabharwal
Allen Institute for AI

Dan Roth
Univ. of Pennsylvania
This work
This work

Q: What is this work about?

Q: What do you mean by “hypothesis”?

Q: Why should I care about hypothesis assessment?
This work

Q: What is this work about?

Different hypothesis assessment algorithms and their comparison

Q: What do you mean by “hypothesis”?

Q: Why should I care about hypothesis assessment?
This work

Q: What is this work about?

Different hypothesis assessment algorithms and their comparison

Q: What do you mean by “hypothesis”?

Q: Why should I care about hypothesis assessment?
This work

Q: What is this work about?
Different hypothesis assessment algorithms and their comparison

Q: What do you mean by “hypothesis”?
it’s a prediction, based on certain assumptions & observations
e.g., classifier-1 is inherently better than classifier-2

Q: Why should I care about hypothesis assessment?
This work

Q: What is this work about?

Different hypothesis assessment algorithms and their comparison

Q: What do you mean by “hypothesis”?

it’s a prediction, based on certain assumptions & observations
e.g., classifier-1 is inherently better than classifier-2

Q: Why should I care about hypothesis assessment?
Q: What is this work about?
Different hypothesis assessment algorithms and their comparison

Q: What do you mean by “hypothesis”?
it’s a prediction, based on certain assumptions & observations
   e.g., classifier-1 is inherently better than classifier-2

Q: Why should I care about hypothesis assessment?
Like any empirical field, in NLP we need to follow scientific principles for drawing conclusions.
Statistical tools considered in this work

- p-value
- Bayes Factor
- Confidence Interval
- Posterior Interval
Contributions
Contributions

• Quantify **usage trends** in NLP community:
  • Annotated ACL’18 papers (~440 papers)
  • Surveyed ~50 random NLP practitioners

• Findings:
  • *Lack of awareness* about various algorithms.
  • *Poor interpretation* of statistical tools – especially the popular ones.
  • *Misleading reporting*, resulting in unintended conclusions.

• A Python **package** for *Bayesian statistical hypothesis assessment*
Contributions

• Quantify usage trends in NLP community:
  • Annotated ACL’18 papers (~440 papers)
  • Surveyed ~50 random NLP practitioners

• Findings:
  • Lack of awareness about various algorithms.
  • Poor interpretation of statistical tools – especially the popular ones.
  • Misleading reporting, resulting in unintended conclusions.

• A Python package for Bayesian statistical hypothesis assessment
Contributions

• Quantify usage trends in NLP community:
  • Annotated ACL’18 papers (~440 papers)
  • Surveyed ~50 random NLP practitioners

• Findings:
  • Lack of awareness about various algorithms.
  • Poor interpretation of statistical tools – especially the popular ones.
  • Misleading reporting, resulting in unintended conclusions.

• A Python package for Bayesian statistical hypothesis assessment
Contributions

• Quantify **usage trends** in NLP community:
  • Annotated ACL’18 papers (~440 papers)
  • Surveyed ~50 random NLP practitioners

• Findings:
  • **Lack of awareness** about various algorithms.
  • **Poor interpretation** of statistical tools – especially the popular ones.
  • **Misleading reporting**, resulting in unintended conclusions.

• A Python **package** for *Bayesian statistical hypothesis assessment*
Contributions

• Quantify **usage trends** in NLP community:
  • Annotated ACL’18 papers (~440 papers)
  • Surveyed ~50 random NLP practitioners

• Findings:
  • **Lack of awareness** about various algorithms.
  • **Poor interpretation** of statistical tools – especially the popular ones.
  • **Misleading reporting**, resulting in unintended conclusions.

• A Python **package** for **Bayesian statistical hypothesis assessment**
Contributions

• Quantify **usage trends** in NLP community:
  • Annotated ACL’18 papers (~440 papers)
  • Surveyed ~50 random NLP practitioners

• Findings:
  • **Lack of awareness** about various algorithms.
  • **Poor interpretation** of statistical tools – especially the popular ones.
  • **Misleading reporting**, resulting in unintended conclusions.

• A Python **package** for **Bayesian statistical hypothesis assessment**
Contributions

• Quantify **usage trends** in NLP community:
  • Annotated ACL’18 papers (~440 papers)
  • Surveyed ~50 random NLP practitioners

• Findings:
  • **Lack of awareness** about various algorithms.
  • **Poor interpretation** of statistical tools – especially the popular ones.
  • **Misleading reporting**, resulting in unintended conclusions.

• **A Python package** for *Bayesian statistical hypothesis assessment*
Contributions

• Quantify usage trends in NLP community:
  • Annotated ACL’18 papers (~440 papers)
  • Surveyed ~50 random NLP practitioners

• Findings:
  • Lack of awareness about various algorithms.
  • Poor interpretation of statistical tools – especially the popular ones.
  • Misleading reporting, resulting in unintended conclusions.

• A Python package for Bayesian statistical hypothesis assessment
  https://github.com/allenai/HyBayes
A Typical AI Experiment

<table>
<thead>
<tr>
<th>System</th>
<th>$\hat{\theta}$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
<td>?</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
<td>?</td>
</tr>
</tbody>
</table>

(Clark et al., 2018) $|D|= 2376$
Empirical performance

<table>
<thead>
<tr>
<th>System</th>
<th>$\hat{\theta}$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>72.4</td>
<td>?</td>
</tr>
<tr>
<td>$B$</td>
<td>68.9</td>
<td>?</td>
</tr>
</tbody>
</table>

(Clark et al., 2018) $|D|= 2376$
A Typical AI Experiment

Empirical performance

<table>
<thead>
<tr>
<th>System</th>
<th>( \hat{\theta} )</th>
<th>( \theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
<td>?</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
<td>?</td>
</tr>
</tbody>
</table>

(Clark et al., 2018) \( |D| = 2376 \)

Inherent performance
A Typical AI Experiment

• The apparent difference in empirical performances be explained simply by random chance.

$$H: \theta_A = \theta_B$$

• We have sufficient evidence to conclude that A is in fact inherently stronger than B.

<table>
<thead>
<tr>
<th>System</th>
<th>$\hat{\theta}$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
<td>?</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
<td>?</td>
</tr>
</tbody>
</table>

(Clark et al., 2018) $|D|= 2376$
A Typical AI Experiment

- The apparent difference in empirical performances be explained simply by random chance.

\[ H: \theta_A = \theta_B \]

- We have sufficient evidence to conclude that \( A \) is in fact inherently stronger than \( B \).

\[ H: \theta_A > \theta_B + \alpha \]

---

<table>
<thead>
<tr>
<th>System</th>
<th>( \hat{\theta} )</th>
<th>( \theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>72.4</td>
<td>?</td>
</tr>
<tr>
<td>( B )</td>
<td>68.9</td>
<td>?</td>
</tr>
</tbody>
</table>

(Clark et al., 2018)  \(|D|= 2376\)
A Typical AI Experiment

- The apparent difference in empirical performances be explained simply by random chance.

\[ \mathcal{H}: \theta_A = \theta_B \]

- We have sufficient evidence to conclude that \( A \) is in fact inherently stronger than \( B \).

\[ \mathcal{H}: \theta_A > \theta_B + \alpha \]

<table>
<thead>
<tr>
<th>System</th>
<th>( \hat{\theta} )</th>
<th>( \theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>72.4</td>
<td>?</td>
</tr>
<tr>
<td>( B )</td>
<td>68.9</td>
<td>?</td>
</tr>
</tbody>
</table>

(Clark et al., 2018) \(|D|= 2376\)
Statistical tools: big picture

- p-value
- Bayes Factor
- Confidence Interval
- Posterior Intervals
Statistical tools: big picture

- p-value
- Bayes Factor
- Confidence Interval
- Posterior Intervals
Suppose I want to assess a hypothesis $H$.

Idea: assuming that an opposite hypothesis is true, compute the likelihood of an observation as “extreme” as what’s observed.

$H$: $\theta_A > \theta_B$
Statistical tools: big picture

• Suppose I want to assess a hypothesis $H$.

• **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an observation as “extreme” as what’s observed.

$H: \theta_A > \theta_B$
Statistical tools: big picture

• Suppose I want to assess a hypothesis $H$.

$$H: \theta_A > \theta_B$$

$$\overline{H}: \theta_A = \theta_B$$

• Idea: assuming that an opposite hypothesis is true, compute the likelihood of an observation as “extreme” as what’s observed.
Statistical tools: big picture

• Suppose I want to assess a hypothesis $H$.

$$\overline{H} : \theta_A = \theta_B$$

• **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an observation as "extreme" as what’s observed.
Statistical tools: big picture

• Suppose I want to assess a hypothesis $H$.

\[ H: \theta_A > \theta_B \]

\[ \bar{H}: \theta_A = \theta_B \]

• Idea: assuming that an opposite hypothesis is true, compute the likelihood of an observation as “extreme” as what’s observed.

\[ P(\text{obs.} > \hat{\theta}_A - \hat{\theta}_B | \bar{H}) \]

the accuracy gap between the two systems
• Suppose I want to assess a hypothesis $H$.

$$H: \theta_A > \theta_B$$

$$\bar{H}: \theta_A = \theta_B$$

• **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an observation as “extreme” as what’s observed.

$$P(\text{obs.} > \hat{\theta}_A - \hat{\theta}_B | \bar{H})$$

**p-value**

---

**the accuracy gap between the two systems**
Statistical tools: big picture

- p-value
- Bayes Factor
- Confidence Interval
- Posterior Intervals
Statistical tools: big picture

- p-value
- Bayes Factor
- Confidence Interval
- Posterior Intervals
• Suppose I want to assess a hypothesis $H$.

• **Idea:** use the Bayes formula to compute a probability for the hypothesis being true.

<table>
<thead>
<tr>
<th>p-value</th>
<th>Bayes Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence Interval</td>
<td>Posterior Intervals</td>
</tr>
</tbody>
</table>
• Suppose I want to assess a hypothesis $H$.

• **Idea:** use the Bayes formula to compute a probability for the hypothesis being true.
Suppose I want to assess a hypothesis $H$.

**Idea:** use the Bayes formula to compute a probability for the hypothesis being true.

$$P(\text{observations} | H)$$

$H: \theta_A > \theta_B + \alpha$
• Suppose I want to assess a hypothesis $H$.

• **Idea:** use the Bayes formula to compute a probability for the hypothesis being true.

$$H: \theta_A > \theta_B + \alpha$$

$$P(H|\text{observations})$$
Statistical tools: big picture

• Suppose I want to assess a hypothesis \( H \).

• Idea: use the Bayes formula to compute a probability for the hypothesis being true.

\[
H: \theta_A > \theta_B + \alpha
\]

\[
P(H|\text{observations})
\]
Usage Patterns

Ease of Interpretation
Usage Patterns

Ease of Interpretation
Trends and Patterns in the field

Study NLP conference papers: ACL’18 papers (439 papers)

How many papers did use significance testing?
Study **NLP conference papers**: ACL’18 papers (439 papers)

*How many papers did use significance testing?*
Study **NLP conference papers**: ACL’18 papers (439 papers)

*How many papers did use significance testing?*

- **73** papers used p-value
- **6** papers used Bayes Factor
- **6** papers used Confidence Interval
- **6** papers used Posterior Intervals
Study **NLP conference papers**: ACL’18 papers (439 papers)

*How many papers did use significance testing?*
Trends and Patterns in the field

Study NLP conference papers: ACL’18 papers (439 papers)

How many papers did use significance testing?

- p-value: 73
- Bayes Factor: 0
- Confidence Interval: 6
- Posterior Intervals: 0
Study **NLP conference papers**: ACL’18 papers (439 papers)

*How many papers did use significance testing?*

- Many papers (~360) did **not** include any hypothesis assessment.
- p-value based tests are the **dominant** choice among NLP practitioners.
Study **NLP conference papers**: ACL’18 papers (439 papers)

*How many papers did use significance testing?*

- Many papers (~360) did **not** include any hypothesis assessment.
- p-value based tests are the **dominant** choice among NLP practitioners.
Study **NLP conference papers**: ACL’18 papers (439 papers)

**How many papers did use significance testing?**

- Many papers (~360) did **not** include any hypothesis assessment.
- p-value based tests are the **dominant** choice among NLP practitioners.

Why?
Lack of exposure to alternative algorithms
Lack of exposure to alternative algorithms

- The imbalance in usage:
  - Is it intentional?

- Many people did not know the definition of “Bayes Factor.” 😐
Lack of exposure to alternative algorithms

• The imbalance in usage:
  • Is it intentional?

• Many people did not know the definition of “Bayes Factor.” 😐

Do you know the definition of "Bayes Factor"?

78.2% Yes
21.8% No
Lack of exposure to alternative algorithms

• The imbalance in usage:
  • Is it intentional?

• Many people did not know the definition of “Bayes Factor.” 😐

Do you know the definition of "Bayes Factor"?
Lack of exposure to alternative algorithms

• The imbalance in usage:
  • Is it intentional?

• Many people did not know the definition of “Bayes Factor.” 😐

Do you know the definition of “Bayes Factor”?

We don’t teach the alternatives in our AI curriculum.
Usage Patterns

Ease of Interpretation
Usage Patterns

Ease of Interpretation

- NLP community is over-using certain techniques.
- One reason could be researchers’ lack of exposure to the alternatives.
Usage Patterns

Ease of Interpretation
Are we good at interpreting the p-values?

\[ P(\text{extreme obs.} \mid \bar{H}) \ll \alpha \]

\[ \text{p-value} \]
Are we good at interpreting the p-values?

\[ P(\text{extreme obs.} \mid \overline{H}) \ll \alpha \]

- Pretty complex notion!
Are we good at interpreting the p-values?

\[ P(\text{extreme obs.} \mid \overline{H}) \ll \alpha \]

- Pretty complex notion!

"The probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null-hypothesis is correct." -- your favorite statistics textbook
A Survey Question: Interpreting P-value (1)
Question 1: do you know p-values and its interpretation?
• **Question 1:** *do you know p-values and its interpretation?*
A Survey Question: Interpreting P-value (1)

• **Question 1:** *do you know p-values and its interpretation?*

86% expressed fair-to-complete confidence in their ability to interpret p-values.
A Survey Question: Interpreting P-value (2)

<table>
<thead>
<tr>
<th>system</th>
<th>$\hat{\theta}$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>classifier-A</td>
<td>38%</td>
<td>?</td>
</tr>
<tr>
<td>classifier-B</td>
<td>45%</td>
<td>?</td>
</tr>
</tbody>
</table>
A Survey Question: Interpreting P-value (2)

The authors claim that the improvement of B over A is "statistically significant" with a significance level of 0.01. Which of the followings is correct?

a) The probability of observing a margin 7% is at most 0.01, assuming that the two classifiers inherently have the same performance.

b) With a probability 99% classifier-2 will have a higher performance than classifier-1.

…
• The authors claim that the improvement of $B$ over $A$ is “statistically significant” with a significance level of 0.01. Which of the followings is correct?

a) The probability of observing a margin 7% is at most 0.01, assuming that the two classifiers inherently have the same performance.

b) With a probability 99% classifier-2 will have a higher performance than classifier-1.

<table>
<thead>
<tr>
<th>system</th>
<th>$\theta$</th>
<th>$\theta'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>classifier-A</td>
<td>38%</td>
<td>?</td>
</tr>
<tr>
<td>classifier-B</td>
<td>45%</td>
<td>?</td>
</tr>
</tbody>
</table>
A Survey Question: Interpreting P-value (2)

• The authors claim that the improvement of B over A is "statistically significant" with a significance level of 0.01. Which of the followings is correct?

- a) The probability of observing a margin 7% is at most 0.01, assuming that the two classifiers inherently have the same performance.  
  
P[\hat{\theta}_B - \hat{\theta}_A > 7 | \theta_A = \theta_B] < 0.01

- b) With a probability 99% classifier-2 will have a higher performance than classifier-1.

  P[\theta_B > \theta_A] > 0.99
The authors claim that the improvement of B over A is "statistically significant" with a significance level of 0.01. Which of the followings is correct?

- a) The probability of observing a margin 7% is at most 0.01, assuming that the two classifiers inherently have the same performance.
  \[ P[\hat{\theta}_B - \hat{\theta}_A > 7 | \theta_A = \theta_B] < 0.01 \]

- b) With a probability 99% classifier-2 will have a higher performance than classifier-1.
  \[ P[\theta_B > \theta_A] > 0.99 \]

Only a small percentage correctly answered a basic p-value interpretation question.
Ease of interpretation: Bayesians vs Freq.

- **Frequentist**: p-value, Confidence Interval
- **Bayesian**: Bayes Factor, Posterior Intervals
Ease of interpretation: Bayesians vs Freq.

Our participants mistakenly interpret frequentist notions in a Bayesian way.
Ease of interpretation: Bayesians vs Freq.

Our participants mistakenly interpret frequentist notions in a Bayesian way.
Ease of interpretation: Bayesians vs Freq.

Our participants mistakenly interpret frequentist notions in a Bayesian way.
Usage Patterns

Ease of Interpretation

• While p-valued based tests are the most popular choice among NLP practitioners, they’re difficult to understand and highly prone to misunderstanding.

• Bayesian Intervals provide results that are more natural to interpret.
Summary

• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • ...

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes
Summary

• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • ...

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes
Summary

• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • ...

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes
• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • ...

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes
Summary

• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • …

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes
Summary

• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • ...

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes
• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • ...

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes
Summary

• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • ...

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes
Summary

• The work surveys four different alternatives for hypothesis assessment:
  • Details in the paper

• We provide comparisons among these algorithms:
  • Whether their easy to interpret / misinterpret
  • ...

• We compare usage patterns:
  • Surveying the field
  • Manual annotation of papers

• HyBayes: https://github.com/allenai/HyBayes