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

arxiv.org/abs/1911.03850

Erfan Sadeqi-Azer (Indiana U → Google)  Ashish Sabharwal (AI2)  Dan Roth (UPenn)
The Cycle of Empirical Research
The Cycle of Empirical Research
The Cycle of Empirical Research
The Cycle of Empirical Research

- Hypothesis, or a General Premise
- Empirical Observations

Research Problem
The Cycle of Empirical Research

- Hypothesis, or a General Premise
- Empirical Observations
- ??? Research Problem
- Assessment/Verification
The Cycle of Empirical Research

Hypothesis, or a General Premise

Empirical Observations

Assessment/Verification

???? Research Problem
Hypotheses
Hypotheses

• A prediction about how the world will behave if our idea is correct
• Worded as an if-then statement
• A hypothesis is a testable prediction
• A hypothesis is a falsifiable statement

• Terminology:
  • A hypothesis is never “proved”
  • But it could be “supported” by the evidence
Hypotheses

• A **prediction** about how the world will behave if our idea is correct
• Worded as an **if-then** statement
• A hypothesis is a **testable** prediction
• A hypothesis is a **falsifiable** statement

• Terminology:
  • A hypothesis is **never “proved”**
  • But it could be “**supported**” by the evidence
Hypotheses

• A prediction about how the world will behave if our idea is correct
• Worded as an if-then statement
• A hypothesis is a testable prediction
• A hypothesis is a falsifiable statement

• Terminology:
  • A hypothesis is never “proved”
  • But it could be “supported” by the evidence
Hypotheses

• A prediction about how the world will behave if our idea is correct
• Worded as an if-then statement
• A hypothesis is a testable prediction
• A hypothesis is a falsifiable statement

• Terminology:
  • A hypothesis is never “proved”
  • But it could be “supported” by the evidence
Hypotheses

• A **prediction** about how the world will behave **if our idea is correct**
• Worded as an **if-then** statement
• A hypothesis is a **testable** prediction
• A hypothesis is a **falsifiable** statement

• Terminology:
  • A hypothesis is **never “proved”**
  • But it could be **“supported”** by the evidence
2020

<

2020

?
Not a good statistical hypothesis
“I can always prepare a nice presentation, if I stay up the night before.”
A Typical AI Experiment

1. Research Problem
2. Empirical Observations
3. Hypothesis, or a General Premise

???? Research Problem

Empirical Observations
A Typical AI Experiment

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

|D| = 2376

(Clark et al., 2018)

| Hypothesis, or a General Premise |

???? Research Problem

| Empirical Observations |

(22)
A Typical AI Experiment

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

|D| = 2376

(Clarke et al., 2018)

Empirical Observations

Hypothesis, or a General Premise

???? Research Problem
A Typical AI Experiment

- Can this apparent difference in performance be explained simply by random chance?
- Do we have sufficient evidence to conclude that A is in fact inherently stronger than B on these datasets?

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

\[ |D| = 2376 \]

(Clarke et al., 2018)
A Typical AI Experiment

- Can this apparent difference in performance be explained simply by random chance?
- Do we have sufficient evidence to conclude that ₠ is in fact inherently stronger than ₡ on these datasets?

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>₠</td>
<td>72.4</td>
</tr>
<tr>
<td>₡</td>
<td>68.9</td>
</tr>
</tbody>
</table>

|D| = 2376

(Clarke et al., 2018)
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>⊙</td>
<td>72.4</td>
</tr>
<tr>
<td>⊙⊙</td>
<td>68.9</td>
</tr>
</tbody>
</table>
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ⓐ</td>
<td>72.4</td>
</tr>
<tr>
<td>Ⓑ</td>
<td>68.9</td>
</tr>
</tbody>
</table>

1. C1*: Ⓐ and Ⓑ are inherently different,

2. C2*: Ⓐ and Ⓑ are inherently different, since with probability at least 95%, the inherent accuracy of Ⓐ exceeds that of Ⓑ by at least $\alpha$%.
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ⓐ</td>
<td>72.4</td>
</tr>
<tr>
<td>ⓑ</td>
<td>68.9</td>
</tr>
</tbody>
</table>

• C1*: ⓐ and ⓑ are inherently different,

* Under some statistical assumptions about sampling of the observations.
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ⓐ</td>
<td>72.4</td>
</tr>
<tr>
<td>Ⓑ</td>
<td>68.9</td>
</tr>
</tbody>
</table>

• **C1**: Ⓐ and Ⓑ are inherently different, in the sense that if they were inherently identical, it would be highly unlikely to witness the observed 3.5% empirical gap.

• **C2**: Ⓐ and Ⓑ are inherently different, since with probability at least 95%, the inherent accuracy of Ⓐ exceeds that of Ⓑ by at least $\alpha$%.

* Under some statistical assumptions about sampling of the observations.
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>72.4</td>
</tr>
<tr>
<td>☐</td>
<td>68.9</td>
</tr>
</tbody>
</table>

- C1*: ☐ and ☐ are inherently different, in the sense that if they were inherently identical, it would be highly unlikely to witness the observed 3.5% empirical gap.

* Under some statistical assumptions about sampling of the observations.
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ⓐ</td>
<td>72.4</td>
</tr>
<tr>
<td>ⓑ</td>
<td>68.9</td>
</tr>
</tbody>
</table>

- **C1**: ⓐ and ⓑ are inherently different, in the sense that if they were inherently identical, it would be highly unlikely to witness the observed 3.5% empirical gap.
- **C2**: ⓐ and ⓑ are inherently different,

* Under some statistical assumptions about sampling of the observations.
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>72.4</td>
</tr>
<tr>
<td>☒</td>
<td>68.9</td>
</tr>
</tbody>
</table>

- **C1**: ☐ and ☒ are inherently different, in the sense that if they were inherently identical, it would be highly unlikely to witness the observed 3.5% empirical gap.
- **C2**: ☐ and ☒ are inherently different, since with probability at least 95%, the inherent accuracy of ☐ exceeds that of ☒.

* Under some statistical assumptions about sampling of the observations.
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ⓐ</td>
<td>72.4</td>
</tr>
<tr>
<td>Ⓗ</td>
<td>68.9</td>
</tr>
</tbody>
</table>

• **C1*: ⓐ and Ⓗ are inherently different, in the sense that if they were inherently identical, it would be highly unlikely to witness the observed 3.5% empirical gap.

• **C2*: ⓐ and Ⓗ are inherently different, since with probability at least 95%, the inherent accuracy of ⓐ exceeds that of Ⓗ by at least \( \alpha \)%.

* Under some statistical assumptions about sampling of the observations.
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ⓟ</td>
<td>72.4</td>
</tr>
<tr>
<td>Ⓠ</td>
<td>68.9</td>
</tr>
</tbody>
</table>

• **C1***: Ⓟ and Ⓠ are inherently different, in the sense that if they were inherently identical, it would be highly unlikely to witness the observed 3.5% empirical gap.

• **C2***: Ⓟ and Ⓠ are inherently different, since with probability at least 95%, the inherent accuracy of Ⓟ exceeds that of Ⓠ by at least α%.

* Under some statistical assumptions about sampling of the observations.
A Typical AI Experiment: Example Hypotheses

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ⓐ</td>
<td>72.4</td>
</tr>
<tr>
<td>ⓑ</td>
<td>68.9</td>
</tr>
</tbody>
</table>

**Spoiler Alert:**
Almost everyone uses C1, even though it is harder to interpret.

- **C1***: ⓐ and ⓑ are inherently different, in the sense that if they were inherently identical, it would be highly unlikely to witness the observed 3.5% empirical gap.
- **C2***: ⓐ and ⓑ are inherently different, since with probability at least 95%, the inherent accuracy of ⓐ exceeds that of ⓑ by at least α%.

* Under some statistical assumptions about sampling of the observations.
A Typical AI Experiment: Example Hypotheses

Spoiler Alert:
Almost everyone uses C1, even though it is harder to interpret.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ⓐ</td>
<td>72.4</td>
</tr>
<tr>
<td>ⓑ</td>
<td>68.9</td>
</tr>
</tbody>
</table>

• C1*: ⓐ and ⓑ are inherently different, in the sense that if they were inherently identical, it would be highly unlikely to witness the observed 3.5% empirical gap.

• C2*: ⓐ and ⓑ are inherently different, since with probability at least 95%, the inherent accuracy of ⓐ exceeds that of ⓑ by at least \( \alpha \)%.

* Under some statistical assumptions about sampling of the observations.
• **Observation 1:** There are many different hypotheses that could address a single research question.
The number of natural hypothesis that can explain any given phenomena is infinite.

— Albert Einstein —
Hypothesis vs Statistical Techniques

- **Observation 2**: Each hypothesis ought to be assessed with an appropriate statistical tool.
Observation 2: Each hypothesis ought to be assessed with an appropriate statistical tool.
Hypothesis vs Statistical Techniques

- Observation 2: Each hypothesis ought to be assessed with an appropriate statistical tool.

- Corollary: Researchers should start with a hypothesis that best serves their goal, followed by an appropriate selection of a statistical approach.
Omission of hypotheses

Statistical Tool A

Hypothesis-1

Research Question

Statistical Tool B

Hypothesis-2

Statistical Tool C

Hypothesis-3
Omission of hypotheses

• **Observation 3:** Somehow, we tend to forget about hypotheses
Omission of hypotheses

The results of these experiments is presented in Table 5. All numbers are reported in percentage accuracy. We perform statistical significance testing on these results using Fisher’s exact test with a p-value of 0.05 and report them in our discussions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Regents Test</th>
<th>Monarch Test</th>
<th>ESSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucene</td>
<td>Regents Tables</td>
<td>37.5</td>
<td>32.6</td>
<td>36.9</td>
</tr>
<tr>
<td></td>
<td>Monarch Tables</td>
<td>28.4</td>
<td>27.3</td>
<td>27.7</td>
</tr>
<tr>
<td></td>
<td>Regents+Monarch Tables</td>
<td>34.8</td>
<td>35.3</td>
<td>37.3</td>
</tr>
<tr>
<td></td>
<td>Waterloo Corpus</td>
<td>55.4</td>
<td>51.8</td>
<td>54.4</td>
</tr>
<tr>
<td>MLN (Khot et al., 2015)</td>
<td></td>
<td>47.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FRETS (Compact)</td>
<td>Regents Tables</td>
<td>60.7</td>
<td>47.2</td>
<td>51.0</td>
</tr>
<tr>
<td></td>
<td>Monarch Tables</td>
<td>56.0</td>
<td>45.6</td>
<td>48.4</td>
</tr>
<tr>
<td></td>
<td>Regents+Monarch Tables</td>
<td>59.9</td>
<td>47.6</td>
<td>50.7</td>
</tr>
<tr>
<td>FRETS</td>
<td>Regents Tables</td>
<td>59.1</td>
<td>52.8</td>
<td>54.4</td>
</tr>
<tr>
<td></td>
<td>Monarch Tables</td>
<td>52.9</td>
<td>49.8</td>
<td>49.5</td>
</tr>
<tr>
<td></td>
<td>Regents+Monarch Tables</td>
<td>59.1</td>
<td>52.4</td>
<td>54.9</td>
</tr>
</tbody>
</table>
Omission of hypotheses

The results of these experiments is presented in Table 5. All numbers are reported in percentage accuracy. We perform statistical significance testing on these results using Fisher’s exact test with a p-value of 0.05 and report them in our discussions.
Omission of hypotheses

The results of these experiments is presented in Table 5. All numbers are reported in percentage accuracy. We perform statistical significance testing on these results using Fisher’s exact test with a p-value of 0.05 and report them in our discussions.
Omission of hypotheses

The results of these experiments is presented in Table 5. All numbers are reported in percentage accuracy. We perform statistical significance testing on these results using Fisher’s exact test with a p-value of 0.05 and report them in our discussions.
Omission of hypotheses

The results of these experiments is presented in Table 5. All numbers are reported in percentage accuracy. We perform statistical significance testing on these results using Fisher’s exact test with a p-value of 0.05 and report them in our discussions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Regents Test</th>
<th>Monarch Test</th>
<th>ESSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucene</td>
<td>Regents Tables</td>
<td>37.5</td>
<td>32.6</td>
<td>36.9</td>
</tr>
<tr>
<td></td>
<td>Monarch Tables</td>
<td>28.4</td>
<td>27.3</td>
<td>27.7</td>
</tr>
<tr>
<td></td>
<td>Regents+Monarch Tables</td>
<td>34.8</td>
<td>35.3</td>
<td>37.3</td>
</tr>
<tr>
<td></td>
<td>Waterloo Corpus</td>
<td>55.4</td>
<td>51.8</td>
<td>54.4</td>
</tr>
<tr>
<td>MLN (Khot et al., 2015)</td>
<td>-</td>
<td>47.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FRETS (Compact)</td>
<td>Regents Tables</td>
<td>60.7</td>
<td>47.2</td>
<td>51.0</td>
</tr>
<tr>
<td></td>
<td>Monarch Tables</td>
<td>56.0</td>
<td>45.6</td>
<td>48.4</td>
</tr>
<tr>
<td></td>
<td>Regents+Monarch Tables</td>
<td>59.9</td>
<td>47.6</td>
<td>50.7</td>
</tr>
<tr>
<td>FRETS</td>
<td>Regents Tables</td>
<td>59.1</td>
<td>52.8</td>
<td>54.4</td>
</tr>
<tr>
<td></td>
<td>Monarch Tables</td>
<td>52.9</td>
<td>49.8</td>
<td>49.5</td>
</tr>
<tr>
<td></td>
<td>Regents+Monarch Tables</td>
<td>59.1</td>
<td>52.4</td>
<td>54.9</td>
</tr>
</tbody>
</table>

**Flawed practice:** Many works use hypothesis assessment tests **without** knowing/stating their hypothesis.
Talk Summary & Statement

• There are several serious malpractices:
  • Incomplete reporting of hypotheses and how they address research questions.
  • Inability to interpret statistical tools or their results.
  • Lack of awareness about various Bayesian hypothesis assessment tools.

• Research works should be explicit about:
  • (a) Their choice of hypothesis and,
  • (b) How selected statistical tool addresses this hypothesis.
Statistical tools in this work . . .

(Kruschke and Liddell, 2018)
Statistical tools in this work . . .

- Frequentist
  - Binary/Categorical Decisions
  - Null-Hypothesis Significance Test
  - Confidence Interval

- Bayesian
  - Uncertainty Estimations
  - Bayes Factor
  - Posterior Intervals

(Kruschke and Liddell, 2018)
Frequentist

Bayesian

Binary/Categorical Decisions

Null-Hypothesis Significance Test

Bayes Factor

Uncertainty Estimations

Confidence Interval

Posterior Intervals

Confidence Interval
Notation
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  - $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  - H1: $\theta_1 > \theta_2$
  - H2: $\theta_1 > \theta_2 + b$
  - H3: $\theta_1 = \theta_2$
  - ...
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S, D)$
  • $\theta_i \neq M(S, D)$

• Several hypotheses:
  • H1: $\theta_1 > \theta_2$
  • H2: $\theta_1 > \theta_2 + b$
  • H3: $\theta_1 = \theta_2$
  • ...

Input instances: $D$
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • H1: $\theta_1 > \theta_2$
  • H2: $\theta_1 > \theta_2 + b$
  • H3: $\theta_1 = \theta_2$
  • ...
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • H1: $\theta_1 > \theta_2$
  • H2: $\theta_1 > \theta_2 + b$
  • H3: $\theta_1 = \theta_2$
  • ...
Notation

- Compare two systems on a set of instances: $D$

- A measure of performance: $M(S_i, D)$
  - $\theta_i \neq M(S_i, D)$

- Several hypotheses:
  - H1: $\theta_1 > \theta_2$
  - H2: $\theta_1 > \theta_2 + b$
  - H3: $\theta_1 = \theta_2$
  - ...
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • H1: $\theta_1 > \theta_2$
  • H2: $\theta_1 > \theta_2 + b$
  • H3: $\theta_1 = \theta_2$
  • ...
Notation

• Compare two systems on a set of instances: \( D \)

• A measure of performance: \( M(S_i, D) \)
  • \( \theta_i \neq M(S_i, D) \)

• Several hypotheses:
  • \( H1: \theta_1 > \theta_2 \)
  • \( H2: \theta_1 > \theta_2 + b \)
  • \( H3: \theta_1 = \theta_2 \)
  • …

Claims about the inherent properties \( \theta_1, \theta_2 \) of the two systems.
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • $H_1$: $\theta_1 > \theta_2$
  • $H_2$: $\theta_1 > \theta_2 + b$
  • $H_3$: $\theta_1 = \theta_2$
  • $\ldots$
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • $H1: \theta_1 > \theta_2$
  • $H2: \theta_1 > \theta_2 + b$
  • $H3: \theta_1 = \theta_2$
  • ...
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • H1: $\theta_1 > \theta_2$
  • H2: $\theta_1 > \theta_2 + b$
  • H3: $\theta_1 = \theta_2$
  • ...
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • $H_1$: $\theta_1 > \theta_2$
  • $H_2$: $\theta_1 > \theta_2 + b$
  • $H_3$: $\theta_1 = \theta_2$
  • ...
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i ≠ M(S_i, D)$

• Several hypotheses:
  • $H1: \theta_1 > \theta_2$
  • $H2: \theta_1 > \theta_2 + b$
  • $H3: \theta_1 = \theta_2$
  • ...

Input instances: $D$

Claims about the inherent properties $\theta_1, \theta_2$ of the two systems.

Hypotheses
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • H1: $\theta_1 > \theta_2$
  • H2: $\theta_1 > \theta_2 + b$
  • H3: $\theta_1 = \theta_2$
  • ...

Input instances: $D$

Claims about the inherent properties $\theta_1, \theta_2$ of the two systems.

Hypotheses
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • H1: $\theta_1 > \theta_2$
  • H2: $\theta_1 > \theta_2 + b$
  • H3: $\theta_1 = \theta_2$
  • ...

Input instances: $D$

Claims about the inherent properties $\theta_1, \theta_2$ of the two systems.

Hypotheses

Hypothesis Assessment
Notation

• Compare two systems on a set of instances: $D$

• A measure of performance: $M(S_i, D)$
  • $\theta_i \neq M(S_i, D)$

• Several hypotheses:
  • H1: $\theta_1 > \theta_2$
  • H2: $\theta_1 > \theta_2 + b$
  • H3: $\theta_1 = \theta_2$
  • ...

Input instances: $D$

Claims about the inherent properties $\theta_1, \theta_2$ of the two systems.

Hypotheses

Hypothesis Assessment

Conclusions validating (or not) the hypotheses.
Null-Hypothesis Significance Testing

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)
Null-Hypothesis Significance Testing

• The goal is to decide whether a particular hypothesis can be rejected.
• Make a hypothesis (that you want it to be rejected): null-hypothesis.
• Assume that null-hypothesis is correct.
• Calculate the probability of getting an outcome as “extreme” or more than the observed outcome.
  • This probability is called a “p-value.”

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)
Null-Hypothesis Significance Testing

• The goal is to decide whether a particular hypothesis can be rejected.
• Make a hypothesis (that you want it to be rejected): null-hypothesis.
• Assume that null-hypothesis is correct.
• Calculate the probability of getting an outcome as “extreme” or more than the observed outcome.
  • This probability is called a “p-value.”

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)
Null-Hypothesis Significance Testing

• The goal is to decide whether a particular hypothesis can be rejected.

• Make a hypothesis (that you want it to be rejected): null-hypothesis.

• Assume that null-hypothesis is correct. \[ H_0: \theta_1 = \theta_2 \]

• Calculate the probability of getting an outcome as “extreme” or more than the observed outcome.
  • This probability is called a “\( p \)-value.”

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)
Null-Hypothesis Significance Testing

• The goal is to decide whether a particular hypothesis can be rejected.
• Make a hypothesis (that you want it to be rejected): null-hypothesis.
• Assume that null-hypothesis is correct. \( H_0: \theta_1 = \theta_2 \)
• Calculate the probability of getting an outcome as “extreme” or more than the observed outcome.
  • This probability is called a “\( p \)-value.”

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)
Null-Hypothesis Significance Testing

• The goal is to decide whether a particular hypothesis can be rejected.
• Make a hypothesis (that you want it to be rejected): null-hypothesis.
• Assume that null-hypothesis is correct. \( H_0: \theta_1 = \theta_2 \)
• Calculate the probability of getting an outcome as “extreme” or more than the observed outcome.
  • This probability is called a “\( p \)-value.”

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)
Null-Hypothesis Significance Testing

• The goal is to decide whether a particular hypothesis can be rejected.
• Make a hypothesis (that you want it to be rejected): null-hypothesis.
• Assume that null-hypothesis is correct. $H_0: \theta_1 = \theta_2$
• Calculate the probability of getting an outcome as “extreme” or more than the observed outcome.
  • This probability is called a “$p$-value.”

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)
P-value, Visualized
P-value, Visualized

• **Null-hypothesis**: a hypothesis you want to reject & assume that it is correct.

• **P-value**: the probability of getting an **outcome** as “extreme” or more than the **observed** outcome.

• A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.
P-value, Visualized

- **Null-hypothesis**: a hypothesis you want to reject & assume that it is correct.
- **P-value**: the probability of getting an **outcome** as “extreme” or more than the observed outcome.
- A small p-value is used as a stronger evidence towards rejecting the **null-hypothesis**.
P-value, Visualized

• **Null-hypothesis**: a hypothesis you **want to reject** & **assume** that it is **correct**.

• **P-value**: the probability of getting an **outcome** as “**extreme**” or more than the **observed** outcome.

• A **small** p-value is used as a **stronger** evidence towards **rejecting** the **null-hypothesis**.
P-value, Visualized

- **Null-hypothesis**: a hypothesis you **want to reject** & **assume** that it is **correct**.
- **P-value**: the probability of getting an **outcome** as “**extreme**” or more than the **observed** outcome.
- A **small** p-value is used as a **stronger** evidence towards **rejecting** the **null-hypothesis**.

*all possible outcomes*
P-value, Visualized

• **Null-hypothesis**: a hypothesis you want to reject & assume that it is correct.

• **P-value**: the probability of getting an **outcome** as “extreme” or more than the observed outcome.

• A **small** p-value is used as a **stronger** evidence towards **rejecting** the null-hypothesis.
P-value, Visualized

• **Null-hypothesis:** a hypothesis you **want to reject** & **assume** that it is **correct**.

• **P-value:** the probability of getting an **outcome** as “**extreme**” or more than the **observed** outcome.

• A small p-value is used as a **stronger** evidence towards **rejecting** the **null-hypothesis**.
P-value, Visualized

- **Null-hypothesis**: a hypothesis you want to reject & **assume** that it is correct.
- **P-value**: the probability of getting an **outcome** as “extreme” or more than the **observed** outcome.
- A **small** p-value is used as a **stronger** evidence towards **rejecting** the null-hypothesis.
P-value, Visualized

• **Null-hypothesis**: a hypothesis you want to reject & assume that it is correct.

• **P-value**: the probability of getting an **outcome** as “extreme” or more than the **observed** outcome.

• A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.
P-value, Visualized

• **Null-hypothesis**: a hypothesis you want to reject & assume that it is correct.
• **P-value**: the probability of getting an observed outcome as “extreme” or more than the outcome.
• A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.
P-value, Visualized

- **Null-hypothesis:** a hypothesis you **want to reject** & **assume** that it is **correct**.
- **P-value:** the probability of getting an **outcome** as **“extreme”** or more than the **observed** outcome.
- A small p-value is used as a **stronger** evidence towards **rejecting** the null-hypothesis.

![Diagram illustrating p-value concept]

\[ p\text{-value} = \frac{\text{the actual outcome}}{\text{all possible outcomes}} \]
P-value, Visualized

• **Null-hypothesis**: a hypothesis you **want to reject** & **assume** that it is **correct**.

• **P-value**: the probability of getting an **outcome** as “**extreme**” or more than the **observed** outcome.

• A small p-value is used as a **stronger** evidence towards **rejecting** the null-hypothesis.

```
the actual outcome
extremer outcomes
all possible outcomes
```

\[
p-value = \frac{\text{all possible outcomes}}{\text{the actual outcome}}
\]
P-value, Visualized

• **Null-hypothesis**: a hypothesis you want to reject & assume that it is correct.

• **P-value**: the probability of getting an *outcome* as “extreme” or more than the observed outcome.

• A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.
P-value, Visualized

- **Null-hypothesis**: a hypothesis you want to reject & assume that it is correct.
- **P-value**: the probability of getting an outcome as “extreme” or more than the observed outcome.
- A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.
P-value, Visualized

- **Null-hypothesis**: a hypothesis you want to reject & assume that it is correct.
- **P-value**: the probability of getting an outcome as “extreme” or more than the observed outcome.
- A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.

![Diagram showing p-value with actual outcome and all possible outcomes]
Null-Hypothesis Significance Testing: Example

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4%</td>
</tr>
<tr>
<td>B</td>
<td>68.9%</td>
</tr>
</tbody>
</table>
Null-Hypothesis Significance Testing: Example

• Hypothesis $H_1$: $\theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>⬅️</td>
<td>72.4%</td>
</tr>
<tr>
<td>⬇️</td>
<td>68.9%</td>
</tr>
</tbody>
</table>
Null-Hypothesis Significance Testing: Example

• Hypothesis $\textbf{H1: } \theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ⓚ</td>
<td>72.4%</td>
</tr>
<tr>
<td>Ⓛ</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

• Null-Hypothesis $\textbf{H0: } \theta_1 = \theta_2$
Null-Hypothesis Significance Testing: Example

• Hypothesis $H_1$: $\theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>☑️</td>
<td>72.4%</td>
</tr>
<tr>
<td>☐️</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

• Null-Hypothesis $H_0$: $\theta_1 = \theta_2$
Null-Hypothesis Significance Testing: Example

- Hypothesis $H_1$: $\theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4%</td>
</tr>
<tr>
<td>B</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

- Null-Hypothesis $H_0$: $\theta_1 = \theta_2$
Null-Hypothesis Significance Testing: Example

• Hypothesis $H_1$: $\theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4%</td>
</tr>
<tr>
<td>B</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

• Null-Hypothesis $H_0$: $\theta_1 = \theta_2$
Null-Hypothesis Significance Testing: Example

- Hypothesis $H_1$: $\theta_1 > \theta_2$
- Null-Hypothesis $H_0$: $\theta_1 = \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4%</td>
</tr>
<tr>
<td>B</td>
<td>68.9%</td>
</tr>
</tbody>
</table>
Null-Hypothesis Significance Testing: Example

- Hypothesis $H_1$: $\theta_1 > \theta_2$
- Null-Hypothesis $H_0$: $\theta_1 = \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4%</td>
</tr>
<tr>
<td>B</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

$S, F, F, S, ...$
Null-Hypothesis Significance Testing: Example

- Hypothesis $H_1$: $\theta_1 > \theta_2$
- Null-Hypothesis $H_0$: $\theta_1 = \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4%</td>
</tr>
<tr>
<td>B</td>
<td>68.9%</td>
</tr>
</tbody>
</table>
Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>📜</td>
<td>72.4%</td>
</tr>
<tr>
<td>📜</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

- Null-Hypothesis **H0**: $\theta_1 = \theta_2$

$\theta_1 = \text{Prob}(S)$  
$\theta_2 = \text{Prob}(S)$  
$S, F, F, S, ...$  
$F, F, F, S, ...$
Null-Hypothesis Significance Testing: Example

- Hypothesis $H_1$: $\theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ⓐ</td>
<td>72.4%</td>
</tr>
<tr>
<td>Ⓑ</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

- Null-Hypothesis $H_0$: $\theta_1 = \theta_2$

\[ \theta_1 = \text{Prob}(S) \]
\[ \hat{\theta}_1 = \text{Avg}[S, F, F, S, ...] \]
\[ \theta_2 = \text{Prob}(S) \]
\[ F, F, F, S, ... \]
Null-Hypothesis Significance Testing: Example

• Hypothesis $\textbf{H1}: \theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>72.4%</td>
</tr>
<tr>
<td>☒</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

• Null-Hypothesis $\textbf{H0}: \theta_1 = \theta_2$

$\theta_1 = \text{Prob}(S)$

$\hat{\theta}_1 = \text{Avg}[S, F, F, S, ... ]$

$\theta_2 = \text{Prob}(S)$

$\hat{\theta}_2 = \text{Avg}[F, F, F, S, ... ]$
Null-Hypothesis Significance Testing: Example

- Hypothesis $H_1$: $\theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ⓞ</td>
<td>72.4%</td>
</tr>
<tr>
<td>Ⓟ</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

- Null-Hypothesis $H_0$: $\theta_1 = \theta_2$

$\theta_1 = \text{Prob}(S)$
$\hat{\theta}_1 = \text{Avg}[S, F, F, S, ... ]$

$\theta_2 = \text{Prob}(S)$
$\hat{\theta}_2 = \text{Avg}[F, F, F, S, ... ]$

$\text{Prob}[\hat{\theta}_1 - \hat{\theta}_2 > 3.5 | \theta_1 = \theta_2]$
Null-Hypothesis Significance Testing: Example

- Hypothesis $\text{H1: } \theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>☀️</td>
<td>72.4%</td>
</tr>
<tr>
<td>☀️</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

- Null-Hypothesis $\text{H0: } \theta_1 = \theta_2$

\[
\begin{align*}
\theta_1 &= \text{Prob}(S) \\
\hat{\theta}_1 &= \text{Avg}[S, F, F, S, ... ] \\
\theta_2 &= \text{Prob}(S) \\
\hat{\theta}_2 &= \text{Avg}[F, F, F, S, ... ]
\end{align*}
\]

$P$-value $= \text{Prob}[\hat{\theta}_1 - \hat{\theta}_2 > 3.5 | \theta_1 = \theta_2]$
Null-Hypothesis Significance Testing: Example

- Hypothesis \( \mathbf{H}_1: \theta_1 > \theta_2 \)

- Null-Hypothesis \( \mathbf{H}_0: \theta_1 = \theta_2 \)

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4%</td>
</tr>
<tr>
<td>B</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\theta_1 &= \text{Prob}(S) \\
\hat{\theta}_1 &= \text{Avg}[S, F, F, S, \ldots] \\
\theta_2 &= \text{Prob}(S) \\
\hat{\theta}_2 &= \text{Avg}[F, F, F, S, \ldots]
\end{align*}
\]

\[
P\text{-value} = \text{Prob}[\hat{\theta}_1 - \hat{\theta}_2 > 3.5 | \theta_1 = \theta_2] \leq \beta \text{ (e.g., 0.005)}
\]
Null-Hypothesis Significance Testing: Example

• Hypothesis $\textbf{H1: } \theta_1 > \theta_2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4%</td>
</tr>
<tr>
<td>B</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

• Null-Hypothesis $\textbf{H0: } \theta_1 = \theta_2$

\[
\theta_1 = \text{Prob}(S) \\
\hat{\theta}_1 = \text{Avg}[S, F, F, S, \ldots ]
\]

\[
\theta_2 = \text{Prob}(S) \\
\hat{\theta}_2 = \text{Avg}[F, F, F, S, \ldots ]
\]

$P$-value = \[\text{Prob}[\hat{\theta}_1 - \hat{\theta}_2 > 3.5 | \theta_1 = \theta_2] \leq \beta \text{ (e.g., 0.005)}\]
Interpreting p-values
Interpreting p-values

• Pretty complex notion!
Interpreting p-values

• 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
Interpreting p-value

If $p < 0.05$, the null-hypothesis has only a 5% chance of being true

(Demsar, 2008; Goodman, 2008)
Interpreting p-value

If $p < 0.05$, the null-hypothesis has only a 5% chance of being true

(Demsar, 2008; Goodman, 2008)
Interpreting p-value

If \( p < 0.05 \), the null-hypothesis has only a 5% chance of being true

• Remember that p-value is defined with the assumption that null-hypothesis is correct.

(Demsar, 2008; Goodman, 2008)
Interpreting p-value

If \( p > 0.05 \), there is no difference between the two systems

(Demsar, 2008; Goodman, 2008)
Interpreting p-value

If $p > 0.05$, there is no difference between the two systems

(Demsar, 2008; Goodman, 2008)
Interpreting p-value

• Having a large p-value only means that the null-hypothesis is consistent with the observations,

• ... but it does not tell anything about the likeliness of the null-hypothesis.

(Demsar, 2008; Goodman, 2008)
Interpreting p-value

If $p > 0.05$, there is no difference between the two systems (Demsar, 2008; Goodman, 2008)

• Having a large p-value only means that the null-hypothesis is consistent with the observations,
• ... but it does not tell anything about the likeliness of the null-hypothesis.
Interpreting p-value

A statistically significant result ($p < 0.05$) indicates a large/notable difference between two systems.

(Demsar, 2008; Goodman, 2008)
Interpreting p-value

A statistically significant result (p < 0.05) indicates a large/notable difference between two systems.
Interpreting p-value

A statistically significant result \((p < 0.05)\) indicates a large/notable difference between two systems.

- P-value only indicates strict superiority and provides no information about the margin of the effect.

(Demsar, 2008; Goodman, 2008)
Remember this?
Remember this?

Important reminder regarding large samples and p-values.

Oren Etzioni <orene@allenai.org>
to team

TL; DR statistical significance on large samples is all-too-easy to achieve and doesn't imply practical significance—use common sense 😊

For more, see the attached paper.

---

You received this message because you are subscribed to the Google Groups "AI2 Team" group.
To unsubscribe from this group and stop receiving emails from it, send an email to team+unsubscribe@allenai.org.
To view this discussion on the web visit https://groups.google.com/a/allenai.org/d/msgid/team/4ee343596d961c28ba90759382e5c876%40mail.gmail.com.
Remember this?

Important reminder regarding large samples and p-values.

Oren Etzioni <orene@allenai.org>
to team

Tue, Aug 20, 2019, 12:40 PM

TL; DR statistical significance on large samples is all-too-easy to achieve and doesn’t imply practical significance---use common sense 😊

For more, see the attached paper.

---

You received this message because you are subscribed to the Google Groups "AI2 Team" group.
To unsubscribe from this group and stop receiving emails from it, send an email to team+unsubscribe@allenai.org.
To view this discussion on the web visit https://groups.google.com/a/allenai.org/d/msgid/team/4ee343596d961c28ba90759382e5c876%40mail.gmail.com.
Remember this?

Important reminder regarding large samples and p-values.

Oren Etzioni <orene@allenai.org>
to team

TL; DR statistical significance on **large samples** is all-too-easy to achieve and **doesn’t imply practical significance**---use common sense 😊

For more, see the attached paper.

---

You received this message because you are subscribed to the Google Groups "AI2 Team" group.
To unsubscribe from this group and stop receiving emails from it, send an email to **team+unsubscribe@allenai.org**.
To view this discussion on the web visit **https://groups.google.com/a/allenai.org/d/msgid/team/4ee343596d961c28ba90759382e5c876%40mail.gmail.com**.
Remember this?

Important reminder regarding large samples and p-values.

Oren Etzioni <orene@allenai.org>

to team

TL; DR statistical significance on large samples is all-too-easy to achieve and doesn’t imply practical significance—use common sense 😊

For more, see the attached paper.

---

You received this message because you are subscribed to the Google Groups "AI2 Team" group.
To unsubscribe from this group and stop receiving emails from it, send an email to team+unsubscribe@allenai.org.
To view this discussion on the web visit https://groups.google.com/a/allenai.org/d/msgid/team/4ee343596d961c28ba90759382e5c876%40mail.gmail.com.

Or just keep listening to Daniel’s presentation!
Intermediate Summary

- **Frequentist**
  - Binary Decision
  - Uncertainty Estimations
  - Null-Hypothesis Significance Test
  - Confidence Interval

- **Bayesian**
  - Bayes Factor
  - Posterior Intervals
Intermediate Summary

• P-values do not provide **probability** estimates on two classifiers being different (or equal).
Intermediate Summary

• P-values do **not** provide *probability* estimates on two classifiers being different (or equal).

• **Statistical significance** is different than practical significance.

• Point-wise null hypotheses could be misused: for big enough data points it is possible to make statistically significant claims.
Intermediate Summary

• P-values do not provide probability estimates on two classifiers being different (or equal).

• Statistical significance is different than practical significance.

• Point-wise null hypotheses could be misused: for big enough data points it is possible to make statistically significant claims.
Frequentist

Bayesian

Binary/Categorical Decisions

Null-Hypothesis Significance Test

Bayes Factor

Uncertainty Estimations

Confidence Interval

Posterior Intervals
Posterior Intervals

• Based on Bayesian inference framework.

(Thomas Bayes 1702-1761)
Posterior Intervals

\[ P(\Theta|Y) = \frac{P(Y|\Theta) \times P(\Theta)}{P(Y)} \]
Posterior Intervals

• Key notions:
  • **Prior**: Assumptions and beliefs about key parameters of a system.
  • **Likelihood**: How the hidden parameters are connected to the observations.
  • **Posterior**: Summary of the inferences about likely values of $\Theta$.

\[
P(\Theta|Y) = \frac{P(Y|\Theta) \times P(\Theta)}{P(Y)}
\]
Posterior Intervals

• Key notions:
  • **Prior**: Assumptions and beliefs about key parameters of a system.
  • **Likelihood**: How the hidden parameters are connected to the observations.
  • **Posterior**: Summary of the inferences about likely values of $\Theta$.

\[
P(\Theta|Y) = \frac{P(Y|\Theta) \times P(\Theta)}{P(Y)}
\]
Posterior Intervals

• Key notions:
  • Prior: Assumptions and beliefs about key parameters of a system.
  • Likelihood: How the hidden parameters are connected to the observations.
  • Posterior: Summary of the inferences about likely values of $\Theta$.

$$P(\Theta|Y) = \frac{P(Y|\Theta) \times P(\Theta)}{P(Y)}$$
Posterior Intervals

\[ P(\text{Hypothesis}|\text{Observations}) \]
Posterior Intervals

- **Goal:** Using Bayes’s Theorem to infer a probability distribution:
  \[ P(\text{Hypothesis}|\text{Observations}) \]

- **Example:** \( H_1: \theta_1 - \theta_2 > \alpha \)
Posterior Intervals

• **Goal:** Using Bayes’s Theorem to infer a probability distribution:

\[ P(\text{Hypothesis} | \text{Observations}) \]

• **Example:** \( H_1: \theta_1 - \theta_2 > \alpha \)
Posterior Intervals

• **Goal:** Using Bayes’s Theorem to infer a probability distribution:

\[ P(\text{Hypothesis} | \text{Observations}) \]

• **Example:** \( H_1: \theta_1 - \theta_2 > \alpha \)
Posterior Intervals

- **Goal:** Using Bayes’s Theorem to infer a probability distribution:

  \[ P(\text{Hypothesis}|\text{Observations}) \]

- **Example:** \( H_1: \theta_1 - \theta_2 > \alpha \)
Posterior Intervals

• **Goal:** Using Bayes’s Theorem to infer a probability distribution:

\[ P(\text{Hypothesis}|\text{Observations}) \]

• **Example:** \( H_1: \theta_1 - \theta_2 > \alpha \)
Posterior Intervals: Example

\[ H_1: \theta_1 - \theta_2 > \alpha \]

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>72.4</td>
</tr>
<tr>
<td>☺</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

\[ H_1: \theta_1 - \theta_2 > \alpha \]

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

\[ H_1: \theta_1 - \theta_2 > \alpha \]

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

\[ H_1: \theta_1 - \theta_2 > \alpha \]
Posterior Intervals: Example

$H_1: \theta_1 - \theta_2 > \alpha$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

\( H_1: \theta_1 - \theta_2 > \alpha \)

\( \theta_1 = \text{Prob}(S) \)
\( S, F, F, S, \ldots \)

\( \theta_2 = \text{Prob}(S) \)
\( F, F, F, S, \ldots \)

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

\[ H_1: \theta_1 - \theta_2 > \alpha \]

\[ \theta_1 = \text{Prob}(S) \]

\[ S, F, F, S, ... \]

\[ \theta_2 = \text{Prob}(S) \]

\[ F, F, F, S, ... \]

\[ P(Y|\theta) \]
Posterior Intervals: Example

\( H_1: \theta_1 - \theta_2 > \alpha \)

\[ \theta_1 = \text{Prob}(S) \]

\[ S, F, F, S, ... \]

\[ \theta_2 = \text{Prob}(S) \]

\[ F, F, F, S, ... \]

\[ P(Y|\theta) \]

\[ \oplus \]

\[ P(\theta) \sim \text{uniform} \]
Posterior Intervals: Example

\[ H_1: \theta_1 - \theta_2 > \alpha \]

\[ \theta_1 = \text{Prob}(S) \]
\[ S, F, F, S, ... \]

\[ \theta_2 = \text{Prob}(S) \]
\[ F, F, F, S, ... \]

\[ P(Y|\Theta) \]
\[ \oplus \]
\[ P(\Theta) \sim \text{uniform} \]

\[ P(\Theta|Y) = \frac{P(Y|\Theta) \times P(\Theta)}{P(Y)} \]

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ⓐ</td>
<td>72.4</td>
</tr>
<tr>
<td>ⓑ</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

$H_1: \theta_1 - \theta_2 > \alpha$

$\theta_1 = \text{Prob}(S)$
$S, F, F, S, ...$

$\theta_2 = \text{Prob}(S)$
$F, F, F, S, ...$

$P(Y|\theta) \oplus P(\theta) \sim \text{uniform}$

$P(\theta|Y) = \frac{P(Y|\theta) \times P(\theta)}{P(Y)}$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>72.4</td>
</tr>
<tr>
<td>☒</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

\( H_1: \theta_1 - \theta_2 > \alpha \)

\[ \begin{align*}
\theta_1 &= \text{Prob}(S) \\
S, F, F, S, ... &\rightarrow \\
\theta_2 &= \text{Prob}(S) \\
F, F, F, S, ... &\rightarrow \\
\end{align*} \]

\[ \begin{align*}
P(Y|\Theta) &\oplus \\
P(\Theta) &\sim \text{uniform} \\
\end{align*} \]

\[ P(\Theta|Y) = \frac{P(Y|\Theta) \times P(\Theta)}{P(Y)} \]

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

$$H_1: \theta_1 - \theta_2 > \alpha$$

$$\theta_1 = \text{Prob}(S)$$

$$\theta_2 = \text{Prob}(S)$$

$$S, F, F, S, \ldots$$

$$F, F, F, S, \ldots$$

$$P(Y | \theta) \oplus \quad P(\theta) \sim \text{uniform}$$

$$P(\theta | Y) = \frac{P(Y | \theta) \times P(\theta)}{P(Y)}$$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

H₁: \( \theta_1 - \theta_2 > \alpha \)

\( \theta_1 = \text{Prob}(S) \)
\( S, F, F, S, ... \)

\( \theta_2 = \text{Prob}(S) \)
\( F, F, F, S, ... \)

\( P(Y|\theta) \)
\( \oplus \)
\( P(\theta) \sim \text{uniform} \)

\[ P(\theta|Y) = \frac{P(Y|\theta) \times P(\theta)}{P(Y)} \]
Posterior Intervals: Example

$H: \theta_1 - \theta_2 > \alpha$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

mode = 0.0356
Posterior Intervals: Example

H: $\theta_1 - \theta_2 > \alpha$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

mode=0.0356
Posterior Intervals: Example

\[ H: \theta_1 - \theta_2 > \alpha \]

- The hypothesis (w/ \( \alpha = 0 \)) holds true ...
  - ... with probability %99.6.

- The hypothesis (w/ \( \alpha = 1 \)) holds true ...
  - ... with probability %94.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

Mode = 0.0356

\( \theta_1 - \theta_2 \)
Posterior Intervals: Example

H: $\theta_1 - \theta_2 > \alpha$

• The hypothesis (w/ $\alpha = 0$) holds true ...
  • ... with probability %99.6.

• The hypothesis (w/ $\alpha = 1$) holds true ...
  • ... with probability %94.

System | Accuracy
---|---
-picker | 72.4
-picker | 68.9
Posterior Intervals: Example

H: $\theta_1 - \theta_2 > \alpha$

• The hypothesis (w/ $\alpha = 0$) holds true ...
  • ... with probability %99.6.

• The hypothesis (w/ $\alpha = 1$) holds true ...
  • ... with probability %94.

System | Accuracy
---|---
Ⓐ 72.4
Ⓑ 68.9

$\theta_1 - \theta_2$
Posterior Intervals: Example

**H:** $\theta_1 - \theta_2 > \alpha$

- The hypothesis (w/ $\alpha = 0$) holds true ...
  - ... with probability %99.6.

- The hypothesis (w/ $\alpha = 1$) holds true ...
  - ... with probability %94.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

$\theta_1 - \theta_2$
Posterior Intervals: Example

H: $\theta_1 - \theta_2 > \alpha$

• The hypothesis (w/ $\alpha = 0$) holds true ...
  • ... with probability %99.6.

• The hypothesis (w/ $\alpha = 1$) holds true ...
  • ... with probability %94.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

$H: \theta_1 - \theta_2 > \alpha$
Posterior Intervals: Example

$H: \theta_1 - \theta_2 > \alpha$

• The hypothesis (w/ $\alpha = 0$) holds true ...
  • ... with probability %99.6.

• The hypothesis (w/ $\alpha = 1$) holds true ...
  • ... with probability %94.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>
Posterior Intervals: Example

H: $\theta_1 - \theta_2 > \alpha$

• The hypothesis (w/ $\alpha = 0$) holds true ...
  • ... with probability %99.6.

• The hypothesis (w/ $\alpha = 1$) holds true ...
  • ... with probability %94.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

$\theta_1 - \theta_2$
Posterior Intervals: Example

\( H: \theta_1 - \theta_2 > \alpha \)

- The hypothesis (w/ \( \alpha = 0 \)) holds true ...
  - ... with probability %99.6.

- The hypothesis (w/ \( \alpha = 1 \)) holds true ...
  - ... with probability %94.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.4</td>
</tr>
<tr>
<td>B</td>
<td>68.9</td>
</tr>
</tbody>
</table>

\( \theta_1 - \theta_2 \)
2\textsuperscript{nd} Intermediate Summary

- **Frequentist**
  - Binary Decision
  - Null-Hypothesis Significance Test
  - Confidence Interval
- **Bayesian**
  - Uncertainty Estimations
  - Bayes Factor
  - Posterior Intervals
2nd Intermediate Summary

• It’s much more intuitive to work with the **probability of hypotheses**.
  • Easier to interpret → less ambiguous.
2nd Intermediate Summary

• It’s much more intuitive to work with the probability of hypotheses.
  • Easier to interpret → less ambiguous.

• Provides a flexible framework
  • E.g., margin of superiority could incorporated in the definition of hypotheses.

• This does not encourage binary decision-making.
It’s much more intuitive to work with the **probability of hypotheses**.
- **Easier** to interpret → less ambiguous.

Provides a **flexible** framework
- E.g., **margin of superiority** could incorporated in the definition of hypotheses.

This does not encourage **binary** decision-making.
Final Section:
Common Practices, Comparisons and Suggestions

<table>
<thead>
<tr>
<th>Binary/Categorical Decisions</th>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null-Hypothesis Significance Test</td>
<td>Confidence Interval</td>
<td>Bayes Factor</td>
</tr>
<tr>
<td>Uncertainty Estimations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Survey of the NLP Community
Survey of the NLP Community

• A questionnaire containing general and specific questions about significance assessment tools
• Sent it to over 400 researchers randomly selected from ACL’18 proceedings
• ~50 individuals responded
Survey of the NLP Community

• A questionnaire containing general and specific questions about significance assessment tools
• Sent it to over 400 researchers randomly selected from ACL’18 proceedings
• ~50 individuals responded
Survey of the NLP Community

• A questionnaire containing general and specific questions about significance assessment tools
• Sent it to over 400 researchers randomly selected from ACL’18 proceedings
• ~50 individuals responded
Survey of the NLP Community

• A questionnaire containing general and specific questions about significance assessment tools
• Sent it to over 400 researchers randomly selected from ACL’18 proceedings
• ~50 individuals responded
Survey of the NLP Community

• “I have learned about statistical hypothesis testing/assessment (via taking classes or reading it from other places).”
Survey of the NLP Community

• “I have learned about statistical hypothesis testing/assessment (via taking classes or reading it from other places).”
Participants in Our Survey

• “I have used "hypothesis testing" in the past (in a homework, a paper, etc.)”
Participants in Our Survey

• “I have used "hypothesis testing" in the past (in a homework, a paper, etc.)”
Participants in Our Survey

• “I am not a robot”
Participants in Our Survey

• “I am not a robot”
Participants in Our Survey

- “I am not a robot”
Participants in Our Survey

• “I am not a robot”

![Bar chart showing response to the statement “I am not a robot.”]
Trends and Patterns in the field

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?

- Frequentist
  - Binary Decision
  - Null-Hypothesis Significance Test
- Bayesian
  - Bayes Factor
  - Uncertainty Estimations
  - Confidence Interval
  - Posterior Intervals

Number of papers using significance testing: 73
Trends and Patterns in the field

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

*How many papers did use significance testing?*

- **73** papers used Frequentist methods
- **6** papers used Bayesian methods
Trends and Patterns in the field

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

How many papers did use significance testing?

- Frequentist: 73 papers
- Bayesian: 0 papers
- Null-Hypothesis Significance Test
- Confidence Interval
- Bayes Factor
- Posterior Intervals
- Binary Decision
- Uncertainty Estimations
Trends and Patterns in the field

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

*How many papers did use significance testing?*

- Frequentist: 73
- Bayesian: 0
- Binary Decision: 6
- Confidence Interval: 0
- Null-Hypothesis Significance Test: 6
- Bayes Factor: 0
- Uncertainty Estimations: 6
- Posterior Intervals: 0
Trends and Patterns in the field

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

*How many papers did use significance testing?*

- **Frequentist**
  - Binary Decision: 73
  - Uncertainty Estimations: 6
- **Bayesian**
  - Null-Hypothesis Significance Test: 0
  - Confidence Interval: 0
  - Bayes Factor: 0
  - Posterior Intervals: 0

Why?
Have you heard about "Bayesian Hypothesis Testing"?
Have you heard about "Bayesian Hypothesis Testing"?
Have you heard about "Bayesian Hypothesis Testing"?

Do you know the definition of "Bayes Factor"?

- Yes: 52.7%
- No: 25.5%
- Not sure: 21.8%
Have you heard about "Bayesian Hypothesis Testing"?

- Yes: 52.7%
- No: 21.8%
- Not sure: 25.5%

Do you know the definition of "Bayes Factor"?

- Yes: 78.2%
- No: 21.8%
Many people did not know the definition of “Bayes Factor” and some only had “heard” about them. 😐
Measures of [Un]Certainty

(Goodman, 2008; Wasserstein et al., 2016)
Measures of [Un]Certainty

- *P-values* do **not** provide probability estimates on validity of hypotheses.

(Goodman, 2008; Wasserstein et al., 2016)
Measures of [Un]Certainty

- *P-values* do **not** provide probability estimates on validity of hypotheses.

(Goodman, 2008; Wasserstein et al., 2016)
Measures of [Un]Certainty

- *P-values* do **not** provide probability estimates on validity of hypotheses.

(Goodman, 2008; Wasserstein et al., 2016)
Measures of [Un]Certainty

• *P-values* do not provide probability estimates on validity of hypotheses.

• Posterior Intervals are interpretable in terms of post-data **probabilities**.

(Goodman, 2008; Wasserstein et al., 2016)
Susceptibility to Misinterpretation

• The complexity of interpreting significance tests could result in ambiguous or misleading conclusions.

• P-values, while being the most common approach, are inherently complex and easy to misinterpret.
Susceptibility to Misinterpretation

- The complexity of interpreting significance tests could result in ambiguous or misleading conclusions.
- P-values, while being the most common approach, are inherently complex and easy to misinterpret.
Participants in Our Survey

• “I know p-values and I know how to interpret them.”
Participants in Our Survey

• "I know p-values and I know how to interpret them."
A Survey Question: Interpreting P-value
A Survey Question: Interpreting P-value

• An NLP paper shows a performance of 38% for a classifier-1. They also show that adding a feature improves the performance to 45% (call this classifier-2). The authors claim that this finding is “statistically significant” with a significance level of 0.01. Which of the following(s) make sense?

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

b) If we repeat the experiment, with a probability 99% classifier-2 will have a higher performance than classifier-1.
A Survey Question: Interpreting P-value

• An NLP paper shows a performance of 38% for a classifier-1. They also show that adding a feature improves the performance to 45% (call this classifier-2). The authors claim that this finding is “statistically significant” with a significance level of 0.01. Which of the following(s) make sense?

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

   b) If we repeat the experiment, with a probability 99% classifier-2 will have a higher performance than classifier-1.
A Survey Question: Interpreting P-value

• An NLP paper shows a performance of 38% for a classifier-1. They also show that adding a feature improves the performance to 45% (call this classifier-2). The authors claim that this finding is “statistically significant” with a significance level of 0.01. Which of the following(s) make sense?

a) The probability of observing a margin 7% is at most 0.01, assuming that the two classifiers inherently have the same performance.  
   23%

b) If we repeat the experiment, with a probability 99% classifier-2 will have a higher performance than classifier-1.  
   30%
Unintended Misleading Result by Iterative Testing
Unintended Misleading Result by Iterative Testing

• Many tests are designed for a single-round experiment.

• In practice researchers perform multiple rounds of experiments.

• This is a major problem when using binary tests.
  • E.g., you can “hack” a p-value test, with enough repetitions.
Unintended Misleading Result by Iterative Testing

• Many tests are designed for a single-round experiment.

• In practice researchers perform multiple rounds of experiments.

• This is a major problem when using binary tests.
  • E.g., you can “hack” a p-value test, with enough repetitions.
Unintended Misleading Result by Iterative Testing

• Many tests are designed for a **single-round** experiment.

• In practice researchers perform **multiple** rounds of experiments.

• This is a major problem when using **binary tests**.
  • E.g., you can “hack” a p-value test, with enough repetitions.
Unintended Misleading Result by Iterative Testing

• Many tests are designed for a **single-round** experiment.

• In practice researchers perform **multiple** rounds of experiments.

• This is a major problem when using **binary tests**.
  • E.g., you can “hack” a p-value test, with enough repetitions.
Unintended Misleading Result by Iterative Testing

- Many tests are designed for a **single-round** experiment.

- In practice researchers perform **multiple** rounds of experiments.

- This is a major problem when using **binary tests**.
  - E.g., you can “hack” a p-value test, with enough repetitions.
The Need for Assumptions

- **Frequentist**
  - Binary Decision
  - Null-Hypothesis Significance Test
  - Confidence Interval

- **Bayesian**
  - Uncertainty Estimations
  - Bayes Factor
  - Posterior Intervals
The Need for Assumptions

• *Which tests have assumptions?*

• Assumptions are necessary to perform any statistical tests.
  • “no free lunch”

• Many of them are questionable!

- Frequentist
- Bayesian
- Binary Decision
- Null-Hypothesis Significance Test
- Bayes Factor
- Uncertainty Estimations
- Confidence Interval
- Posterior Intervals
Ambiguity problem in interpreting “significance”
Ambiguity problem in interpreting “significance”
Ambiguity problem in interpreting “significance”

Abstract

Multi-hop reasoning is an effective approach for query answering (QA) over incomplete knowledge graphs (KGs). The problem can be formulated in a reinforcement learning (RL) setup, where a policy-based agent sequentially extends its inference path until it reaches a target. However, in an incomplete KG environment, the agent receives low-quality rewards corrupted by false negatives in the training data, which harms generalization at test time. Furthermore, since no golden action sequence is used for training, the agent can be misled by spurious search trajectories that incidentally lead to the correct answer. We propose two modeling advances to address both issues: (1) we reduce the impact of false negative supervision by adopting a pretrained one-hop embedding model to estimate the reward of unobserved facts; (2) we counter the sensitivity to spurious paths of on-policy RL by forcing the agent to explore a diverse set of paths using randomly generated edge masks. Our approach significantly improves over existing path-based KGQA models on several benchmark datasets and is comparable or better than embedding-based models.
Ambiguity problem in interpreting “significance”

Abstract

Multi-hop reasoning is an effective approach for query answering (QA) over incomplete knowledge graphs (KGs). The problem can be formulated in a reinforcement learning (RL) setup, where a policy-based agent sequentially extends its inference path until it reaches a target. However, in an incomplete KG environment, the agent receives low-quality rewards corrupted by false negatives in the training data, which harms generalization at test time. Furthermore, since no golden action sequence is used for training, the agent can be misled by spurious search trajectories that incidentally lead to the correct answer. We propose two modeling advances to address both issues: (1) we reduce the impact of false negative supervision by adopting a pretrained one-hop embedding model to estimate the reward of unobserved facts; (2) we counter the sensitivity to spurious paths of on-policy RL by forcing the agent to explore a diverse set of paths using randomly generated edge masks. Our approach significantly improves over existing path-based KGQA models on several benchmark datasets and is comparable or better than embedding-based models.

Abstract

Most social media platforms grant users freedom of speech by allowing them to freely express their thoughts, beliefs, and opinions. Although this represents incredible and unique communication opportunities, it also presents important challenges. Online racism is such an example. In this study, we present a supervised learning strategy to detect racist language on Twitter based on word embedding that incorporate demographic (Age, Gender, and Location) information. Our methodology achieves reasonable classification accuracy over a gold standard dataset ($F_1=76.3\%$) and significantly improves over the classification performance of demographic-agnostic models.
Ambiguity problem in interpreting “significance”
Ambiguity problem in interpreting “significance”

• An NLP paper presents system-1 and it compares it with a baseline system-2. In its “abstract” it writes: “... system-1 significantly improves over system-2.” What are the right way(s) to interpret this (select all that applies)

  • It is expected that authors have performed some type of “hypothesis testing.”

  • It is expected that the authors have reported the performances of two systems on a dataset where system-1 has a higher performance than system-2 with a notable margin in the dataset.
Ambiguity problem in interpreting “significance”

• An NLP paper presents system-1 and it compares it with a baseline system-2. In its “abstract” it writes: “... system-1 significantly improves over system-2.” What are the right way(s) to interpret this (select all that applies)

  • It is expected that authors have performed some type of “hypothesis testing.”

  • It is expected that the authors have reported the performances of two systems on a dataset where system-1 has a higher performance than system-2 with a notable margin in the dataset.
Ambiguity problem in interpreting “significance”

• An NLP paper presents system-1 and it compares it with a baseline system-2. In its “abstract” it writes: “… system-1 significantly improves over system-2.” What are the right way(s) to interpret this (select all that applies)

  • It is expected that authors have performed some type of “hypothesis testing.”
  • It is expected that the authors have reported the performances of two systems on a dataset where system-1 has a higher performance than system-2 with a notable margin in the dataset.
Ambiguity problem in interpreting “significance”

• An NLP paper presents system-1 and it compares it with a baseline system-2. In its “abstract” it writes: “... system-1 significantly improves over system-2.” What are the right way(s) to interpret this (select all that applies)

  • It is expected that authors have performed some type of “hypothesis testing.” 83%

  • It is expected that the authors have reported the performances of two systems on a dataset where system-1 has a higher performance than system-2 with a notable margin in the dataset. 53%
The Usage of “Significance”: Our Recommendation

• When referring to performing some type of “hypothesis testing,” use prefixes like “statistical”

• When referring to big empirical improvements, use alternative terms like: “notable” or “remarkable.”
Define the research hypothesis you are after:

- **C1**: ☼ and ⚫ are **inherently different**, in the sense that if they were inherently **identical**, it would be highly **unlikely** to witness the observed 3.5% empirical gap.
- **C2**: ☼ and ⚫ are **inherently different**, since with **probability** at least 95%, the inherent accuracy of ☼ **exceeds** that of ⚫ by at least \( \alpha \)%.
- ...
Tips and Suggestions

• If using frequentist tests:
  • The statements reporting p-value and confidence interval need to be precise.
  • ... so that the results are not misinterpreted.
    • The term “significant” should be used with caution and clear purpose in order to not cause any misinterpretations.
      
      *better under a significance test* !≠ *significantly better*
    • One way to achieve this is by using adjectives “statistical” or “practical” before any (possibly inflected) usage of “significance.”
Tips and Suggestions

The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing

Rotem Dror, Gili Baumer, Segev Shlomov, Roi Reichart
Faculty of Industrial Engineering and Management, Technion, IIT
{rtmdrr@campus|sgbaumer@campus|segevs@campus|roiri}.technion.ac.il

Abstract

Statistical significance testing is a standard statistical tool designed to ensure that experimental results are not coincidental. In this opinion/theoretical paper we discuss the role of statistical significance testing in Natural Language Processing (NLP) research. We establish the fundamental rules for selecting the right "test" and how to report your results.

Lots of good tips about:
- Selecting the right “test”
- How to report your results.
Tips and Suggestions

• If using Bayesian tests:
  • Be clear about your hierarchical model, any parameters in the model and the choice of priors.
  • Comment on the certainty (or the lack of) of your inference.
Not All Claims are Created Equal: Choosing the Right Approach to Assess Your Hypotheses

Erfan Sadeqi Azer\textsuperscript{1}  Daniel Khashabi\textsuperscript{2,*}  Ashish Sabharwal\textsuperscript{2}  Dan Roth\textsuperscript{3}

\textsuperscript{1}Indiana University  \textsuperscript{2}Allen Institute for Artificial Intelligence  \textsuperscript{3}University of Pennsylvania

esadeqia@indiana.edu, \{danielk, ashishs\}@allenai.org danroth@cis.upenn.edu

Abstract

Empirical research in Natural Language Processing (NLP) has adopted a narrow set of principles for assessing hypotheses, relying mainly on \(p\)-value computation, which suffers from several known issues. While alternative proposals have been well-debated and adopted in other fields, they remain rarely discussed or used within the NLP community. We address

\begin{table}[h]
\centering
\begin{tabular}{|l|l|c|c|c|c|}
\hline
System ID & Description & ARC-easy & ARC-\textsuperscript{challenge} \\
& & \#Correct Acc. & \#Correct Acc. & \\
\hline
\text{S}_1 & BERT & 1721 & 72.4 & 566 & 48.3 \\
\text{S}_2 & Reading Strategies & 1637 & 68.9 & 496 & 42.3 \\
\hline
\end{tabular}
\caption{Performance of two systems (Devlin et al., 2019; Sun et al., 2018) on the ARC question-answering dataset (Clark et al., 2018). ARC-easy & ARC-challenge have 2376 & 1172 instances, respectively. Acc.: accuracy as a percentage.}
\end{table}
That’s it!
Participants in our Survey

- 41.8%
- 16.4%
- 10.9%
- 25.5%

Legend:
- <1
- 1-5
- 5-10
- >10
- I am still a PhD student or I have not started a PhD problem.
- BSc student
- MSc student
- PhD student
- Postdoc
- University professor
- Researcher (industry or academia)
- Other
Participants in our Survey

What venues do you usually publish in?

52 responses

- ACL, EMNLP, NAACL, TACL and similar "n..." (49 responses, 94.2%)
- AAAI, IJCAI, and similar "artificial in..." (7 responses, 13.5%)
- KDD, ICDM, WSDM, other "data mining" ve... (3 responses, 5.8%)
- STOC, FOCS, SODA or other "theory" venu... (0 responses, 0%)
Participants in Our Survey

• “I can understand almost all the "statistical" terms I encounter in papers.”
Participants in Our Survey

- “I can understand almost all the "statistical" terms I encounter in papers.”

[Bar chart showing survey responses]

1 (1.8%) 7 (12.7%) 16 (29.1%) 23 (41.8%) 8 (14.5%)