Not All Claims are Created Equal:

Choosing the Right Statistical Approach to Assess Hypotheses

Me

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?

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?

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

System	$\hat{\theta}$	θ
A	72.4	?
B	68.9	?

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

A Typical AI Experiment

Empirical performance

System	$\hat{\theta}$	θ
A	72.4	?
B	68.9	?

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

A Typical AI Experiment

Empirical performance	Inherent performance

System	$\hat{\theta}$	θ
A	72.4	?
B	68.9	?

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

A Typical AI Experiment

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

$$\boldsymbol{H}\text{: } \theta_A = \theta_B$$

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

Empirical performance

Inherent performance

System	$\hat{\theta}$	θ
A	72.4	?
B	68.9	?

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

A Typical AI Experiment

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

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

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

$$\boldsymbol{H}:\ \theta_A > \theta_B + \alpha$$

Empirical performance

Inherent performance

System	$\hat{\theta}$	θ
A	72.4	?
B	68.9	?

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

A Typical AI Experiment

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

$$\boldsymbol{H}\colon\ \theta_A = \theta_B$$

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

$$\boldsymbol{H}\colon\ \theta_A > \theta_B + \alpha$$

Empirical performance

Inherent performance

System	$\hat{\theta}$	θ
A	72.4	?
B	68.9	?

(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

Statistical tools: big picture

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.

$$\boldsymbol{H}: \theta_A > \theta_B$$

- **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an **observation as "extreme"** as what's observed.

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Statistical tools: big picture

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.

$$\boldsymbol{H}:\ \theta_A > \theta_B$$

- **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an **observation as "extreme"** as what's observed.

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Statistical tools: big picture

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.

$$\boldsymbol{H}: \theta_A > \theta_B$$

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

- **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an **observation as "extreme"** as what's observed.

p-value

Bayes Factor

Confidence Interval

Posterior Intervals

Statistical tools: big picture

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.

$$\boldsymbol{H}:\ \theta_A > \theta_B$$

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

- **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an **observation as "extreme"** as what's observed.

the accuracy gap between the two systems

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Statistical tools: big picture

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.

$$\boldsymbol{H}:\ \theta_A > \theta_B$$

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

- **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an **observation as "extreme"** as what's observed.

the accuracy gap between the two systems

$$P\left(\text{obs.} > \hat{\theta}_A - \hat{\theta}_B \,|\, \overline{\boldsymbol{H}}\right)$$

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Statistical tools: big picture

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.

$$\boldsymbol{H}:\ \theta_A > \theta_B$$

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

- **Idea:** assuming that an opposite hypothesis is true, compute the likelihood of an **observation as "extreme"** as what's observed.

the accuracy gap between the two systems

$$\underbrace{P\left(\text{obs.} > \hat{\theta}_A - \hat{\theta}_B \,|\, \overline{\boldsymbol{H}}\right)}_{\textit{p-value}}$$

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Statistical tools: big picture

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

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 $\boldsymbol{H}$.
- **Idea:** use the Bayes formula to compute a probability for the hypothesis being true.

Statistical tools: big picture

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.
- **Idea:** use the Bayes formula to compute a probability for the hypothesis being true.

Statistical tools: big picture

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.

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

$$\boldsymbol{H}:\ \theta_A > \theta_B + \alpha$$

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

Statistical tools: big picture

	Bayesian
p-value	Bayes Factor
Confidence Interval	Posterior Intervals

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.
- **Idea:** use the Bayes formula to compute a probability for the hypothesis being true.

$$\boldsymbol{H}\colon\ \theta_A > \theta_B + \alpha$$

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

Statistical tools: big picture

Frequentist	Bayesian
p-value	Bayes Factor
Confidence Interval	Posterior Intervals

- Suppose I want to assess a hypothesis $\boldsymbol{H}$.

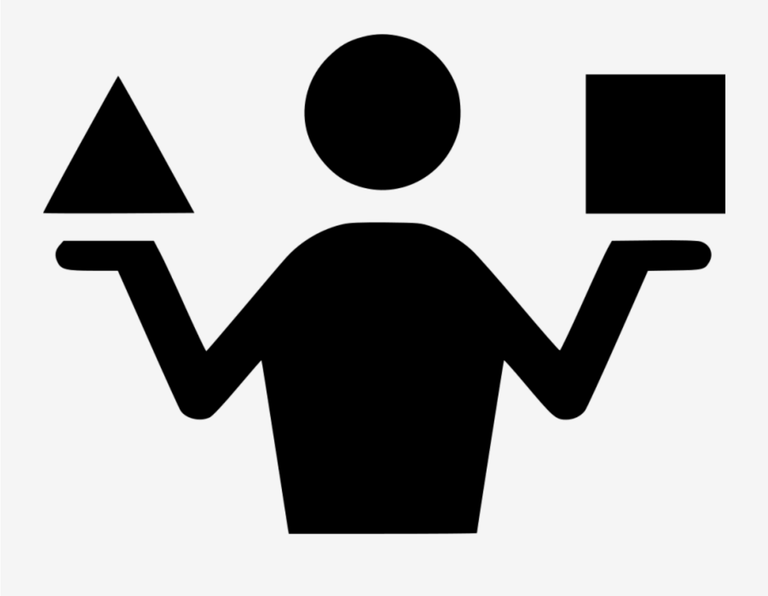
- **Idea:** use the Bayes formula to compute a probability for the hypothesis being true.

$$\boldsymbol{H}\colon\ \theta_A > \theta_B + \alpha$$

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

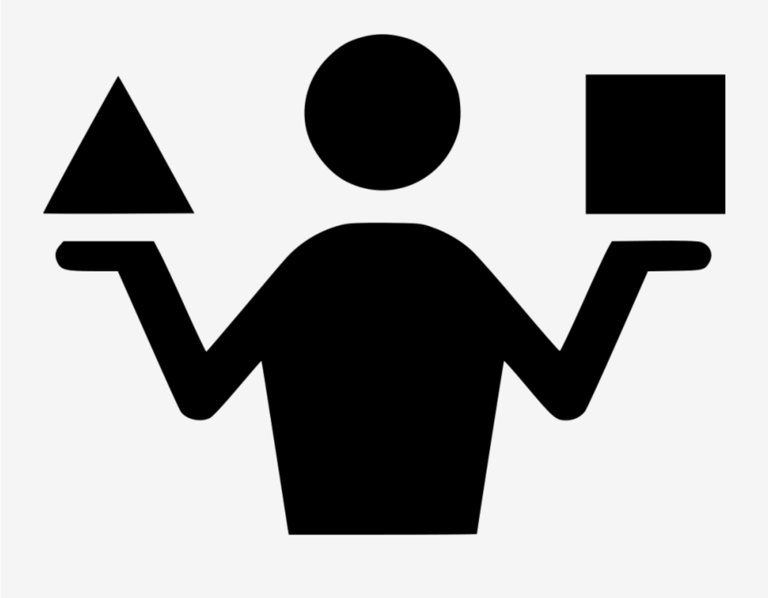
Usage Patterns

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?

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Trends and Patterns in the field

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

How many papers did use significance testing?

73

p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Trends and Patterns in the field

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

How many papers did use significance testing?

73 p-value	Bayes Factor
Confidence Interval **6**	Posterior Intervals

Trends and Patterns in the field

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

How many papers did use significance testing?

73 p-value	Bayes Factor 0
Confidence Interval 6	Posterior Intervals

Trends and Patterns in the field

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

How many papers did use significance testing?

73 p-value	Bayes Factor **0**
6 Confidence Interval	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?

- Many papers (~360) did **not** include any hypothesis assessment.

- p-value based tests are the **dominant** choice among NLP practitioners.

p-value: 73	Bayes Factor: 0
Confidence Interval: 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?

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

73 p-value	Bayes Factor 0
6 Confidence Interval	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?

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

Why?

p-value: 73	Bayes Factor: 0
Confidence Interval: 6	Posterior Intervals: 0

Lack of exposure to alternative algorithms

Lack of exposure to alternative algorithms

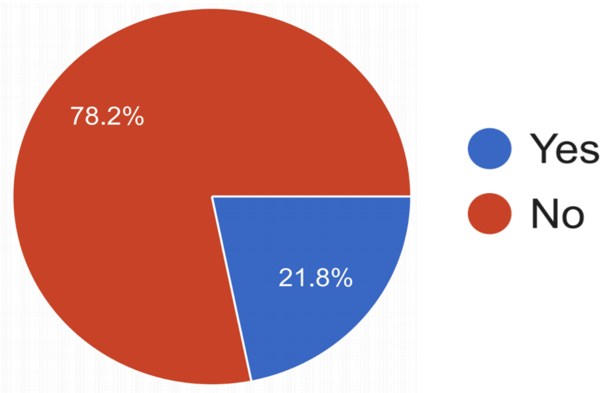
- The imbalance in usage:
  - Is it intentional?

- 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.” 🤔

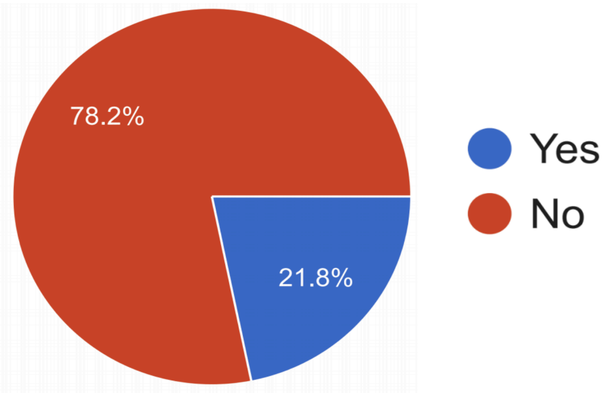


Do you know the definition of "Bayes Factor"?

Lack of exposure to alternative algorithms

- The imbalance in usage:
  - Is it intentional?

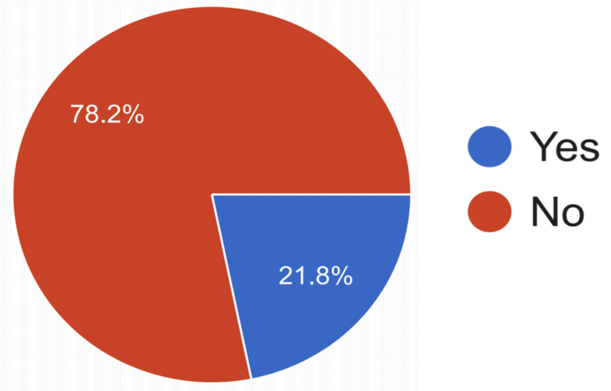
- Many people did not know the definition of “Bayes Factor.” 🤔



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

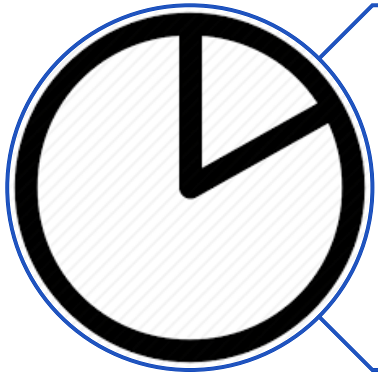
Ease of Interpretation

Usage Patterns

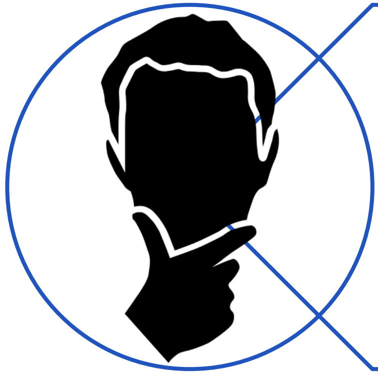
Ease of Interpretation

- 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?

$$\underbrace{P(\text{extreme obs.} \mid \bar{H})}_{p\text{-value}} \ll \alpha$$

Are we good at interpreting the p-values?

$$\underbrace{P(\text{extreme obs.} \mid \bar{H})}_{\textit{p-value}} \ll \alpha$$

- Pretty complex notion!

Are we good at interpreting the p-values?

$$\underbrace{P(\text{extreme obs.} \mid \bar{H})}_{p\text{-value}} \ll \alpha$$

"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*

- Pretty complex notion!

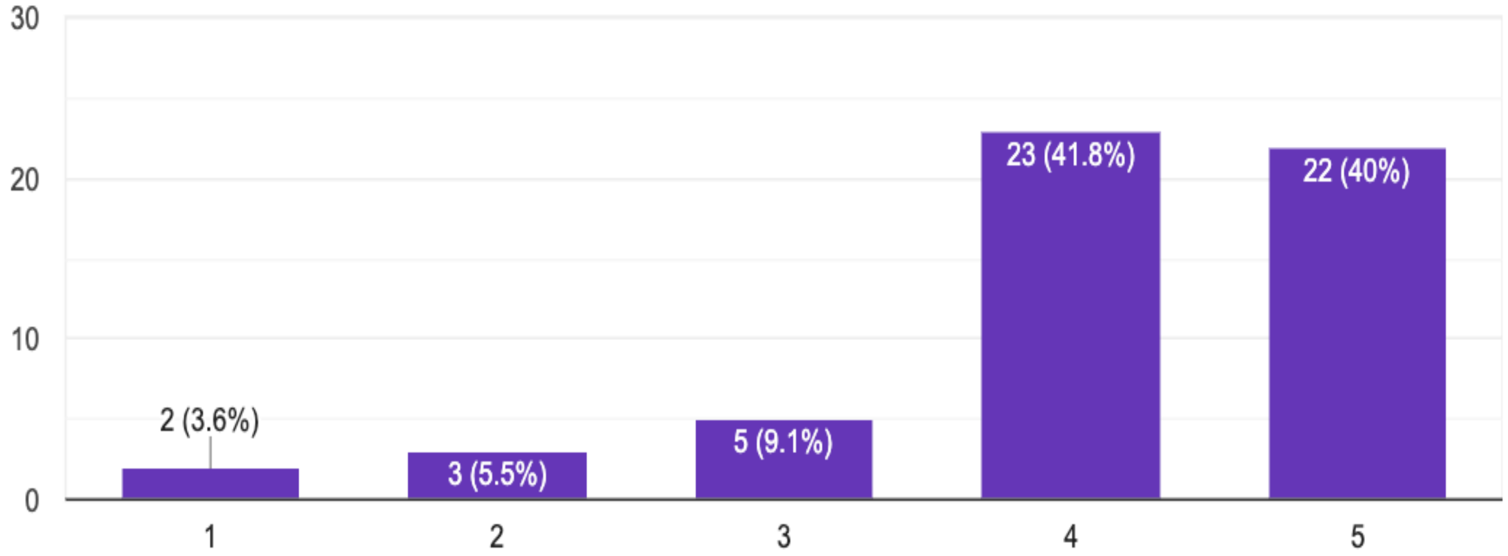
A Survey Question: Interpreting P-value (1)

A Survey Question: Interpreting P-value (1)

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

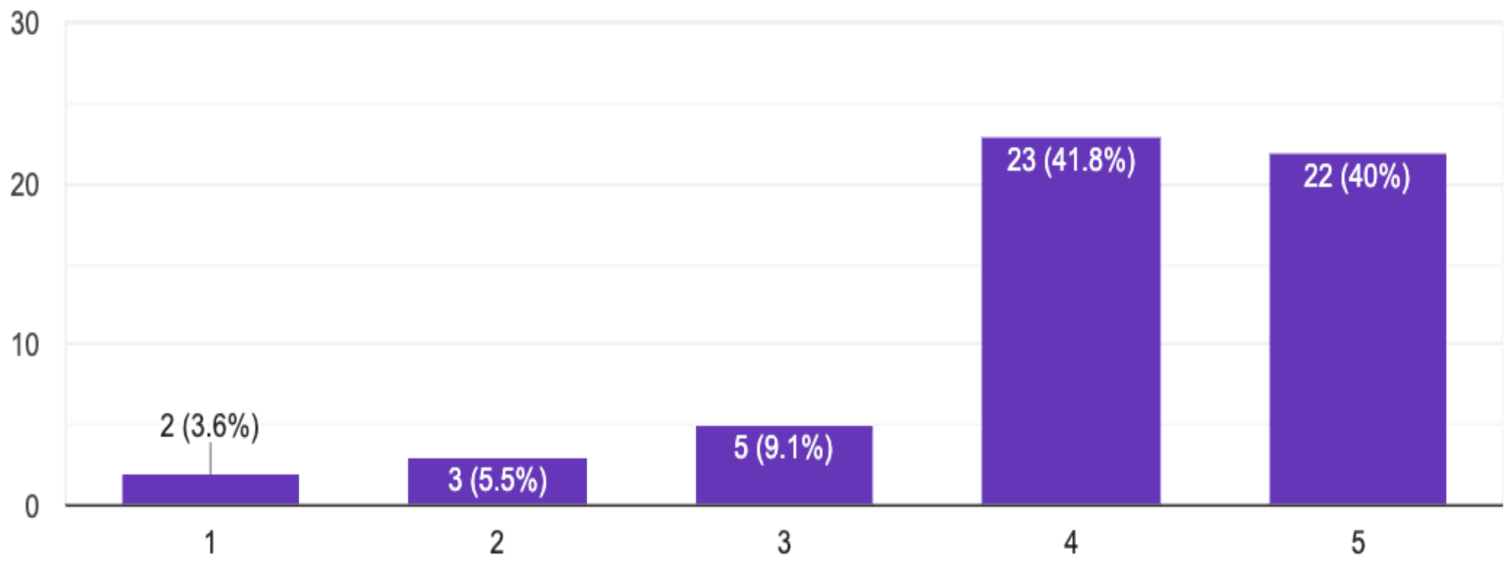
A Survey Question: Interpreting P-value (1)

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



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)

system	$\hat{\theta}$	θ
classifier-A	*38%*	?
classifier-B	*45%*	?

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?*

system	$\hat{\theta}$	θ
classifier-A	*38%*	?
classifier-B	*45%*	?

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.

...

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?*

system	$\hat{\theta}$	θ
classifier-A	*38%*	?
classifier-B	*45%*	?

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.

...

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?*

system	$\hat{\theta}$	θ
classifier-A	*38%*	?
classifier-B	*45%*	?

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

$$\mathbf{P}\left[\hat{\theta}_B - \hat{\theta}_A > 7 \middle| \theta_A = \theta_B\right] < 0.01$$

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

$$\mathbf{P}[\theta_B > \theta_A] > 0.99$$

...

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?*

system	$\hat{\theta}$	θ
classifier-A	*38%*	?
classifier-B	*45%*	?

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

$$\mathbf{P}\left[\hat{\theta}_B - \hat{\theta}_A > 7 \middle| \theta_A = \theta_B\right] < 0.01$$

30% ✗ b) With a probability 99% classifier-2 will have a higher performance than classifier-1.

$$\mathbf{P}[\theta_B > \theta_A] > 0.99$$

...

Only a small percentage correctly answered a basic p-value interpretation question.

Ease of interpretation: Bayesians vs Freq.

Frequentist	Bayesian
p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Ease of interpretation: Bayesians vs Freq.

Frequentist	Bayesian
p-value	Bayes Factor
Confidence Interval	Posterior Intervals

*Our participants mistakenly interpret **frequentist** notions in a **Bayesian** way.*

Ease of interpretation: Bayesians vs Freq.

$$\underbrace{P\left(\text{obs.} > \hat{\theta}_A - \hat{\theta}_B \mid \overline{\boldsymbol{H}}\right)}_{p\text{-value}}$$

Frequentist	Bayesian
p-value	Bayes Factor
Confidence Interval	Posterior Intervals

Our participants mistakenly interpret ***frequentist*** *notions in a* ***Bayesian*** *way.*

Ease of interpretation: Bayesians vs Freq.

$$\underbrace{P\left(\text{obs.} > \hat{\theta}_A - \hat{\theta}_B \mid \overline{\boldsymbol{H}}\right)}_{p\text{-value}}$$

Frequentist	Bayesian
p-value	Bayes Factor
Confidence Interval	Posterior Intervals

$$P(\boldsymbol{H} \mid \text{observations})$$

*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

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`

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`

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`