Allen Institute for AI

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

ACL 2020

Erfan Sadeqi-Azer (Indiana U → Google) Ashish Sabharwal (AI2) Dan Roth (UPenn).

About me

- Join in 2013
- Graduated in early 2019
- Now: AI2, Seattle

This talk

- Hypothesis testing/assessment:

  A topic we're [kind of] familiar with, by virtue of working in an empirical field.

  There are holes in our understanding of these concepts and their usage.

- Mix of new ideas and known stuff.

The Cycle of Empirical Research

The Cycle of Empirical Research

Empirical Observations

The Cycle of Empirical Research

????
Research Problem

Empirical Observations

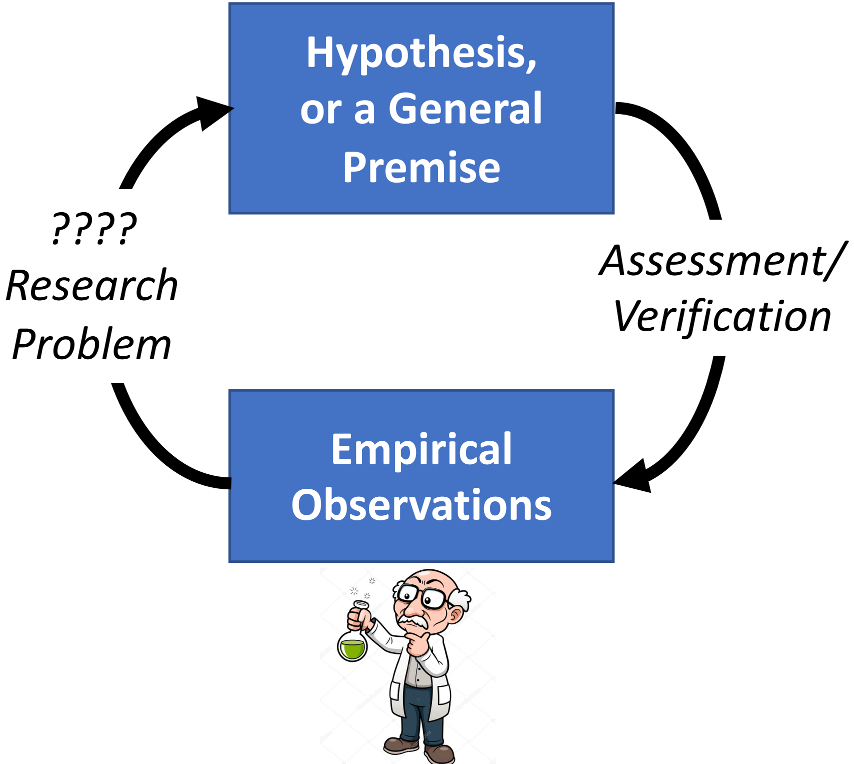
The Cycle of Empirical Research

Hypothesis, or a General Premise

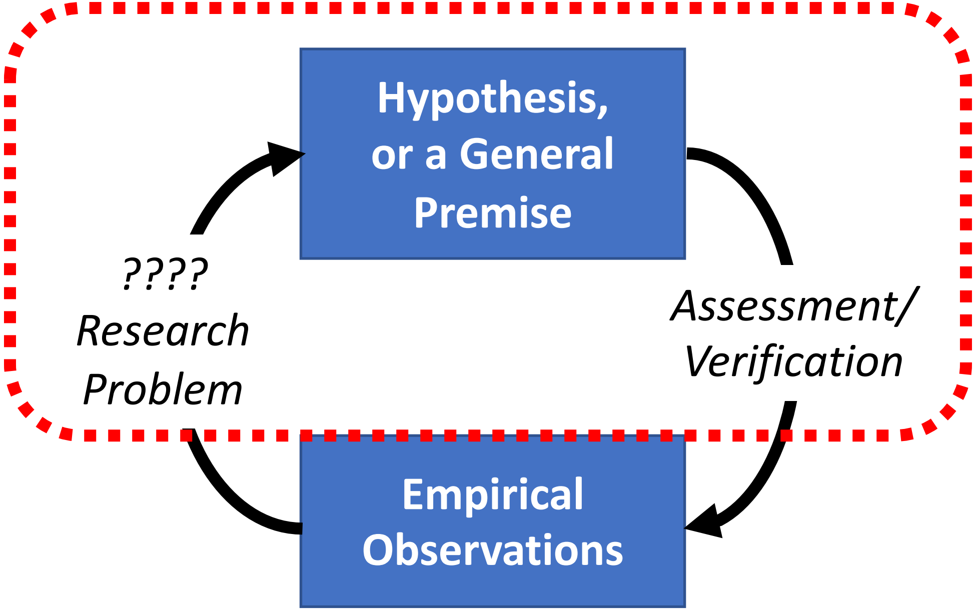
????
Research Problem

Empirical Observations

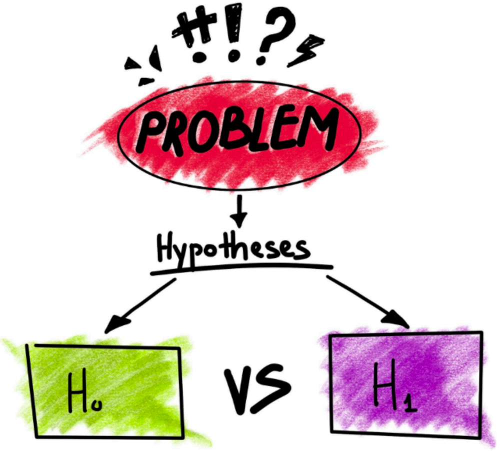
The Cycle of Empirical Research



The Cycle of Empirical Research



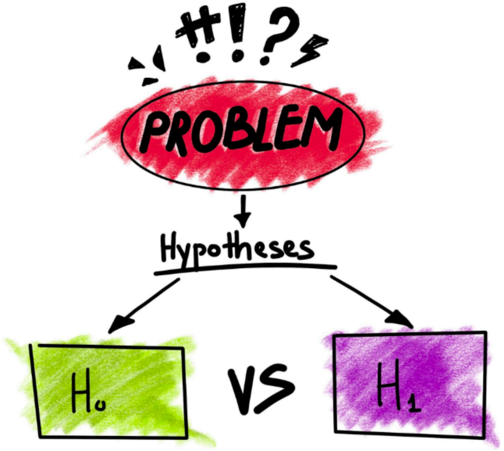
Hypotheses



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

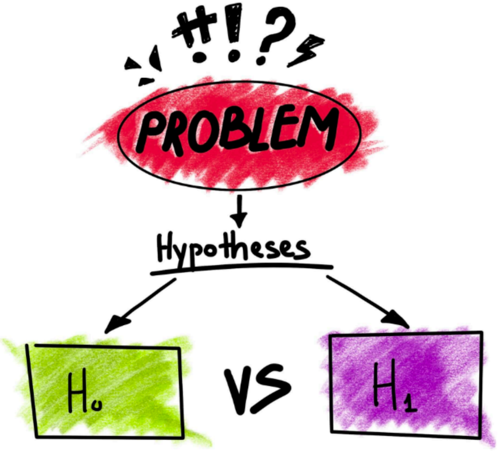
- Terminology:
  - A hypothesis is **never "proved"**
  - But it could be **"supported"** by the evidence



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

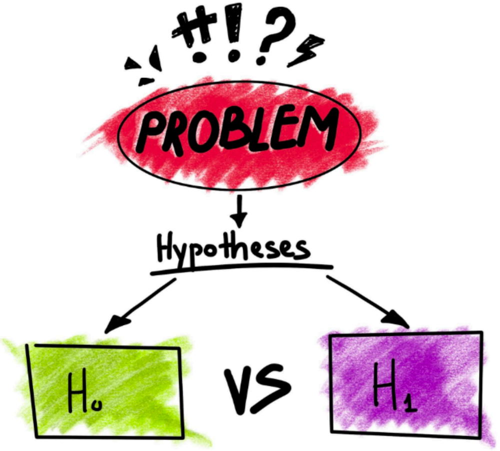
- Terminology:
  - A hypothesis is **never "proved"**
  - But it could be **"supported"** by the evidence



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

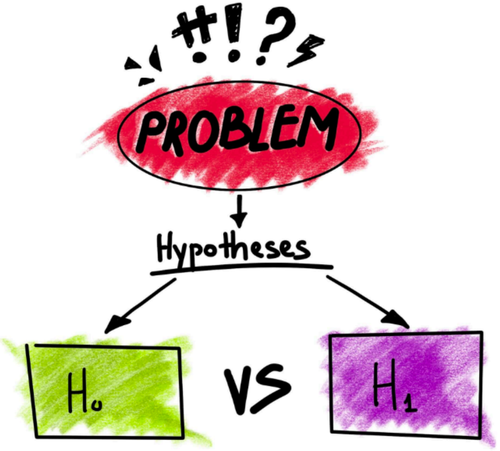
- Terminology:
  - A hypothesis is **never "proved"**
  - But it could be **"supported"** by the evidence



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

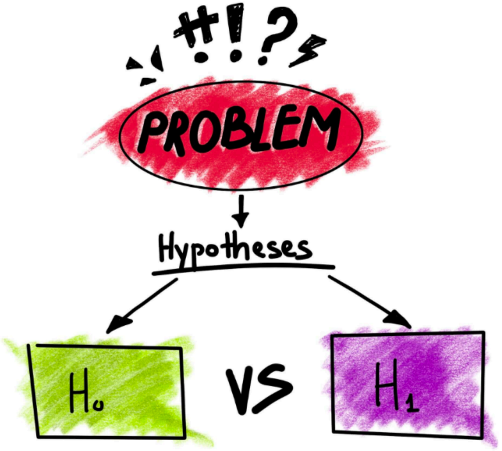
- Terminology:
  - A hypothesis is **never "proved"**
  - But it could be **"supported"** by the evidence



Hypotheses

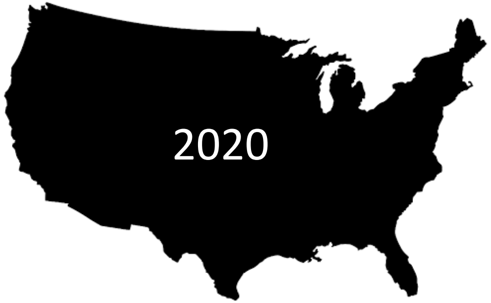
- A **prediction** about how the world will behave **if our idea is correct**
- Worded as an **if-then** statement
- A hypothesis is a **testable** prediction
- A hypothesis is a **falsifiable** statement

- Terminology:
  - A hypothesis is **never "proved"**
  - But it could be **"supported"** by the evidence

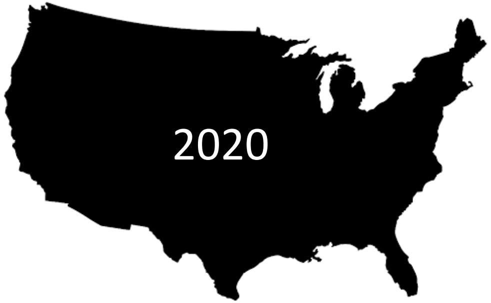


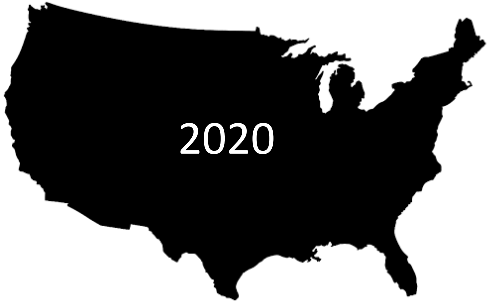
Not a good statistical hypothesis

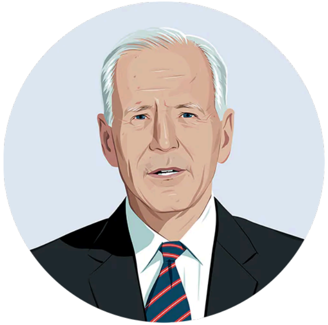
Not a good statistical hypothesis



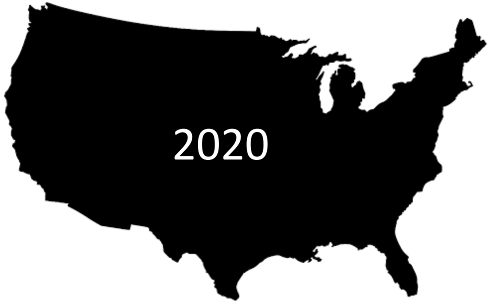
Not a good statistical hypothesis

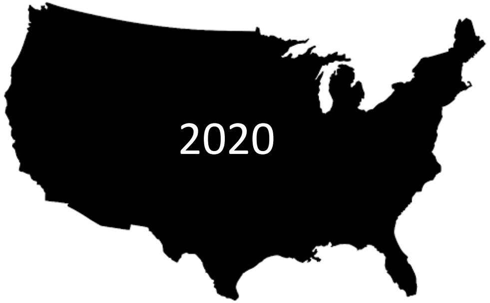






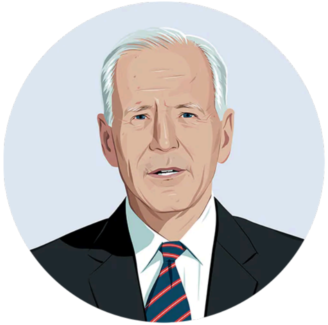
Not a good statistical hypothesis





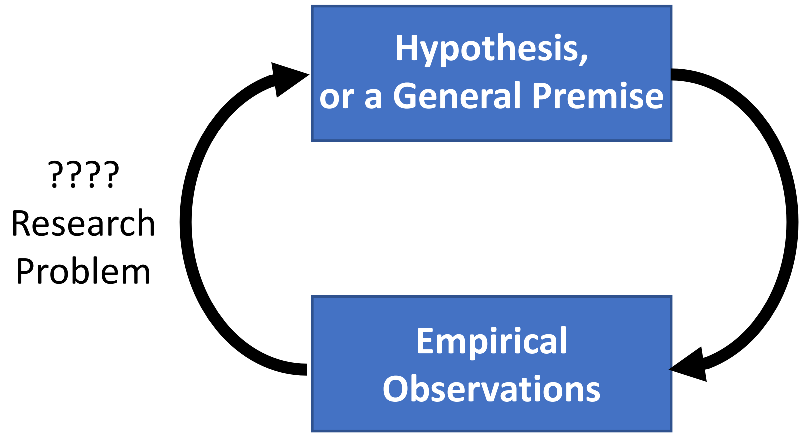
$<$

?



“I can always prepare a nice presentation, if I stay up the night before.”

A Typical AI Experiment

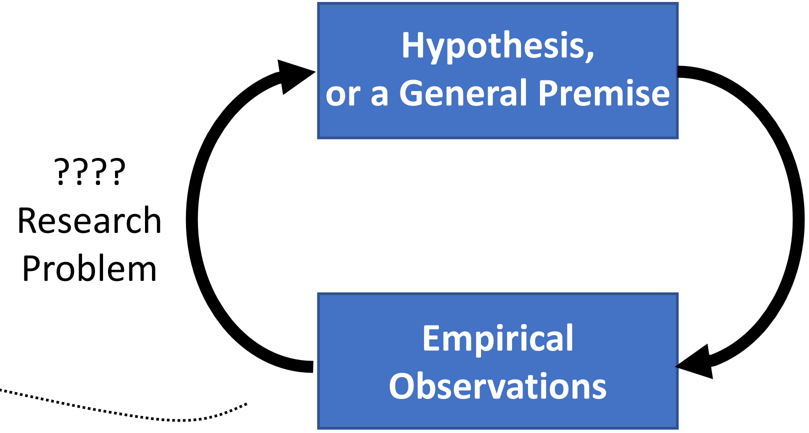


A Typical AI Experiment

(Clark et al., 2018)

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

|D|= 2376



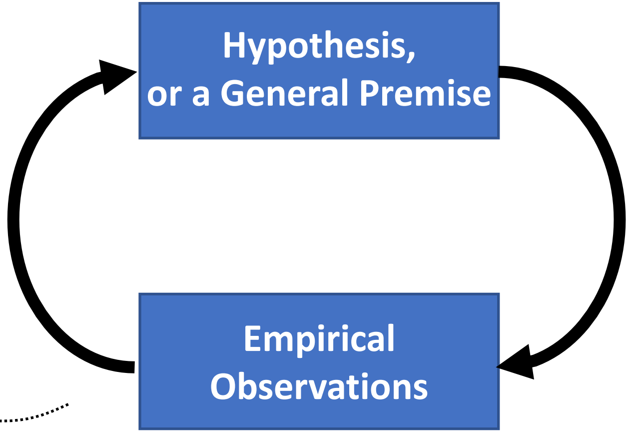
A Typical AI Experiment

(Clark et al., 2018)

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

|D|= 2376

????
Research
Problem



A Typical AI Experiment

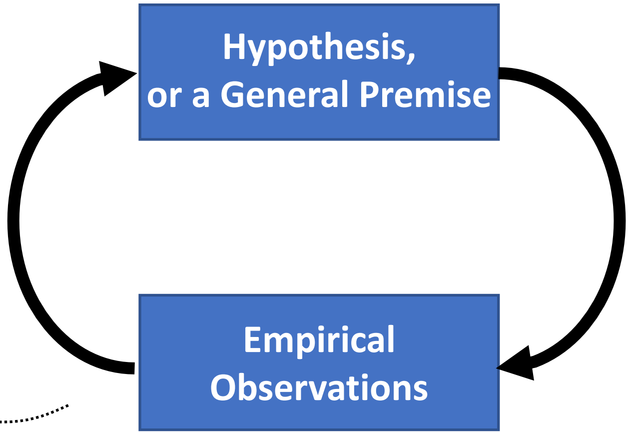
- Can this apparent difference in performance be explained simply by **random chance**?
- Do we have sufficient evidence to conclude that Ⓐ is in fact **inherently** stronger than Ⓑ on these datasets?

(Clark et al., 2018)

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

|D|= 2376

????
Research
Problem



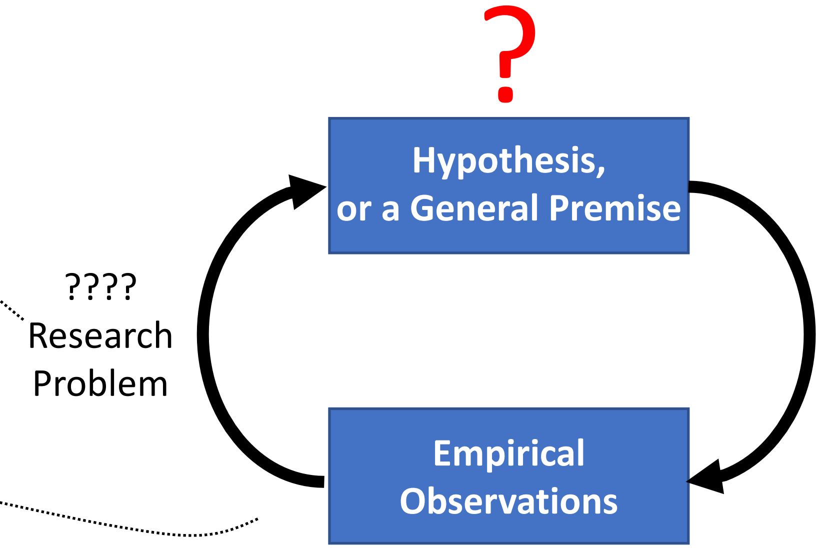
A Typical AI Experiment

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

(Clark et al., 2018)

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

|D|= 2376



A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ and Ⓑ are **inherently different**,

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ and Ⓑ are **inherently different**,

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ and Ⓑ are **inherently different**, in the sense that **if** they were inherently **identical**,

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ 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

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ 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.
- H2*: Ⓐ and Ⓑ are **inherently different**,

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ 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.
- H2*: Ⓐ 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

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ 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.
- H2*: Ⓐ 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:

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ 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.
- H2*: Ⓐ 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 **H1**, even though it is harder to interpret.*

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ 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.
- H2*: Ⓐ 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:

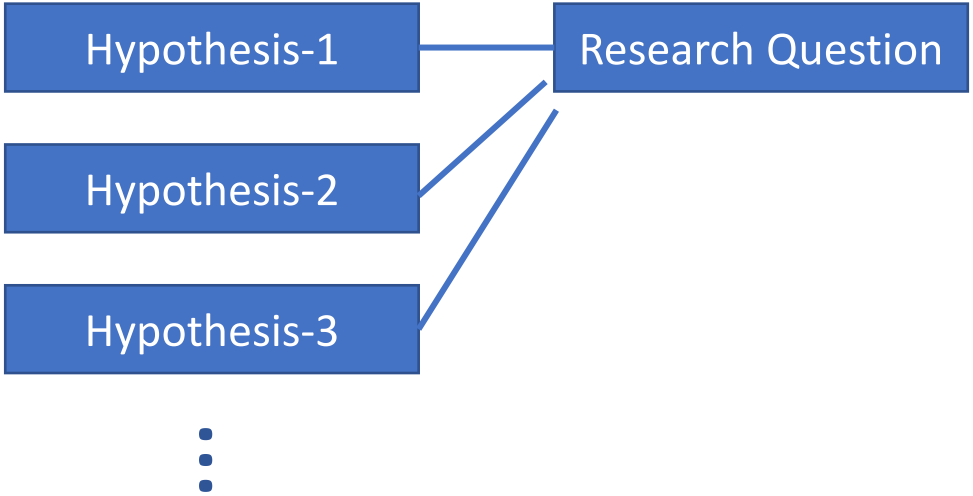
*Almost everyone uses **H1**, even though it is harder to interpret.*

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- H1*: Ⓐ 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.
- H2*: Ⓐ 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.

And many more . . .



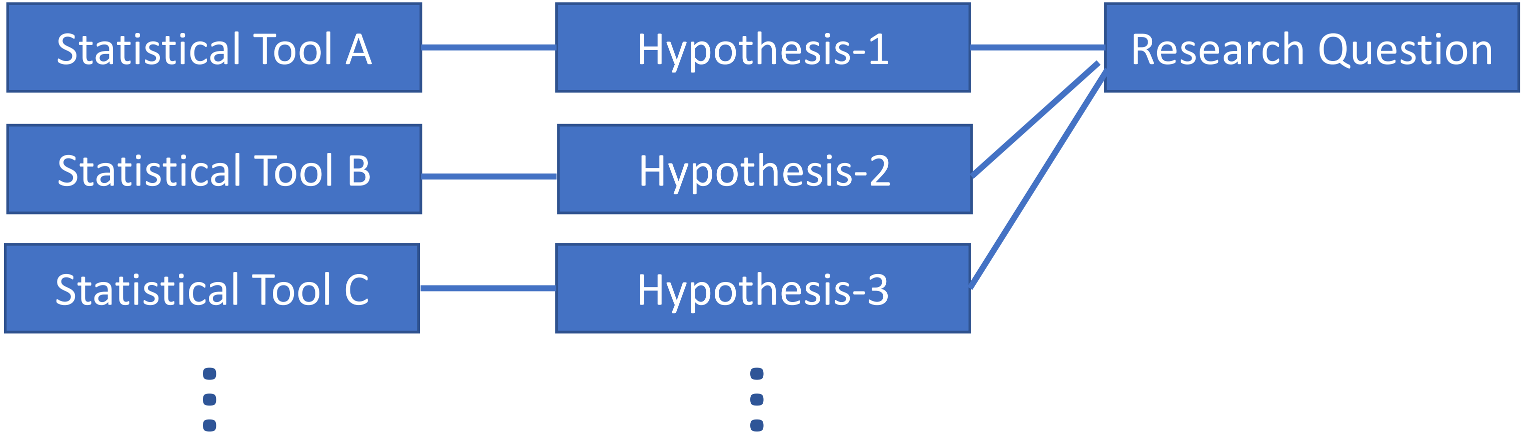
- Observation 1: There are **many different hypotheses** that could address a **single research question**.

Hypothesis vs Statistical Techniques

Research Question

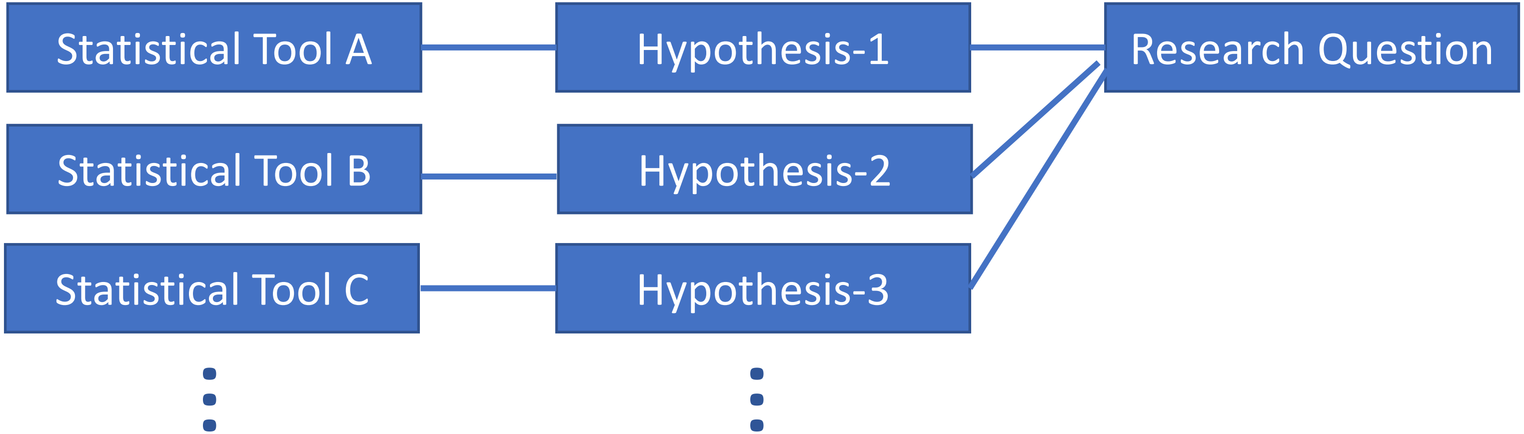
- Observation 2: Each hypothesis ought to be assessed with an **appropriate** statistical tool.

Hypothesis vs Statistical Techniques



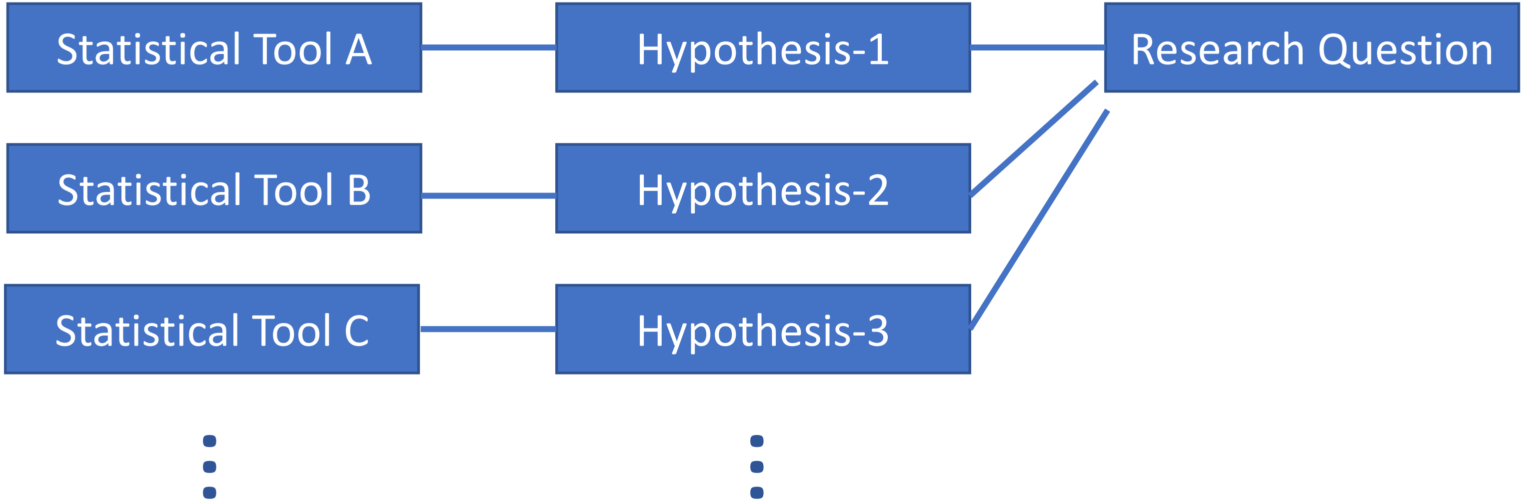
- Observation 2: Each hypothesis ought to be assessed with an **appropriate** statistical tool.

Hypothesis vs Statistical Techniques



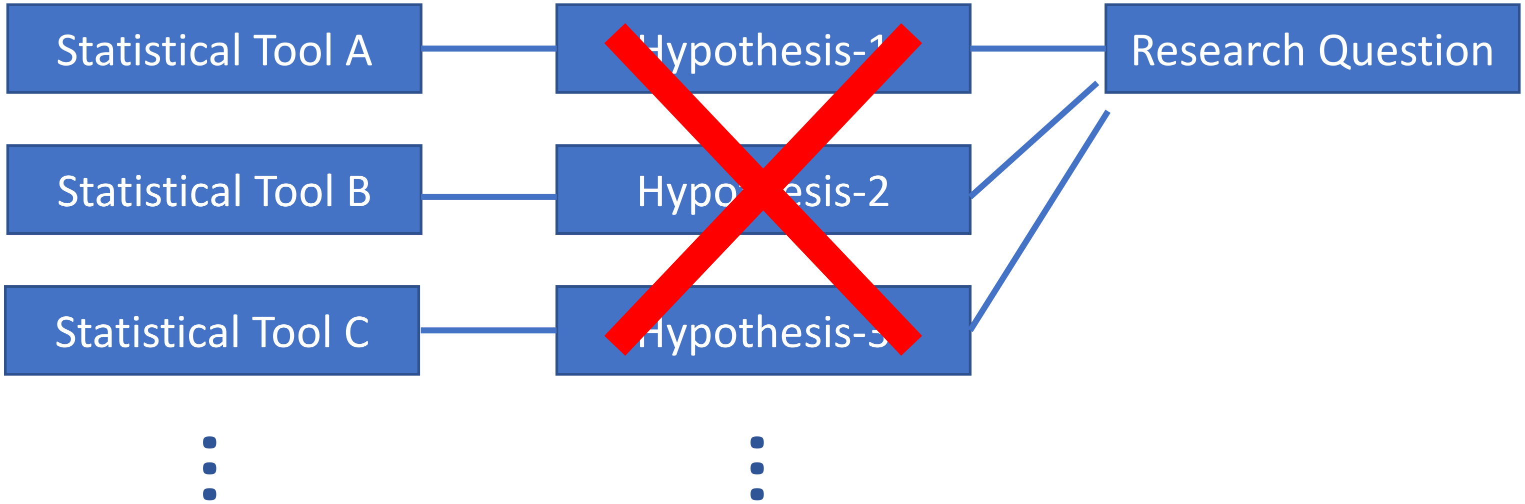
- Observation 2: Each hypothesis ought to be assessed with an **appropriate** statistical tool.
- Corollary: Researchers should **start with a hypothesis** that best serves their goal and choose an appropriate statistical assessment accordingly.

Omission of hypotheses



Omission of hypotheses

- Observation 3: Somehow, we tend to forget about hypotheses



Omission of hypotheses

(EMNLP 2018)

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.

Model	Data	Regents Test
MLN (Khot et al., 2015)	-	47.5
FRETS (Compact)	Regents Tables	**60.7**
	Monarch Tables	56.0
	Regents+Monarch Tables	59.9

Statistical Tool — Hypothesis — Research Question

Omission of hypotheses

(EMNLP 2018)

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.

Model	Data	Regents Test
MLN (Khot et al., 2015)	-	47.5
FRETS (Compact)	Regents Tables	**60.7**
	Monarch Tables	56.0
	Regents+Monarch Tables	59.9

Ⓐ >> Ⓑ

Statistical Tool — Hypothesis — Research Question

Omission of hypotheses

(EMNLP 2018)

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.

Model	Data	Regents Test
MLN (Khot et al., 2015)	-	47.5
FRETS (Compact)	Regents Tables	**60.7**
	Monarch Tables	56.0
	Regents+Monarch Tables	59.9

Statistical Tool — Hypothesis — Research Question

Ⓐ >> Ⓑ

Omission of hypotheses

(EMNLP 2018)

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.

Model	Data	Regents Test
MLN (Khot et al., 2015)	-	47.5
FRETS (Compact)	Regents Tables	**60.7**
	Monarch Tables	56.0
	Regents+Monarch Tables	59.9

Statistical Tool — Hypothesis — ? — Research Question

Ⓐ >> Ⓑ

Omission of hypotheses

(EMNLP 2018)

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.

Model	Data	Regents Test
MLN (Khot et al., 2015)	-	47.5
FRETS (Compact)	Regents Tables	**60.7**
	Monarch Tables	56.0
	Regents+Monarch Tables	59.9

Statistical Tool — Hypothesis — ? — Research Question

Ⓐ >> Ⓑ

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

Talk Summary & Statement

- Motivated by several serious **malpractices**:
  - **Under-reporting** of hypotheses and how they address research questions.
  - Inability to **interpret** statistical tools or their results.
  - Lack of **awareness** about various alternatives; e.g., **Bayesian** 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 . . .

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

(Kruschke and Liddell, 2018)

Statistical tools in this work . . .

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

(Kruschke and Liddell, 2018)

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

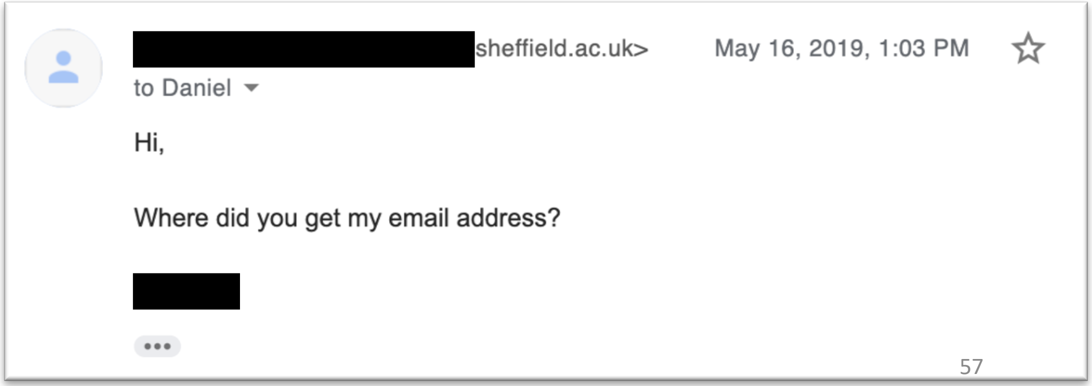
- A questionnaire containing general and specific questions about significance assessment tools
- Sent it to over 400 researchers randomly selected from ACL'18 proceedings
- ~50 individuals responded

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



[redacted]sheffield.ac.uk> May 16, 2019, 1:03 PM

to Daniel

Hi,

Where did you get my email address?

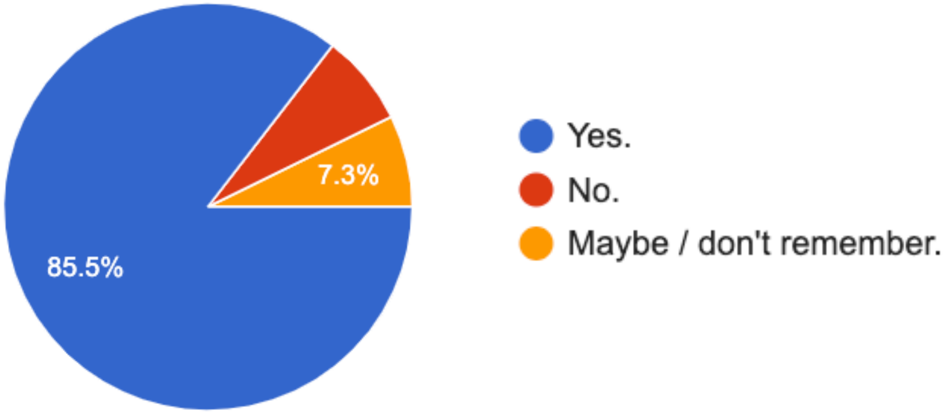
[redacted]

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)."*

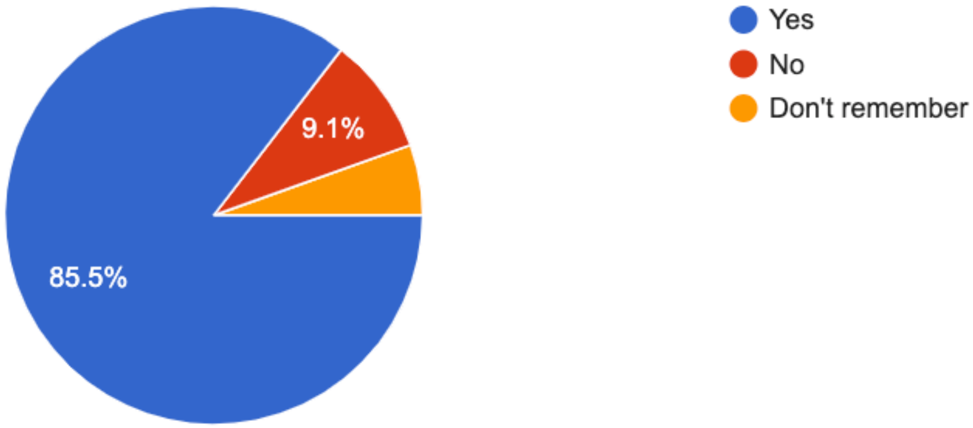


Participants in Our Survey

- *“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.)”*

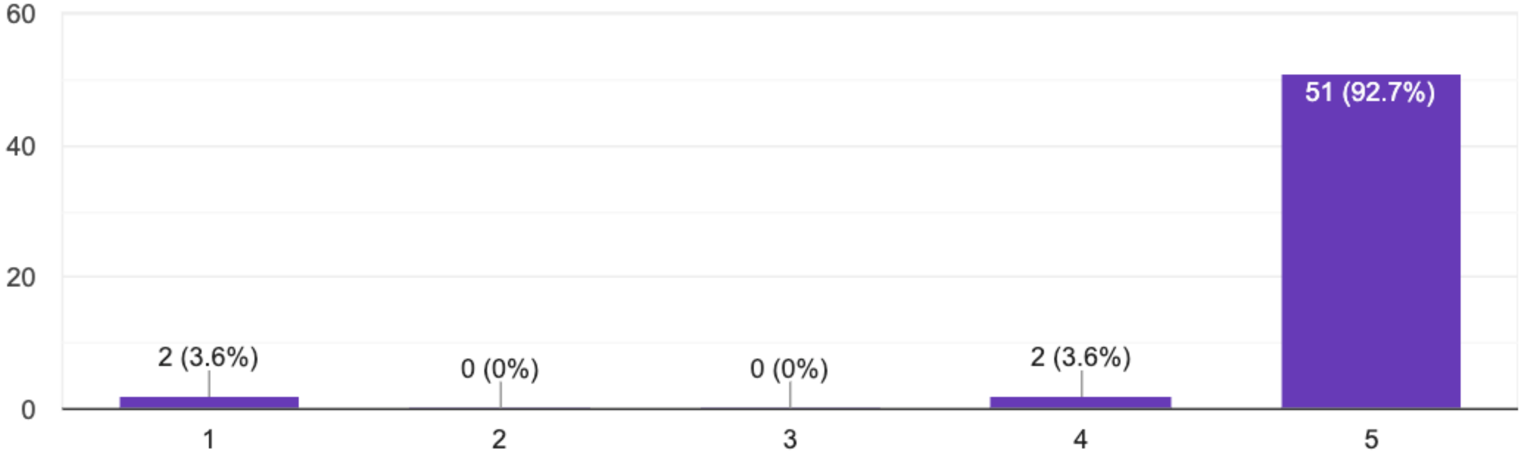


Participants in Our Survey

- *“I am not a robot”*

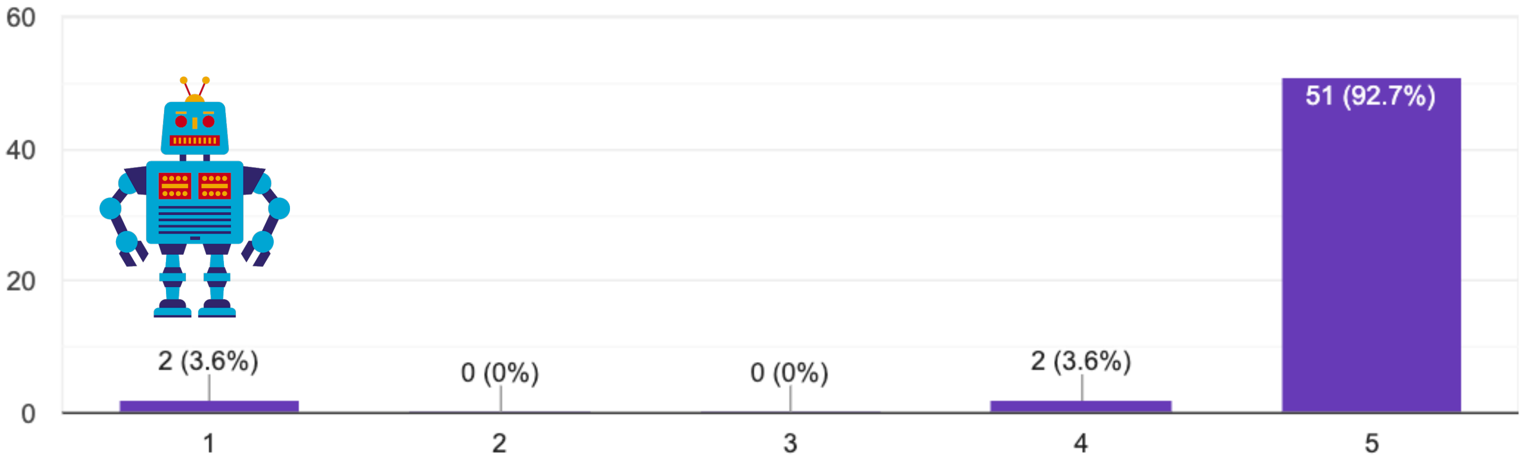
Participants in Our Survey

- *"I am not a robot"*



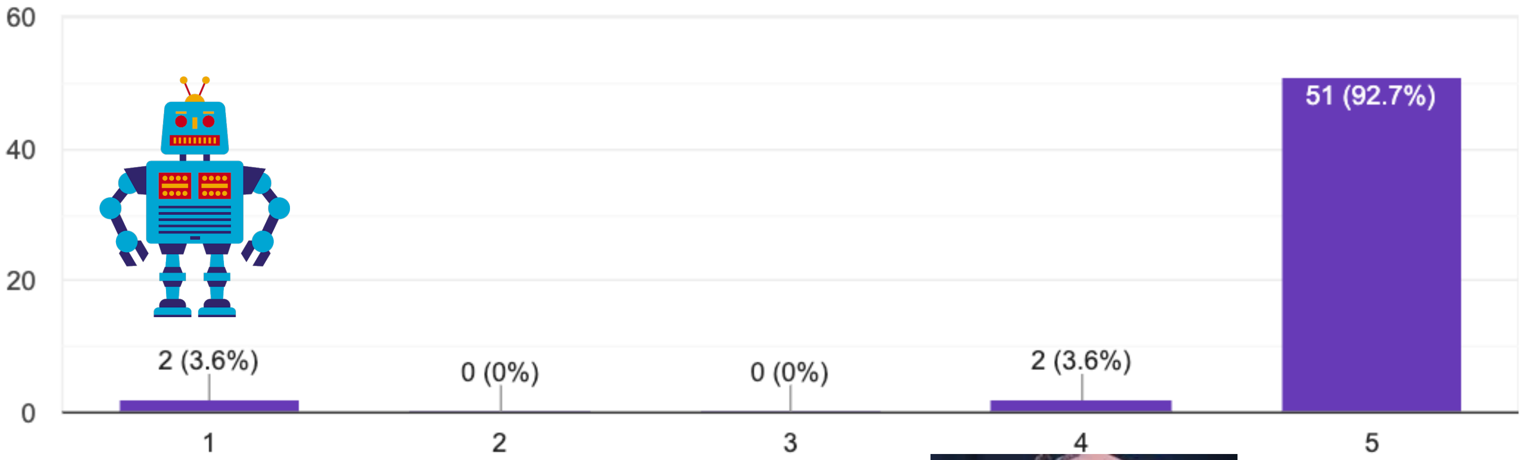
Participants in Our Survey

- *“I am not a robot”*



Participants in Our Survey

- *“I am not a robot”*



Trends and Patterns in the field

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

How many papers did use significance testing?

	Frequentist	Bayesian
Binary Decision	Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	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

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

Trends and Patterns in the field

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

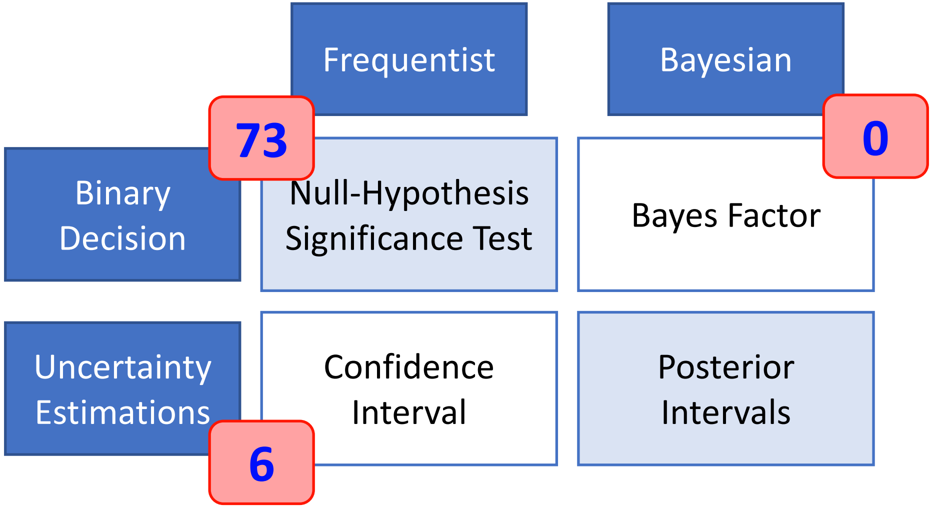
How many papers did use significance testing?

	Frequentist	Bayesian
Binary Decision	**73** Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	**6** 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?



Trends and Patterns in the field

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

How many papers did use significance testing?

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

Trends and Patterns in the field

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

How many papers did use significance testing?

- The overuse of NHST is why we focus on its issues.
- All techniques have their own limitations and ought to be used with this in mind.

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

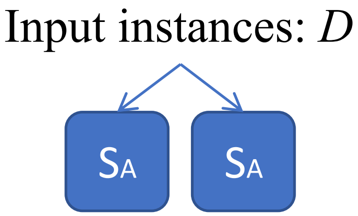
	Frequentist	Bayesian
Binary/Categorical Decisions	Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	Confidence Interval	Posterior Intervals

Notation

Notation

- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(S_i, D)$
  - $\theta_i \neq \boldsymbol{M}(S_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …

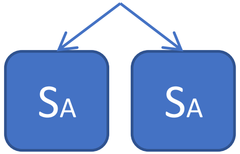
Notation



- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …

Notation

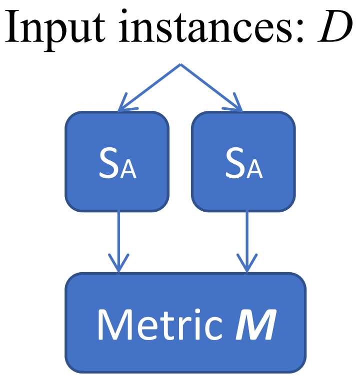
Input instances: D



- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …

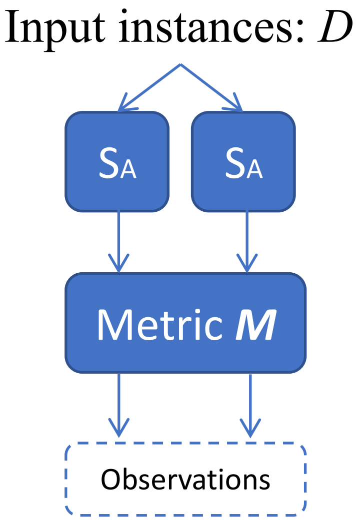
Notation

- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …



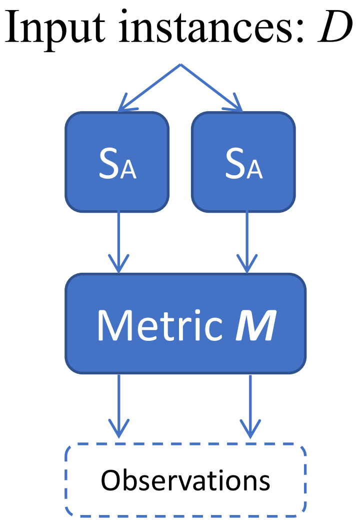
Notation

- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …



Notation

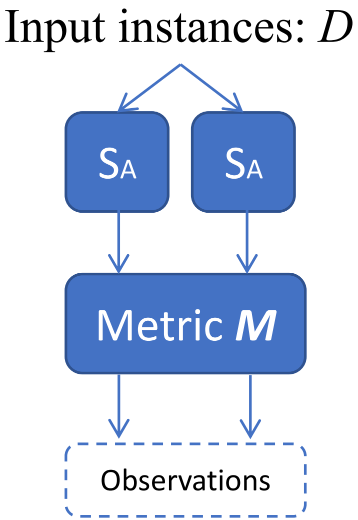
- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …



Claims about the inherent properties $\boldsymbol{\theta_A}, \boldsymbol{\theta_B}$ of the two systems.

Notation

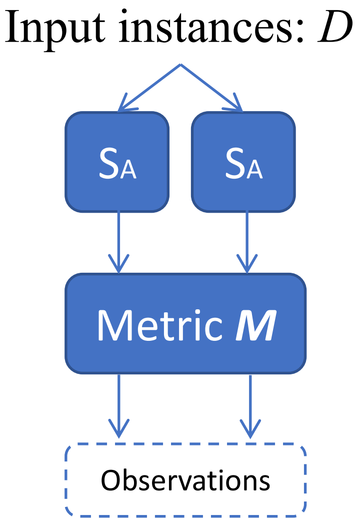
- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(S_i, D)$
  - $\theta_i \neq \boldsymbol{M}(S_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …

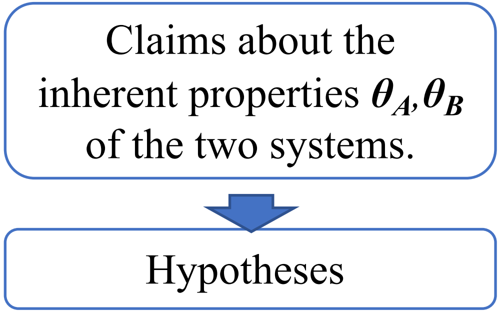


Claims about the inherent properties $\boldsymbol{\theta_A}, \boldsymbol{\theta_B}$ of the two systems.

Notation

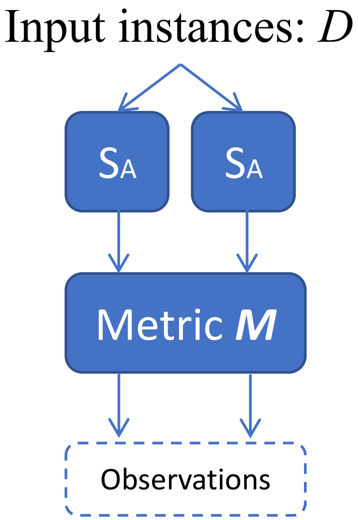
- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …

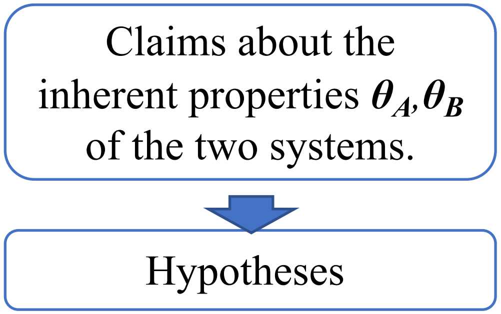




Notation

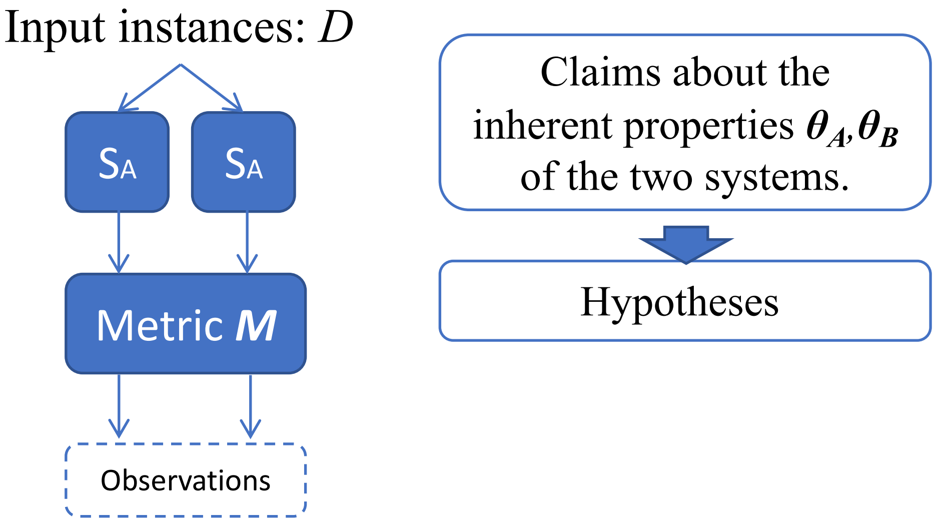
- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …





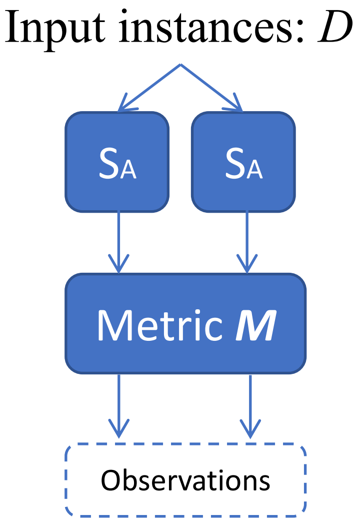
Notation

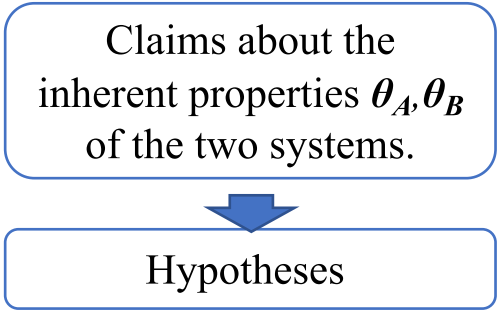
- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …



Notation

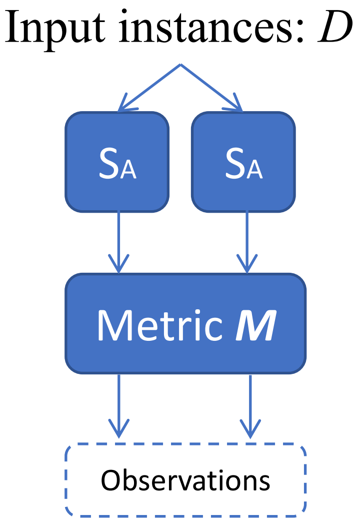
- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …

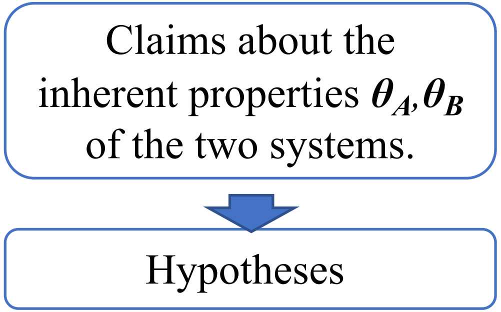




Notation

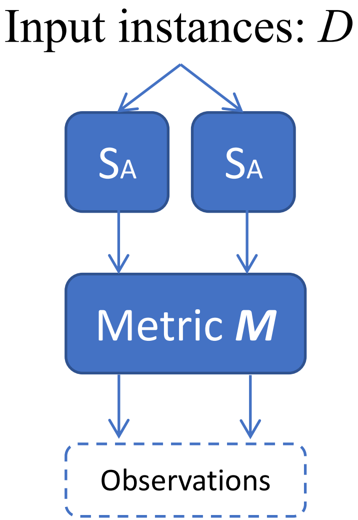
- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …

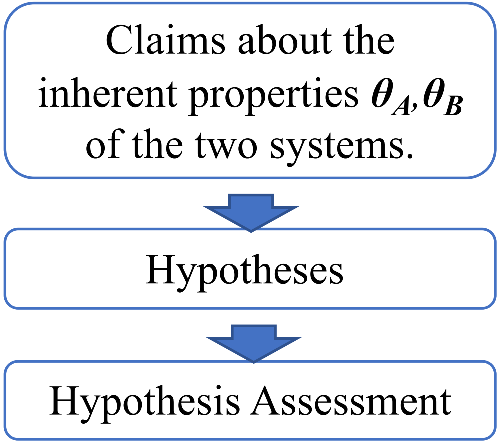




Notation

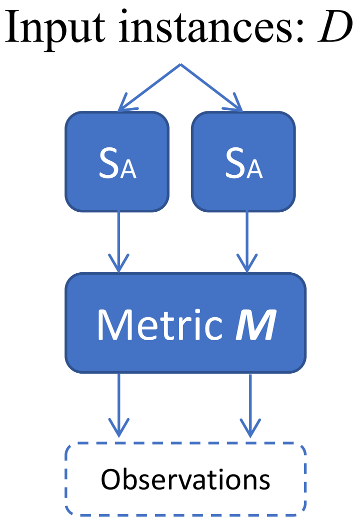
- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …

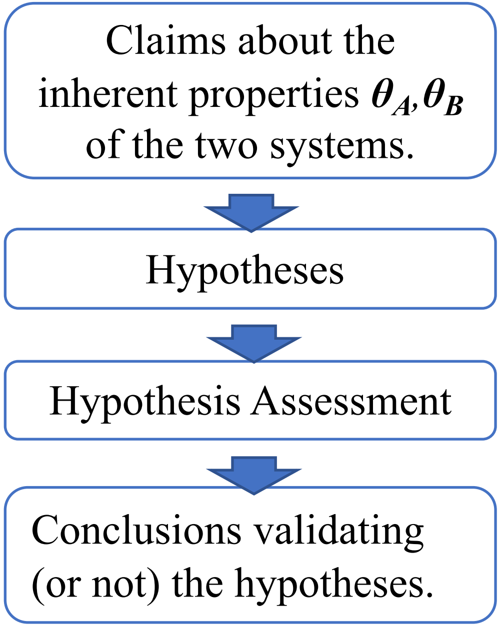




Notation

- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(S_i, D)$
  - $\theta_i \neq \boldsymbol{M}(S_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$
  - H2: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B} + b$
  - …





Null-Hypothesis Significance Testing

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)

Null-Hypothesis Significance Testing

- The goal is to decide whether the inverse of your claim 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 the inverse of your claim 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 the inverse of your claim can be rejected.
- Make a hypothesis (that you want it to be rejected): null-hypothesis.
- Assume that null-hypothesis is correct. $H_0: \boldsymbol{\theta}_A = \boldsymbol{\theta}_B$
- 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 the inverse of your claim can be rejected.
- Make a hypothesis (that you want it to be rejected): null-hypothesis.
- Assume that null-hypothesis is correct. $H_0\colon \boldsymbol{\theta_A} = \boldsymbol{\theta_B}$
- 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 the inverse of your claim can be rejected.
- Make a hypothesis (that you want it to be rejected): null-hypothesis.
- Assume that null-hypothesis is correct. H_0: $\boldsymbol{\theta_A} = \boldsymbol{\theta_B}$
- Calculate the probability of getting an outcome as "extreme" or more than the observed outcome.
  - This probability is called a "*p*-value."

e.g., a bigger accuracy improvement.

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)

Null-Hypothesis Significance Testing

- The goal is to decide whether the inverse of your claim can be rejected.
- Make a hypothesis (that you want it to be rejected): null-hypothesis.
- Assume that null-hypothesis is correct. H_0: $\boldsymbol{\theta_A} = \boldsymbol{\theta_B}$
- Calculate the probability of getting an outcome as "extreme" or more than the observed outcome.
  - This probability is called a "*p*-value."

e.g., a bigger accuracy improvement.

(Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018)

Null-Hypothesis Significance Testing: Example

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$

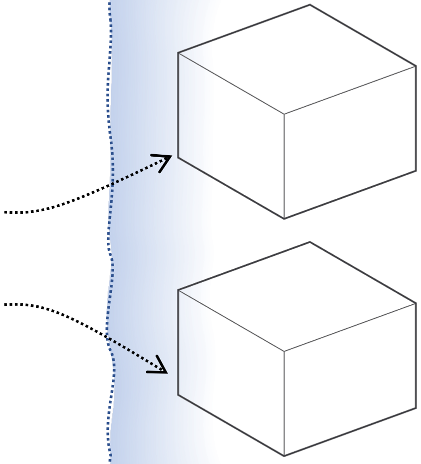
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\theta_A > \theta_B$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\theta_A = \theta_B$

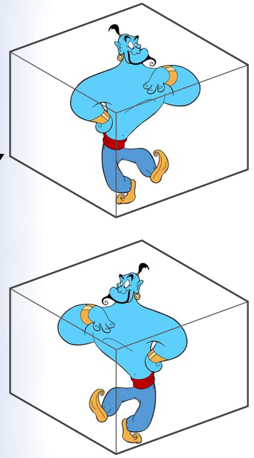


Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\theta_A > \theta_B$

- Null-Hypothesis **H0**: $\theta_A = \theta_B$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

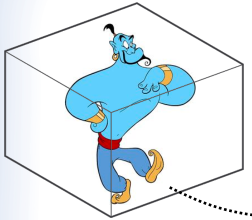


Null-Hypothesis Significance Testing: Example

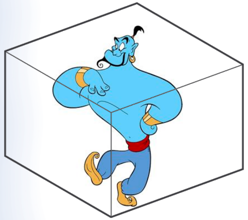
- Hypothesis **H1**: $\boldsymbol{\theta_A} > \boldsymbol{\theta_B}$

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A} = \boldsymbol{\theta_B}$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%



1, 0, 0, 0, ...

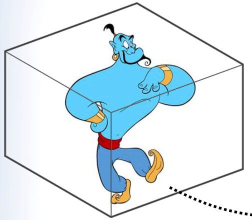


Null-Hypothesis Significance Testing: Example

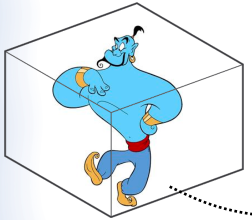
- Hypothesis **H1**: $\theta_A > \theta_B$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\theta_A = \theta_B$



1, 0, 0, 0, ...



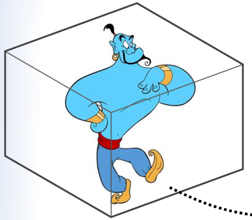
0, 0, 0, 1, ...

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

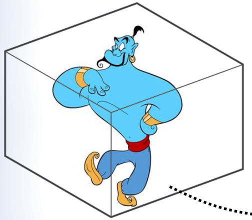
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$



$\theta_A = \mathrm{Prob}(X_A = 1)$

$1, 0, 0, 0, \ldots$



$\theta_B = \mathrm{Prob}(X_B = 1)$

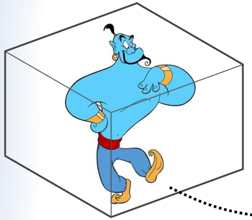
$0, 0, 0, 1, \ldots$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

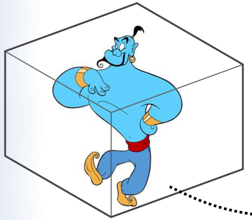
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$



$$\theta_A = \text{Prob}(X_A = 1)$$

$$\hat{\theta}_A = \text{Avg}[1, 0, 0, 0, \ldots\ ]$$



$$\theta_B = \text{Prob}(X_B = 1)$$

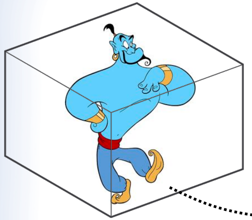
$$0, 0, 0, 1, \ldots$$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

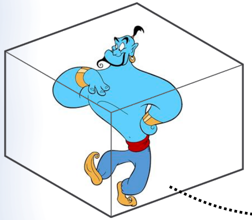
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$



$\theta_A = \mathrm{Prob}(X_A = 1)$

$\hat{\theta}_A = \mathrm{Avg}[1, 0, 0, 0, \ldots\ ]$



$\theta_B = \mathrm{Prob}(X_B = 1)$

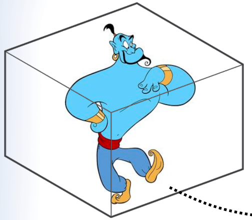
$\hat{\theta}_B = \mathrm{Avg}[0, 0, 0, 1, \ldots\ ]$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

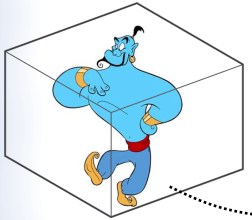
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$



$$\theta_A = \text{Prob}(X_A = 1)$$

$$\hat{\theta}_A = \text{Avg}[1, 0, 0, 0, \dots]$$



$$\theta_B = \text{Prob}(X_B = 1)$$

$$\hat{\theta}_B = \text{Avg}[0, 0, 0, 1, \dots]$$

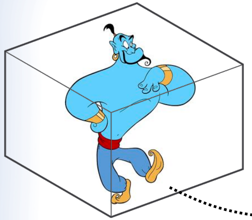
P-value=

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

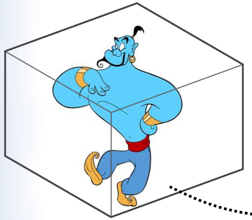
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$



$$\theta_A = \text{Prob}(X_A = 1)$$

$$\hat{\theta}_A = \text{Avg}[1, 0, 0, 0, \ldots]$$



$$\theta_B = \text{Prob}(X_B = 1)$$

$$\hat{\theta}_B = \text{Avg}[0, 0, 0, 1, \ldots]$$

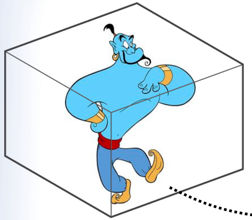
P-value= $\textbf{Prob}\left[\hat{\theta}_A - \hat{\theta}_B > 3.5 \middle| \boldsymbol{\theta_A = \theta_B}\right]$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

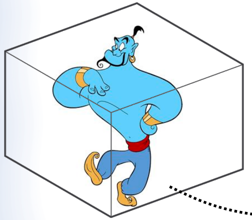
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$



$$\theta_A = \text{Prob}(X_A = 1)$$

$$\hat{\theta}_A = \text{Avg}[1, 0, 0, 0, \ldots]$$



$$\theta_B = \text{Prob}(X_B = 1)$$

$$\hat{\theta}_B = \text{Avg}[0, 0, 0, 1, \ldots]$$

P-value= $\mathbf{Prob}\left[\hat{\theta}_A - \hat{\theta}_B > 3.5 \middle| \boldsymbol{\theta_A = \theta_B}\right] <_? \beta$ (e.g., 0.05)

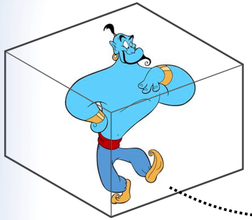
Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_A > \theta_B}$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

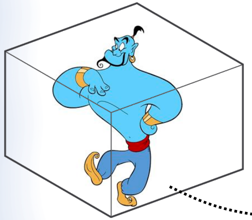
One-sided z-test

- Null-Hypothesis **H0**: $\boldsymbol{\theta_A = \theta_B}$



$$\theta_A = \text{Prob}(X_A = 1)$$

$$\hat{\theta}_A = \text{Avg}[1, 0, 0, 0, \dots\ ]$$



$$\theta_B = \text{Prob}(X_B = 1)$$

$$\hat{\theta}_B = \text{Avg}[0, 0, 0, 1, \dots\ ]$$

P-value= $\textbf{Prob}\left[\hat{\theta}_A - \hat{\theta}_B > 3.5 \middle| \boldsymbol{\theta_A = \theta_B}\right] <_? \beta$ (e.g., 0.05)

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

(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**.
- … but it does not tell anything about the likeliness of the null-hypothesis.

(Demsar, 2008; Goodman, 2008)

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.

(Demsar, 2008; Goodman, 2008)

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.**

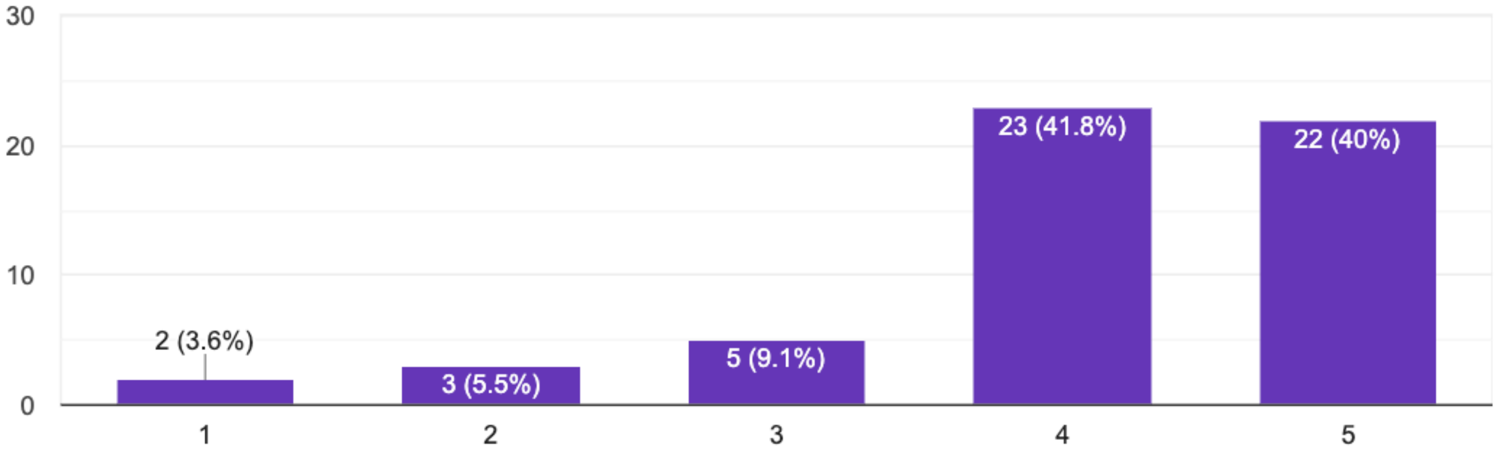
(Demsar, 2008; Goodman, 2008)

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

classifier-A	*38%*
classifier-B	*45%*

A Survey Question: Interpreting P-value

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

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

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

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

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

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.

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

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

$$\mathbf{P}[\boldsymbol{\theta}_B > \boldsymbol{\theta}_A] > 0.99$$

A Survey Question: Interpreting P-value

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

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.

$$\mathbf{P}[\hat{\theta}_B - \hat{\theta}_A > 7 \,|\, \boldsymbol{\theta}_A = \boldsymbol{\theta}_B] < 0.01$$

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

$$\mathbf{P}[\boldsymbol{\theta}_B > \boldsymbol{\theta}_A] > 0.99$$

🤦

Intermediate Summary

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

Intermediate Summary

- Null-Hypothesis Significance Tests are **the most popular choice** among NLP practitioners. Meanwhile, they're **difficult to understand** and highly **prone to misunderstanding**.

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

Intermediate Summary

- Null-Hypothesis Significance Tests are **the most popular choice** among NLP practitioners. Meanwhile, they're **difficult to understand** and highly **prone to misunderstanding**.

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

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

Intermediate Summary

- Null-Hypothesis Significance Tests are **the most popular choice** among NLP practitioners. Meanwhile, they're **difficult to understand** and highly **prone to misunderstanding**.

- P-values do <u>not</u> provide **probability** estimates on two classifiers being different (or equal).

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

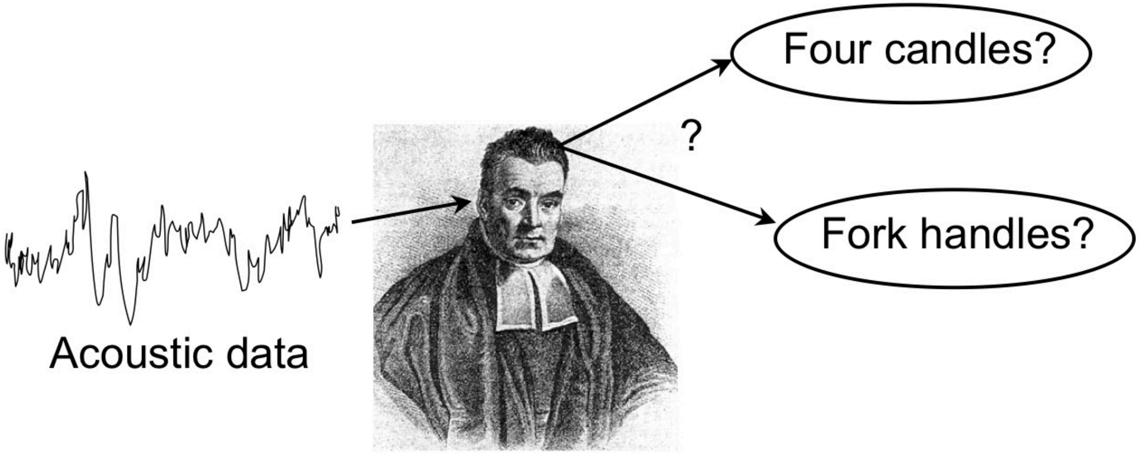
	Frequentist	Bayesian
Binary Decision	Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	Confidence Interval	Posterior Intervals

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

	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

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

Posterior Intervals

- Key notions:
  - **Prior:** Assumptions and beliefs about key parameters of a system Θ.
  - **Likelihood:** How the hidden parameters Θ are connected to the observations Y.
  - **Posterior:** Summary of the inferences about likeliness of Θ.

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

Goal: use the posterior $\boldsymbol{P}(\Theta|Y)$ to to calculate:

$\boldsymbol{P}(\text{Hypothesis}|Y)$ e.g., $\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$

Posterior Intervals

- Key notions:
  - **Prior:** Assumptions and beliefs about key parameters of a system Θ.
  - **Likelihood:** How the hidden parameters Θ are connected to the observations Y.
  - **Posterior:** Summary of the inferences about likeliness of Θ.

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

Goal: use the posterior $\boldsymbol{P}(\Theta|Y)$ to to calculate:

$$\boldsymbol{P}(\text{Hypothesis}|Y) \quad \text{e.g.,} \quad \text{H}_1\text{: } \theta_A - \theta_B > \alpha$$

Posterior Intervals

- Key notions:
  - **Prior:** Assumptions and beliefs about key parameters of a system Θ.
  - **Likelihood:** How the hidden parameters Θ are connected to the observations Y.
  - **Posterior:** Summary of the inferences about likeliness of Θ.

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

Goal: use the posterior $\boldsymbol{P}(\Theta|Y)$ to to calculate:

$$\boldsymbol{P}(\text{Hypothesis}|Y) \quad \text{e.g.,} \quad \mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$$

Posterior Intervals

- Key notions:
  - **Prior:** Assumptions and beliefs about key parameters of a system Θ.
  - **Likelihood:** How the hidden parameters Θ are connected to the observations Y.
  - **Posterior:** Summary of the inferences about likeliness of Θ.

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

Goal: use the posterior $\boldsymbol{P}(\Theta|Y)$ to to calculate:

$\boldsymbol{P}(\mathrm{Hypothesis}|Y)$ e.g., H_1: $\theta_A - \theta_B > \alpha$

Posterior Intervals: Example

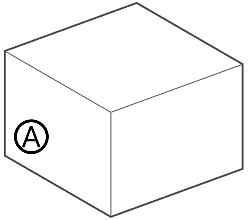
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

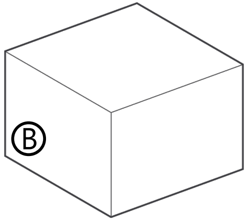
$$\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$$

Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1 : \theta_A - \theta_B > \alpha$$

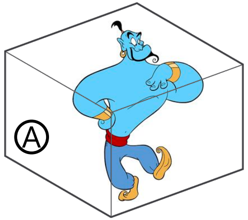


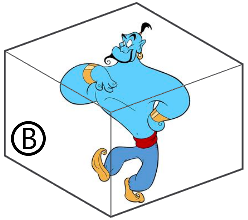


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$$

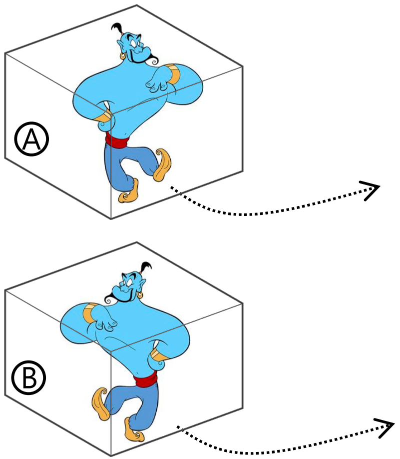




Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

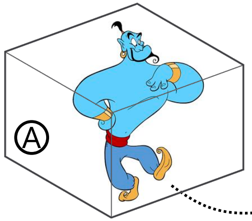
$$\mathrm{H}_1\colon \theta_A - \theta_B > \alpha$$



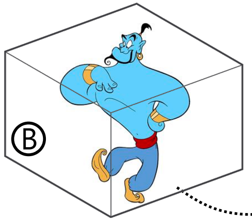
Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1: \theta_A - \theta_B > \alpha$$



$0, 1, 1, 0, \ldots$

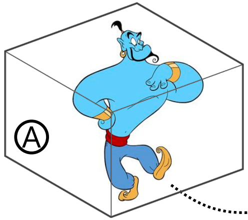


$0, 0, 0, 1, \ldots$

Posterior Intervals: Example

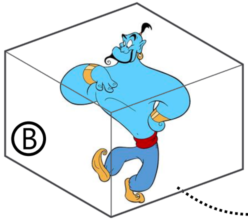
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$$



$$\theta_A = \mathrm{Prob}(Y = 1)$$

$$0, 1, 1, 0, \ldots$$



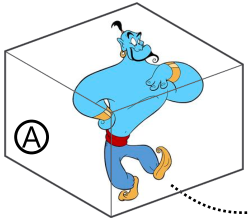
$$\theta_B = \mathrm{Prob}(Y = 1)$$

$$0, 0, 0, 1, \ldots$$

Posterior Intervals: Example

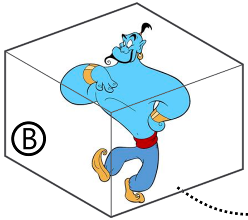
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\colon\ \theta_A - \theta_B > \alpha$$



$$\theta_A = \mathrm{Prob}(Y = 1)$$

$$0, 1, 1, 0, \ldots$$



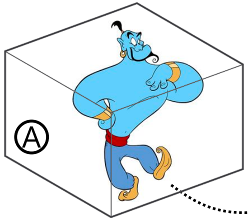
$$\theta_B = \mathrm{Prob}(Y = 1)$$

$$\underbrace{0, 0, 0, 1, \ldots}_{\boldsymbol{P}(Y|\Theta)}$$

Posterior Intervals: Example

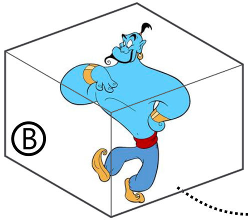
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$$



$$\theta_A = \mathrm{Prob}(Y = 1)$$

$$0, 1, 1, 0, \ldots$$



$$\theta_B = \mathrm{Prob}(Y = 1)$$

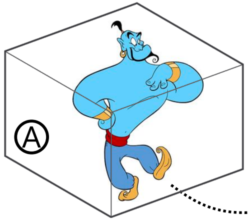
$$0, 0, 0, 1, \ldots$$

$$\boldsymbol{P}(Y|\Theta)$$
$$\oplus$$
$$\boldsymbol{P}(\Theta) \sim \mathrm{uniform}$$

Posterior Intervals: Example

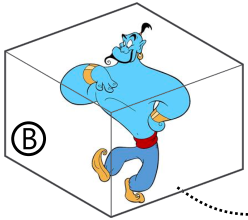
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$$



$$\theta_A = \mathrm{Prob}(Y = 1)$$

$$0, 1, 1, 0, \ldots$$



$$\theta_B = \mathrm{Prob}(Y = 1)$$

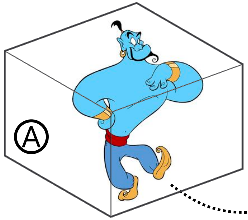
$$0, 0, 0, 1, \ldots$$

$$\boldsymbol{P}(Y|\Theta) \oplus \boldsymbol{P}(\Theta) \sim \text{uniform} \Rightarrow \boldsymbol{P}(\Theta|Y) = \frac{\boldsymbol{P}(Y|\Theta) \times \boldsymbol{P}(\Theta)}{\boldsymbol{P}(Y)}$$

Posterior Intervals: Example

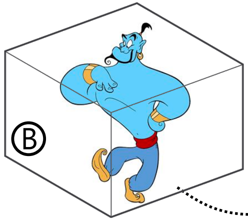
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$$



$\theta_A = \text{Prob}(Y = 1)$

$0, 1, 1, 0, \ldots$



$\theta_B = \text{Prob}(Y = 1)$

$0, 0, 0, 1, \ldots$

$\boldsymbol{P}(Y|\Theta)$

$\oplus$

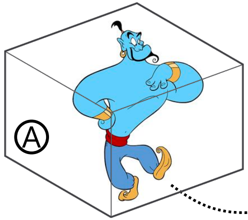
$\boldsymbol{P}(\Theta) \sim \text{uniform}$

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

Posterior Intervals: Example

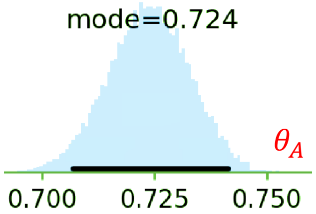
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

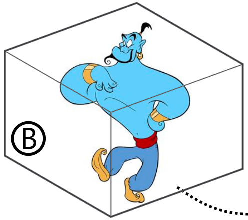
$\mathrm{H}_1\colon\ \theta_A - \theta_B > \alpha$



$\theta_A = \mathrm{Prob}(Y = 1)$

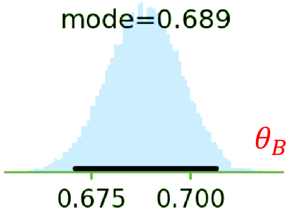
$0, 1, 1, 0, \ldots$





$\theta_B = \mathrm{Prob}(Y = 1)$

$\underbrace{0, 0, 0, 1, \ldots}$



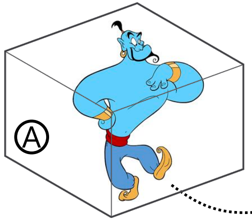
$\boldsymbol{P}(Y|\Theta)$
$\oplus$
$\boldsymbol{P}(\Theta) \sim \mathrm{uniform}$

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

Posterior Intervals: Example

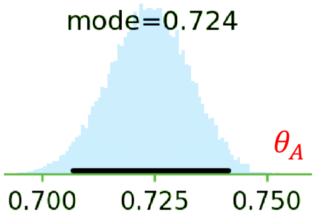
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

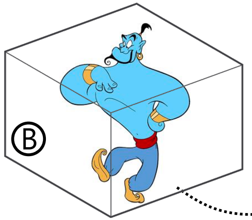
$\mathrm{H}_1\colon\ \theta_A - \theta_B > \alpha$



$\theta_A = \mathrm{Prob}(Y = 1)$

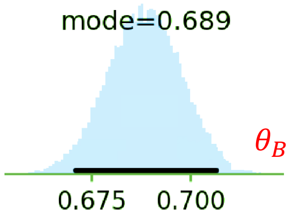
$0, 1, 1, 0, \ldots$





$\theta_B = \mathrm{Prob}(Y = 1)$

$\underbrace{0, 0, 0, 1, \ldots}$



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

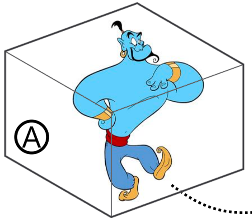
$$\boldsymbol{P}(\Theta|Y) = \frac{\boldsymbol{P}(Y|\Theta) \times \boldsymbol{P}(\Theta)}{\boldsymbol{P}(Y)}$$

$$\boldsymbol{P}(\mathrm{H}_1\colon\ \theta_A - \theta_B > \alpha\ |Y)$$

Posterior Intervals: Example

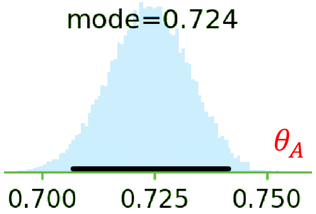
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

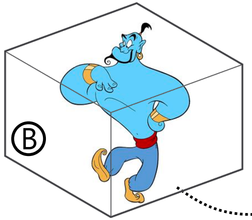
$$\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha$$



$\theta_A = \mathrm{Prob}(Y = 1)$

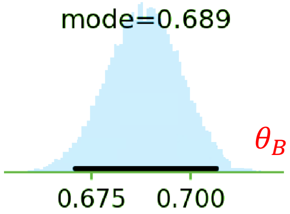
$0, 1, 1, 0, \ldots$

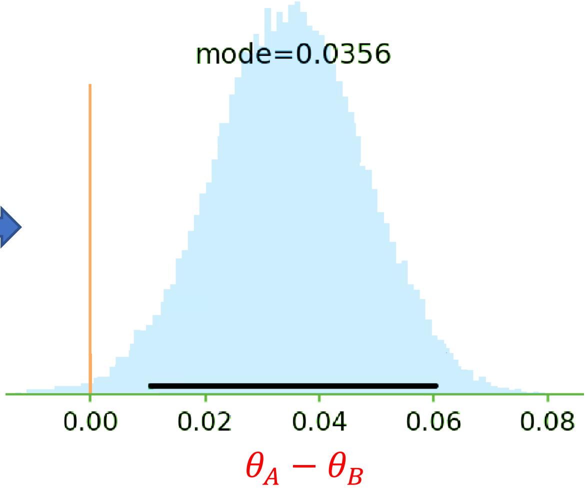




$\theta_B = \mathrm{Prob}(Y = 1)$

$\underbrace{0, 0, 0, 1, \ldots}_{\substack{\boldsymbol{P}(Y|\Theta) \\ \oplus \\ \boldsymbol{P}(\Theta)\sim\text{uniform}}}$





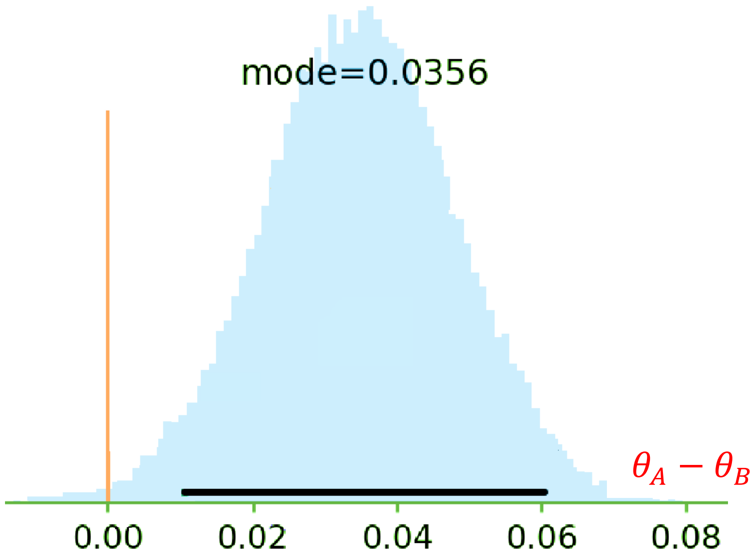
$$\boldsymbol{P}(\Theta|Y) = \frac{\boldsymbol{P}(Y|\Theta) \times \boldsymbol{P}(\Theta)}{\boldsymbol{P}(Y)}$$

$$\boldsymbol{P}(\mathrm{H}_1\text{: } \theta_A - \theta_B > \alpha \,|Y)$$

Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

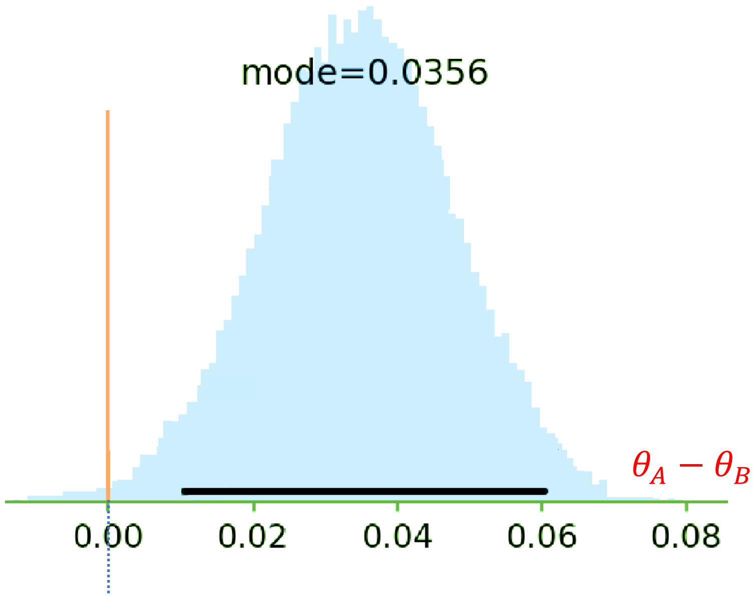
H: $\theta_A - \theta_B > \alpha$



Posterior Intervals: Example

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

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

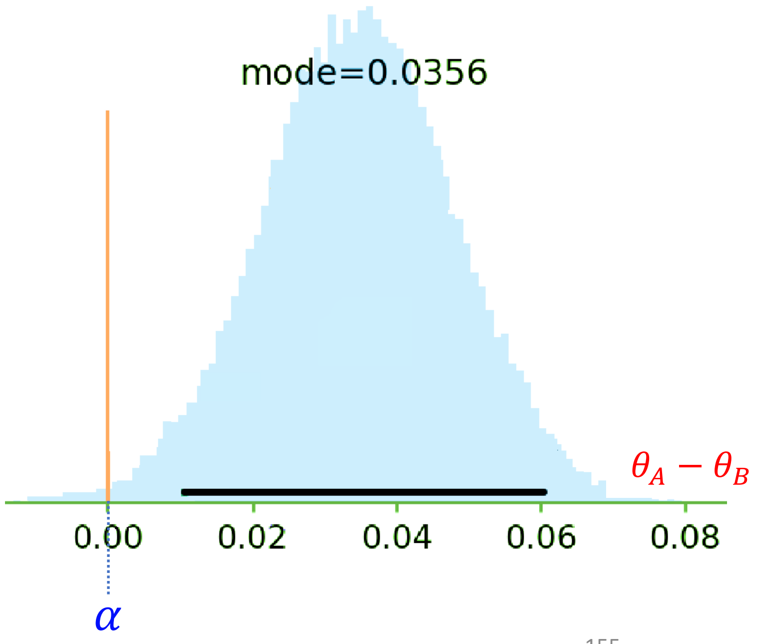


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

- The hypothesis (w/ $\alpha = 0$) holds true …
    - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
    - … with probability %94.

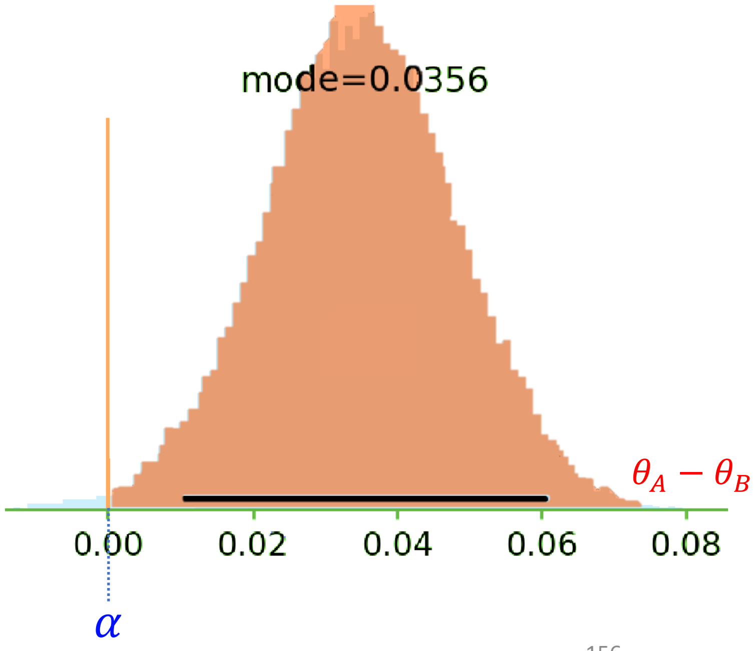


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

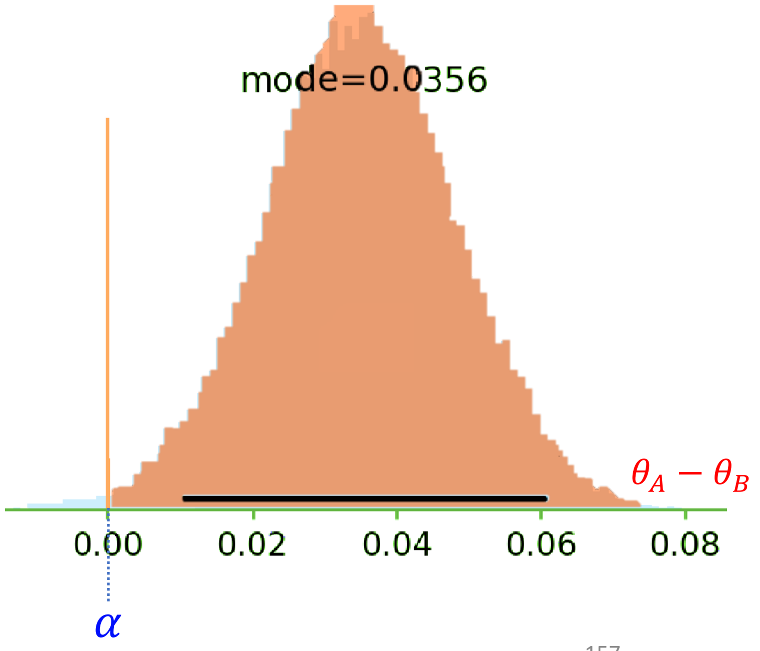
- The hypothesis (w/ $\alpha = 0$) holds true …
  - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.



Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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



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

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

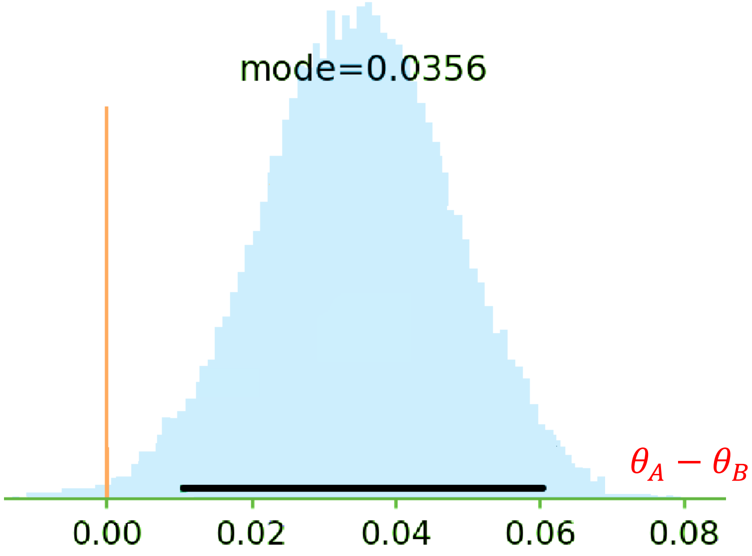
Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

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

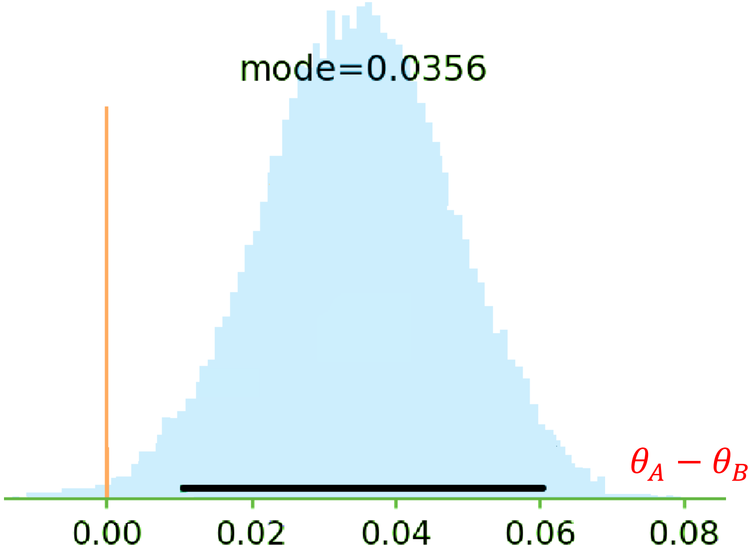
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.



Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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



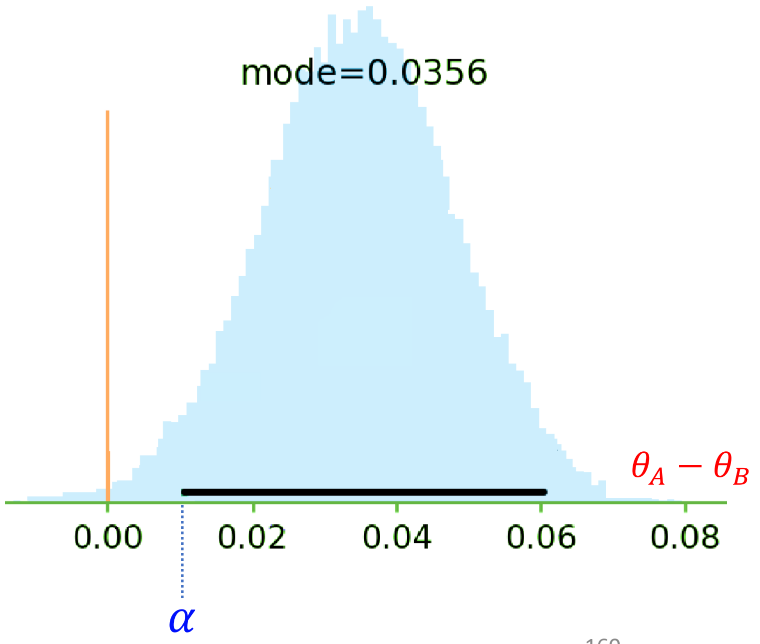
- The hypothesis (w/ $\alpha = 0$) holds true …
  - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.

Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

- The hypothesis (w/ $\alpha = 0$) holds true …
  - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.

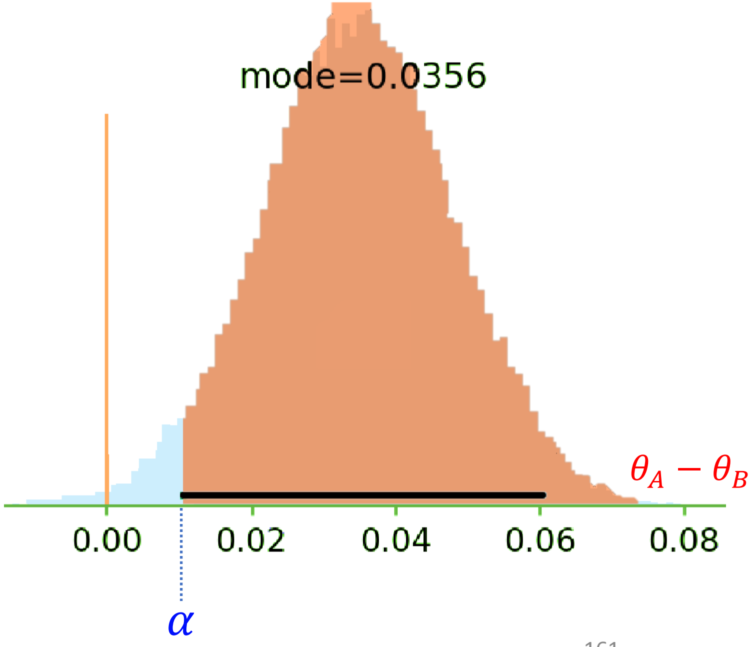


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

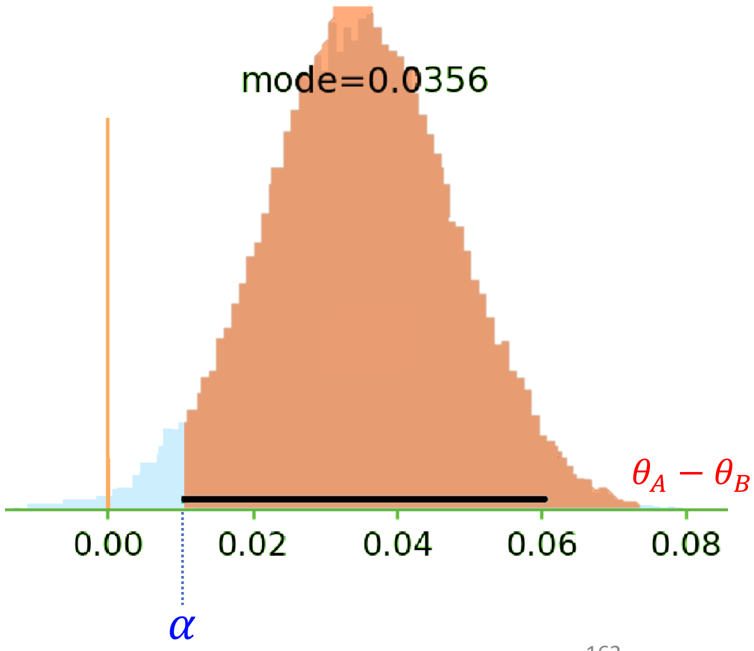
- The hypothesis (w/ $\alpha = 0$) holds true …
  - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.



Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$\mathrm{H}\text{:}\ \theta_A - \theta_B > \alpha$



- The hypothesis (w/ $\alpha = 0$) holds true …
  - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.

2^{nd} Intermediate Summary

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

2^{nd} Intermediate Summary

- Provides **probability estimates over** hypothesis of interest.
  - **Easier** to **interpret** → less ambiguous.

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

2nd Intermediate Summary

- Provides **probability estimates over** hypothesis of interest.
  - **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.

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

2nd Intermediate Summary

- Provides **probability estimates over** hypothesis of interest.
  - **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.

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

Measures of [Un]Certainty

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

(Goodman, 2008; Wasserstein et al., 2016)

Measures of [Un]Certainty

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

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

(Goodman, 2008; Wasserstein et al., 2016)

Measures of [Un]Certainty

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

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

(Goodman, 2008; Wasserstein et al., 2016)

Measures of [Un]Certainty

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

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

(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**.

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

(Goodman, 2008; Wasserstein et al., 2016)

Trends and Patterns in the field

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

How many papers did use significance testing?

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

Trends and Patterns in the field

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

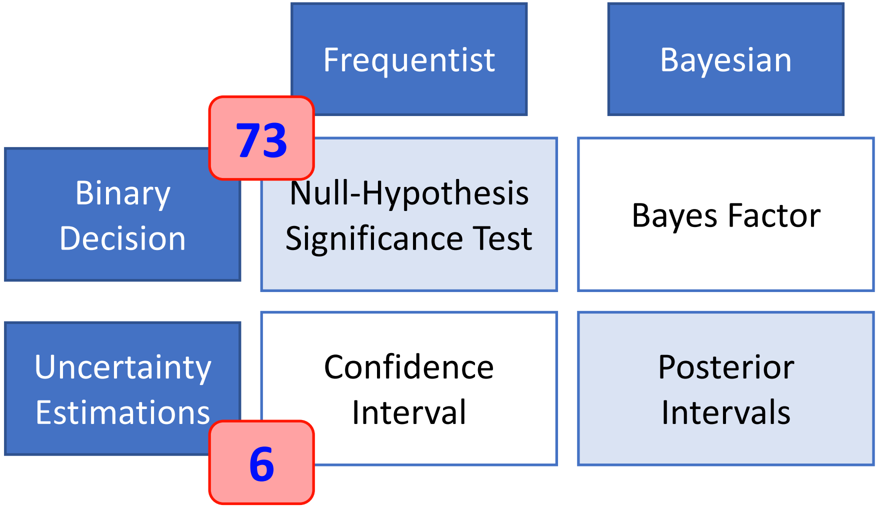
How many papers did use significance testing?

	Frequentist	Bayesian
Binary Decision	73 Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	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?



Trends and Patterns in the field

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

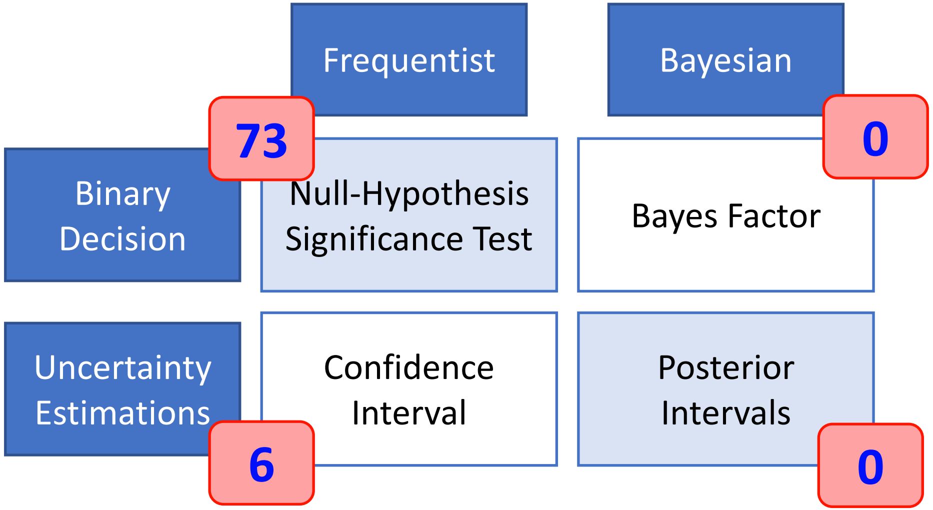
How many papers did use significance testing?

	Frequentist	Bayesian
Binary Decision	**73** Null-Hypothesis Significance Test	Bayes Factor **0**
Uncertainty Estimations	**6** 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?



Trends and Patterns in the field

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

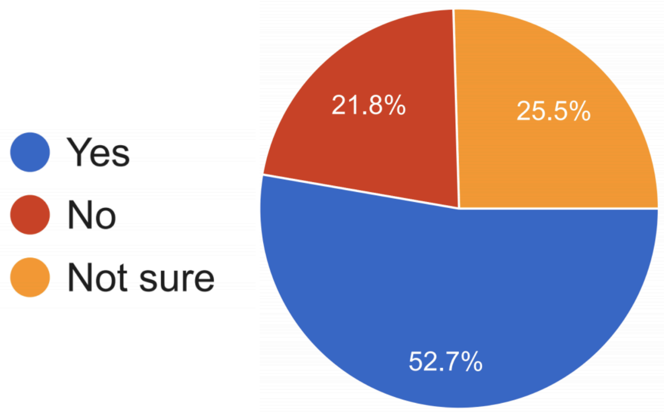
How many papers did use significance testing?

Why?

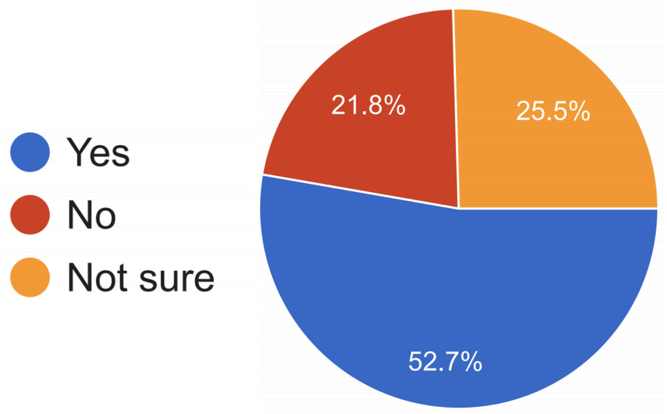
	Frequentist	Bayesian
Binary Decision	Null-Hypothesis Significance Test **73**	Bayes Factor **0**
Uncertainty Estimations	Confidence Interval **6**	Posterior Intervals **0**

Have you heard about "Bayesian Hypothesis Testing"?

Have you heard about "Bayesian Hypothesis Testing"?

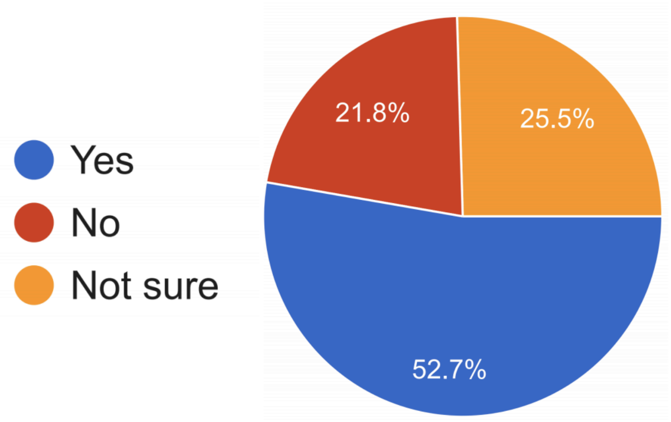


Have you heard about "Bayesian Hypothesis Testing"?

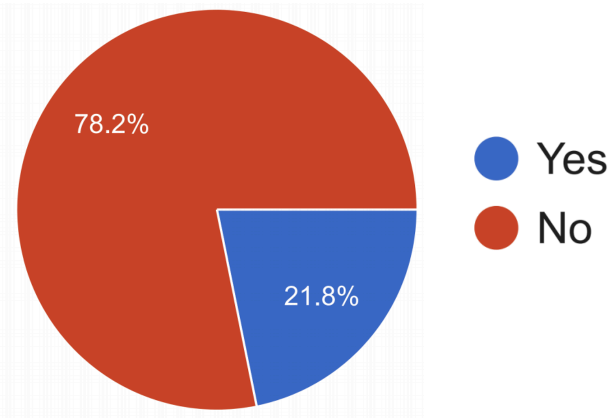


Do you know the definition of "Bayes Factor"?

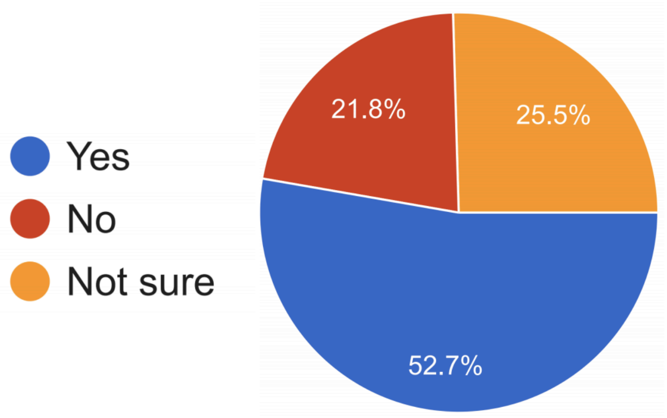
Have you heard about "Bayesian Hypothesis Testing"?



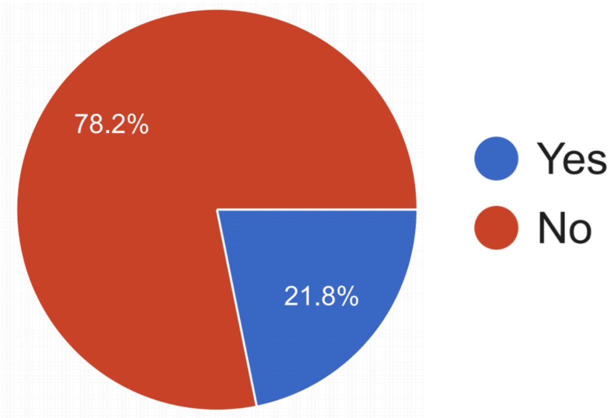
Do you know the definition of "Bayes Factor"?



Have you heard about "Bayesian Hypothesis Testing"?



Do you know the definition of "Bayes Factor"?



- Many people did not know the definition of “Bayes Factor” and some only had “heard” about them. 🤔

Final Section:

Malpractices & Suggestions

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

Ambiguous reporting

The White House <info@mail.whitehouse.g… May 5, 2020, 8:40 AM (4 days ago)

to me

WEST WING

READS

China's Coronavirus Lies Pile Up

“A Department of Homeland Security analysis has concluded that China hid the early spread of the coronavirus so it could hoard medical equipment, keeping it from other countries that would have bought it if they had known of the danger that was coming their way from Wuhan,” the *Washington Examiner* editorial board writes.

“Specifically, DHS found, with 95% statistical confidence, that changes to China's personal protective equipment import and export behavior were highly abnormal and not random.”

[Click here to read more.]

Ambiguous reporting

When referring to the results of significance testing, one should be mindful of **how others are going to interpret it**.

The White House <info@mail.whitehouse.g… May 5, 2020, 8:40 AM (4 days ago)
to me

WEST WING
READS

China's Coronavirus Lies Pile Up

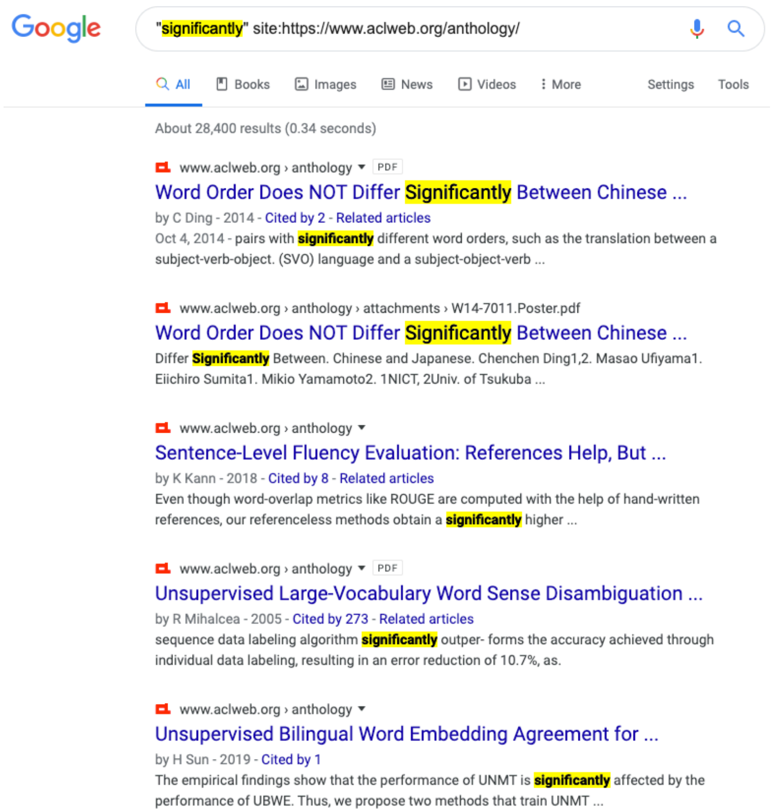
"A Department of Homeland Security analysis has concluded that China hid the early spread of the coronavirus so it could hoard medical equipment, keeping it from other countries that would have bought it if they had known of the danger that was coming their way from Wuhan," the *Washington Examiner* editorial board writes.

"Specifically, DHS found, with 95% statistical confidence, that changes to China's personal protective equipment import and export behavior were highly abnormal and not random."

Click here to read more.

Ambiguity problem in interpreting “significance”

Ambiguity problem in interpreting "significance"



"significantly" site:https://www.aclweb.org/anthology/

All | Books | Images | News | Videos | More | Settings | Tools

About 28,400 results (0.34 seconds)

www.aclweb.org › anthology PDF
Word Order Does NOT Differ Significantly Between Chinese ...
by C Ding - 2014 - Cited by 2 - Related articles
Oct 4, 2014 - pairs with significantly different word orders, such as the translation between a subject-verb-object. (SVO) language and a subject-object-verb ...

www.aclweb.org › anthology › attachments › W14-7011.Poster.pdf
Word Order Does NOT Differ Significantly Between Chinese ...
Differ Significantly Between. Chinese and Japanese. Chenchen Ding1,2. Masao Ufiyama1. Eiichiro Sumita1. Mikio Yamamoto2. 1NICT, 2Univ. of Tsukuba ...

www.aclweb.org › anthology
Sentence-Level Fluency Evaluation: References Help, But ...
by K Kann - 2018 - Cited by 8 - Related articles
Even though word-overlap metrics like ROUGE are computed with the help of hand-written references, our referenceless methods obtain a significantly higher ...

www.aclweb.org › anthology PDF
Unsupervised Large-Vocabulary Word Sense Disambiguation ...
by R Mihalcea - 2005 - Cited by 273 - Related articles
sequence data labeling algorithm significantly outper- forms the accuracy achieved through individual data labeling, resulting in an error reduction of 10.7%, as.

www.aclweb.org › anthology
Unsupervised Bilingual Word Embedding Agreement for ...
by H Sun - 2019 - Cited by 1
The empirical findings show that the performance of UNMT is significantly affected by the performance of UBWE. Thus, we propose two methods that train UNMT ...

Ambiguity problem in interpreting “significance”

Google

"significantly" site:https://www.aclweb.org/anthology/

All | Books | Images | News | Videos | More | Settings | Tools

About 28,400 results (0.34 seconds)

www.aclweb.org › anthology PDF
Word Order Does NOT Differ Significantly Between Chinese ...
by C Ding - 2014 - Cited by 2 - Related articles
Oct 4, 2014 - pairs with significantly different word orders, such as the translation between a subject-verb-object. (SVO) language and a subject-object-verb ...

www.aclweb.org › anthology › attachments › W14-7011.Poster.pdf
Word Order Does NOT Differ Significantly Between Chinese ...
Differ Significantly Between. Chinese and Japanese. Chenchen Ding1,2. Masao Ufiyama1. Eiichiro Sumita1. Mikio Yamamoto2. 1NICT, 2Univ. of Tsukuba ...

www.aclweb.org › anthology
Sentence-Level Fluency Evaluation: References Help, But ...
by K Kann - 2018 - Cited by 8 - Related articles
Even though word-overlap metrics like ROUGE are computed with the help of hand-written references, our referenceless methods obtain a significantly higher ...

www.aclweb.org › anthology PDF
Unsupervised Large-Vocabulary Word Sense Disambiguation ...
by R Mihalcea - 2005 - Cited by 273 - Related articles
sequence data labeling algorithm significantly outper- forms the accuracy achieved through individual data labeling, resulting in an error reduction of 10.7%, as.

www.aclweb.org › anthology
Unsupervised Bilingual Word Embedding Agreement for ...
by H Sun - 2019 - Cited by 1
The empirical findings show that the performance of UNMT is significantly affected by the performance of UBWE. Thus, we propose two methods that train UNMT ...

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"

"significantly" site:https://www.aclweb.org/anthology/

All | Books | Images | News | Videos | More | Settings | Tools

About 28,400 results (0.34 seconds)

www.aclweb.org › anthology PDF
Word Order Does NOT Differ Significantly Between Chinese ...
by C Ding - 2014 - Cited by 2 - Related articles
Oct 4, 2014 - pairs with significantly different word orders, such as the translation between a subject-verb-object. (SVO) language and a subject-object-verb ...

www.aclweb.org › anthology › attachments › W14-7011.Poster.pdf
Word Order Does NOT Differ Significantly Between Chinese ...
Differ Significantly Between. Chinese and Japanese. Chenchen Ding1,2. Masao Ufiyama1. Eiichiro Sumita1. Mikio Yamamoto2. 1NICT, 2Univ. of Tsukuba ...

www.aclweb.org › anthology
Sentence-Level Fluency Evaluation: References Help, But ...
by K Kann - 2018 - Cited by 8 - Related articles
Even though word-overlap metrics like ROUGE are computed with the help of hand-written references, our referenceless methods obtain a significantly higher ...

www.aclweb.org › anthology PDF
Unsupervised Large-Vocabulary Word Sense Disambiguation ...
by R Mihalcea - 2005 - Cited by 273 - Related articles
sequence data labeling algorithm significantly outper- forms the accuracy achieved through individual data labeling, resulting in an error reduction of 10.7%, as.

www.aclweb.org › anthology
Unsupervised Bilingual Word Embedding Agreement for ...
by H Sun - 2019 - Cited by 1
The empirical findings show that the performance of UNMT is significantly affected by the performance of UBWE. Thus, we propose two methods that train UNMT ...

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-A and it compares it with a baseline system-B. In its “abstract” it writes: “... system-A significantly improves over system-B.” 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-A has a higher performance than system-B with a notable margin in the dataset.

Ambiguity problem in interpreting “significance”

- *An NLP paper presents system-A and it compares it with a baseline system-B. In its “abstract” it writes: “... system-A significantly improves over system-B.” 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-A has a higher performance than system-B with a notable margin in the dataset.

Ambiguity problem in interpreting “significance”

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

83%

- 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-A has a higher performance than system-B with a notable margin in the dataset.

Ambiguity problem in interpreting "significance"

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

83%

- It is expected that authors have performed some type of "hypothesis testing."

53%

- It is expected that the authors have reported the performances of two systems on a dataset where system-A has a higher performance than system-B with a notable margin in the dataset.

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.”

Tips and Suggestions

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

Define the research hypothesis you are after:

- H1: Ⓐ 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.
- H2: Ⓐ and Ⓑ are **inherently different**, since with **probability** at least 95%, the inherent accuracy of Ⓐ **exceeds** that of Ⓑ by at least α%.
- …

Tips and Suggestions

- 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

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

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 funda

The extended reach of NLP algorithms has also resulted in NLP papers giving much more emphasis to the experiment and result sections by showing comparisons between multiple algorithms on various datasets from different languages and domains. This emphasis on empirical results highlights the role of statistical significance testing in NLP research: if we rely on empirical evaluation to validate our hypotheses and reveal the cor

Lots of good tips about:

- Selecting the right "test"
- How to report your results.

Tips and Suggestions

- **If using Bayesian tests:** `https://github.com/allenai/HyBayes/`

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

Erfan Sadeqi Azer[1] **Daniel Khashabi**[2*] **Ashish Sabharwal**[2] **Dan Roth**[3]
[1]Indiana University [2]Allen Institute for Artificial Intelligence [3]University of Pennsylvania
esadeqia@indiana.edu {danielk,ashishs}@allenai.org danroth@cis.upenn.edu

That’s it!

The Need for Assumptions

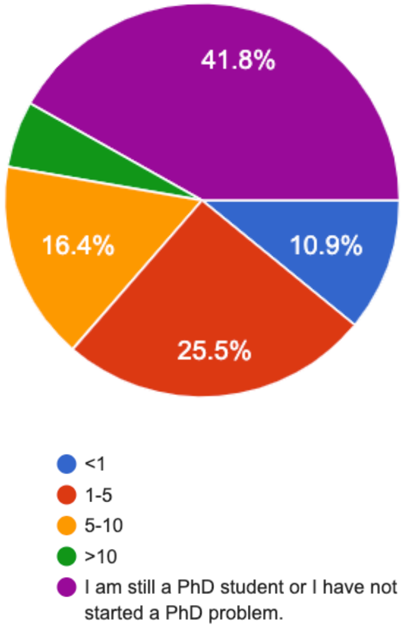
	Frequentist	Bayesian
Binary Decision	Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	Confidence Interval	Posterior Intervals

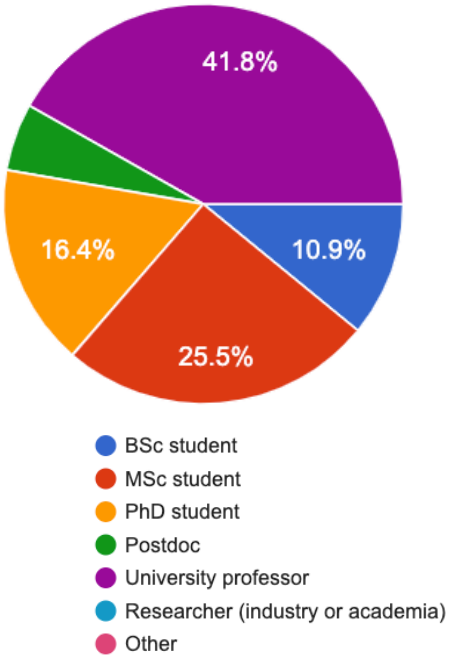
The Need for Assumptions

- *Which tests have assumptions?*
- Assumptions are necessary to perform any statistical tests.
  - “no free lunch”
- Many of them are questionable!

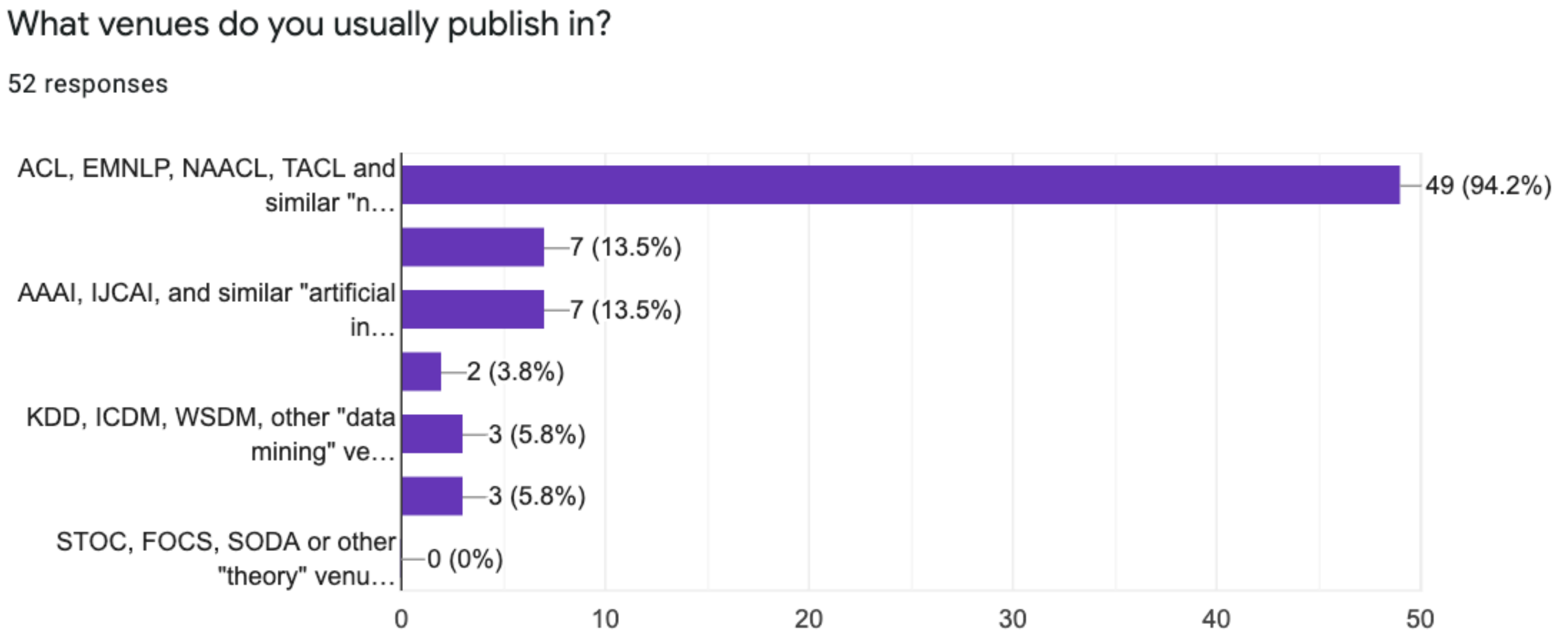
	Frequentist	Bayesian
Binary Decision	Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	Confidence Interval	Posterior Intervals

Participants in our Survey





Participants in our Survey

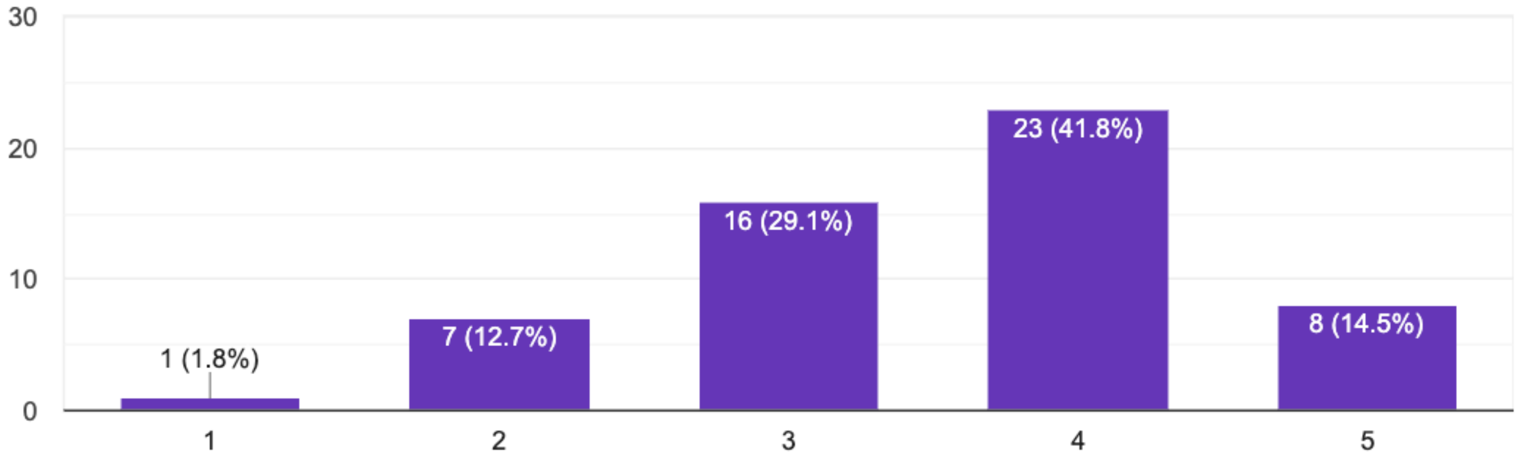


Participants in Our Survey

- *“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.”*



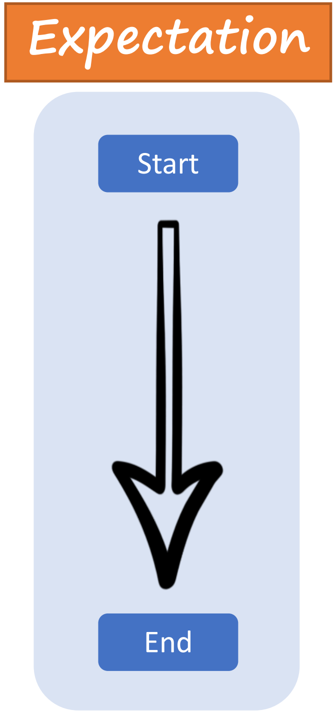
Unintended Misleading Result by Iterative Testing

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.

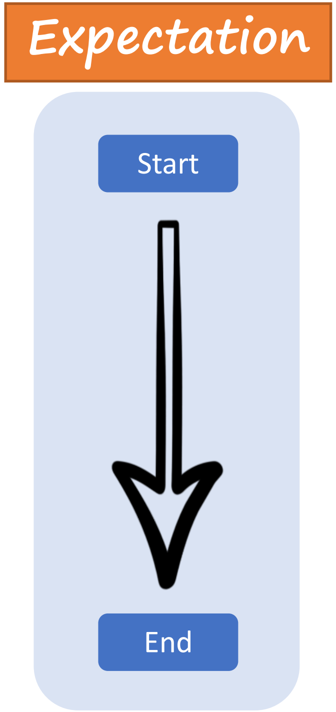


Unintended Misleading Result by Iterative Testing

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

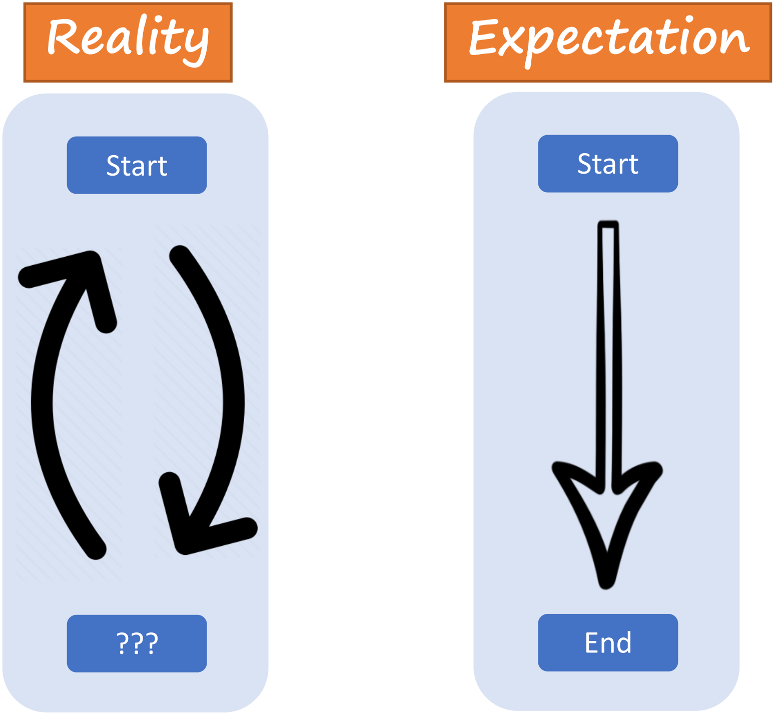
- In practice researchers perform **multiple** rounds of experiments.

- 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.