Allen Institute for AI

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

arxiv.org/abs/1911.03850



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

The Cycle of Empirical Research

The Cycle of Empirical Research

Empirical Observations

The Cycle of Empirical Research

????
Research
Problem

Empirical Observations

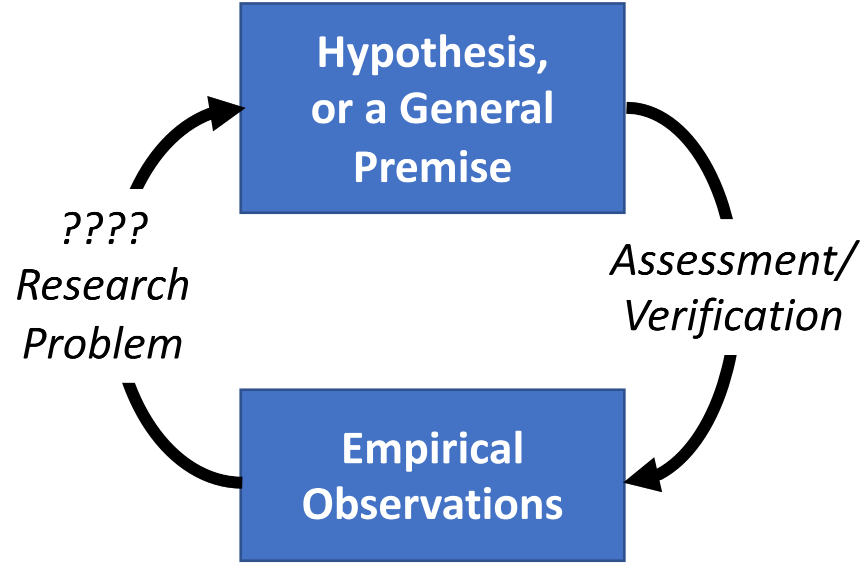
The Cycle of Empirical Research

Hypothesis, or a General Premise

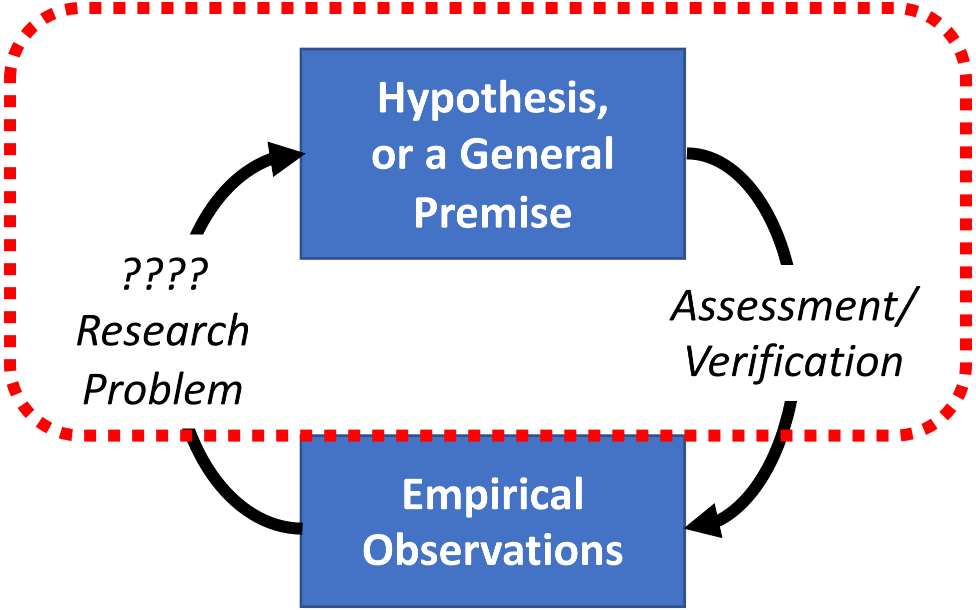
????
Research Problem

Empirical Observations

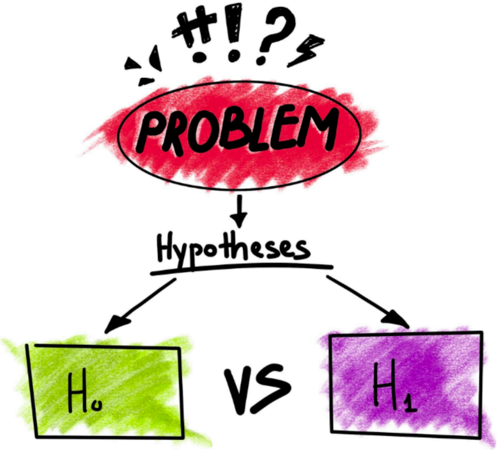
The Cycle of Empirical Research



The Cycle of Empirical Research



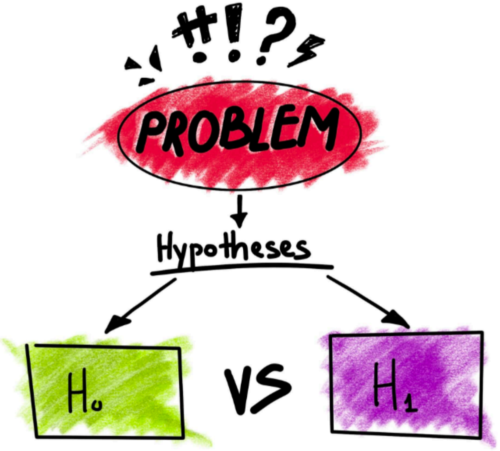
Hypotheses



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

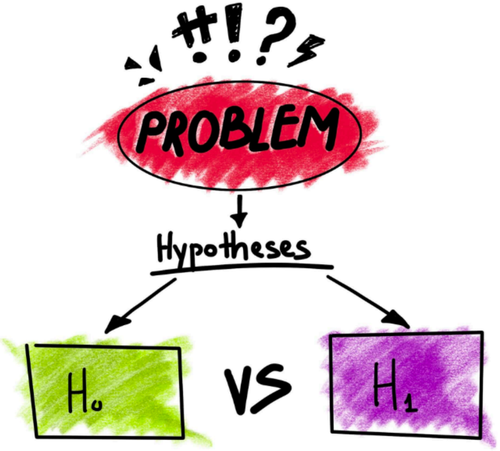
- Terminology:
  - A hypothesis is **never “proved”**
  - But it could be **“supported”** by the evidence



Hypotheses

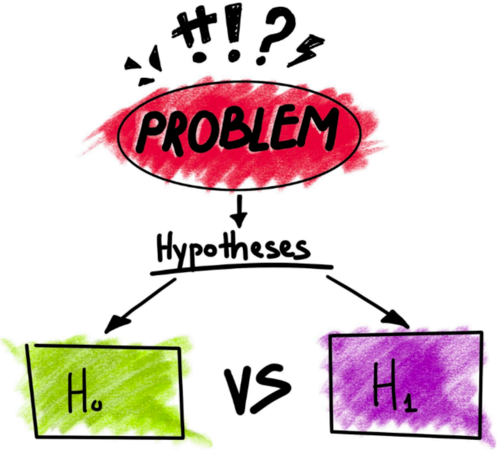
- A **prediction** about how the world will behave **if our idea is correct**
- Worded as an **if-then** statement
- A hypothesis is a **testable** prediction
- A hypothesis is a **falsifiable** statement

- 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

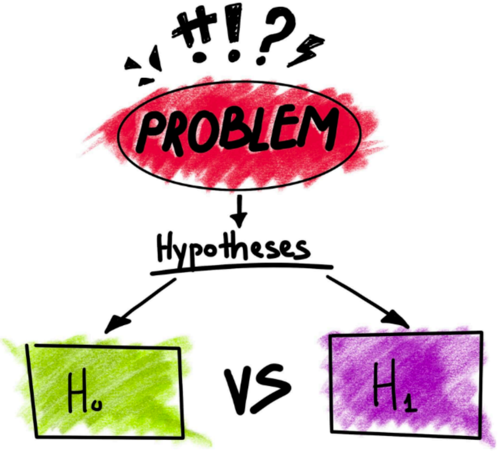


- Terminology:
  - A hypothesis is **never "proved"**
  - But it could be **"supported"** by the evidence

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

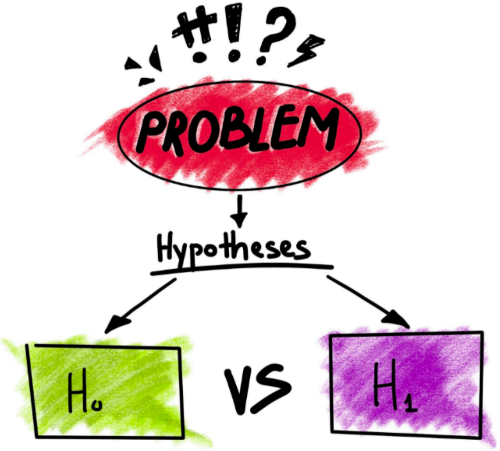
- Terminology:
  - A hypothesis is **never “proved”**
  - But it could be **“supported”** by the evidence

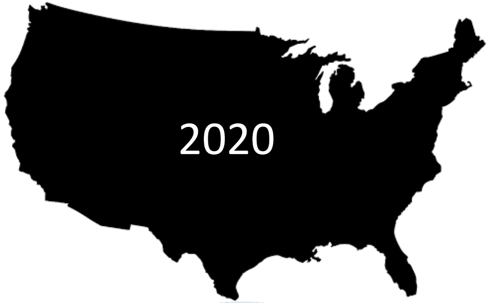


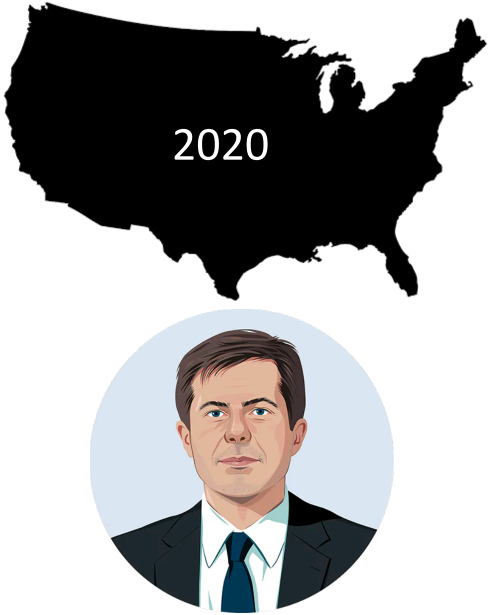
Hypotheses

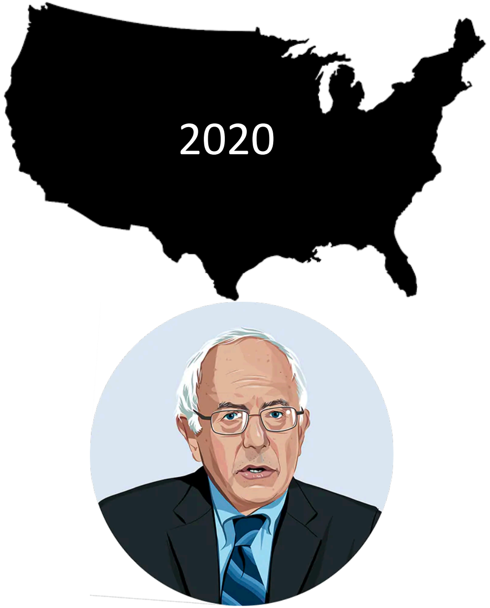
- A **prediction** about how the world will behave **if our idea is correct**
- Worded as an **if-then** statement
- A hypothesis is a **testable** prediction
- A hypothesis is a **falsifiable** statement

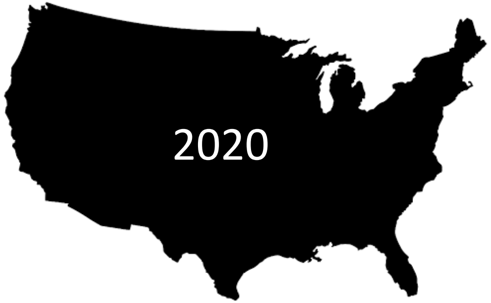
- Terminology:
  - A hypothesis is **never "proved"**
  - But it could be **"supported"** by the evidence




2020

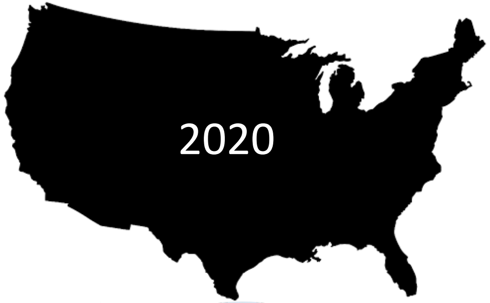

2020


2020

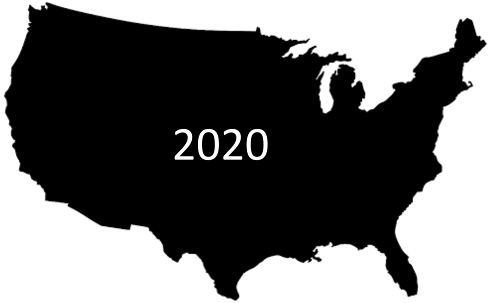

2020

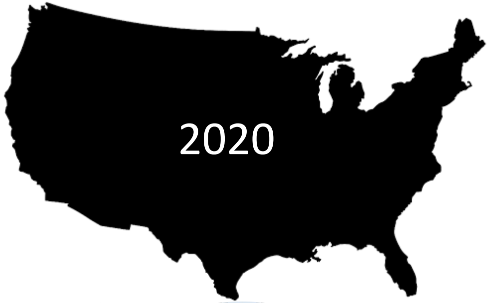
<

?


2020

Not a good statistical hypothesis



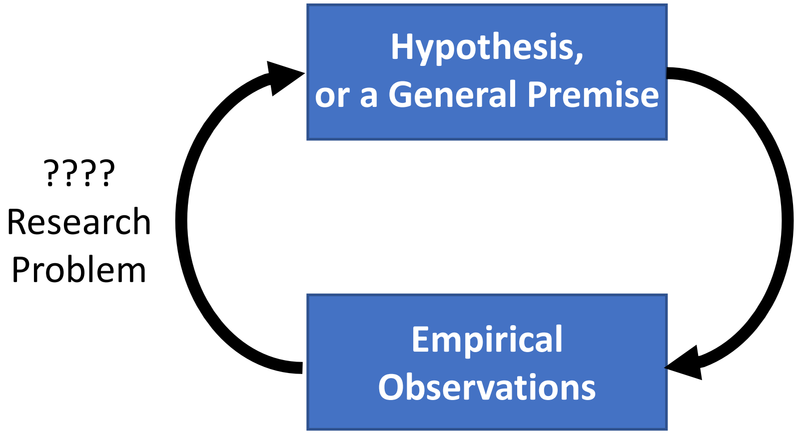


<

?

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

A Typical AI Experiment

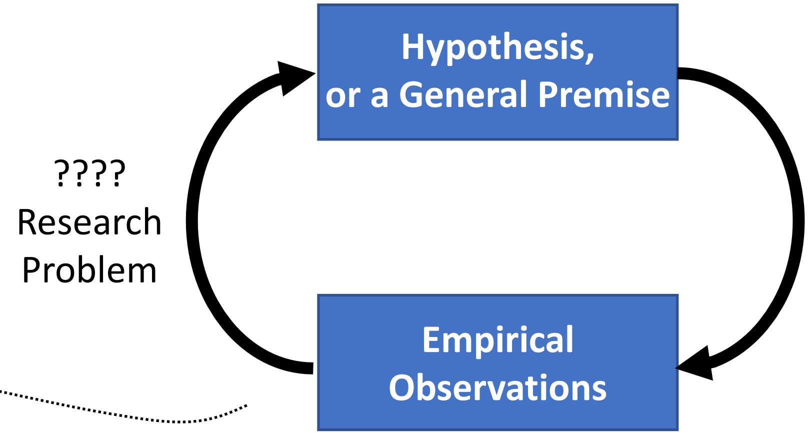


A Typical AI Experiment

(Clark et al., 2018)

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

|D|= 2376



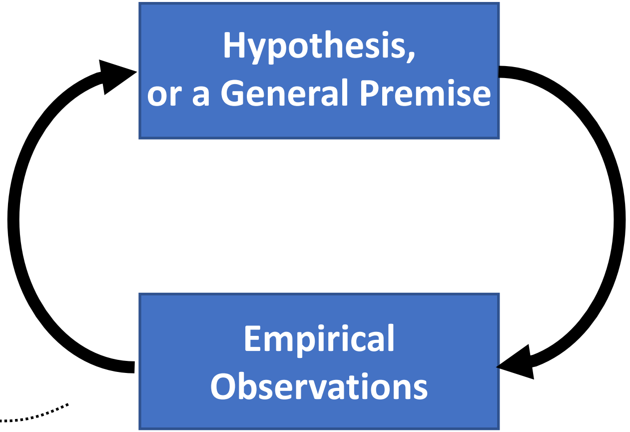
A Typical AI Experiment

(Clark et al., 2018)

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

|D|= 2376

????
Research
Problem



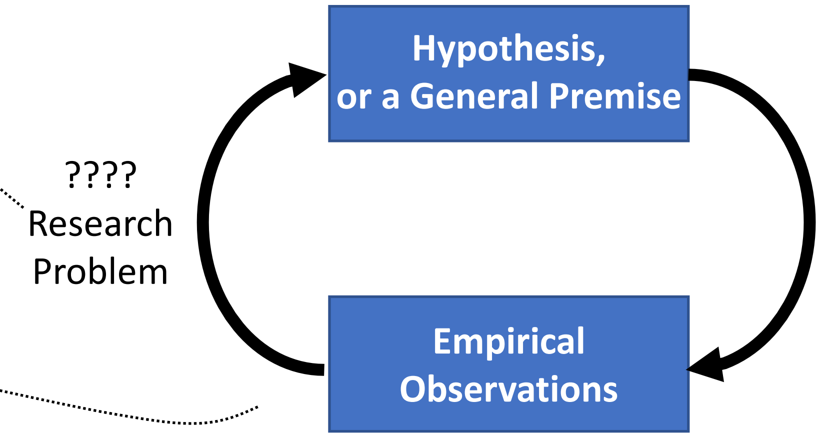
A Typical AI Experiment

- 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



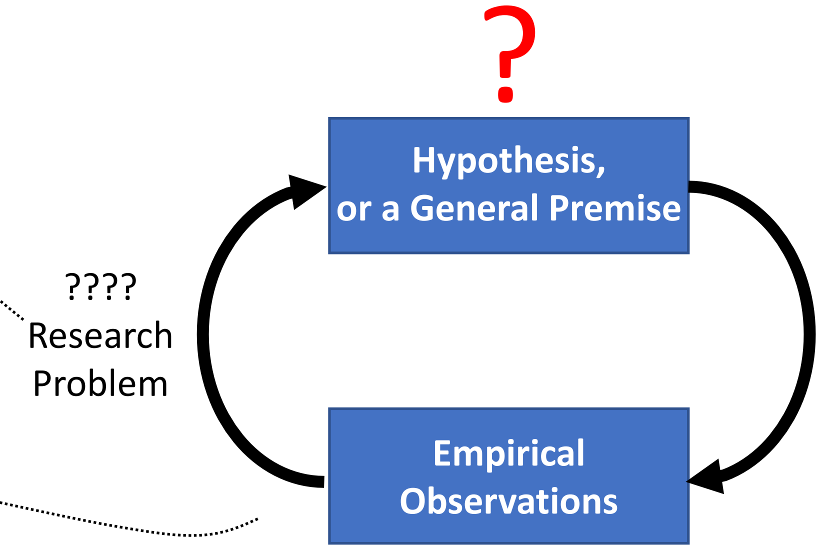
A Typical AI Experiment

- 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

- C1*: Ⓐ and Ⓑ are **inherently different,**

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

- C1*: Ⓐ 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

- C1*: Ⓐ 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

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

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

Spoiler Alert:

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

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

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

* Under some statistical assumptions about sampling of the observations.

A Typical AI Experiment: Example Hypotheses

Spoiler Alert:

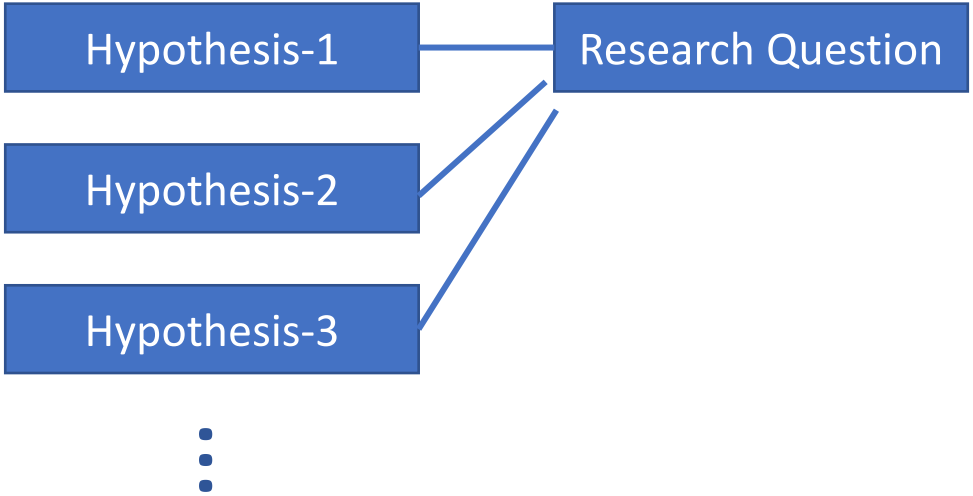
*Almost everyone uses **C1**, even though it is harder to interpret.*

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

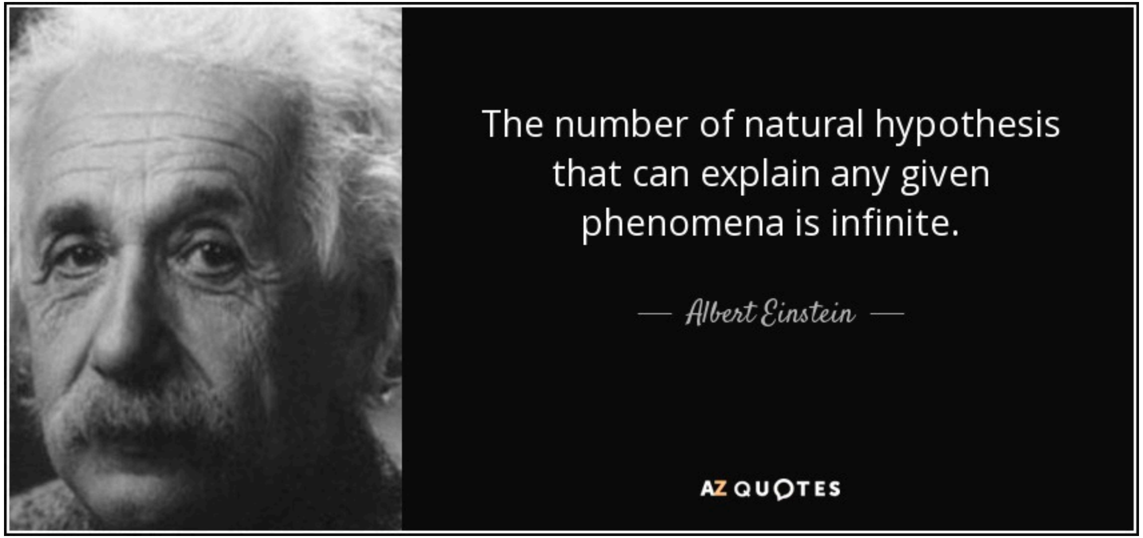
- C1*: Ⓐ and Ⓑ are **inherently different**, in the sense that **if** they were inherently **identical**, it would be highly **unlikely** to witness the observed 3.5% empirical gap.
- C2*: Ⓐ and Ⓑ are **inherently different**, since with **probability** at least 95%, the inherent accuracy of Ⓐ **exceeds**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

* Under some statistical assumptions about sampling of the observations.

And many more . . .



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

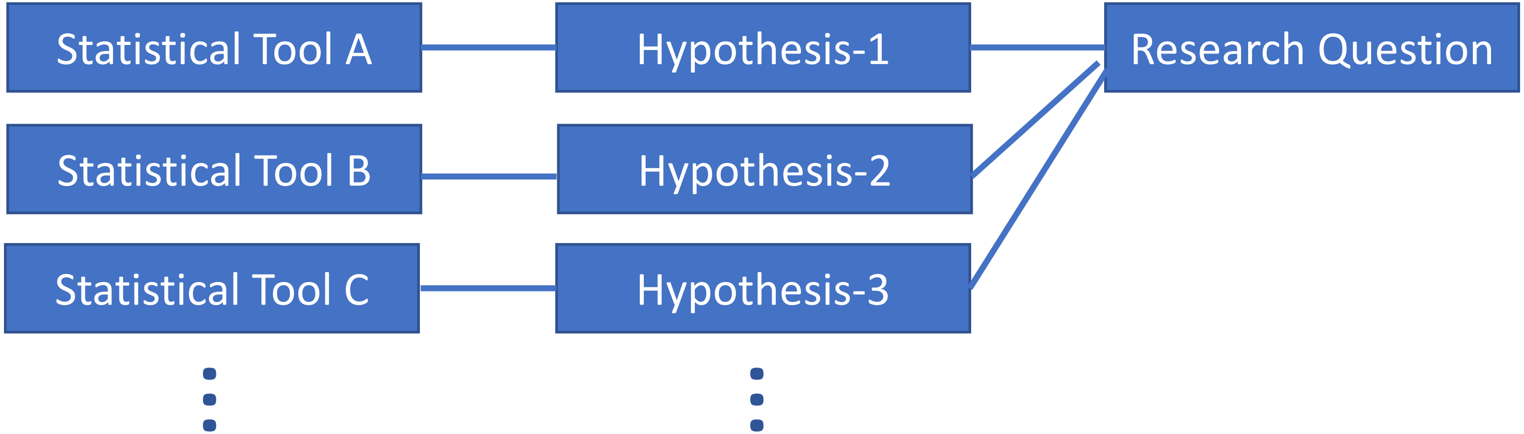

The number of natural hypothesis
that can explain any given
phenomena is infinite.
— Albert Einstein —
AZ QUOTES

Hypothesis vs Statistical Techniques

Research Question

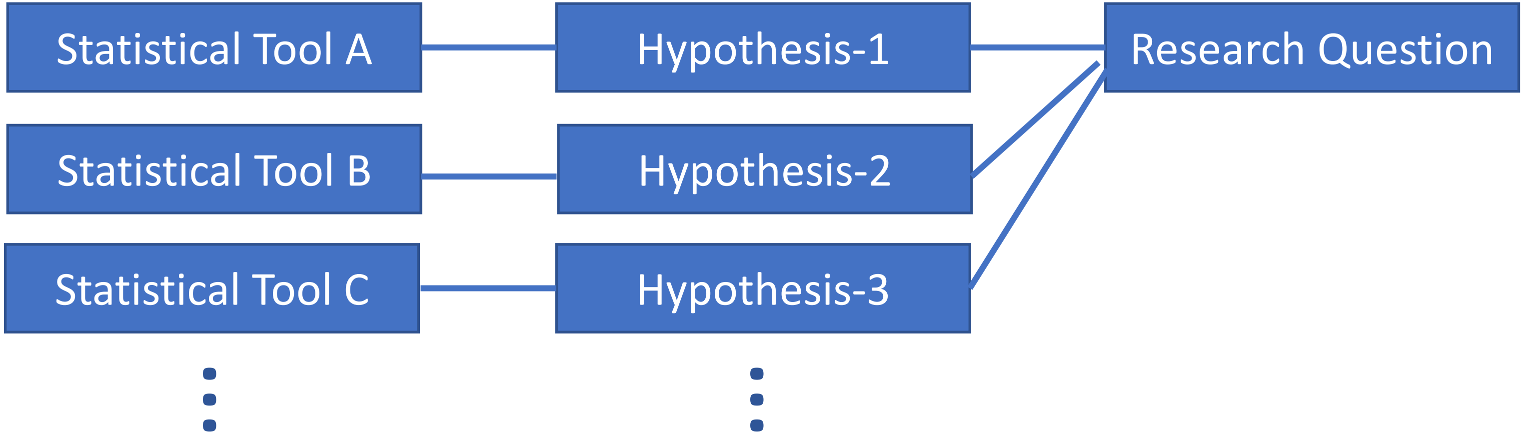
- Observation 2: Each hypothesis ought to be assessed with an **appropriate** statistical tool.

Hypothesis vs Statistical Techniques



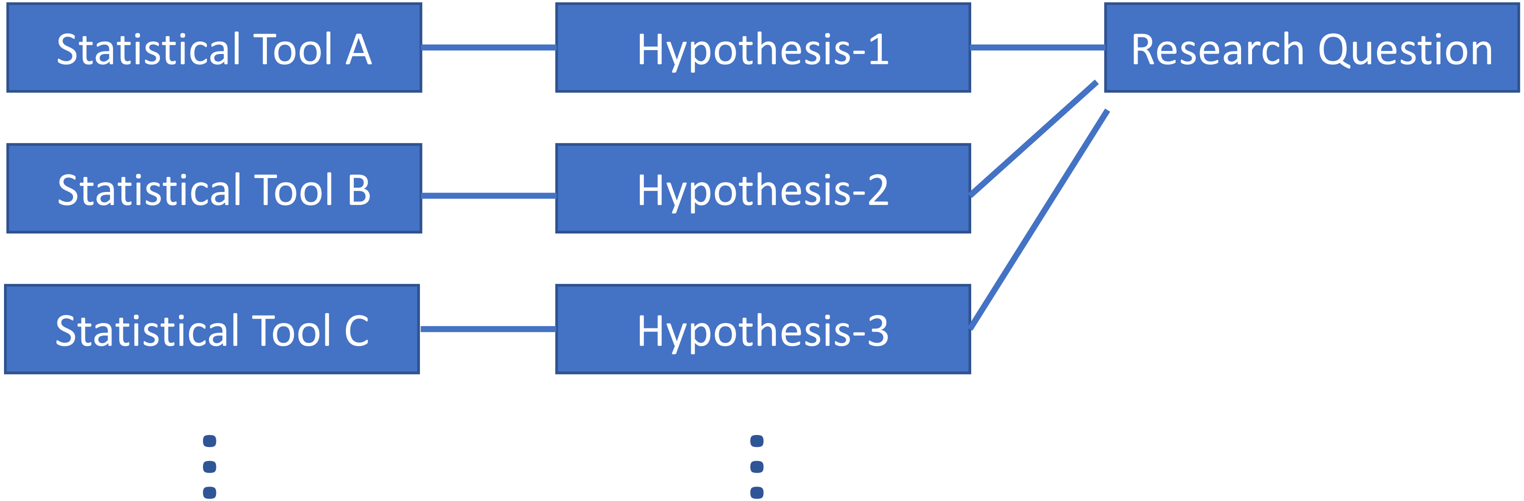
- Observation 2: Each hypothesis ought to be assessed with an **appropriate** statistical tool.

Hypothesis vs Statistical Techniques



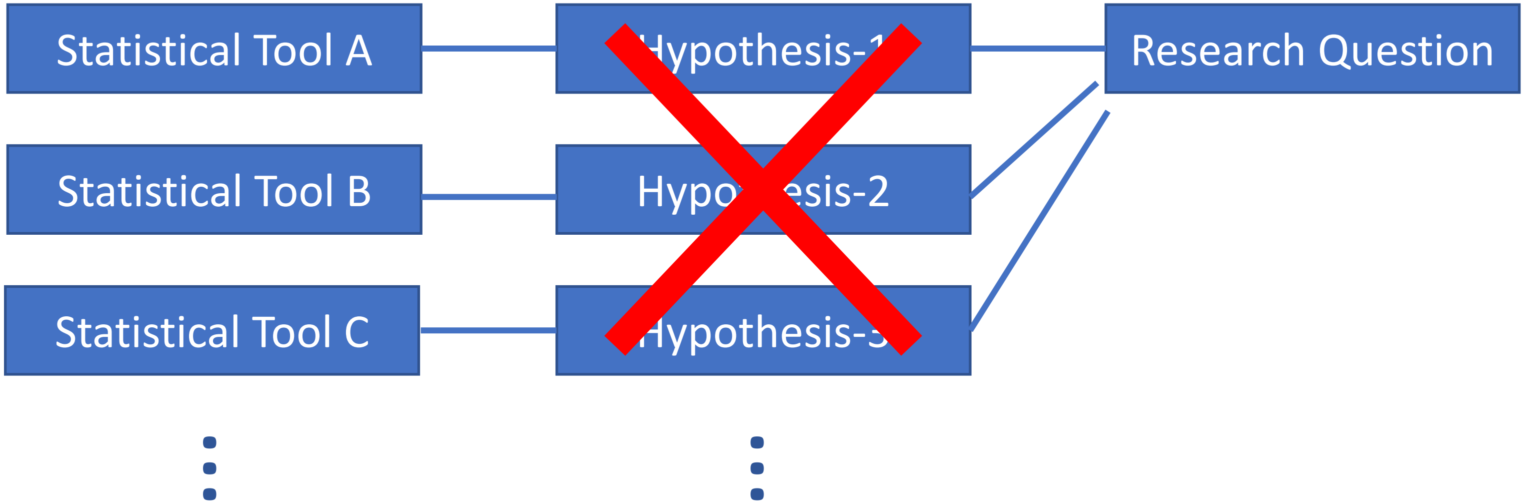
- Observation 2: Each hypothesis ought to be assessed with an **appropriate** statistical tool.
- Corollary: Researchers should **start with a hypothesis** that best serves their goal, followed by an appropriate selection of a statistical approach.

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

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

S1 >> S2

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

S1 >> S2

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

Statistical Tool — Hypothesis — ? — Research Question

S1 >> S2

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

Statistical Tool — Hypothesis — ? — Research Question

S1 >> S2

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

Talk Summary & Statement

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

- Research works should be **explicit** about:
  - (a) Their choice of **hypothesis** and,
  - (b) How selected **statistical tool** addresses this hypothesis.

Statistical tools in this work . . .

	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)

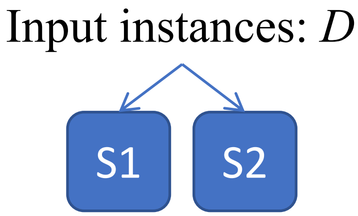
	Frequentist	Bayesian
Binary/Categorical Decisions	Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	Confidence Interval	Posterior Intervals

Notation

Notation

- Compare two systems on a set of instances: D
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$
- Several hypotheses:
  - H1: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …

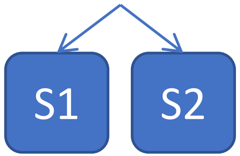
Notation



- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …

Notation

Input instances: D



- Compare two systems on a set of instances: D

- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$

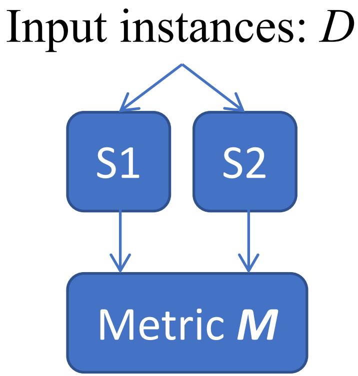
- Several hypotheses:
  - H1: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …

Notation

- Compare two systems on a set of instances: D

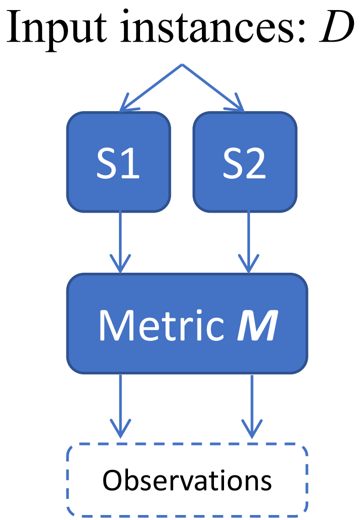
- A measure of performance: $\boldsymbol{M}(\mathrm{S}_i, D)$
  - $\theta_i \neq \boldsymbol{M}(\mathrm{S}_i, D)$

- Several hypotheses:
  - H1: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



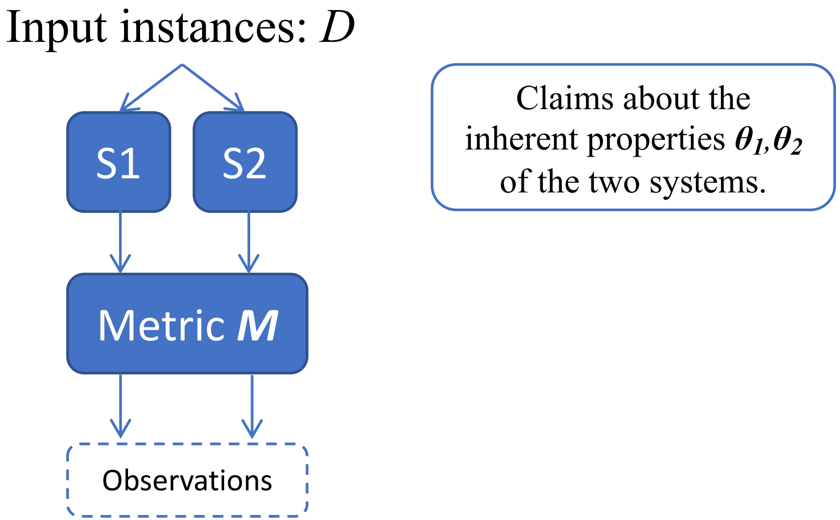
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



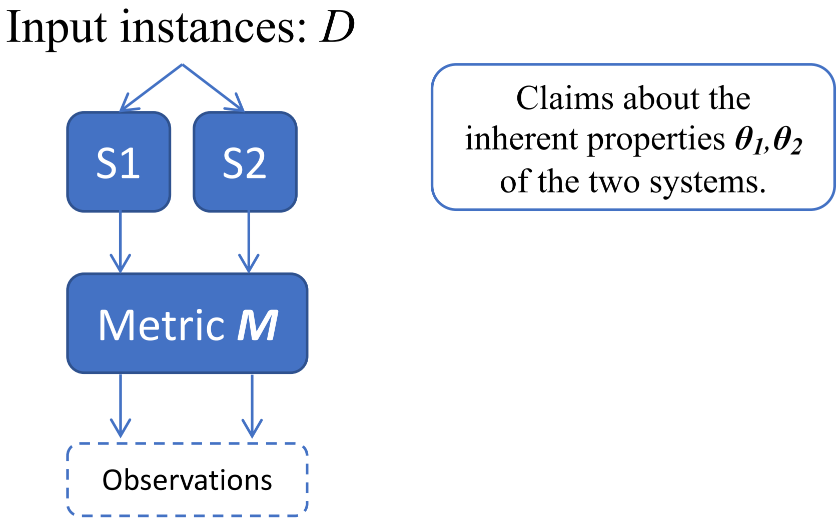
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



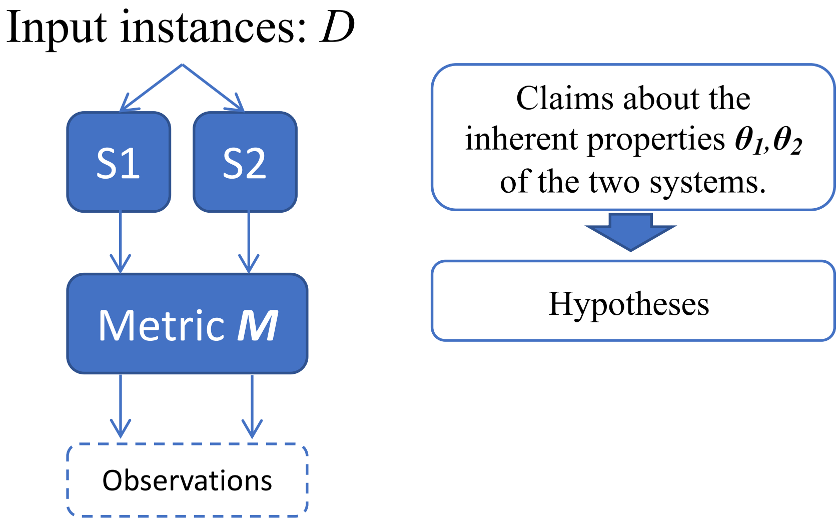
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



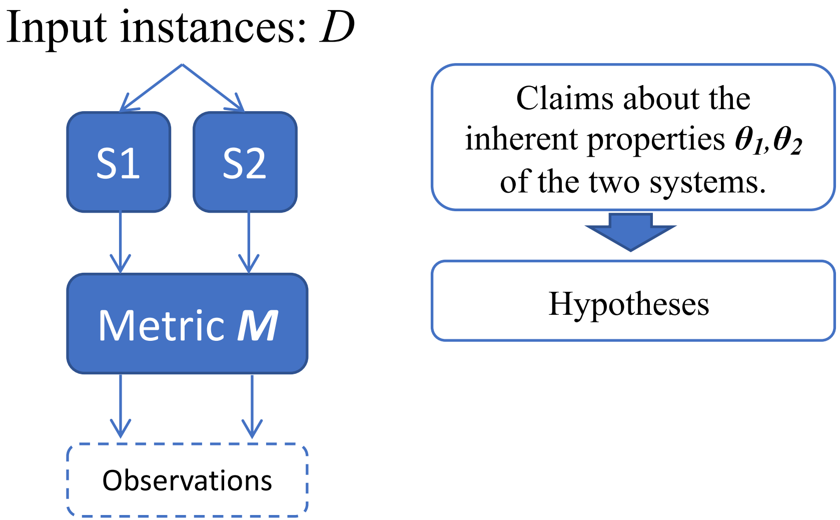
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



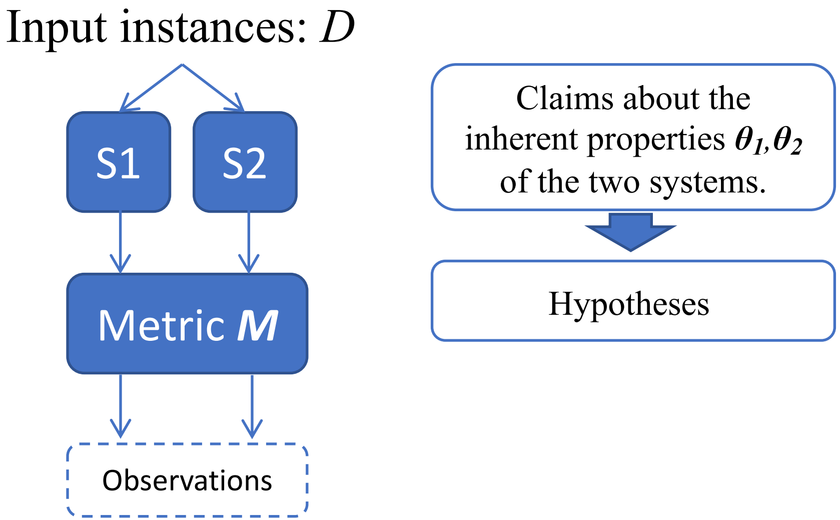
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



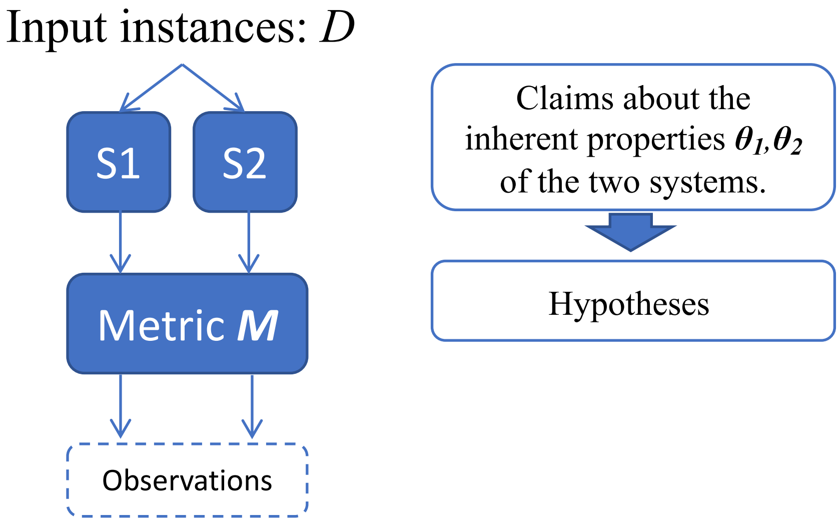
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



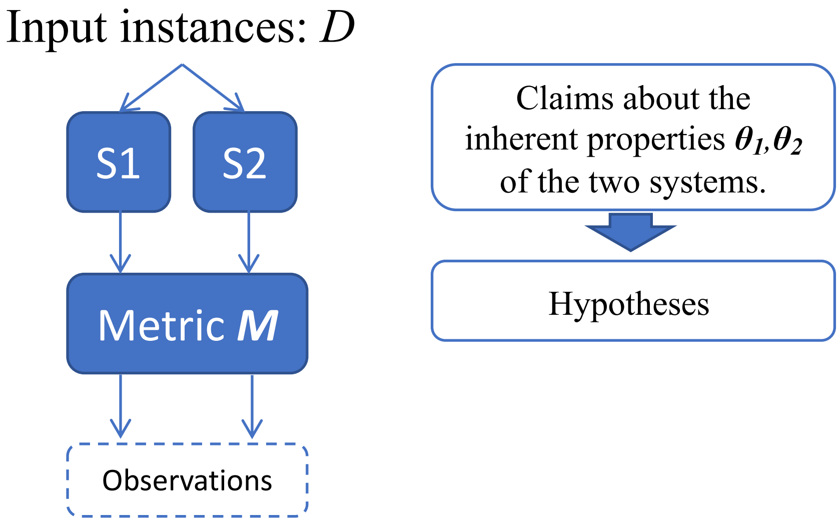
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



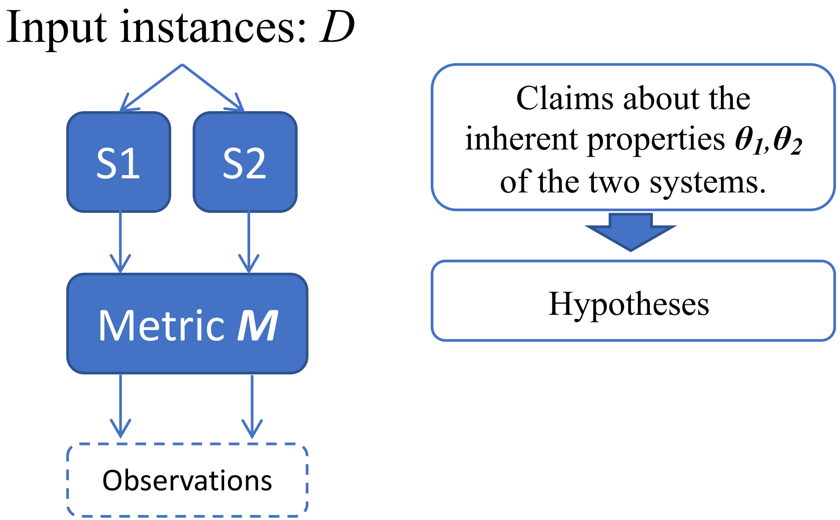
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



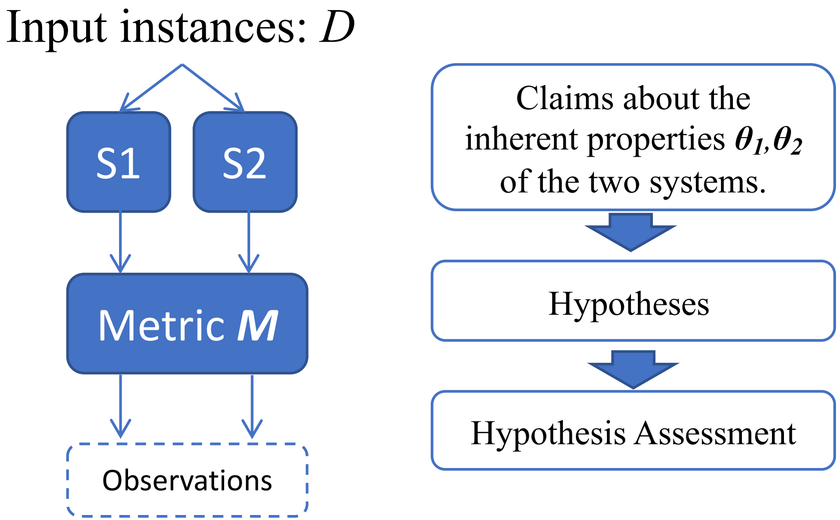
Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …

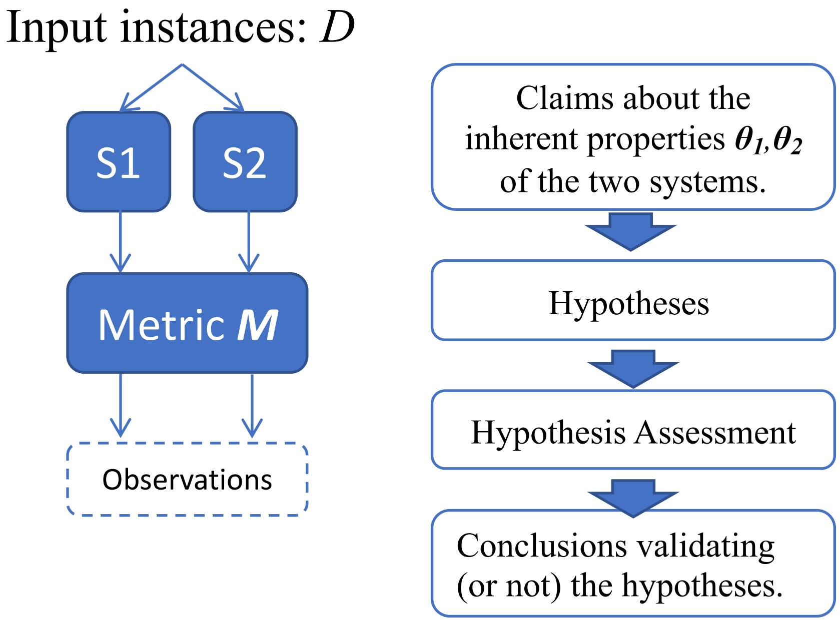


Notation

- 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: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$
  - H2: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2} + b$
  - H3: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
  - …



Null-Hypothesis Significance Testing

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

Null-Hypothesis Significance Testing

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

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

Null-Hypothesis Significance Testing

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

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

Null-Hypothesis Significance Testing

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

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

Null-Hypothesis Significance Testing

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

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

Null-Hypothesis Significance Testing

- The goal is to decide whether a particular hypothesis can be rejected.
- Make a hypothesis (that you want it to be rejected): null-hypothesis.
- <u>Assume</u> that null-hypothesis is correct. H_0: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
- 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 a particular hypothesis can be rejected.
- Make a hypothesis (that you want it to be rejected): null-hypothesis.
- Assume that null-hypothesis is correct. H_0: $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$
- 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)

P-value, Visualized

P-value, Visualized

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

P-value, Visualized

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

P-value, Visualized

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

P-value, Visualized

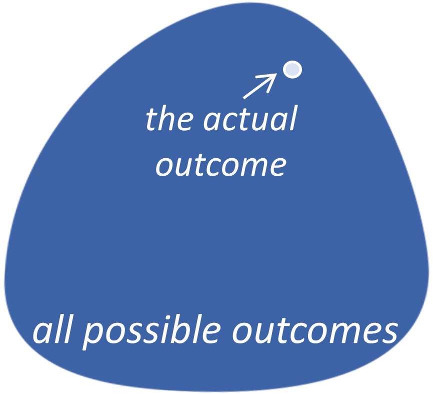
- Null-hypothesis: a hypothesis you want to reject & <u>assume</u> that it is correct.
- **P-value:** the probability of getting an **outcome** as **“extreme”** or more than the **observed** outcome.
- A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.

P-value, Visualized

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

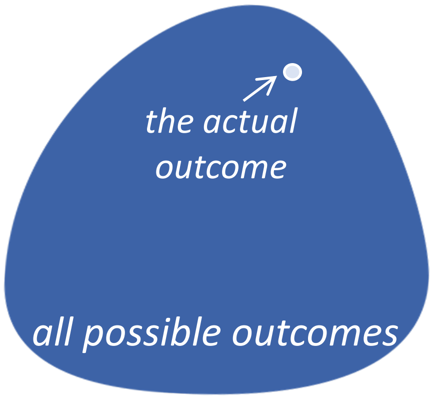
P-value, Visualized

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



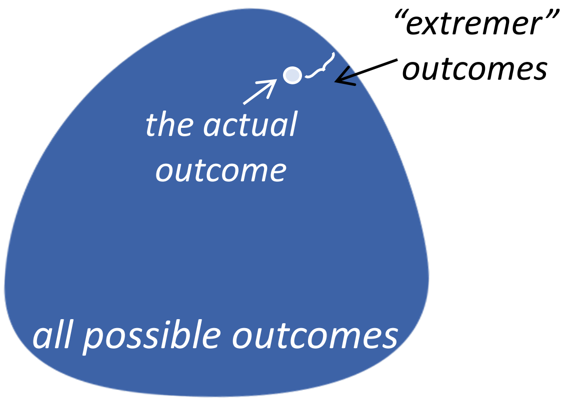
P-value, Visualized

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



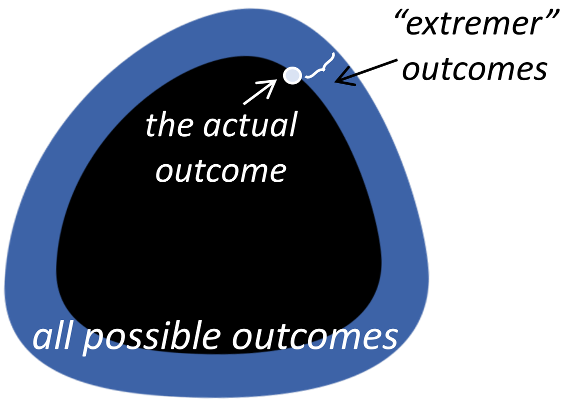
P-value, Visualized

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



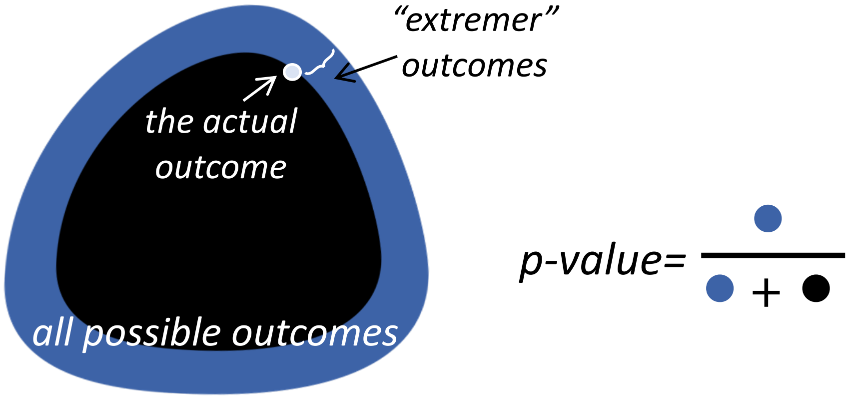
P-value, Visualized

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



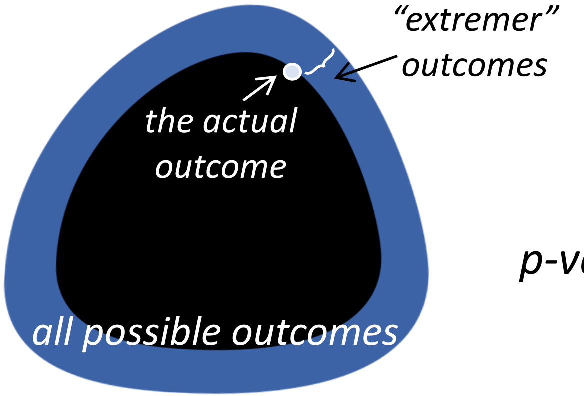
P-value, Visualized

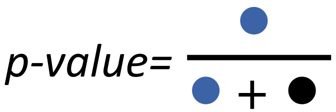
- Null-hypothesis: a hypothesis you want to reject & <u>assume</u> that it is correct.
- **P-value:** the probability of getting an **outcome** as **“extreme”** or more than the **observed** outcome.
- A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.



P-value, Visualized

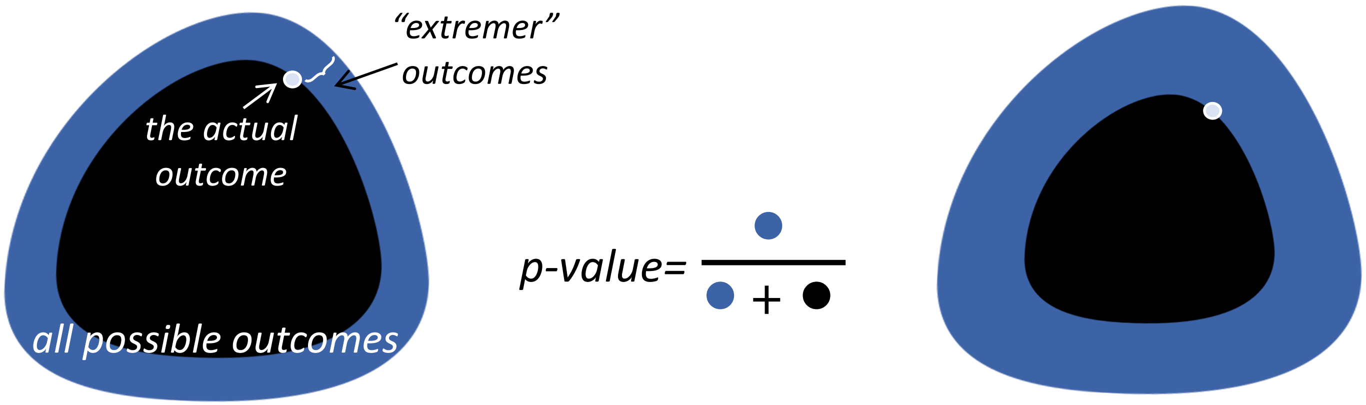
- Null-hypothesis: a hypothesis you want to reject & assume that it is correct.
- **P-value:** the probability of getting an **outcome** as **“extreme”** or more than the **observed** outcome.
- A small p-value is used as a stronger evidence towards rejecting the null-hypothesis.





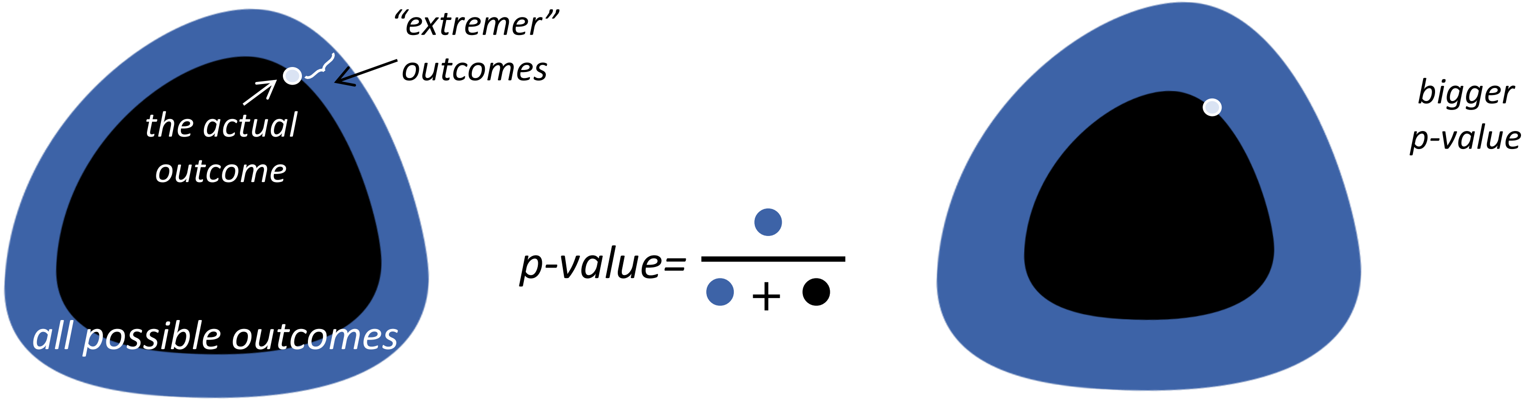
P-value, Visualized

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



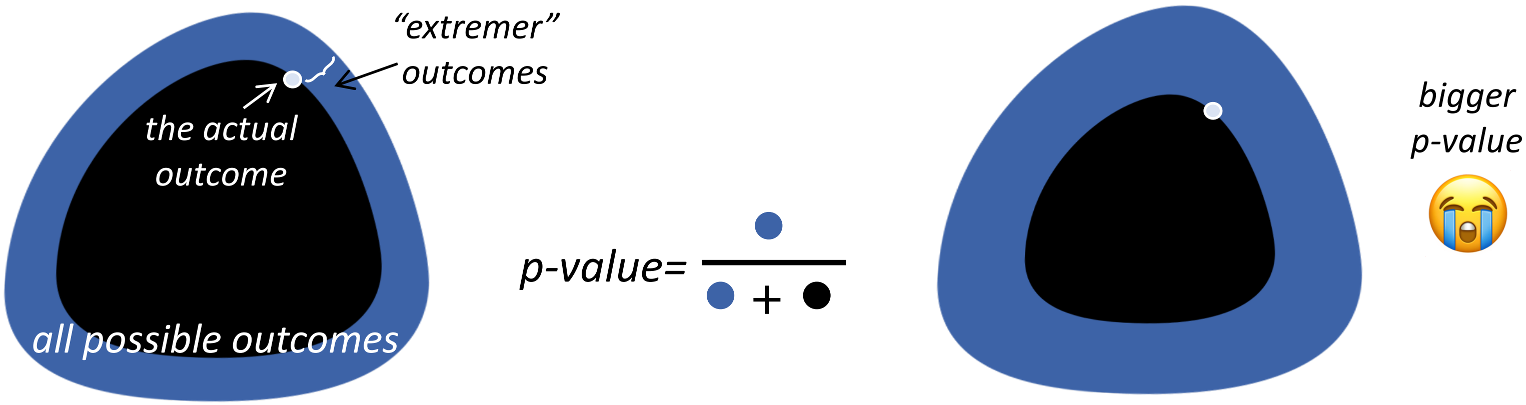
P-value, Visualized

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



P-value, Visualized

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



Null-Hypothesis Significance Testing: Example

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

Null-Hypothesis Significance Testing: Example

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

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\theta_1 > \theta_2$
- Null-Hypothesis **H0**: $\theta_1 = \theta_2$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

Null-Hypothesis Significance Testing: Example

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

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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

Null-Hypothesis Significance Testing: Example

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

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

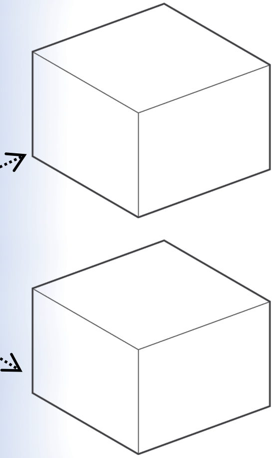
- Null-Hypothesis **H0**: $\theta_1 = \theta_2$

Null-Hypothesis Significance Testing: Example

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

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

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

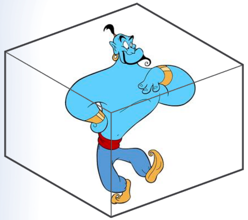


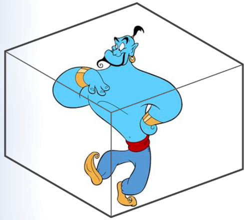
Null-Hypothesis Significance Testing: Example

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

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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



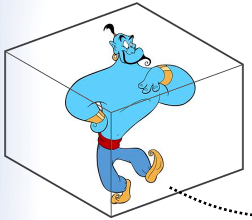


Null-Hypothesis Significance Testing: Example

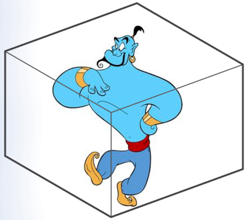
- Hypothesis **H1**: $\theta_1 > \theta_2$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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



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

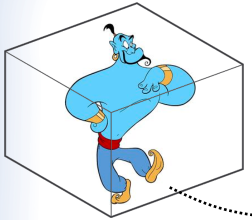


Null-Hypothesis Significance Testing: Example

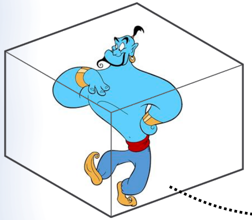
- Hypothesis **H1**: $\theta_1 > \theta_2$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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



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



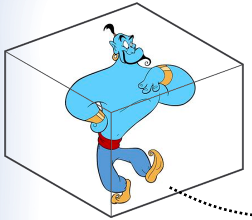
$F, F, F, S, \ldots$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_1 > \theta_2}$

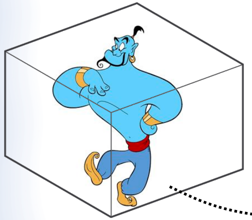
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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



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

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



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

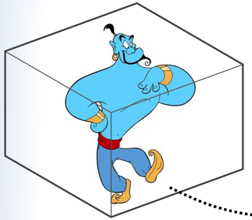
$F, F, F, S, \ldots$

Null-Hypothesis Significance Testing: Example

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

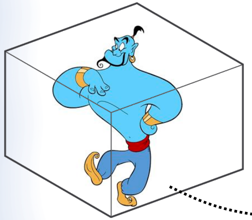
- Null-Hypothesis **H0**: $\theta_1 = \theta_2$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%



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

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



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

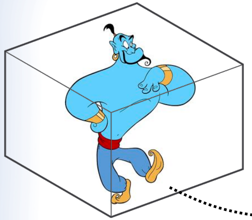
$F, F, F, S, \ldots$

Null-Hypothesis Significance Testing: Example

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

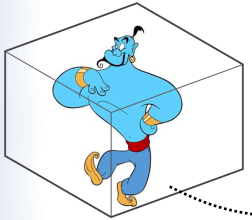
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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



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

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



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

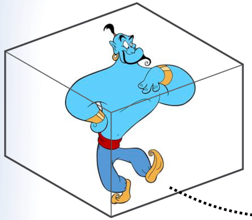
$\hat{\theta}_2 = \text{Avg}[F, F, F, S, \ldots\ ]$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_1 > \theta_2}$

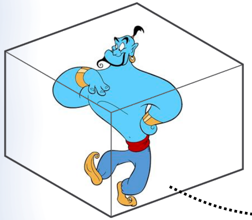
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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



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

$\hat{\theta}_1 = \mathrm{Avg}[S, F, F, S, \ldots\ ]$



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

$\hat{\theta}_2 = \mathrm{Avg}[F, F, F, S, \ldots\ ]$

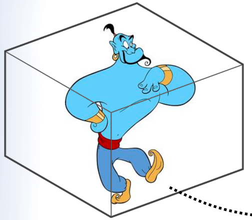
$\mathbf{Prob}\left[\hat{\theta}_1 - \hat{\theta}_2 > 3.5 \middle| \boldsymbol{\theta_1 = \theta_2}\right]$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_1 > \theta_2}$

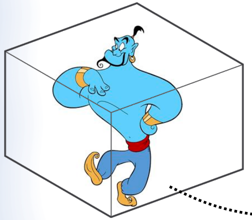
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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



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

$$\hat{\theta}_1 = \mathrm{Avg}[S, F, F, S, \ldots\ ]$$



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

$$\hat{\theta}_2 = \mathrm{Avg}[F, F, F, S, \ldots\ ]$$

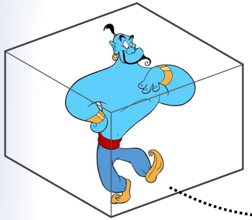
P-value= $\mathbf{Prob}\left[\hat{\theta}_1 - \hat{\theta}_2 > 3.5 \middle| \boldsymbol{\theta_1 = \theta_2}\right]$

Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_1 > \theta_2}$

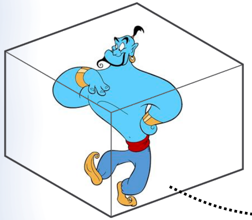
System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

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



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

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



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

$\hat{\theta}_2 = \text{Avg}[F, F, F, S, \ldots\ ]$

P-value= $\textbf{Prob}\left[\hat{\theta}_1 - \hat{\theta}_2 > 3.5 \middle| \boldsymbol{\theta_1 = \theta_2}\right] <_? \beta$ (e.g., 0.005)

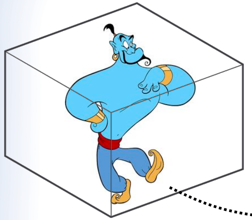
Null-Hypothesis Significance Testing: Example

- Hypothesis **H1**: $\boldsymbol{\theta_1} > \boldsymbol{\theta_2}$

System	Accuracy
Ⓐ	72.4%
Ⓑ	68.9%

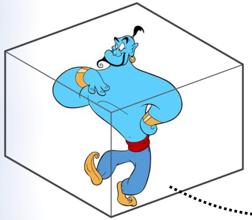
One-sided z-test

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



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

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



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

$\hat{\theta}_2 = \text{Avg}[F, F, F, S, \ldots\ ]$

P-value= $\textbf{Prob}\left[\hat{\theta}_1 - \hat{\theta}_2 > 3.5 \middle| \boldsymbol{\theta_1} = \boldsymbol{\theta_2}\right] \ <_? \beta$ (e.g., 0.005)

Interpreting p-values

Interpreting p-values

- Pretty complex notion!

Interpreting p-values

- Pretty complex notion!

"The probability of obtaining test results **at least as extreme as the results** actually observed during the test, **assuming** that the **null-hypothesis is correct**."
--*your favorite statistics textbook*

Interpreting p-value

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

(Demsar, 2008; Goodman, 2008)

Interpreting p-value

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

(Demsar, 2008; Goodman, 2008)

Interpreting p-value

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

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

(Demsar, 2008; Goodman, 2008)

Interpreting p-value

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

(Demsar, 2008; Goodman, 2008)

Interpreting p-value

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

(Demsar, 2008; Goodman, 2008)

Interpreting p-value

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

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

(Demsar, 2008; Goodman, 2008)

Interpreting p-value

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

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

(Demsar, 2008; Goodman, 2008)

Interpreting p-value

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)

Remember this?

Remember this?

Important reminder regarding large samples and p-values. Inbox × AI2 ×

Oren Etzioni <orene@allenai.org> Tue, Aug 20, 2019, 12:40 PM

to team

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

For more, see the attached paper.

--

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

p-values paper.pdf

Remember this?

Important reminder regarding large samples and p-values. Inbox × AI2 ×

Oren Etzioni <orene@allenai.org> Tue, Aug 20, 2019, 12:40 PM

to team

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

For more, see the attached paper.

--

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


Published online ahead of print April 12, 2013

Information Systems Research

Too Big to Fail: Large Samples and the p-Value Problem

Mingfeng Lin

Henry C. Lucas, Jr.

Galit Shmueli

p-values paper.pdf

Remember this?

Important reminder regarding large samples and p-values.

Inbox × AI2 ×

Oren Etzioni <orene@allenai.org>
to team

Tue, Aug 20, 2019, 12:40 PM

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

For more, see the attached paper.

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

p-values paper.pdf

Remember this?

Important reminder regarding large samples and p-values. Inbox × AI2 ×

Oren Etzioni <orene@allenai.org> Tue, Aug 20, 2019, 12:40 PM

to team

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

Or just keep listening to Daniel's presentation!

For more, see the attached paper.

--

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


Published online ahead of print April 12, 2013

Information Systems Research

Too Big to Fail: Large Samples and the p-Value Problem

Mingfeng Lin

Henry C. Lucas, Jr.

Galit Shmueli

p-values paper.pdf

Intermediate Summary

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

Intermediate Summary

- 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

- P-values do <u>not</u> provide **probability** estimates on two classifiers being different (or equal).
- **Statistical significance** is different than **practical significance.**
- Point-wise null hypotheses could be misused: **for big enough data points it is possible to make statistically significant claims.**

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

Intermediate Summary

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

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

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

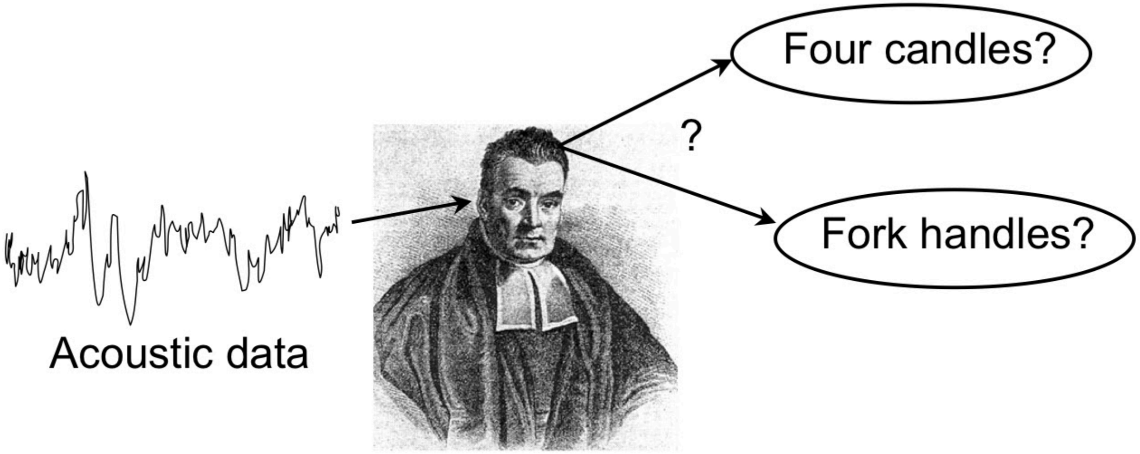
	Frequentist	Bayesian
Binary Decision	Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	Confidence Interval	Posterior Intervals

	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.
  - **Likelihood:** How the hidden parameters are connected to the observations.
  - **Posterior:** Summary of the inferences about likely values of Θ.

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

Posterior Intervals

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

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

Posterior Intervals

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

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

Posterior Intervals

$$\boldsymbol{P}(\text{Hypothesis}|\text{Observations})$$

Posterior Intervals

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

$$\boldsymbol{P}(\text{Hypothesis}|\text{Observations})$$

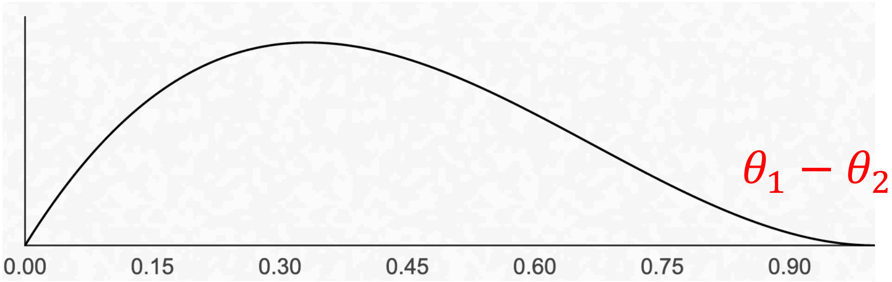
- **Example:** H_1: $\theta_1 - \theta_2 > \alpha$

Posterior Intervals

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

$$\boldsymbol{P}(\text{Hypothesis}|\text{Observations})$$

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

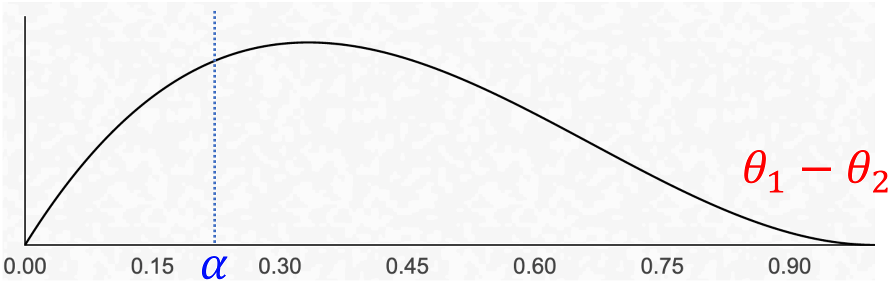


Posterior Intervals

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

$$\boldsymbol{P}(\text{Hypothesis}|\text{Observations})$$

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

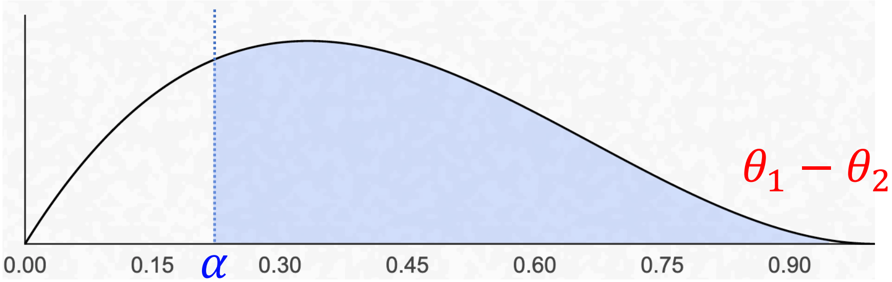


Posterior Intervals

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

$$\boldsymbol{P}(\text{Hypothesis}|\text{Observations})$$

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

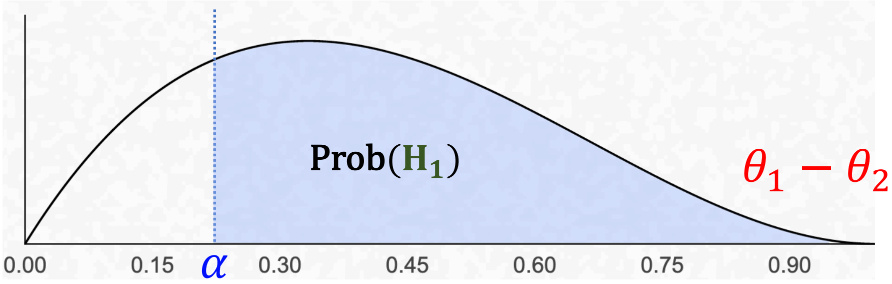


Posterior Intervals

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

$$\boldsymbol{P}(\text{Hypothesis}|\text{Observations})$$

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



Posterior Intervals: Example

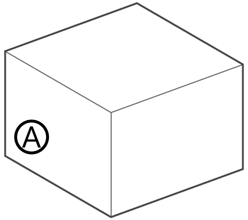
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

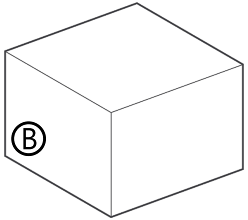
$$\mathrm{H}_1:\ \theta_1 - \theta_2 > \alpha$$

Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_1 - \theta_2 > \alpha$$

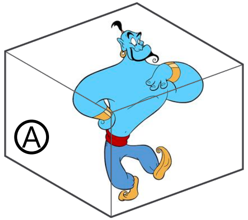


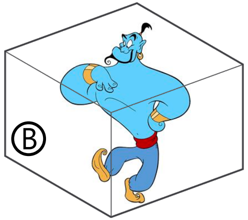


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\colon \theta_1 - \theta_2 > \alpha$$

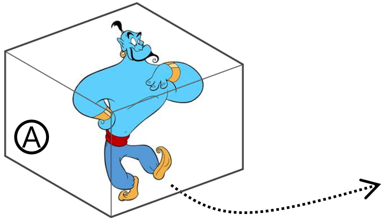


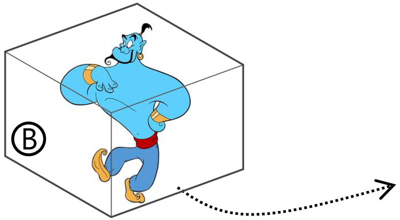


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\colon \theta_1 - \theta_2 > \alpha$$

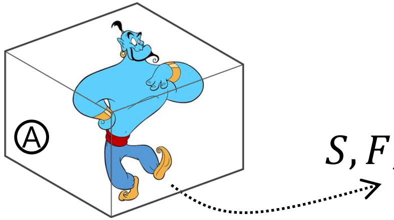




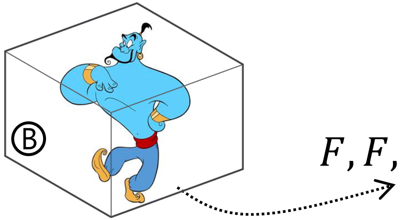
Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\colon \theta_1 - \theta_2 > \alpha$$



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

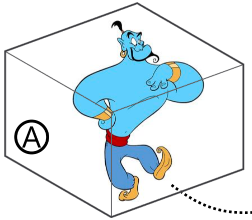


$F, F, F, S, \ldots$

Posterior Intervals: Example

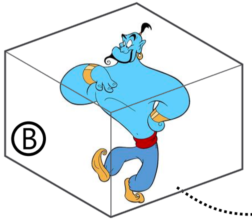
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\colon \theta_1 - \theta_2 > \alpha$$



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

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



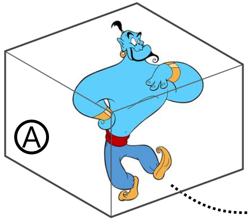
$$\theta_2 = \mathrm{Prob}(S)$$

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

Posterior Intervals: Example

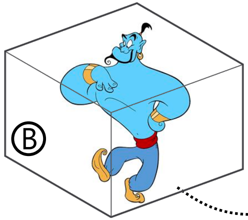
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_1 - \theta_2 > \alpha$$



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

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



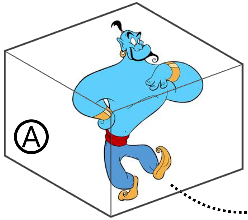
$$\theta_2 = \mathrm{Prob}(S)$$

$$\underbrace{F, F, F, S, \ldots}_{\boldsymbol{P}(Y|\Theta)}$$

Posterior Intervals: Example

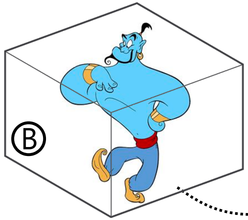
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_1 - \theta_2 > \alpha$$



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

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



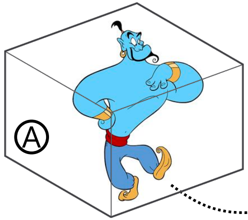
$$\theta_2 = \mathrm{Prob}(S)$$

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

Posterior Intervals: Example

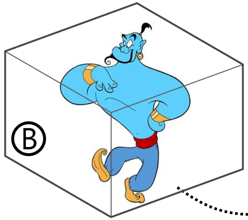
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_1 - \theta_2 > \alpha$$



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

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



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

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

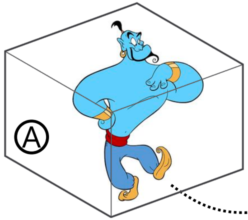
$$\boldsymbol{P}(Y|\Theta)$$
$$\oplus$$
$$\boldsymbol{P}(\Theta) \sim \text{uniform}$$

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

Posterior Intervals: Example

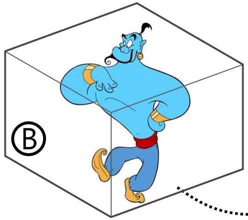
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}_1\text{: } \theta_1 - \theta_2 > \alpha$$



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

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



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

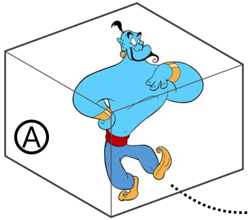
$$\underbrace{F, F, F, S, \ldots}_{\substack{\boldsymbol{P}(Y|\Theta) \\ \oplus \\ \boldsymbol{P}(\Theta)\sim\mathrm{uniform}}}$$

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

Posterior Intervals: Example

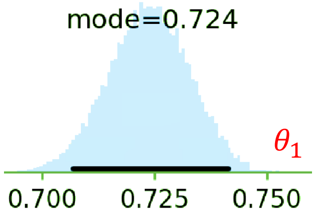
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

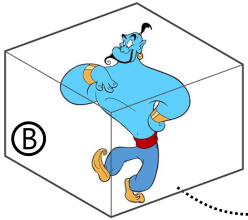
$$\mathrm{H}_1\text{: } \theta_1 - \theta_2 > \alpha$$



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

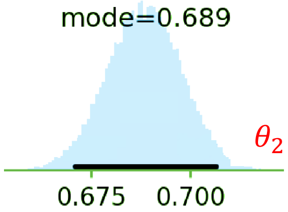
$S, F, F, S, \ldots$





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

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



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

$\oplus$

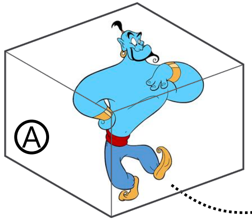
$\boldsymbol{P}(\Theta) \sim \text{uniform}$

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

Posterior Intervals: Example

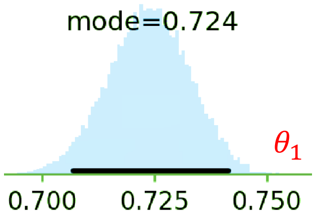
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

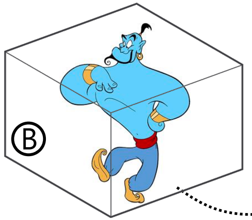
$$\mathrm{H}_1\text{: } \theta_1 - \theta_2 > \alpha$$



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

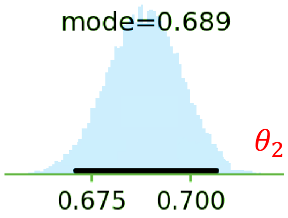
$S, F, F, S, \ldots$





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

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



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

$\oplus$

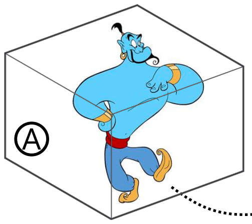
$\boldsymbol{P}(\Theta) \sim \mathrm{uniform}$

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

Posterior Intervals: Example

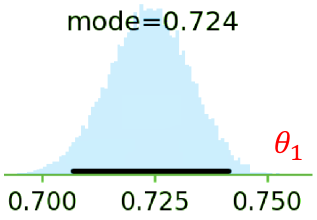
System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

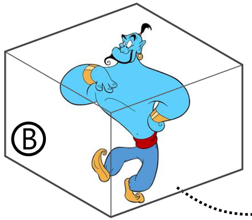
$$\mathrm{H}_1\!: \theta_1 - \theta_2 > \alpha$$



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

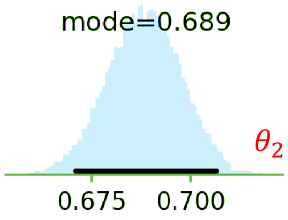
$S, F, F, S, \ldots$

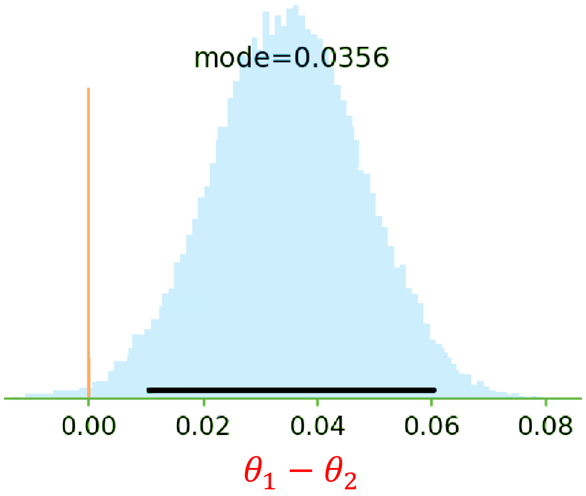




$\theta_2 = \mathrm{Prob}(S)$

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





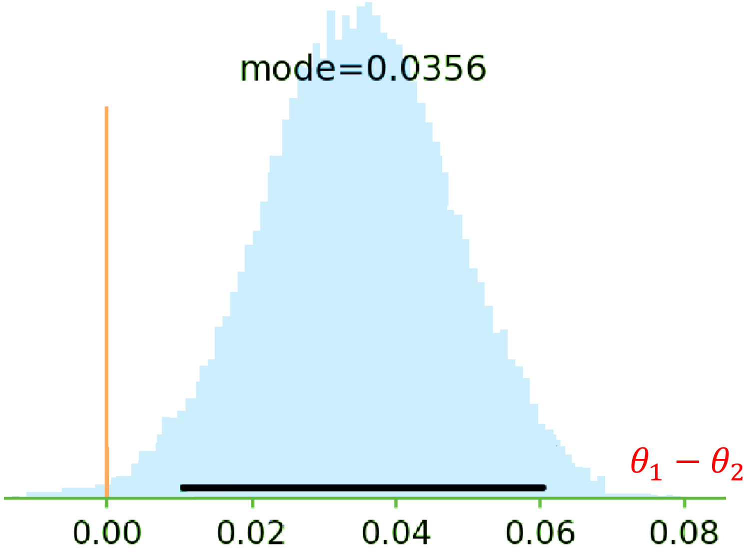
$$\boldsymbol{P}(Y|\Theta) \oplus \boldsymbol{P}(\Theta) \sim \text{uniform}$$

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

Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

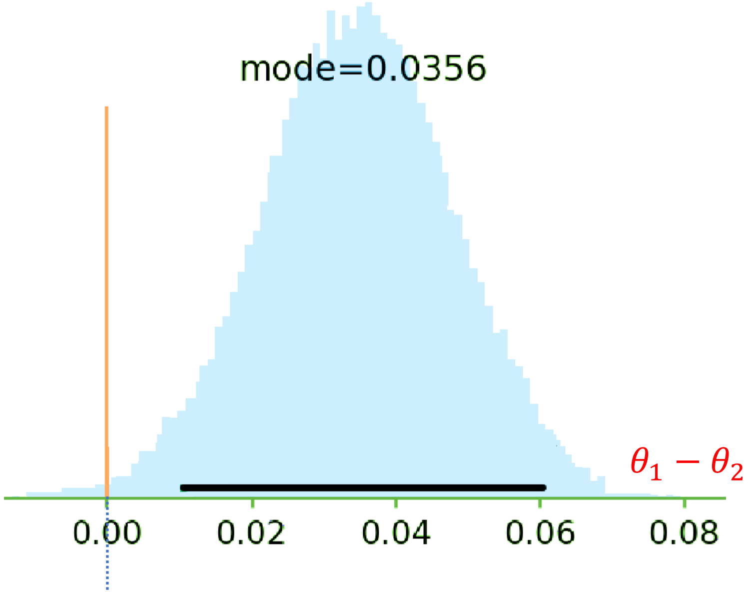
H: $\theta_1 - \theta_2 > \alpha$



Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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



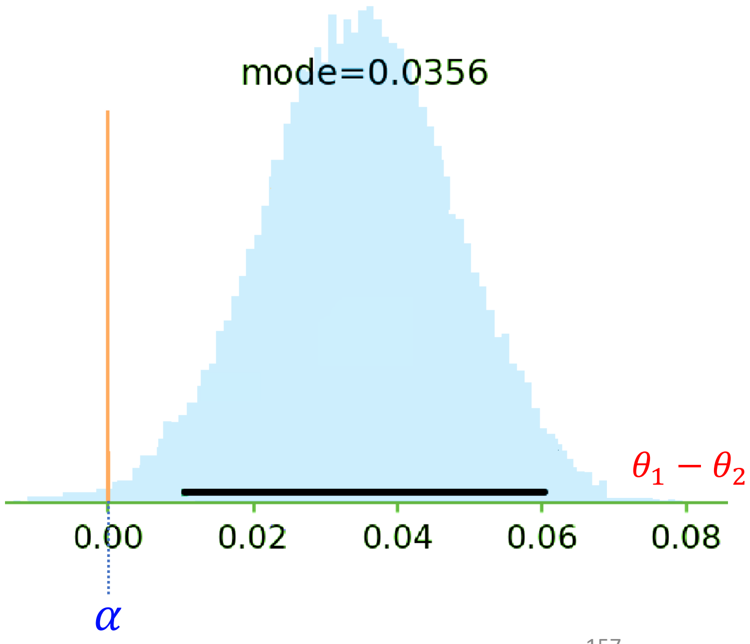
Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

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

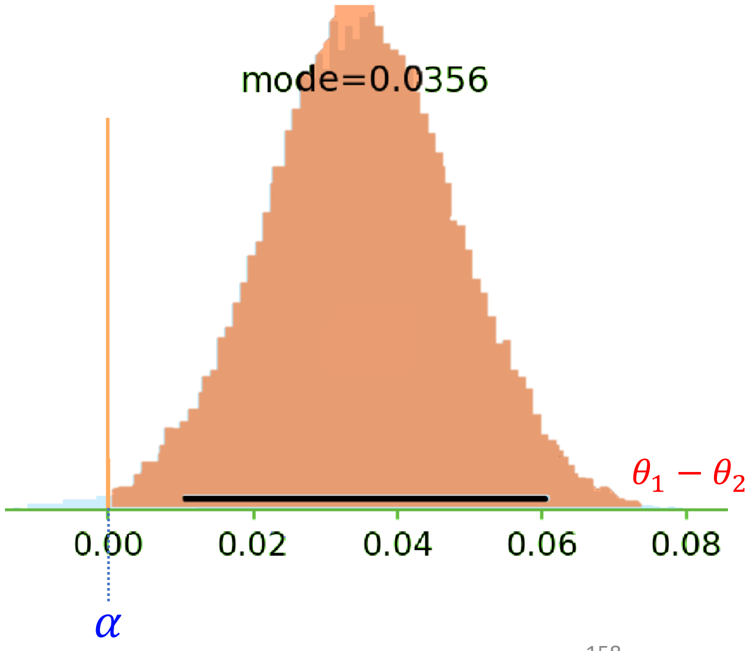
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.



Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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



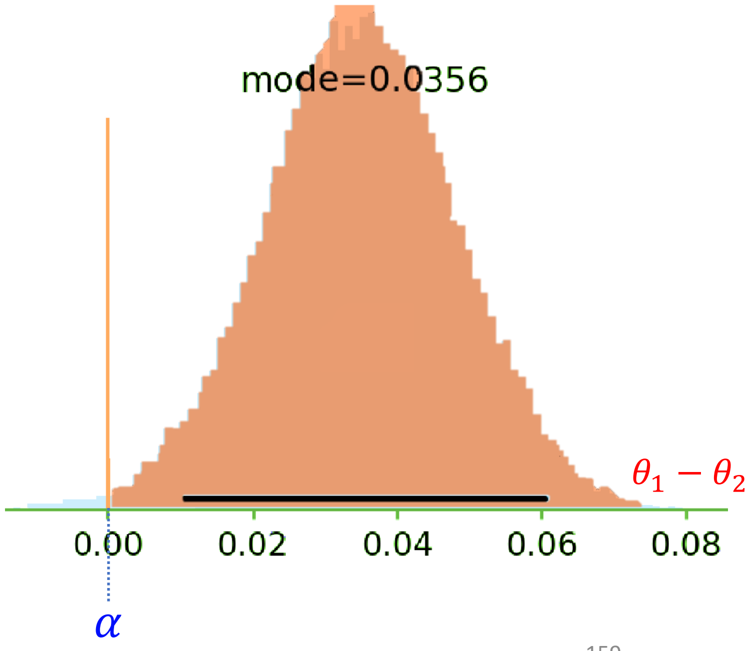
- The hypothesis (w/ $\alpha = 0$) holds true …
    - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
    - … with probability %94.

Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

$$\mathrm{H}:\ \theta_1 - \theta_2 > \alpha$$

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

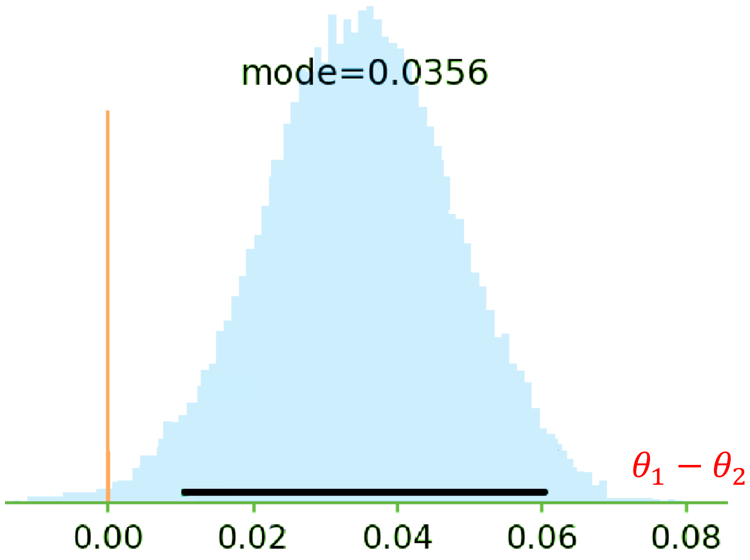


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

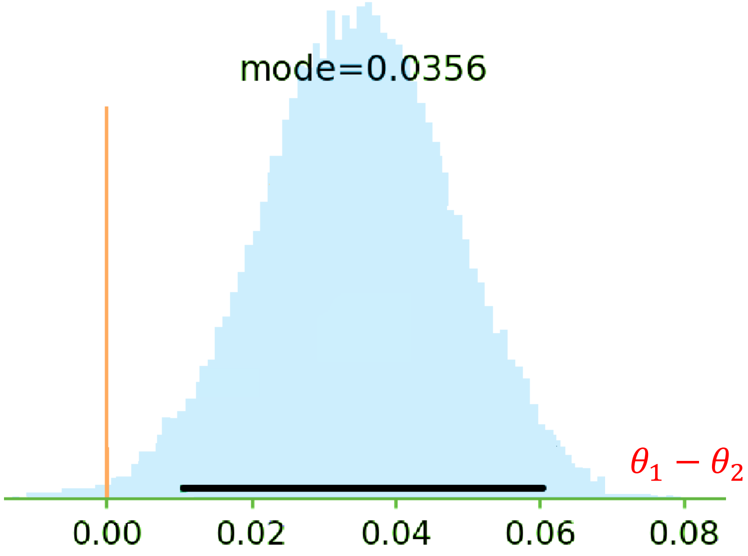
- The hypothesis (w/ $\alpha = 0$) holds true …
  - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.



Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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



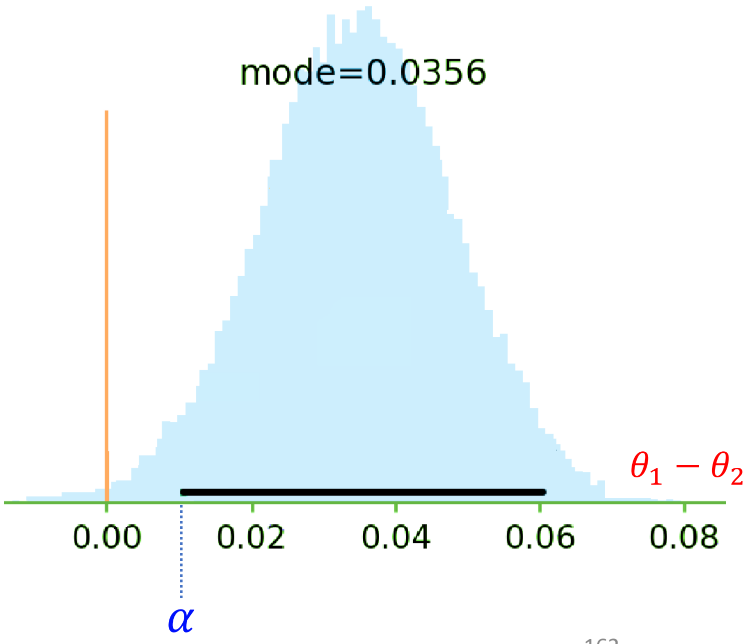
- The hypothesis (w/ $\alpha = 0$) holds true …
  - … with probability %99.6.
- The hypothesis (w/ $\alpha = 1$) holds true …
  - … with probability %94.

Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

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

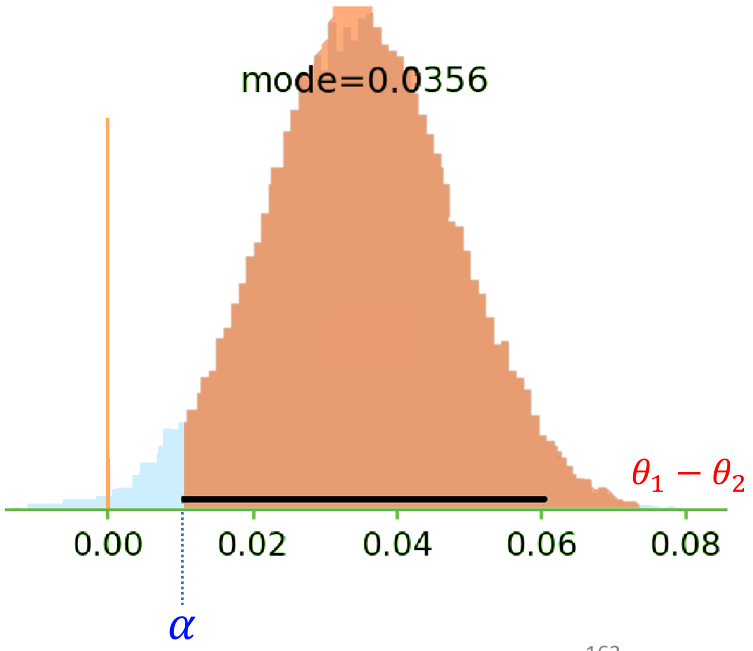


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

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

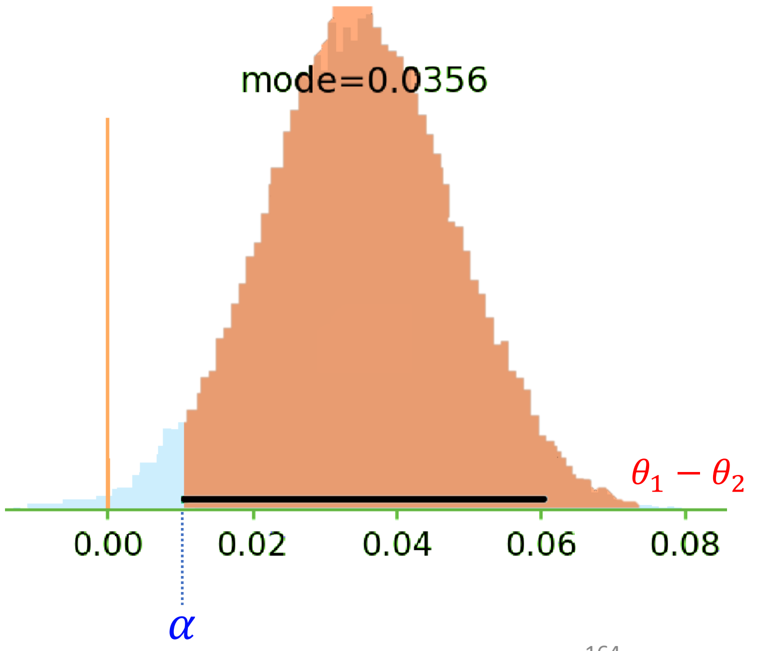


Posterior Intervals: Example

System	Accuracy
Ⓐ	72.4
Ⓑ	68.9

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

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



2nd Intermediate Summary

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

2nd Intermediate Summary

- It's much more intuitive to work with the **probability of hypotheses**.
  - **Easier** to **interpret** → less ambiguous.

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

2nd Intermediate Summary

- It's much more intuitive to work with the **probability of hypotheses**.
  - **Easier** to **interpret** → less ambiguous.

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

- This does not encourage **binary** decision-making.

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

2^{nd} Intermediate Summary

- It's much more intuitive to work with the **probability of hypotheses**.
  - **Easier** to **interpret** → less ambiguous.
- Provides a **flexible** framework
  - E.g., **margin of superiority** could incorporated in the definition of hypotheses.
- This does not encourage **binary** decision-making.

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

Final Section:

Common Practices, Comparisons and Suggestions

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

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

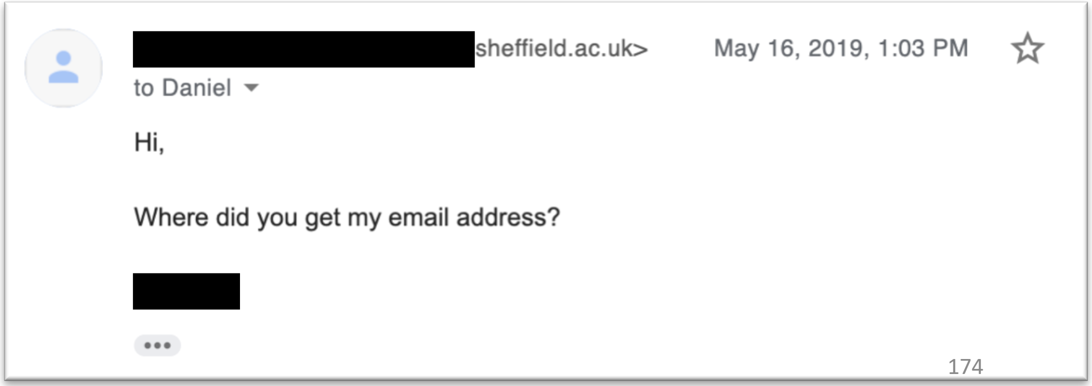
- A questionnaire containing general and specific questions about significance assessment tools
- Sent it to over 400 researchers randomly selected from ACL'18 proceedings
- ~50 individuals responded

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?

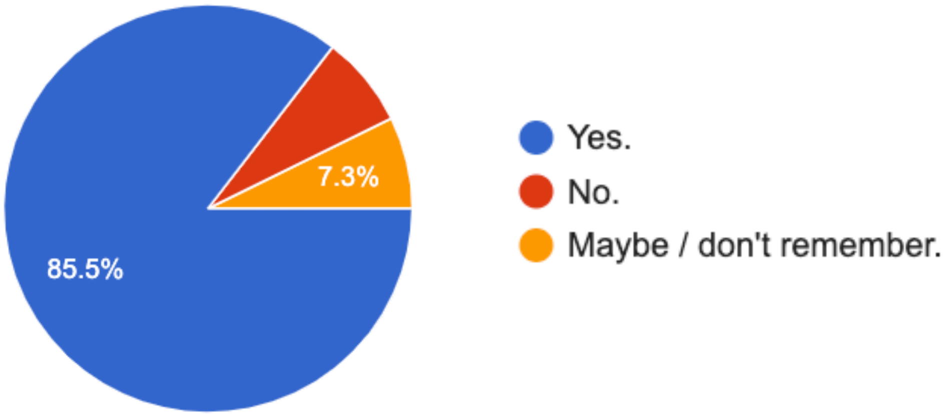
[redacted]

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)."*

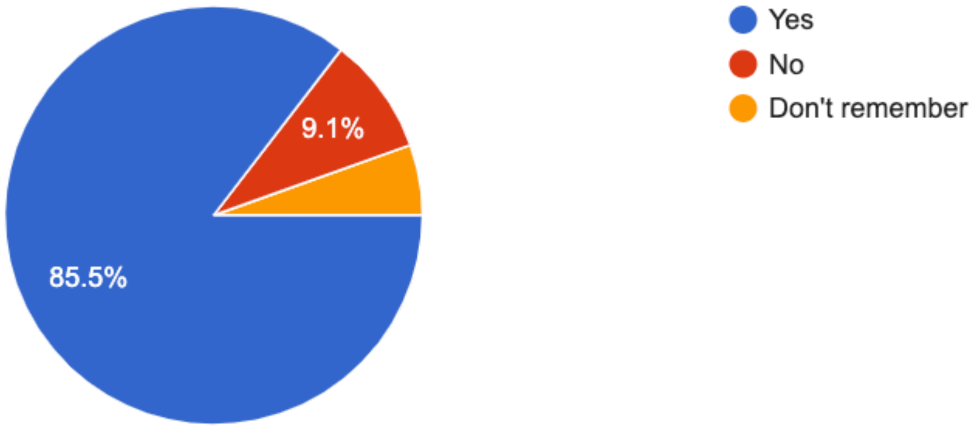


Participants in Our Survey

- *“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.)”*

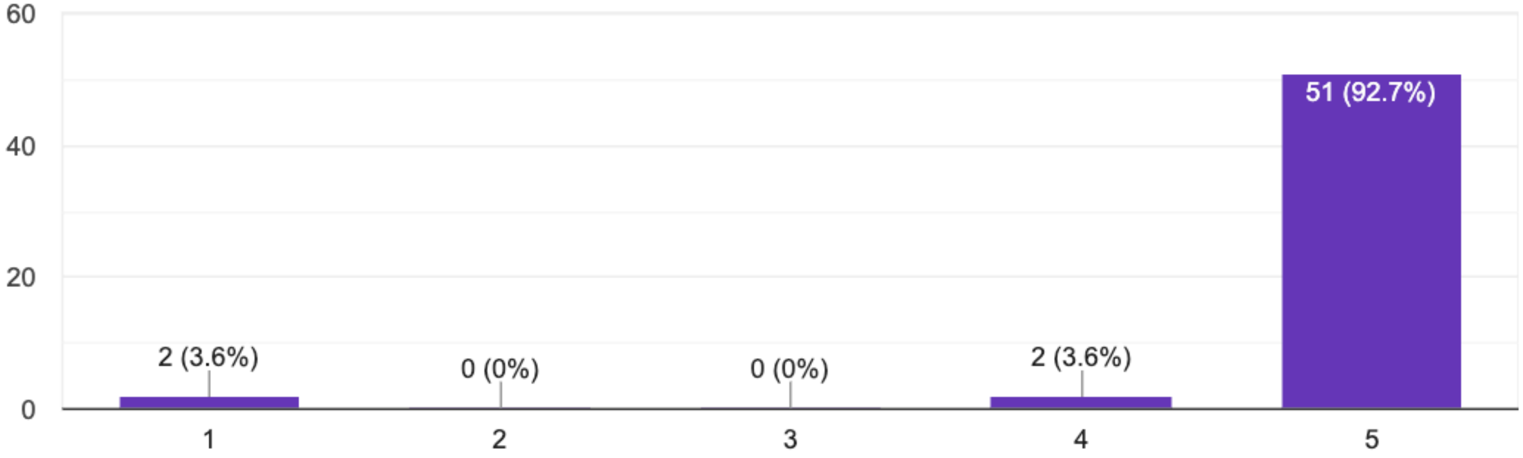


Participants in Our Survey

- *“I am not a robot”*

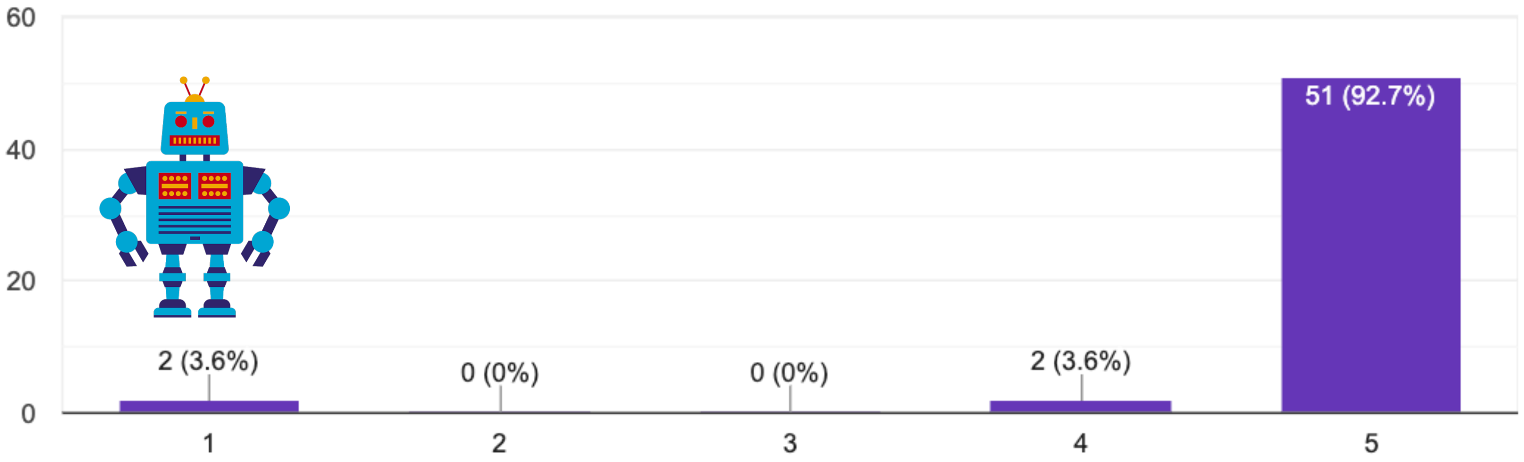
Participants in Our Survey

- *"I am not a robot"*



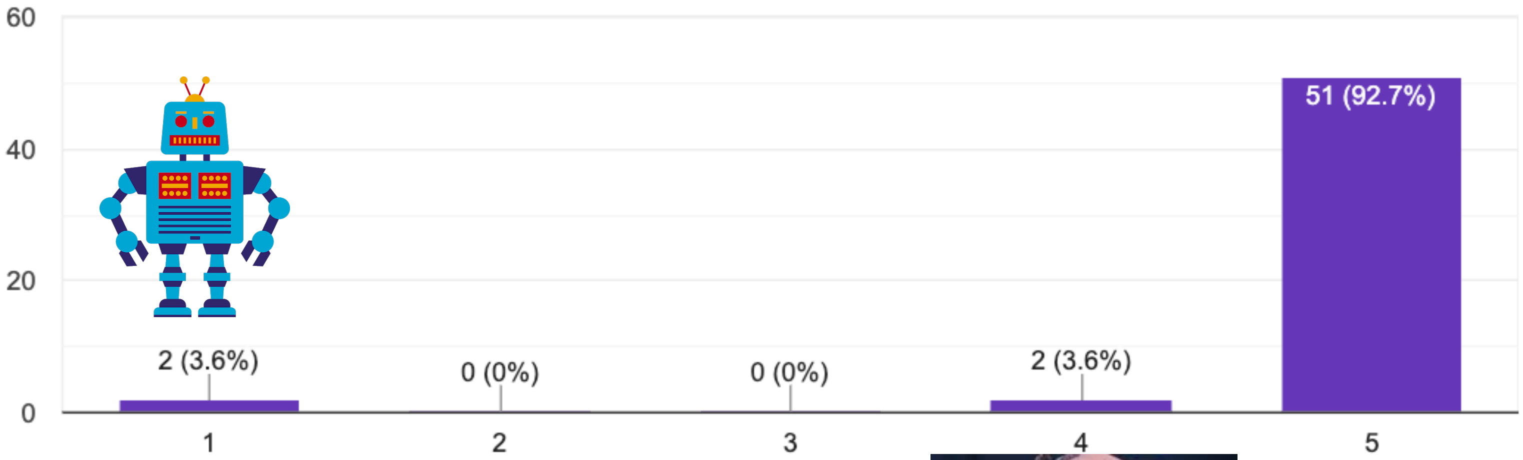
Participants in Our Survey

- *"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?

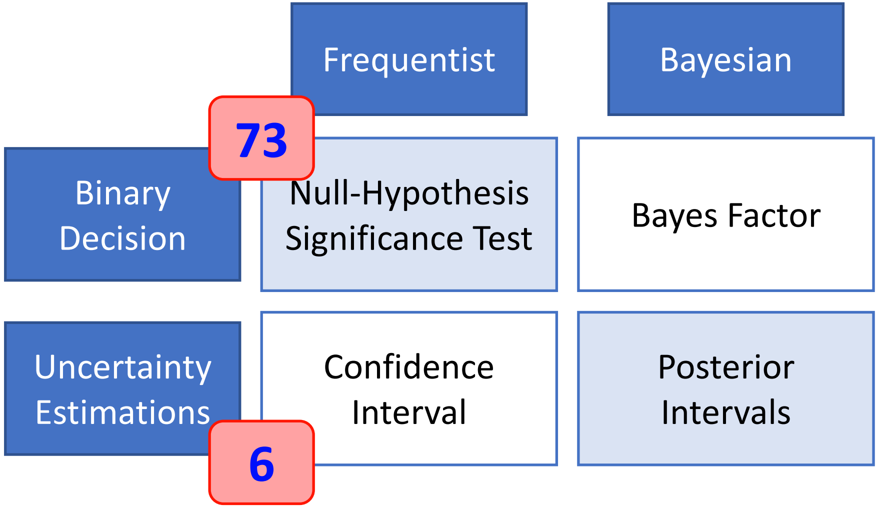
73

	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)

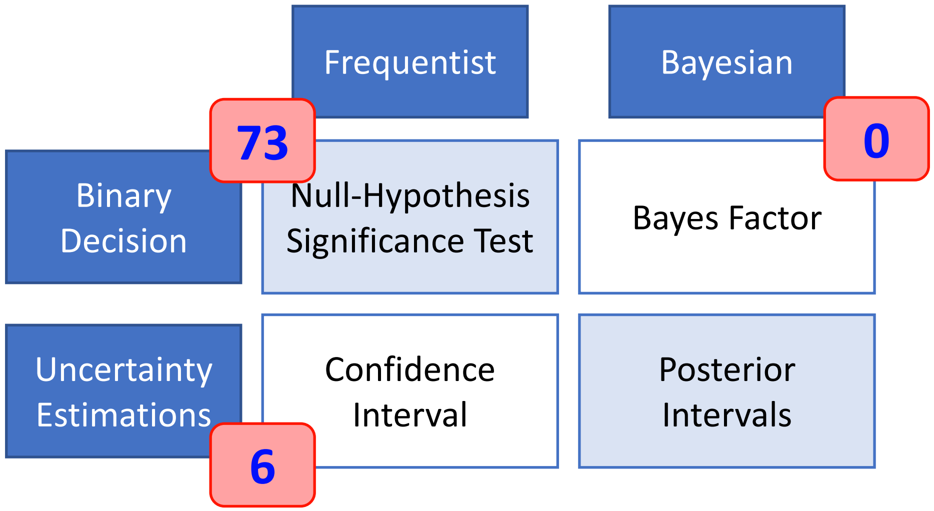
How many papers did use significance testing?



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	Confidence Interval 6	Posterior Intervals 0

Trends and Patterns in the field

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

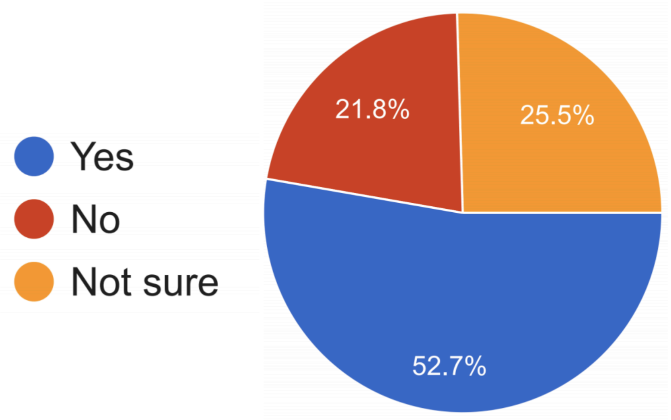
How many papers did use significance testing?

Why?

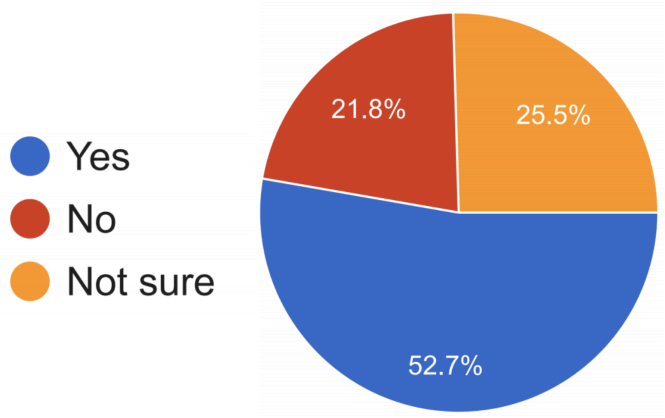
	Frequentist	Bayesian
Binary Decision	73 Null-Hypothesis Significance Test	0 Bayes Factor
Uncertainty Estimations	6 Confidence Interval	Posterior Intervals 0

Have you heard about "Bayesian Hypothesis Testing"?

Have you heard about "Bayesian Hypothesis Testing"?

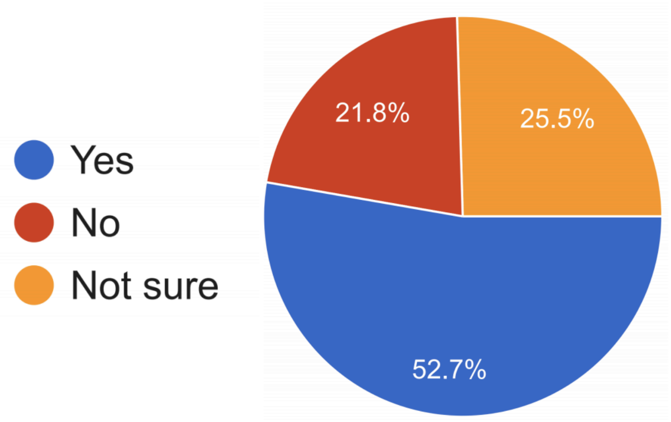


Have you heard about "Bayesian Hypothesis Testing"?

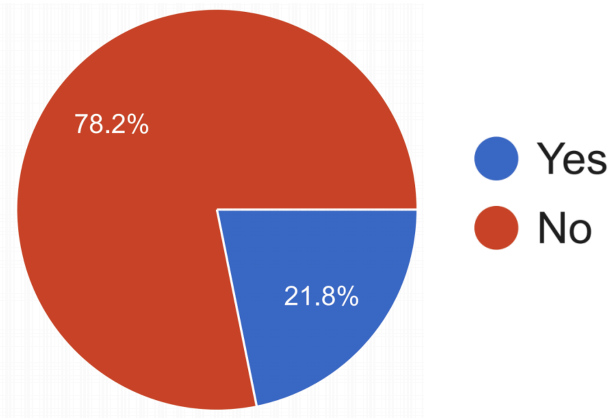


Do you know the definition of "Bayes Factor"?

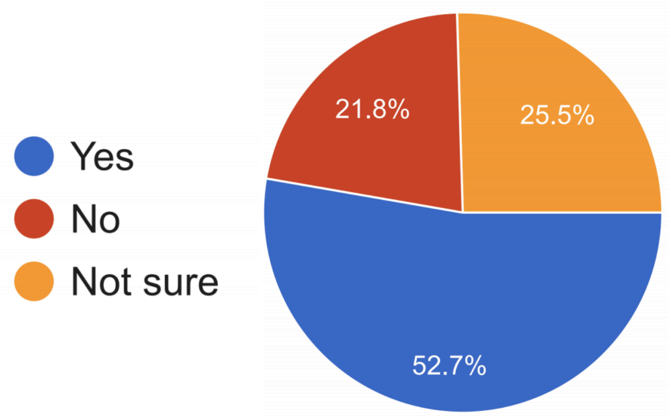
Have you heard about "Bayesian Hypothesis Testing"?



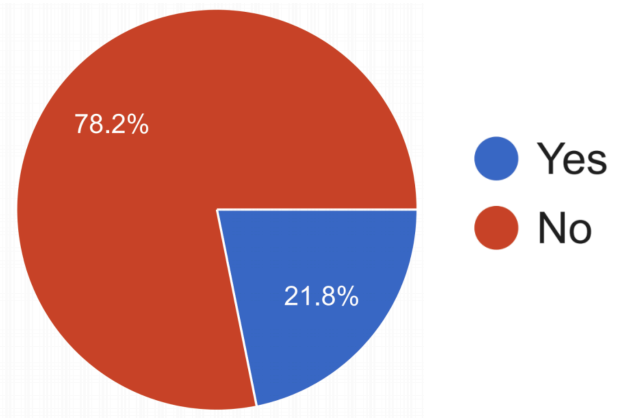
Do you know the definition of "Bayes Factor"?



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. 🤔

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)

Susceptibility to Misinterpretation

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

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

Susceptibility to Misinterpretation

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

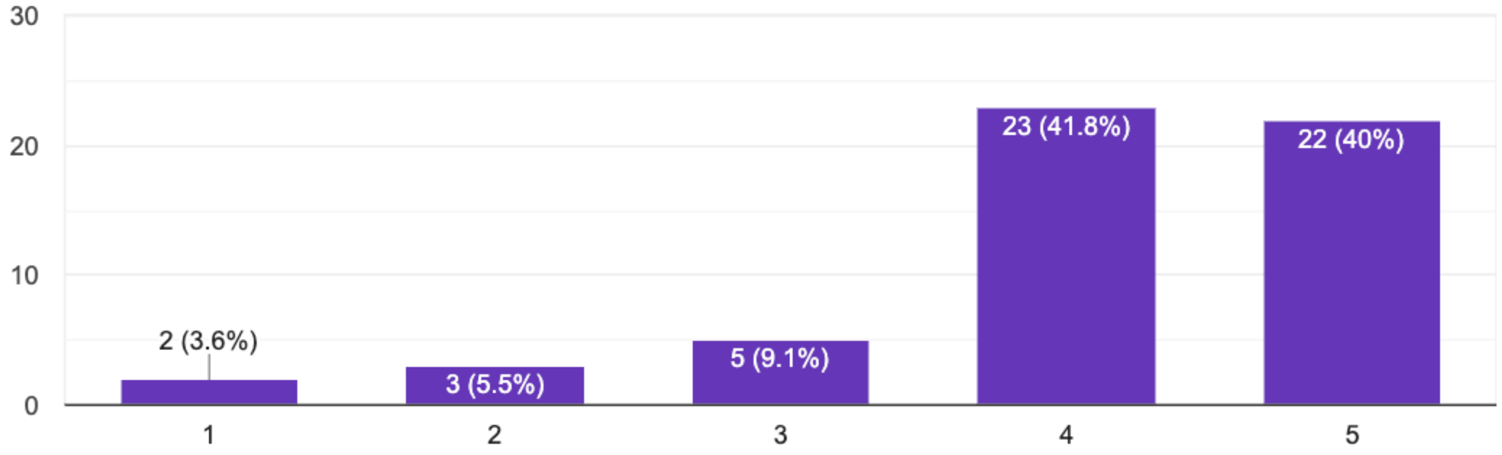
	Frequentist	Bayesian
Binary Decision	👎 Null-Hypothesis Significance Test	Bayes Factor
Uncertainty Estimations	Confidence Interval	Posterior Intervals 😑

Participants in Our Survey

- *"I know p-values and I know how to interpret them."*

Participants in Our Survey

- *"I know p-values and I know how to interpret them."*



A Survey Question: Interpreting P-value

A Survey Question: Interpreting P-value

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

  a) The probability of observing a margin 7% is at most 0.01, assuming that the two classifiers inherently have the same performance.
  b) If we repeat the experiment, with a probability 99% classifier-2 will have a higher performance than classifier-1.

A Survey Question: Interpreting P-value

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

a) The probability of observing a margin 7% is at most 0.01, assuming that the two classifiers inherently have the same performance.
b) If we repeat the experiment, with a probability 99% classifier-2 will have a higher performance than classifier-1.

A Survey Question: Interpreting P-value

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

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

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

Unintended Misleading Result by Iterative Testing

Unintended Misleading Result by Iterative Testing

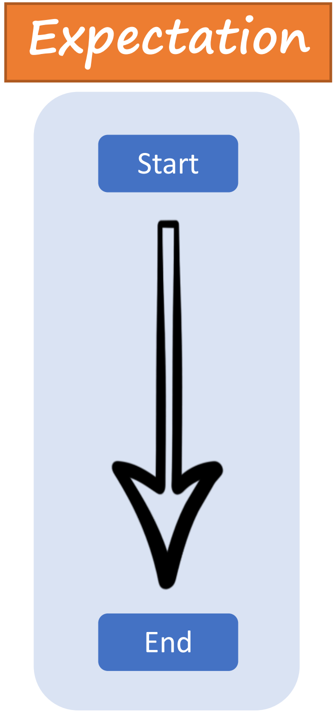
- Many tests are designed for a **single-round** experiment.
- In practice researchers perform **multiple** rounds of experiments.
- This is a major problem when using **binary tests.**
  - E.g., you can "hack" a p-value test, with enough repetitions.

Unintended Misleading Result by Iterative Testing

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

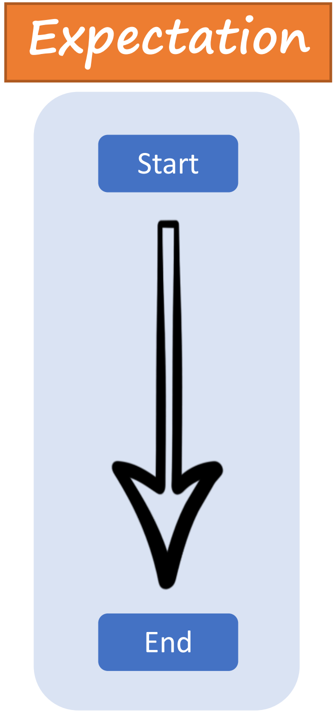
- In practice researchers perform **multiple** rounds of experiments.

- This is a major problem when using **binary tests.**
  - E.g., you can "hack" a p-value test, with enough repetitions.



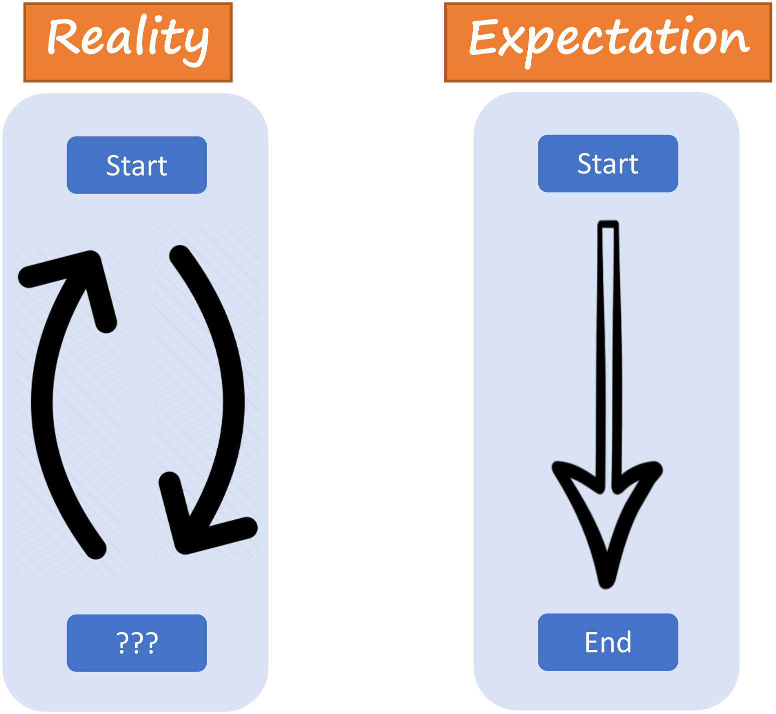
Unintended Misleading Result by Iterative Testing

- Many tests are designed for a **single-round** experiment.
- In practice researchers perform **multiple** rounds of experiments.
- This is a major problem when using **binary tests.**
  - E.g., you can "hack" a p-value test, with enough repetitions.



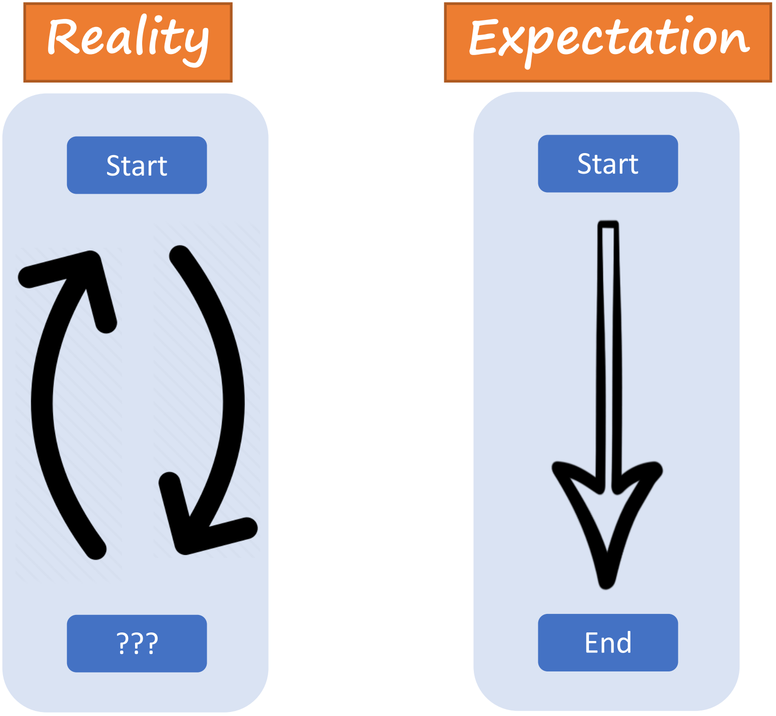
Unintended Misleading Result by Iterative Testing

- Many tests are designed for a **single-round** experiment.
- In practice researchers perform **multiple** rounds of experiments.
- This is a major problem when using **binary tests.**
    - E.g., you can "hack" a p-value test, with enough repetitions.



Unintended Misleading Result by Iterative Testing

- Many tests are designed for a **single-round** experiment.
- In practice researchers perform **multiple** rounds of experiments.
- This is a major problem when using **binary tests.**
  - E.g., you can “hack” a p-value test, with enough repetitions.



The Need for Assumptions

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

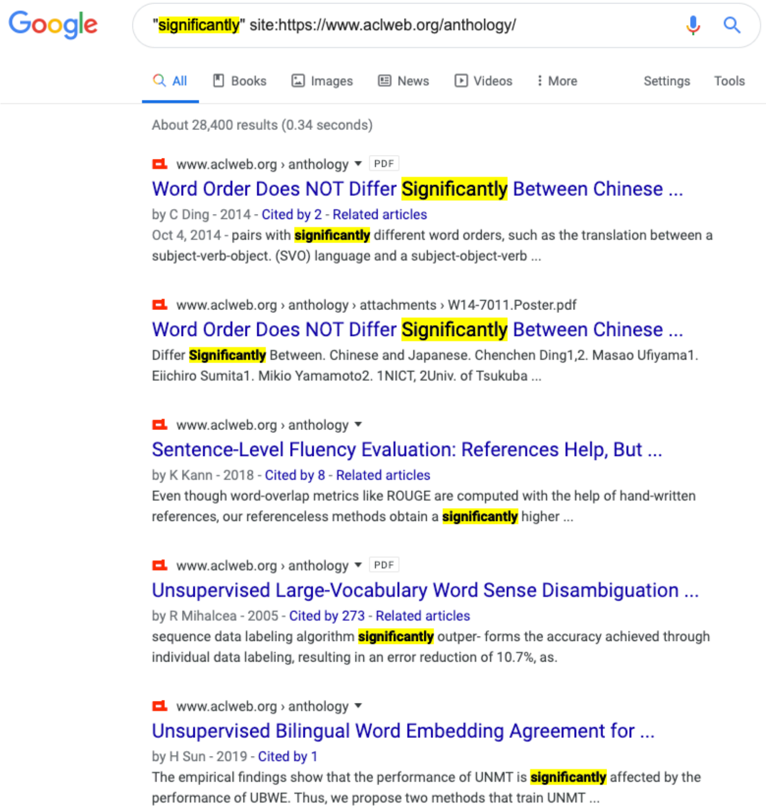
The Need for Assumptions

- *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

Ambiguity problem in interpreting “significance”

Ambiguity problem in interpreting "significance"



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”

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.

Abstract

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

Ambiguity problem in interpreting “significance”

Ambiguity problem in interpreting “significance”

- *An NLP paper presents system-1 and it compares it with a baseline system-2. In its “abstract” it writes: “... system-1 significantly improves over system-2.” What are the right way(s) to interpret this (select all that applies)*
  - It is expected that authors have performed some type of “hypothesis testing.”
  - It is expected that the authors have reported the performances of two systems on a dataset where system-1 has a higher performance than system-2 with a notable margin in the dataset.

Ambiguity problem in interpreting “significance”

- *An NLP paper presents system-1 and it compares it with a baseline system-2. In its “abstract” it writes: “... system-1 significantly improves over system-2.” What are the right way(s) to interpret this (select all that applies)*
  - It is expected that authors have performed some type of “hypothesis testing.”
  - It is expected that the authors have reported the performances of two systems on a dataset where system-1 has a higher performance than system-2 with a notable margin in the dataset.

Ambiguity problem in interpreting “significance”

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

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-1 has a higher performance than system-2 with a notable margin in the dataset.

Ambiguity problem in interpreting “significance”

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

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-1 has a higher performance than system-2 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.”

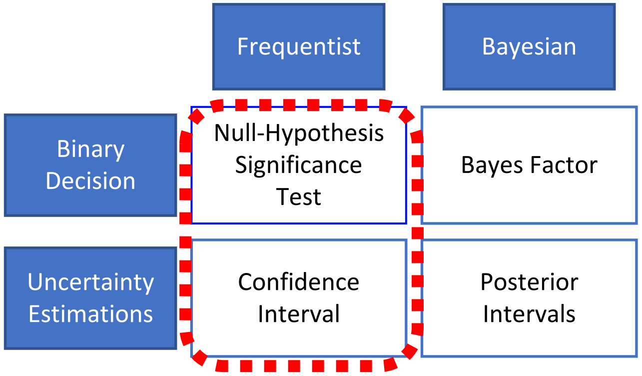
Tips and Suggestions

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

Define the research hypothesis you are after:

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

Tips and Suggestions



- **If using frequentist tests:**
  - The statements reporting p-value and confidence interval need to be precise.
  - … so that the results are not misinterpreted.
    - The term “significant” should be used with caution and clear purpose in order to not cause any misinterpretations.

      better under a significance test != *significantly better*
    - One way to achieve this is by using adjectives “statistical” or “practical” before any (possibly inflected) usage of “significance.”

Tips and Suggestions

	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 Pro-

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 evalua-

Lots of good tips about:

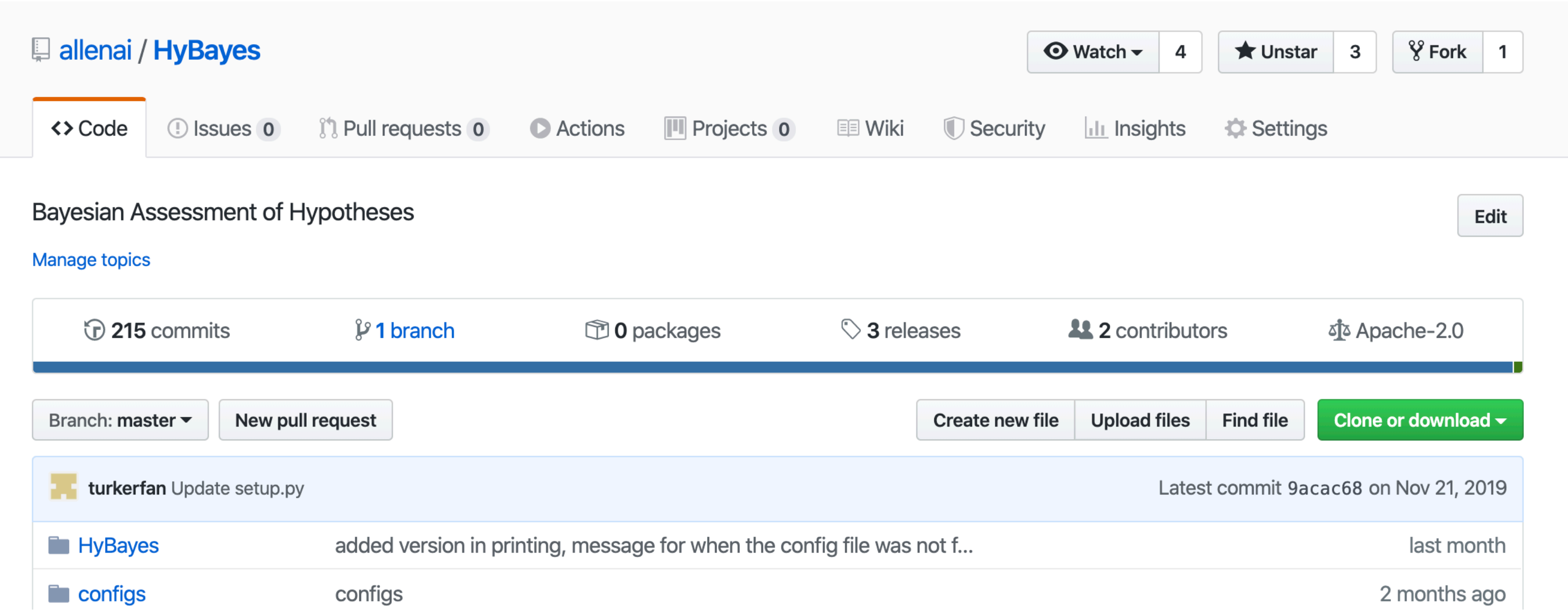
- Selecting the right "test"
- How to report your results.

Tips and Suggestions

- **If using Bayesian tests:**
  - Be clear about your hierarchical model, any parameters in the model and the choice of priors.
  - Comment on the certainty (or the lack of) of your inference.

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

HyBayes Package



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

Nov 2019

Abstract

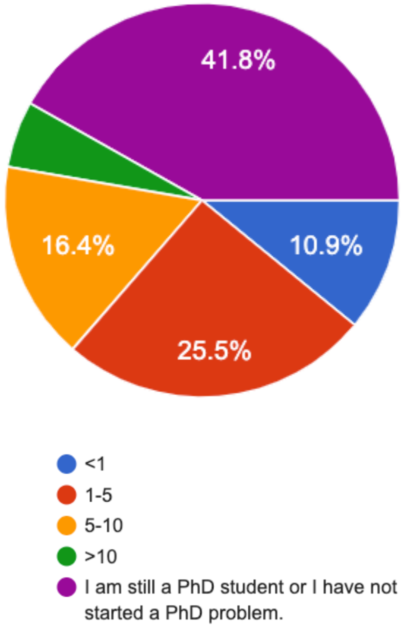
Empirical research in Natural Language Processing (NLP) has adopted a narrow set of principles for assessing hypotheses, relying mainly on p-value computation, which suffers from several known issues. While alternative proposals have been well-debated and adopted in other fields, they remain rarely discussed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

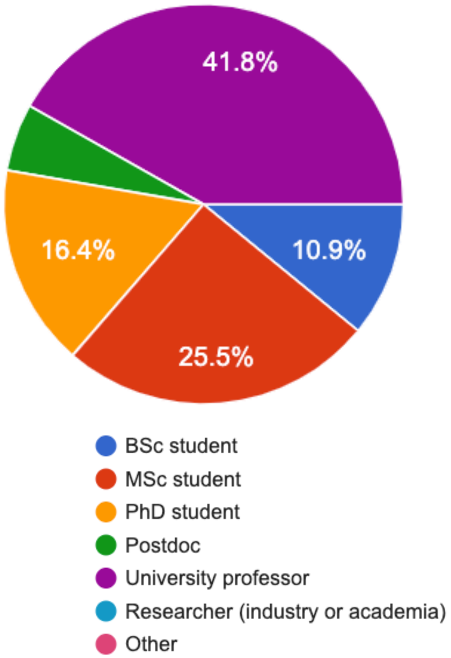
System ID	Description	ARC-easy #Correct	ARC-easy Acc.	ARC-challenge #Correct	ARC-challenge Acc.
S_1	BERT	1721	72.4	566	48.3
S_2	Reading Strategies	1637	68.9	496	42.3

Table 1: Performance of two systems (Devlin et al., 2019; Sun et al., 2018) on the ARC question-answering dataset (Clark et al., 2018). ARC-easy & ARC-challenge have 2376 & 1172 instances, respectively. Acc.: accuracy as a percentage.

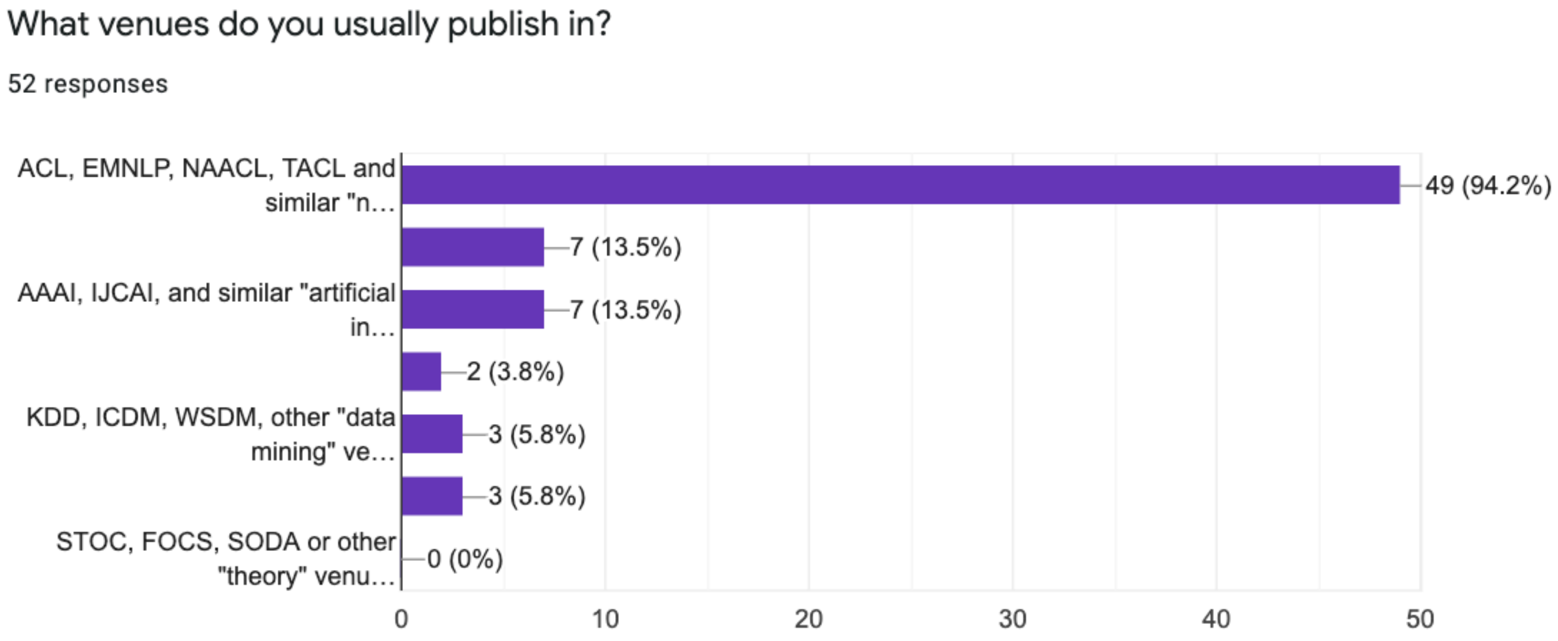
That’s it!

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.”*