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

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

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 this gap by contrasting various hypothesis assessment techniques, especially those not commonly used in the field (such as evaluations based on Bayesian inference). Since these statistical techniques differ in the hypotheses they can support, we argue that practitioners should first decide their target hypothesis before choosing an assessment method. This is crucial because common fallacies, misconceptions, and misinterpretation surrounding hypothesis assessment methods often stem from a discrepancy between what one would like to claim versus what the method used actually assesses. Our survey reveals that these issues are omnipresent in the NLP research community. As a step forward, we provide best practices and guidelines tailored towards NLP research, as well as an easy-to-use package called **`HyBayes`** for Bayesian assessment of hypotheses,[1] complementing existing tools.

1 Introduction

Empirical fields, such as Natural Language Processing (NLP), must follow scientific principles for assessing hypotheses and drawing conclusions from experiments. For instance, suppose we come across the results in Table 1, summarizing the accuracy of two question-answering (QA) systems S_1 and S_2 on some datasets. What is the correct way to interpret this empirical observation in terms of the superiority of one system over another? While S_1 has higher accuracy than S_2 in both cases, the gap is moderate and the datasets are of limited size. Can this apparent difference in performance be explained simply by random chance, or do we have sufficient evidence to conclude that S_1 is in fact *inherently different* (in particular, inherently stronger) than S_2 on these datasets? If the latter, can we quantify this gap in inherent strength while accounting for random fluctuation?

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.

Such fundamental questions arise in one form or another in every empirical NLP effort. Researchers often wish to draw conclusions such as:

> ***(Ca)*** *I'm 95% confident that S_1 and S_2 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 for ARC-easy.*
>
> ***(Cb)*** *With probability at least 95%, the inherent accuracy of S_1 exceeds that of S_2 by at least 1% for ARC-easy.*

These two conclusions differ in two respects. First, ***Ca*** claims the two systems are inherently different, while ***Cb*** goes further to claim a margin of at least 1% between their inherent accuracies. The second, more subtle difference lies in the interpretation of the 95% figure: the 95% confidence expressed in ***Ca*** is in terms of the space of empirical observations we could have made, given some underlying truth about how the inherent accuracies of S_1 and S_2 relate; while the 95% probability expressed in ***Cb*** is directly over the space of possible

*Work done while the second author was affiliated with the University of Pennsylvania.

[1]https://github.com/allenai/HyBayes

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5715–5725
July 5 - 10, 2020.

inherent accuracies of the two systems.

To support such a claim, one must turn it into a proper mathematical statement that can be validated using a statistical calculation. This in turn brings in additional choices: we can make at least four statistically *distinct* hypotheses here, each supported by a different statistical evaluation:

> ***(H1)*** *Assuming S_1 and S_2 have inherently identical accuracy, the probability (**p-value**) of making a hypothetical observation with an accuracy gap at least as large as the empirical observation (here, 3.5%) is at most 5% (making us 95% confident that the above assumption is false).*
>
> ***(H2)*** *Assuming S_1 and S_2 have inherently identical accuracy, the empirical accuracy gap (here, 3.5%) is larger than the maximum possible gap (**confidence interval**) that could hypothetically be observed with a probability of over 5% (making us 95% confident that the above assumption is false).*
>
> ***(H3)*** *Assume a prior belief (a probability distribution) w.r.t. the inherent accuracy of typical systems. Given the empirically observed accuracies, the probability (**posterior interval**) that the inherent accuracy of S_1 exceeds that of S_2 by a margin of 1% is at least 95%.*
>
> ***(H4)*** *Assume a prior belief (a probability distribution) w.r.t. the inherent accuracies of typical systems. Given the empirically observed accuracies, the odds increase by a factor of 1.32 (**Bayes factor**) in favor of the hypothesis that the inherent accuracy of S_1 exceeds that of S_2 by a margin of 1%.*

As this illustrates, there are multiple ways to formulate empirical hypotheses and support empirical claims. Since each hypothesis starts with a different assumption and makes a (mathematically) different claim, it can only be tested with a certain set of statistical methods. Therefore, *NLP practitioners ought to define their target hypothesis before choosing an assessment method.*

The most common statistical methodology used in NLP is null-hypothesis significance testing (NHST) which uses p-values (Søgaard et al., 2014; Koehn, 2004; Dror and Reichart, 2018). Hypotheses ***H1&H2*** can be tested with p-value-based methods, which include confidence intervals and operate over the *probability space of observations*[2] (§2.1 and §2.2). On the other hand, there are often overlooked approaches, based on Bayesian inference (Kruschke and Liddell, 2018), that can be used to assess hypotheses ***H3&H4*** (§2.3 and §2.4) and have two broad strengths: they can deal more naturally with accuracy margins and they operate directly over the *probability space of inherent accuracy* (rather than of observations).

For each technique reviewed in this work, we discuss how it compares with alternatives and summarize common misinterpretations surrounding it (§3). For example, a common misconception about p-value is that it represents *a probability of the validity of a hypothesis*. While desirable, p-values in fact do not provide such a probabilistic interpretation (§3.2). It is instead through a Bayesian analysis of the posterior distribution of the test statistic (inherent accuracy in the earlier example) that one can make claims about the probability space of that statistic, such as ***H3***.

We quantify and demonstrate related common malpractices in the field through a manual annotation of 439 ACL-2018 conference papers,[3] and a survey filled out by 55 NLP researchers (§4). We highlight surprising findings from the survey, such as the following: While 86% expressed fair-to-complete confidence in the interpretation of p-values, only a small percentage of them correctly answered a basic p-value interpretation question.

Contributions. This work seeks to inform the NLP community about crucial distinctions between various statistical hypotheses and their corresponding assessment methods, helping move the community towards well-substantiated empirical claims and conclusions. Our exposition covers a broader range of methods (§2) than those included in recent related efforts (§1.1), and highlights that these methods achieve different goals. Our surveys of NLP researchers reveals problematic trends (§4), emphasizing the need for increased scrutiny and clarity. We conclude by suggesting guidelines for better testing (§5), as well as providing a toolkit called `HyBayes` (cf. Footnote 1) tailored towards commonly used NLP metrics. We hope this work will encourage an improved understanding of statistical assessment methods and effective reporting practices with measures of uncertainty.

1.1 Related Work

While there is an abundant discussion of significance testing in other fields, only a handful of NLP efforts address it. For instance, Chinchor (1992) defined the principles of using hypothesis testing in the context of NLP problems. Most-notably, there are works studying various randomized tests (Koehn, 2004; Ojala and Garriga, 2010; Graham et al., 2014), or metric-specific tests (Evert, 2004). More recently, Dror et al. (2018) and Dror and Reichart (2018) provide a thorough review of

[2] More precisely, over the probability space of an aggregation function over observations, called test statistics.

[3] https://www.aclweb.org/anthology/events/acl-2018/

frequentist tests. While an important step in better informing the community, it covers a subset of statistical tools. Our work complements this effort by pointing out alternative tests.

With increasing over-reliance on certain hypothesis testing techniques, there are growing troubling trends of misuse or misinterpretation of such techniques (Goodman, 2008; Demšar, 2008). Some communities, such as statistics and psychology, even have published guidelines and restrictions on the use of p-values (Trafimow and Marks, 2015; Wasserstein et al., 2016). In parallel, some authors have advocated for using alternate paradigms such as Bayesian evaluations (Kruschke, 2010).

NLP is arguably an equally empirical field, yet with a rare discussion of proper practices of scientific testing, common pitfalls, and various alternatives. In particular, while limitations of p-values are heavily discussed in statistics and psychology, only a few NLP efforts approach them: over-estimation of significance by model-based tests (Riezler and Maxwell, 2005), lack of independence assumption in practice (Berg-Kirkpatrick et al., 2012), and sensitivity to the choice of the significance level (Søgaard et al., 2014). Our goal is to provide a unifying view of the pitfalls and best practices, and equip NLP researchers with Bayesian hypothesis assessment approaches as an important alternative tool in their toolkit.

2 Assessment of Hypotheses

We often wish to draw qualitative inferences based on the outcome of experiments (for example, inferring the relative inherent performance of systems). To do so, we usually formulate a hypothesis that can be *assessed* through some analysis.

Suppose we want to compare two systems on a dataset of instances $\boldsymbol{x} = [x_1, \ldots, x_n]$ with respect to a measure $\mathcal{M}(S, x)$ representing the performance of a system S on an instance x. Let $\mathcal{M}(S, \boldsymbol{x})$ denote the vector $[\mathcal{M}(S, x_i)]_{i=1}^n$. Given systems S_1, S_2, define $\boldsymbol{y} \triangleq [\mathcal{M}(S_1, \boldsymbol{x}), \mathcal{M}(S_2, \boldsymbol{x})]$ as a vector of observations.[4]

In a typical NLP experiment, the goal is to infer some *inherent* and *unknown* properties of systems. To this end, a practitioner assumes a probability distribution on the observations $\boldsymbol{y}$, parameterized by $\boldsymbol{\theta}$, the properties of the systems. In other words, $\boldsymbol{y}$ is assumed to have a distribution[5] with unknown parameters $\boldsymbol{\theta}$. In this setting, a *hypothesis* H is a condition on $\boldsymbol{\theta}$. Hypothesis assessment is a way of evaluating the degree to which the observations $\boldsymbol{y}$ are compatible with H. The overall process is depicted in Figure 1.

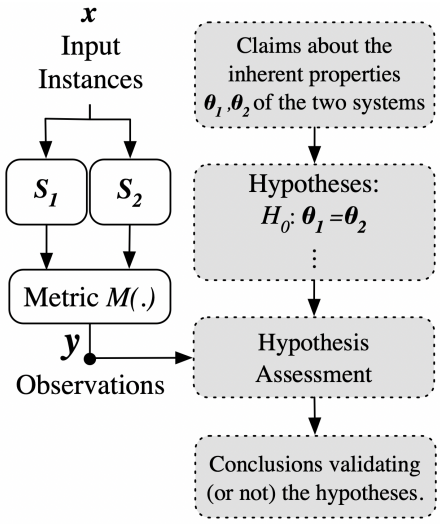

Figure 1: Progression of steps taken during a scientific assessment of claims from empirical observations.

Following our running example, we use the task of answering natural language questions (Clark et al., 2018). While our examples are shown for this particular task, all the ideas are applicable to more general experimental settings.

For this task, the performance metric $\mathcal{M}(S, x)$ is defined as a binary function indicating whether a system S answers a given question x correctly or not. The performance vector $\mathcal{M}(S, \boldsymbol{x})$ captures the system's accuracy on the entire dataset (cf. Table 1). We assume that each system S_i has an unknown *inherent accuracy* value, denoted θ_i. Let $\boldsymbol{\theta} = [\theta_1, \theta_2]$ denote the unknown inherent accuracy of two systems. In this setup, one might, for instance, be interested in assessing the credibility of the hypothesis H that $\theta_1 < \theta_2$.

Table 2 shows a categorization of statistical tools developed for the assessment of such hypotheses. The two tools on the left are based on *frequentist* statistics, while the ones on the right are based on *Bayesian* inference (Kruschke and Liddell, 2018). A complementary categorization of these tools is based on the nature of the results that they provide: the ones on the top encourage *binary* decision mak-

[4] For simplicity of exposition, we assume the performances of two systems are on a single dataset. However, the discussion also applies to observations on multiple different datasets.

[5] *Parametric* tests assume this distribution, while *non-parametric* tests do not.

	Frequentist	*Bayesian*
Binary Decision	***(Section 2.1)*** *Null-Hypothesis Significance Test*	***(Section 2.4)*** *Bayes Factor*
Uncertainty estimation	***(Section 2.2)*** *Confidence Intervals*	***(Section 2.3)*** *Posterior Intervals*

Table 2: Various classes of methods for statistical assessment of hypotheses.

ing, while those on the bottom provide uncertainty around estimates. We discuss all four classes of tests in the following sub-sections.

2.1 Null-Hypothesis Significance Testing

In frequentist hypothesis testing, there is an asymmetric relationship between two hypotheses. The hypothesis formulated to be rejected is usually called the *null-hypothesis* H_0. For instance, in our example H_0: $\theta_1 = \theta_2$. A decision procedure is devised by which, depending on $\boldsymbol{y}$, the null-hypothesis will either be *rejected* in favor of H_1, or the test will stay *undecided*.

A key notion here is p-value, the probability, under the null-hypothesis H_0, of observing an outcome at least equal to or *extreme* than the empirical observations $\boldsymbol{y}$. To apply this notion on a set of observations $\boldsymbol{y}$, one has to define a function that maps $\boldsymbol{y}$ to a numerical value. This function is called the *test statistic* $\delta(.)$ and it formalizes the interpretation of *extremeness*. Concretely, p-value is defined as,

$$\mathbb{P}(\delta(Y) \geq \delta(\boldsymbol{y})|H_0) \qquad (1)$$

In this notation, Y is a random variable over possible observations and $\delta(\boldsymbol{y})$ is the empirically observed value of the test statistic.

A large p-value implies that the data could easily have been observed under the null-hypothesis. Therefore, a lower p-value is used as evidence towards rejecting the null-hypothesis.

Example 1 (Assessment of *H1*) *We form a null-hypothesis using the accuracy of the two systems (Table 1) using a one-sided z-test [a] with $\delta(\boldsymbol{y}) \triangleq (1/n)\sum_{i=1}^{n}\left[\mathcal{M}(S_1, x_i) - \mathcal{M}(S_2, x_i)\right]$. We formulate a null-hypothesis against the claim of S_1 having strictly better accuracy than S_2. This results in a p-value of 0.0037 (details in §A.1) and can be interpreted as the following: if the systems have inherently identical accuracy values, the probability of observing a superiority at least as extreme as our observations is 0.0037. For a significance level of 0.05 (picked before the test) this p-value is small enough to reject the null-hypothesis.*

[a] The choice of this test is based on an implicit assumption that two events corresponding to answering two distinct questions, are independent with identical probability, i.e., equal to the inherent accuracy of the system. Hence, the number of correct answers follows a binomial distribution. Since, the total number of questions is large, i.e., 2376 in ARC-easy, this distribution can be approximated with a normal distribution. It is possible to use other tests with less restrictive assumptions (see Dror et al. (2018)), but for the sake of simplicity we use this test to illustrate core ideas of "p-value" analysis.

This family of the tests is thus far the most widely used tool in NLP research. Each variant of this test is based on some assumptions about the distribution of the observations, under the null-hypothesis, and an appropriate definition of the test statistics $\delta(.)$. Since a complete exposition of such tests is outside the scope of this work, we encourage interested readers to refer to the existing reviews, such as Dror et al. (2018).

2.2 Confidence Intervals

Confidence Intervals (CIs) are used to express the uncertainty of estimated parameters. In particular, the 95% CI is the range of values for parameter $\boldsymbol{\theta}$ such that the corresponding test based on p-value is not rejected:

$$\mathbb{P}(\delta(Y) \geq \delta(\boldsymbol{y})|H_0(\boldsymbol{\theta})) \geq 0.05. \qquad (2)$$

In other words, the confidence interval merely asks which values of the parameter $\boldsymbol{\theta}$ could be used, before the test is rejected.

Example 2 (Assessment of *H2*) *Consider the same setting as in Example 1. According to Table 1, the estimated value of the accuracy differences (maximum-likelihood estimates) is $\theta_1 - \theta_2 = 0.035$. A 95% CI of this quantity provides a range of values that are not rejected under the corresponding null-hypothesis. In particular, a 95% CI gives $\theta_1 - \theta_2 \in [0.0136, 0.057]$ (details in §A.2). The blue bar in Figure 2 (right) shows the corresponding CI. Notice that the conclusion of Example 1 is compatible with this CI; the null-hypothesis $\theta_1 = \theta_2$ which got rejected is not included in the CI.*

2.3 Posterior Intervals

Bayesian methods focus on prior and posterior distributions of $\boldsymbol{\theta}$. Recall that in a typical NLP experiment, these parameters can be, e.g., the *actual* mean or standard deviation for the performance of a system, as its inherent and unobserved property.

In Bayesian inference frameworks, a priori assumptions and beliefs are encoded in the form of a *prior* distribution $\mathbb{P}(\boldsymbol{\theta})$ on parameters of the model.[6] In other words, a prior distribution describes the common belief about the parameters of the model. It also implies a distribution over possible observations. For assessing hypotheses ***H3*** and ***H4*** in our running example, we will simply use the uniform prior, i.e., the inherent accuracy is uniformly distributed over $[0, 1]$. This corresponds to having no prior belief about how high or low the inherent accuracy of a typical QA system may be.

In general, the choice of this prior can be viewed as a compromise between the beliefs of the analyzer and those of the audience. The above uniform prior, which is equivalent to the Beta(1,1) distribution, is completely non-committal and thus best suited for a broad audience who has no reason to believe an inherent accuracy of 0.8 is more likely than 0.3. For a moderately informed audience that already believes the inherent accuracy is likely to be widely distributed but centered around 0.67, the analyzer may use a Beta(3,1.5) prior to evaluate a hypothesis. Similarly, for an audience that already believes the inherent accuracy to be highly peaked around 0.75, the analyzer may want to use a Beta(9,3) prior. Formally, one incorporates $\boldsymbol{\theta}$ in a hierarchical model in the form of a *likelihood function* $\mathbb{P}(\boldsymbol{y}|\boldsymbol{\theta})$. This explicitly models the underlying process that connects the latent parameters to the observations. Consequently, a *posterior* distribution is inferred using the Bayes rule and conditioned on the observations: $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\mathbb{P}(\boldsymbol{y}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\boldsymbol{y})}$.

The posterior distribution is a combined summary of the data and prior information, about likely values of $\boldsymbol{\theta}$. The mode of the posterior (maximum a posteriori) can be seen as an estimate for $\boldsymbol{\theta}$. Additionally, the posterior can be used to describe the *uncertainty* around the mode.

While the posterior distribution can be analytically calculated for simple models, it is not so straightforward for general models. Fortunately, recent advances in hardware, Markov Chain Monte Carlo (MCMC) techniques (Metropolis et al., 1953; Gamerman and Lopes, 2006), and probabilistic programming[7] allow sufficiently-accurate numerical approximations of posteriors.

One way to summarize the uncertainty around the point estimate of parameters is by marking the span of values that cover $\alpha\%$ of the most-credible density in the posterior distribution (e.g., $\alpha = 95\%$). This is called *Highest Density Intervals* (HDIs) or Bayesian Confidence Intervals (Oliphant, 2006) (not to be confused with CI, in §2.2).

Recall that a hypothesis H is a condition on $\boldsymbol{\theta}$ (see Figure 1). Therefore, given the posterior $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{y})$, one can calculate the probability of H, as a probabilistic event, conditioned on $\boldsymbol{y}$: $\mathbb{P}(H|\boldsymbol{y})$.

For example in an unpaired t-test, H_0 is the event that the means of two groups are equal. Bayesian statisticians usually relax this strict equality $\theta_1 = \theta_2$ and instead evaluate the credibility of $|\theta_1 - \theta_2| < \varepsilon$ for some small value of ε. The intuition is that when θ_1 and θ_2 are close enough they are *practically* equivalent. This motivates the definition of *Region Of Practical Equivalence* (ROPE): An interval around zero with "negligible" radius. The boundaries of ROPE depend on the application, the meaning of the parameters and its audience. In our running example, a radius of one percent for ROPE implies that improvements less than 1 percent are not considered notable. For a discussion on setting ROPE see Kruschke (2018).

These concepts give researchers the flexibility to define and assess a wide range of hypotheses. For instance, we can address ***H3*** (from Introduction) and its different variations that can be of interest depending on the application. The analysis of ***H3*** is depicted in Figure 2 and explained next.[8]

Example 3 (Assessment of *H3*) *Recall the setting from previous examples. The left panel of Figure 2 shows the prior on the latent accuracy of the systems and their differences (further details on the hierarchical model in §A.3.) We then obtain the posterior distribution (Figure 2, right), in this case via numerical methods).*

Notice that one can read the following conclusion: with probability 0.996*, the hypothe-*

[6] We use $\mathbb{P}(x)$ in its most general form, to denote the Probability Mass Function for discrete variables and the Probability Density Function for continuous variables.

[7] Pymc3 (in Python) and JAGS & STAN (in R) are among the commonly-used packages for this purpose.

[8] Figure 2 can be readily reproduced via the accompanying software, `HyBayes`.

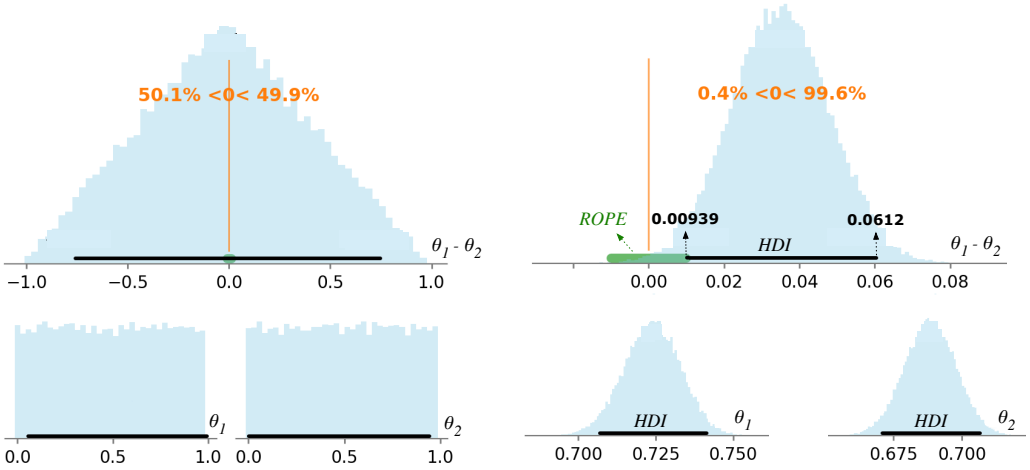

Figure 2: Left: Prior distributions of two systems (bottom row) and their difference (top row). Right: Posterior distributions of two systems (bottom row) and their difference (top row) after observing the performances on ARC-easy dataset. Note the posterior HDI estimate, $(0.00939, 0.0612)$. Here we assume at least one percent accuracy difference to be considered practically different. Hence, we indicate the interval $(-0.01, 0.01)$ as *ROPE* (§2.3.)

sis **H3** *(with a margin of 0%) holds true. As explained in §C.2, this statement does not imply any difference with a notable margin. In fact, the posterior in Figure 2 implies that this experiment is not sufficient to claim the following: with probability at least* 0.95*, hypothesis* **H3** *(with a margin of 1%) holds true. This is the case since ROPE* $(0.01, 0.01)$ *overlaps with* 95% *HDI* $(0.00939, 0.0612)$.

2.4 Bayes Factor

A common tool among Bayesian frameworks is the notion of *Bayes Factor*.[9] Intuitively, it compares how the observations $\boldsymbol{y}$ shift the credibility from prior to posterior of the two competing hypothesis:

$$BF_{01} = \frac{\mathbb{P}(H_0|\boldsymbol{y})}{\mathbb{P}(H_1|\boldsymbol{y})} \Big/ \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$$

If the BF_{01} equals to 1 then the data provide equal support for the two hypotheses and there is no reason to change our a priori opinion about the relative likelihood of the two hypotheses. A smaller Bayes Factor is an indication of rejecting the null-hypothesis H_0. If it is greater than 1 then there is support for the null-hypothesis and we should infer that the odds are in favor of H_0.

Notice that the symmetric nature of Bayes Factor allows all the three outcomes of "accept", "reject", and "undecided," as opposed to the definition of p-value that cannot accept a hypothesis.

[9] *"Bayesian Hypothesis Testing"* usually refers to the arguments based on *"Bayes Factor."* However, as shown in §2.3, there are other Bayesian approaches for assessing hypotheses.

Example 4 (Assessment of *H4*) *Here we want to assess the null-hypothesis* H_0*:* $|\theta_1 - \theta_2| < 0.01$ *against* H_1*:* $|\theta_1 - \theta_2| \geq 0.01$ *(*$x = 0.01$*). Substituting posterior and prior values, one obtains:*

$$BF_{01} = \frac{0.027}{0.980} \Big/ \frac{0.019}{0.972} = 1.382$$

. This value is very close to 1 *which means that this observation does not change our prior belief about the two systems difference.*

3 Comparisons

Many aspects influence the choice of an approach to assess significance of hypotheses. This section provides a comparative summary, with details in Appendix C and an overall summary in Table 3.

3.1 Susceptibility to Misinterpretation

The complexity of interpreting significance tests combined with insufficient reporting could result in ambiguous or misleading conclusions. This ambiguity can not only confuse authors but also cause confusion among readers of the papers.

While p-values (§2.1) are the most common approach, they are inherently complex, which makes them easier to misinterpret (see examples in §C.1). Interpretation of confidence intervals (§2.2) can also be challenging since it is an extension of p-value (Hoekstra et al., 2014). Approaches that provide measures of uncertainty directly in the hypothesis space (like the ones in §2.3) are often more

Method	Paradigm	Ease of interpretation (1 =easy) (§3.1)	Encourages binary-thinking (3.2)	Depends on stopping intention (3.3)	Dependence on prior (3.4)	Decision rule	# of papers using this test in ACL'18
(§2.1) p-value	frequentist	3	Yes	Yes	No	Acceptable p-value	73
(§2.2) CI	frequentist	4	No	Yes	No	Acceptable confidence margin	6
(§2.3) HDI	Bayesian	1	No	No	Not sensitive but takes it into account	HDI relative to ROPE	0
(§2.4) BF	Bayesian	2	Yes	No	Highly sensitive	Acceptable BF	0

Table 3: A comparison of different statistical methods for evaluating the credibility of a hypothesis given a set of observations. The total number of published papers in at the ACL-2018 conference is 439.

natural choices for reporting the results of experiments (Kruschke and Liddell, 2018).

3.2 Measures of Certainty

A key difference is that not all methods studied here provide a measure of uncertainty over the hypothesis space. For instance, p-values (§2.1) do *not* provide probability estimates on two systems being different (or equal) (Goodman, 2008). On the contrary, they encourage *binary* thinking (Gelman, 2013), that is, confidently concluding that one system is better than another, without taking into account the extent of the difference between the systems. CIs (§2.2) provide a range of values for the target parameter. However, this range also does not have any *probabilistic* interpretation in the hypothesis space (du Prel et al., 2009). On the other hand, posterior intervals (§2.3) generally provide a useful summary as they capture probabilistic estimates of the correctness of the hypothesis.

3.3 Dependence on Stopping Intention

The process by which samples in the test are collected can affect the outcome of a test. For instance, the sample size n (whether it is determined before the process of gathering information begins, or is a random variable itself) can change the result. Once observations are recorded, this distinction is usually ignored. Hence, the testing algorithms that do not depend on the distribution of n are more desirable. Unfortunately, the definition of p-value (§2.1) depends on the distribution of n. For instance, Kruschke (2010, §11.1) provides examples where this subtlety can change the outcome of a test, even when the final set of observations is identical.

3.4 Sensitivity to the Choice of Prior

The choice of the prior can change the outcome of Bayesian approaches (§2.3 & §2.4). Decisions of Bayes Factor (§2.4) are known to be sensitive to the choice of prior, while posterior estimates (§2.3) are less so. For further discussion, see C.4 or refer to discussions by Sinharay and Stern (2002); Liu and Aitkin (2008) or Dienes (2008).

4 Current Trends and Malpractices

This section highlights common practices relevant to the our target approaches. To better understand the common practices or misinterpretations in the field, we conducted a survey. We shared the survey among ∼450 NLP researchers (randomly selected from ACL'18 Proceedings) from which 55 individuals filled out the survey. While similar surveys have been performed in other fields (Windish et al., 2007), this is the first in the NLP community, to the best of our knowledge. Here we review the main highlights (see Appendix for more details and charts).

Interpreting p-values. While the majority of the participants have a self-claimed ability to interpret p-values (Figure 9f), many choose its imprecise interpretation *"The probability of the observation this extreme happening due to pure chance"* (the popular choice) vs. a more precise statement *"Conditioned on the null hypothesis, the probability of the observation this extreme happening."* (see *Q1* & *Q2* in Appendix B.)

The use of CIs. Even though 95% percent of the participants self-claimed the knowledge of CIs (Figure 9e), it is rarely used in practice. In an annotation done on ACL'18 papers by two of the authors, only 6 (out of 439) papers were found to use CIs.

The use of Bayes Factors. A majority of the participants had "heard" about "Bayesian Hypothesis Testing" but did not know the definition of "Bayes Factor" (Figure 3). HDIs (discussed in §2.3) were the least known. We did not find any papers in ACL'18 that use Bayesian tools.

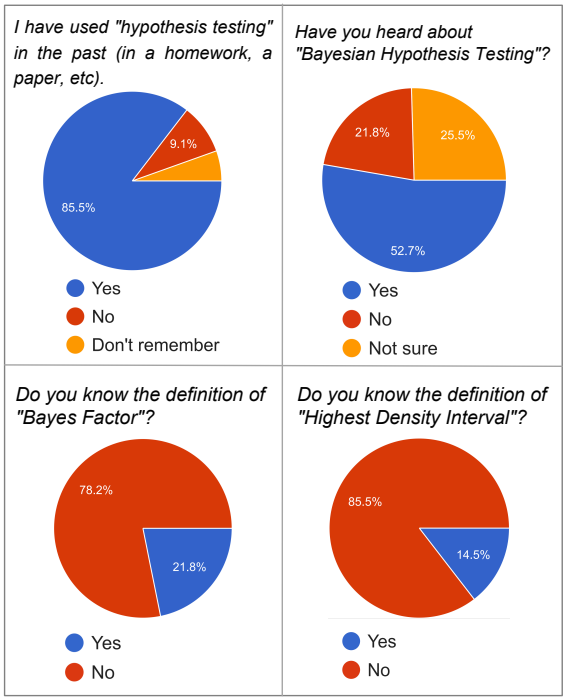

Figure 3: Select results from our survey.

The use of "significan*". A notable portion of NLP papers express their findings by using the term "significant" (e.g., *"our approach significantly improves over X."*) Almost all ACL'18 papers use the term "significant"[10] somewhere. Unfortunately, there is no single universal interpretation of such phrases across readers. In our survey, we observe that when participants read "X significantly improves Y" in the abstract of a hypothetical paper:

1. About 82% expect the claim to be backed by "hypothesis testing"; however, only 57% expect notable empirical improvement (see *Q3* in Appendix B);
2. About 35% expect the paper to test "practical significance", which is not generally assessed by popular tests (see §C.2);
3. A few also expect a theoretical argument.

Recent trends. Table 3 provides a summary of the techniques studied here. We make two key observations: (i) many papers don't use any hypothesis assessment method and would benefit from one; (ii) from the final column, p-value based techniques clearly dominate the field, a clear disregard to the advantages that the bottom two alternatives offer.

[10] Or other variants "significantly", "significance", etc.

5 Recommended Practices

Having discussed common issues, we provide a collection of recommendations (in addition to the prior recommendations, such as by Dror et al. (2018)).

The first step is to define your goal. Each of the tools in §2 provides a distinct set of information. Therefore, one needs to formalize a hypothesis and consequently the question you intend to answer by assessing this hypothesis. Here are four representative questions, one for each method:

1. *Assuming that the null-hypothesis is true, is it likely to witness observations this extreme? (§2.1)*
2. *How much my null-hypothesis can deviate from the mean of the observations until a p-value argument rejects it. (§2.2)*
3. *Having observed the observations, how probable is my claimed hypothesis?(§2.3)*
4. *By observing the data how much do the odds increase in favor of the hypothesis?(§2.4)*

If you decide to use **frequentist tests**:

- Check if your setting is compatible with the assumptions of the test. In particular, investigate if the meaning of null-hypothesis and sampling distribution match the experimental setting.
- Include a summary of the above investigation. Justify unresolved assumption mismatches.
- Statements reporting p-value and confidence interval must be precise enough so that the results are not misinterpreted (see §3.1).
- The term "significant" should be used with caution and clear purpose to avoid misinterpretations (see §4). One way to achieve this is by using adjectives "statistical" or "practical" before any (possibly inflected) usage of "significance."
- Often times, the desired conclusion is a notable margin in the superiority of one system over another (see §3). In such cases, a pointwise p-value argument is *not* sufficient; a confidence interval analysis is needed. If CI is inapplicable for some reason, this should be mentioned.

If you decide to use **Bayesian approaches**:

- Since Bayesian tests are less known, it is better to provide a short motivation for the usage.
- Familiarize yourself with packages that help you decide a hierarchical model, e.g., the software provided here. If necessary, customize these models for your specific problem.
- Be clear about your hierarchical model, including model parameters and priors. In most cases, these choices should be justified (see §2.3.)

Statistical Model	Observation Type	Hierarchical Model	Assumptions	Parameters	Common settings / metrics	Common Frequentist test (Parametric)	Common Frequentist test (Non-Parametric)
Binary model	binary output	Bernoulli distribution with Beta prior	2	For each group: p ∈ [0,1] (success probablity)	correct vs incorrect predictions	Binomial test	bootstrap / permutation
Binomial model	binomial output	Binomial distribution with Beta prior	2,3,6	For each group: p ∈ [0,1] (success probablity)	Exact match, Accuracy, Recall, UAS (sentencelevel), LAS (sentencelevel)*	Binomial test	bootstrap / permutation
Metric model	metric observations	T-Student distribution with muliple priors *	1,2,4	For each group: mu ∈ **R** and sigma ∈ **R+** Shared between groups: nu ∈ **R+** (normality parameter)	Exact match, Accuracy, Recall, UAS (sentencelevel), LAS (sentencelevel), running time. energy usage, L2 error	t-test	bootstrap / permutation
Count model	counts	Negative Binomial distribution with Normal prior	2,5	For each group: mu ∈ **R+** (rate parameter) alpha ∈ **R+** (shape parameter)	The count of certain patterns an algorithm could find in a big pool, in a fixed amount of time. Notice that you can't convert this into a ratio form, since there is no well-defined denominator. Ex: measuring how many of questions could be answered correctly (from an infinite pool of questions) by a particular QA systems, in a limited minute (the system is allowed to skip the questions too)		bootstrap / permutation
Ordinal model	ordinals	Normal distribution with parameterized tresholds	2	For each group: mu ∈ **R** and sigma ∈ **R+** Shared between groups: thresholds between possibe levels	Collection of objects/labels arranged in a certain ordering, not necessarily with a metric distance between them; for example sentiment labels (https://www.aclweb.org/anthology/S16-1001.pdf), product review categories, grammaticality of sentences		bootstrap / permutation

Assumption 1: The observations are distributed as a t-student with unknown normality parameter (a normal distribution with potentially longer tales).
Assumption 2: The observations from each group are assumed to be i.i.d, conditioned on the inherent characterstics of two systems
Assumption 3: The total number of instances (the denominators) is known.
Assumption 4: The variable is inherently continuous, or the granularity (the denominator) is high enough to treat the variable as continuous.
Assumption 5: The observations follow a Negative-Binomail / Poisson distribution.
Assumption 6: The observations follow a binomial-distribution.
Assumption 7: The observations follow a normal distribution.
* In this model (unlike frequentist t-test) outliers don't need to be discarded manually to realize the strict normality assumption.

Table 4: Select models supported by our package **HyBayes** at the time of this publication.

- Comment on the certainty (or the lack of) of your inference in terms of HDI and ROPE: (I) is HDI completely inside ROPE, (II) they are completely disjoint, (III) HDI contains values both inside and outside ROPE (see §2.3.)
- For reproducibility, include further details about your test: MCMC traces, convergence plots, etc. (Our **HyBayes** package provides all of this.)
- Be wary that Bayes Factor is highly sensitive to the choice of prior (see §3.4). See Appendix §C.4 for possible ways to mitigate this.

5.1 Package HyBayes

We provide an accompanying package, **HyBayes**, to facilitate comparing systems using the two Bayesian hypothesis assessment approaches discussed earlier: (a) posterior probabilities and (b) Bayes Factors. (Several packages are already available for frequentist assessments.)

Table 4 summarizes common settings in which **HyBayes** can be employed[11] in NLP research, including typical use cases, underlying data assumptions, recommended hierarchical model, metrics (accuracy, exact match, etc.), and frequentist tests generally used in these cases. These settings cover several typical assumptions on observed NLP data. However, if a user has specific information on observations or can capitalize on other assumptions, we recommend adding a custom model, which can be done relatively easily.

[11]These settings are available at the time of this publication, with more options likely to be added in the future.

6 Conclusion

Using well-founded mechanisms for assessing the validity of hypotheses is crucial for any field that relies on empirical work. Our survey indicates that the NLP community is not fully utilizing scientific methods geared towards such assessment, with only a relatively small number of papers using such methods, and most of them relying on p-value.

Our goal was to review different alternatives, especially a few often ignored in NLP. We surfaced various issues and potential dangers of careless use and interpretations of different approaches. We do not recommend a particular approach. Every technique has its own weaknesses. Hence, a researcher should pick the right approach according to their needs and intentions, with a proper understanding of the techniques. Incorrect use of any technique can result in misleading conclusions.

We contribute a new toolkit, **HyBayes**, to make it easy for NLP practitioners to use Bayesian assessment in their efforts. We hope that this work provides a *complementary* picture of hypothesis assessment techniques for the field and encourages more rigorous reporting trends.

Acknowledgments

The authors would like to thank Rotem Dror, Jordan Kodner, and John Kruschke for invaluable feedback on an early version of this draft. This work was partly supported by a gift from the Allen Institute for AI and by DARPA contracts FA8750-19-2-1004 and FA8750-19-2-0201.

References

Valentin Amrhein, Fränzi Korner-Nievergelt, and Tobias Roth. 2017. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ*, 5:e3544.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of EMNLP*, pages 995–1005.

James O Berger and Thomas Sellke. 1987. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–122.

Nancy Chinchor. 1992. The statistical significance of the MUC-4 results. In *Proceedings of the 4th conference on Message understanding*, pages 30–50. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *CoRR*, abs/1803.05457.

Janez Demšar. 2008. On the appropriateness of statistical tests in machine learning. In *Workshop on Evaluation Methods for Machine Learning in conjunction with ICML*, page 65.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Zoltan Dienes. 2008. *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhikers guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, pages 1383–1392.

Rotem Dror and Roi Reichart. 2018. Recommended statistical significance tests for NLP tasks. *arXiv preprint arXiv:1809.01448*.

Stefan Evert. 2004. Significance tests for the evaluation of ranking methods. In *Proceedings of COLING*.

Dani Gamerman and Hedibert F Lopes. 2006. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Andrew Gelman. 2013. The problem with p-values is how they're used.

Steven Goodman. 2008. A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, 45(3):135–140.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274.

Rink Hoekstra, Richard D Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review*, 21(5):1157–1164.

Jeehyoung Kim and Heejung Bang. 2016. Three common misuses of p values. *Dental hypotheses*, 7(3):73.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

John K Kruschke. 2010. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5):658–676.

John K Kruschke. 2018. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280.

John K Kruschke and Torrin M Liddell. 2018. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206.

Charles C Liu and Murray Aitkin. 2008. Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52(6):362–375.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087.

Markus Ojala and Gemma C Garriga. 2010. Permutation tests for studying classifier performance. *JMLR*, 11(Jun):1833–1863.

Travis E Oliphant. 2006. A Bayesian perspective on estimating mean, variance, and standard-deviation from data. Technical report, Brigham Young University. https://scholarsarchive.byu.edu/facpub/278/.

Jean-Baptist du Prel, Gerhard Hommel, Bernd Röhrig, and Maria Blettner. 2009. Confidence interval or p-value?: Part 4 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 106(19):335.

Stefan Riezler and John T Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 57–64.

Sandip Sinharay and Hal S Stern. 2002. On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, 56(3):196–201.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Héctor Martínez Alonso. 2014. What's in a p-value in NLP? In *Proceedings of CoNLL*, pages 1–10.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2018. Improving machine reading comprehension with general reading strategies. In *Proceedings of NAACL*.

David Trafimow and Michael Marks. 2015. Editorial. *Basic and Applied Social Psychology*, 37(1):1–2.

Ronald L Wasserstein, Nicole A Lazar, et al. 2016. The ASAs statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.

Donna M Windish, Stephen J Huot, and Michael L Green. 2007. Medicine residents' understanding of the biostatistics and results in the medical literature. *Jama*, 298(9):1010–1022.