Scientific research heavily relies on statistical hypothesis testing—for instance, to evaluate theories, or to assess the effectiveness of public policies. Such tests are informative when their size is properly controlled, so the frequency of rejecting true null hypotheses (type I error) stays below a prespecified nominal level. Publication bias, however, tends to exaggerate test sizes. Since scientists are typically only able to publish significant results, they have an incentive to conduct studies until reaching significance. Such a process leads to test sizes inflated above nominal significance levels, because the critical values used to determine significance assume that test statistics are constructed from a single study. This paper addresses this problem by constructing critical values that are compatible with scientists’ behavior given their incentives. We assume that researchers conduct studies until their test statistic exceeds the critical value, or until the expected benefit from conducting an extra study falls below the cost. When such incentive-compatible critical value (ICCV) is used to assess significance, researchers may conduct multiple studies; but readers know that a significant result would not occur more often than the nominal significance level when the null hypothesis is true. ICCVs are larger than classical critical values because researchers report the best result from multiple studies. Yet, for a broad range of research processes and calibrations, ICCVs stay in a fairly narrow range.
1. Introduction

Statistical hypothesis testing is a key tool for scientific investigation and discovery. It is used to evaluate existing theories and paradigms. In particular it allows for the identification of anomalies: instances where the theory does not accord well with empirical observations. It is also used to assess the effectiveness of new medical treatments, public policies, designs, processes, and other potential remedies to real-world problems. In that context it allows for the detection of robust positive effects.

Scientists conduct many hypothesis tests; but they are typically only able to publish those that reject the null hypothesis under investigation. This preference for “positive” or “significant” results is clearly visible in scientific journals: first identified in psychology journals (Sterling 1959; Bozarth and Roberts 1972), it has since been observed across the social sciences (Christensen, Freese, and Miguel 2019, chap. 3), medical sciences (Begg and Berlin 1988; Song et al. 2000; Ioannidis and Trikalinos 2007; Dwan et al. 2008), biological sciences (Csada, James, and Espie 1996; Jennions and Moeller 2002), and many other sciences (Head et al. 2015; Fanelli, Costas, and Ioannidis 2017). This preference reflects the presumption that “positive” or “significant” results are more valuable to scientific progress than “negative” or “insignificant” results.

Publications, however, are the main output of scientific research, and the main marker of scientific productivity. A scientist’s publications determine her career path: whether she will receive tenure; whether she will be promoted further; and at each step what salary she will command (Gibson, Anderson, and Tressler 2014). The publications also determine the scientist’s recognition and status in the profession. And the publications influence the resources made available to the scientist: research funds, grants, fellowships, and so on. The resources are required to pursue new projects and obtain additional publications.

Given that publishing yields significant rewards but requires significant results, scientists have the incentive to collect and analyze data until they are able to reject the null hypothesis they are investigating (Nosek, Spies, and Motyl 2012). Such behavior is well-documented: researchers inflate reported $t$-statistics by choosing appropriate specifications (Brodeur et al. 2016); they omit to write up experimental results that fail to attain significance (Franco, Malhotra, and Simonovits 2014); and they favor reporting statistically significant outcomes at the expense of non-significant outcomes (Dwan et al. 2008). And indeed, some flexibility in data collection and analysis affords researchers many opportunities to obtain significant results (Lovell 1983; Simmons, Nelson, and Simonsohn 2011).

Scientists’ behavior, however, invalidates the assumptions of classical statistics on which hypothesis testing is based. One key assumption is that scientists report whatever they observe
in a single data sample that is representative of an underlying population of interest. Thus, the use of standard critical values based upon this assumption to determine statistical significance in scientific publications leads to the over-rejection of true null hypotheses. For instance, when using a standard critical value of 1.96 for a two-sided \( t \)-test of nominal level 5%, a true null hypothesis will be rejected much more often than 5% of the time. Such statistical bias is referred to as publication bias: “the systematic error induced in a statistical inference by conditioning on the achievement of publication status” (Begg and Berlin 1988, p. 422).

Publication bias is problematic because the informativeness of hypothesis tests relies upon properly controlling the size of the test: in order to limit spurious findings of statistical significance, one must be confident that the frequency of rejecting true null hypotheses does not exceed a pre-specified nominal level. When using data to test theory, if the size of a test is too large, too many anomalies will accumulate and perfectly sound paradigms may become disfigured or abandoned. In applied science, if the size is too large, ineffective methods, remedies, or policies will be implemented in the real world, and the problems they were designed to alleviate may persist or even worsen.

To address such size distortions in hypothesis testing, we construct critical values that are compatible with the incentives faced by scientists. Our incentive-compatible critical values (ICCVs) therefore control test size by the nominal significance level. To construct ICCVs, we model the strategic behavior of researchers. Researchers face costs and benefits from conducting studies, and these incentives determine how many studies researchers conduct.

Imagine for example that the researcher conducts two-sided \( t \)-tests at a nominal significance level of 5%. The usual critical value for such tests based upon the large sample standard normal distribution of a \( t \)-statistic under the null hypothesis is 1.96. Now imagine that the research collects a random sample of data, performs the test, and fails to reject the null hypothesis because the observed absolute \( t \)-statistic is below 1.96. Given that such a result is unlikely to be published, if the relative cost is low enough, the researcher may collect another sample of data and analyze it, in isolation or pooled with the previous sample, in a further attempt to reject the null hypothesis. This implies that the critical value of 1.96, based upon the large sample distribution of a \( t \)-statistic from a single study in isolation, is generally not incentive-compatible: it is built upon the assumption that researchers collect one data sample, but it generally incentivizes them to collect more than one (that is, so long as the benefits from publication are large enough).

Indeed, for a given critical value, researchers have the incentive to continue collecting data until either their test statistic exceeds the critical value, or the expected benefit from data collection falls below the collection cost. A critical value that is incentive-compatible needs to take this behavior into account to control tests size by the nominal significance level. When the ICCV is
used, researchers may still conduct several studies. Nevertheless, using information about the
cost of conducting research, the rewards to publication, and data-collection process, we can infer
the maximum number of studies that the researcher has the incentive to conduct at any given
critical value. From this, we can compute an upper bound on the distribution of the reported
test statistic, and therefore make sure that the critical value achieves the desired significance
level. When using our ICCVs, readers can be confident that a published rejection of a true null
hypothesis occurs no more frequently than the nominal level of the test.

Because researchers often conduct more than one study before submitting their results to a
journal, our ICCVs are larger than standard critical values. In particular, when scientific rewards
are higher or data-collection costs are lower, the ICCVs are larger. Using both theoretical and
numerical results, we obtain ICCVs for a broad range of researcher behavior, research costs,
publication rewards, and researcher prior knowledge; for two-sided tests with 5% nominal
significance level, we find ICCVs between 2 and 3.5 instead of the standard 1.96. For fields with
better estimates of the cost and benefits of research and of prior knowledge, it is possible to
obtain more precise ICCVs; however, our results indicate that ICCVs are fairly insensitive to these
inputs, yielding convenient rules-of-thumb for practical application.

Other methods to control test size: modeling researcher behavior. The issue of publication bias is of
course well-known, and several methods have been developed to address various aspects of it
(Christensen, Freese, and Miguel 2019; Wasserstein, Schirm, and Lazar 2019). In this paper we
focus on one type of publication bias: test size distortions. Such distortions mean that nominal
significance levels (say, 5%) understate the actual probability of rejecting a true null hypothesis,
such that positive results may be due much more to chance—and much less to the null hypothesis
being false—than nominally stated.

To address size distortions, we opt to model researcher behavior and that this behavior into
account to undo the distortions. In spirit our approach is close to that of Glaeser (2006): he
assumes that researchers do not report the result of one study, but the most extreme result
from \( n > 1 \) studies, and subsequently proposes corrections to debias estimates arising from
researchers’ behavior. The problem is that in practice, such technique is bound to unravel: when
 Corrections based upon this assumed, fixed behavior are implemented, researchers may then
conduct additional studies in search of results that overwhelm the initial corrections. Because our
method takes into account researchers’ behavioral response to critical values, it is not subject to
this limitation and does indeed properly control the size of the published test results. Our method
factors in a certain behavior that is compatible with the incentives arising from the critical value.
Once the critical value is in place, researchers will adapt their behaviors to the new incentive
structure, but the change would already have been accounted for by our critical value.

Another seemingly related method is to lower the nominal significance level from 5% down to 0.5% (Benjamin et al. 2018). Such change would make it harder to reach significance, but it would not help control the size of hypothesis tests. In the same way that the standard critical value for 5% significance level (1.96 for a two-sided test) does not deliver a size of 5%, the standard critical value for 0.5% significance level (2.81 for a two-sided test) will not deliver a size of 0.5%. Although researchers will reject fewer null hypotheses, readers would remain unable to know how much of a rejection is due to chance and how much is due to the null hypothesis being false.

**Other methods to control test size: constraining researcher behavior.** A difference approach to control test size consist in constraining researchers’ behavior through a pre-analysis plan (PAP), so that they indeed only conduct one study and report the results from that one unique study (Miguel et al. 2014; Olken 2015; Coffman and Niederle 2015). With such constraints, researchers’ behavior conforms to the assumptions from classical statistics and publication bias disappears. In exchange for abiding by a PAP, journals could promise to publish results irrespective of significance as long as the research design is of sufficient quality (Christensen, Freese, and Miguel 2019, pp. 110–112). Results-blind review would eliminate journals’ current preference for significance and thus researchers’ incentive to work within the boundaries of the PAP to obtain significance.

Christensen, Freese, and Miguel (2019, pp. 107–117) discuss the strengths but also the limitations of PAPs. Among the main limitations are that PAPs prevent scientists from engaging in exploratory analysis, although it is often a source of new ideas and scientific discoveries. With observational data, an additional problem is that it would difficult to ensure that the PAP is written before the data are observed.\(^1\) The ICCV and PAP approaches therefore appear complementary. Relying on ICCVs could be particularly helpful in two settings. First, because the ICCV approach is much more flexible than the PAP approach, it could be useful for more innovative, less explored research questions, for which it is surely fruitful to let researchers explore the data without any constraints. Second, because the ICCV approach does not require to have never seen the data, it could be useful with observational data, which are often familiar to researchers (for instance, the time series of unemployment and nominal interest rates in macroeconomics).

\(^1\)Gelman and Loken (2014) concur. They note that “For most of our own research projects [a PAP] hardly seems possible: In our many applied research projects, we have learned so much by looking at the data. Our most important hypotheses could never have been formulated ahead of time…. In any case, as applied social science researchers we are often analyzing public data on education trends, elections, the economy, and public opinion that have already been studied by others many times before, and it would be close to meaningless to consider preregistration for data with which we are already so familiar…. The most valuable statistical analyses often arise only after an iterative process involving the data. Preregistration may be practical in some fields and for some types of problems, but it cannot realistically be a general solution.”
**Methods to correct other facets of publication bias.** The focus of this paper is on controlling the size of hypothesis tests, which is inflated by publication bias. Publication bias also takes other forms, which our methodology cannot address, and which require additional corrections. For example, our methodology does not address the fact that only significant results are published, and almost all non-significant results remain unpublished, or even unwritten. Such selectivity is a problem for meta-analyses of literatures: a large subset of all studies may never appear in print, biasing the evidence included in meta-analyses (Rosenthal 1979). This bias is very different from ours; it would occur even if scientists were not strategic and did not respond to incentives: even then the sample of published studies would be truncated at the significance level. Numerous methods have been developed to address this bias, and they would continue to be useful in meta-analyses even if ICCV replaced standard critical values.²

2. **Research and publication process**

We start by modeling the process of research and publication.

We assume that researchers in a scientific community are interested in inferring the true value of a scalar parameter \( \beta \in \mathbb{R} \). In particular, we focus on problems for which scientists wish to test a particular null hypothesis for the value that \( \beta \) takes. In economics, the parameter \( \beta \) often takes the form of a causal or treatment effect, regression coefficient or parameter of interest in a structural economic model.

For concreteness, we focus on the most common form of hypotheses used in economics, two-sided hypotheses of the form

\[
H_0 : \beta = \beta_0 \quad \text{versus} \quad H_1 : \beta \neq \beta_0,
\]

where \( H_0 \) and \( H_1 \) constitute the null and alternative hypotheses of interest. In the common application of statistical significance testing, \( \beta_0 \) takes the value of zero. The analysis here extends straightforwardly to one-sided hypotheses.

After performing statistical tests of (1), researchers communicate their findings in journals. We assume that journal editors only wish to publish papers that reject the prevailing null hypothesis characterized by \( H_0 \). We characterize journals as wishing to publish “significant” or “non-null” results, as is common practice in economics and many other disciplines. Researchers gather data to form a \( t \)-statistic for testing (1) and report the value of this statistic to a journal.

Common practice in applied work is to then deem $H_0$ as rejected or the result of the research as “statistically significant” if the absolute value of the statistic exceeds an appropriate critical value (CV). Researchers receive an expected payoff $v > 0$ for producing a study that rejects $H_0$. The value that expected payoff $v$ takes will depend upon the journal the researcher intends to submit to and the research question itself. For example, a researcher receives a higher expected payoff for submitting their work to a more prestigious journal but a lower expected payoff if that journal does not find the research question very interesting.

It is common practice to derive the CV that determines rejection of $H_0$ from the upper quantiles of a standard normal distribution as this is the approximate large-sample distribution of a $t$-statistic under standard assumptions when $H_0$ holds. One of these standard assumptions is that only a single model and/or dataset is used to form the $t$-statistic. However, this assumption is violated in the majority of applied research. Typically, a researcher will look through many model specifications or sets of data to form statistics. Given a researcher’s positive payoff from publishing their work, we work under the assumption that the researcher actually reports the maximum absolute value of a set of $t$-statistics, generated across different sets of data or model specifications, to maximize his chances of rejecting $H_0$. Formally, we model the researcher as constructing a sequence of $n$ $t$-statistics $X_1^*, \ldots, X_n^*$ from $n$ “studies” that are unobservable to the editor. As is standard, we assume the latent $t$-statistics take the form

$$X_i^* = \frac{\hat{\beta}_i - \beta_0}{\text{se}(\hat{\beta}_i)},$$

where $\hat{\beta}_i$ denotes the estimator of $\beta$ in study $i$ and $\text{se}(\hat{\beta}_i)$ denotes a consistent standard error estimator for $\hat{\beta}_i$. After having conducted $n$ studies, the researcher reports $X_n = \max\{|X_1^*|, \ldots, |X_n^*|\}$ to the journal upon submission of his research study. By standard central limit theorem and standard error consistency arguments, each of the latent $t$-statistics $X_i^*$ is approximately normally distributed with unit variance. In particular under $H_0$, $X_1^* \sim \mathcal{N}(0, 1)$. Unless $n = 1$, the use of a standard CV equal to the $1 - \alpha/2$ quantile of a standard normal distribution will result in over-rejection of a true null hypothesis as the reported statistic is distributed as the maximum absolute value of a sequence of normal random variables. That is, the probability of rejecting $H_0$ when it is true is greater than $\alpha$. This over-rejection is precisely the form of publication bias we aim to correct in this paper.

The researcher will determine when to stop conducting additional studies based upon a marginal cost-benefit analysis. After having conducted $n - 1$ studies, the researcher receives the expected payoff of $v$ if and only if $X_n$ exceeds the CV $z$ used by the editor to determine a statistically significant rejection of $H_0$. If $X_{n-1} < z$, the researcher will choose to conduct the $n^{th}$
study if the expected marginal payoff of conducting this additional study exceeds the expected cost of doing so. Formally, after having conducted $n$ studies, we model the expected benefit of conducting the $n^{th}$ study as

$$v \mathbb{P}(X_n > z \mid X_{n-1}^{*}) \mathbb{I}(X_{n-1} < z),$$

where $X_n^* = (X_1^*, \ldots, X_n^*)$. We may wish to allow the expected cost of a study to depend upon the specific nature of that study. The costs of studies can vary due factors such as different costs of acquiring data or running experiments or differences in the opportunity costs of different studies. In the most general version of our model, we thus assume the researcher incurs an expected marginal cost of $c(n) \geq 0$ for conducting the $n^{th}$ study. Having already conducted $n-1$ studies, the researcher’s expected marginal profit from conducting an additional study is thus equal to

$$\mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid X_{n-1}^*) = v \mathbb{P}(X_n > z \mid X_{n-1}^*) \mathbb{I}(X_{n-1} < z) - c(n)$$

(3)

$$= v \mathbb{P}(\max\{|X_n^*|, X_{n-1}^*\} > z \mid X_{n-1}^*) \mathbb{I}(X_{n-1} < z) - c(n),$$

where $\mathbb{P}$ is the researcher’s subjective probability measure that incorporates his beliefs on the true value of $\beta$.

Since the editor cannot observe the number of latent studies conducted by the researcher, the best the editor can do is to obtain the maximum number of studies $N(z)$ it would be profitable for the researcher to engage in given that the CV threshold is set to $z$. In other words, we seek to characterize the maximum $n = N(z)$ for which $\mathbb{E}(\pi(X_{n-1}, X_n; z, v, c) \mid X_{n-1}^*) > 0$ with positive probability. This maximum number of studies will only be obtained in the case that no test statistics $|X_1^*|$ through $|X_{N(z)-1}^*|$ cross the threshold $z$ but since the editor does not observe these statistics, he cannot know whether this has occurred. However upon submission of research to the journal, the editor then knows that the value of the reported test statistic can be no larger than $X_{N(z)}$. If the editor chooses the CV threshold by looking at the upper $\alpha$-quantile of $X_{N(z)}$ under $H_0$, he can therefore be confident that when $H_0$ holds he will receive a submission with at most probability $\alpha$. Since $N(z)$ endogenously depends upon $z$, the editor should then wish to find the value of $z$ at which the probability that $X_{N(z)}$ exceeds $z$ is no larger than $\alpha$ when $H_0$ is true. This will guarantee that over repeated submissions of results in favor of $H_1$, less than $100 \cdot \alpha\%$ of these results will be false when $H_0$ is the truth.

Our goal in this paper is to thus characterize the value of the CV $z = z^*$ that, based upon the researcher’s incentives, does not lead to over-rejection of a true $H_0$. That is, we seek to find the
fixed point $z^*$ such that

\begin{equation}
P_{H_0}(X_{N(z^*)} > z^*) = \alpha,
\end{equation}

where $P_{H_0}$ denotes the objective probability measure given that the null hypothesis holds. We define this $z^*$ as the incentive-compatible critical value since the researcher has no incentive to conduct additional studies that would result in rejecting a true $H_0$ with probability larger than the nominal level of $\alpha$.

We now study different settings that give rise to different forms of cost functions, subjective probability measures and joint distributions of latent studies, leading us to different conclusions on how $z^*$ ought to be chosen by the journal editor so that (4) holds.

### 3. Illustrative calibration

In the next sections we construct ICCVs under different assumptions about researchers and the research process. In these different cases we derive some theoretical results and compute the ICCVs in Monte Carlo simulations. Here, as a preamble, we calibrate the parameters of our model (table 1). We will use these calibrated values in the simulations. The calibration also illustrates how the model can be mapped to the actual research process.

This illustrative calibration is based on Pedro Dal Bo’s paper: Dal Bo (2005). There are several advantages to the paper. First, it is based on laboratory experiments, so it is easy to define each experiment as one iid study. Second, everything in the paper is very well documented, which will be helpful to calibrate the parameters. Third, the paper is based on a specific question, with a well delineated literature, which will be key to calibrate reasonable prior beliefs for Pedro.

Conceptually, the null hypothesis explored by Pedro is $H_0$: people do not take the future into account in strategic decisions; the alternative hypothesis is $H_1$: people do take the future into account in strategic decisions.

The hypothesis is explored in 8 experiments. Pedro compares the level of cooperation in

---

**TABLE 1. Parameter values in simulations**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.73</td>
<td>Mean of $t$-statistic prior</td>
<td>Roth and Murnighan (1978); Murnighan and Roth (1983)</td>
</tr>
<tr>
<td>0.30</td>
<td>Standard deviation of $t$-statistic prior</td>
<td>Roth and Murnighan (1978); Murnighan and Roth (1983)</td>
</tr>
<tr>
<td>$v = 5000$</td>
<td>Expected value of a publication</td>
<td>Gibson, Anderson, and Tressler (2014)</td>
</tr>
<tr>
<td>$c = 933$</td>
<td>Cost of a study</td>
<td>Dal Bo (2005)</td>
</tr>
<tr>
<td>$e = 1$</td>
<td>Cost elasticity</td>
<td>–</td>
</tr>
</tbody>
</table>
repeated prisoner’s dilemma games that have different probabilities of continuation. In some treatments, the prisoner’s dilemma games are one-shot. The only equilibrium is no cooperation, so players are not expected to cooperate. In other treatments, the probability of continuation—which governs the probability of future interaction between players—is 1/2 or 3/4. In these treatments, while no cooperation remains an equilibrium outcome, cooperation equilibria also appear. Then, the likelihood of cooperation between players is expected to increase.

Formally, the statistical parameter of interest $\beta$ is the increase in the cooperation probability between games in which theory predicts no cooperation and games in which theory predicts possible cooperation. The null and alternative hypotheses become

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta > 0.$$  

The first parameters to calibrate are the mean and standard deviation of Pedro’s prior belief about the estimate of $\beta$ obtained from the experiments, $\hat{\beta}$. Before Pedro’s study, there were two studies on the same topic: Roth and Murnighan (1978) and Murnighan and Roth (1983). Pedro would have used these two precursors to form a belief about what the mean and standard deviation of $\hat{\beta}$ could be. In fact, Pedro reports the results from these studies early in his paper (Dal Bo 2005, table 1). Four values of $\hat{\beta}$ are obtained, from two treatments in the two studies. Taking the weighted average of these values, we obtain the mean of the prior belief about the estimated parameter: $E(\hat{\beta}) = 14.99$. Then, measuring the dispersion of the four values of $\hat{\beta}$, we obtain the standard deviation of the prior belief about the estimated parameter: $sd(\hat{\beta}) = 3.99$. From these, we obtain the mean and standard deviation of the $t$-statistic.

The laboratory where the experiments were conducted was available at no cost, so the marginal cost of an experiment is the cost to pay the participants and the research assistant monitoring them. On average, there are 48.75 subjects in each experiment, and each subject earns $18.94, so paying the participants costs $48.75 \times $18.94 = $923.3. In addition each experiment lasts one hour, so the research assistant must be paid for that time, which we assume costs $10. The average cost of an experiment therefore is $c = $923.3 + $10 = $933.3.4

Pedro’s objective is to reject the null hypothesis that people are not forward-looking. Such result opens the door to a publication in a top economics journal—and indeed Pedro’s work was published in the American Economic Review (AER). What is the value of such publication? Using data from the University of California system, Gibson, Anderson, and Tressler (2014) find that the

---

3The weights account for the fact that one study has 121 subjects while the other is much larger, with 252 subjects.

4There is also a fixed cost to the project, incurred before any experiment, which includes the time spent designing the experiments and researching the literature. But since this is a fixed cost, it does not influence the marginal decision of conducting an extra experiment or not.
publication of 10 AER-quality pages increases academic earnings by 1.3%, which translates to a net present value over an entire career of $28,466. Since the length of Pedro’s article is 14 pages, its value is $28,466 \times 1.4 = $39,853. The same would be true of articles published in the Quarterly Journal of Economics, Econometrica, and Journal of Political Economy. The value of publishing in other journals can be inferred from their relative ranking compared to these top journals (see Gibson, Anderson, and Tressler 2017, table A1).

Of course obtaining a significant result is not sufficient to publish in the top journals: each of them publishes less than 5% of the articles they receive (Card and DellaVigna 2013; Zheng and Kaiser 2016). Depending on the correlation structure between the decisions of these top journals, as well as the correlations with decisions of lower-ranked journals, we could compute the expected net present value of submitting a significant result. As a preliminary step, we set the value to $v = $5,000.

4. Baseline: an impossibility result

To build intuition, we begin with the simplest version of our model of research and publication.

Consider a researcher who may conduct a sequence of independent studies to determine the value of $\beta$, where the estimators $\hat{\beta}_i$ are constructed from the data gathered in each study. These estimators are statistically identical and, appealing to a standard central limit theorem, approximately distributed as $N(\beta, \text{var}(\hat{\beta}))$ in large samples. This case would arise for example when the studies are experiments, the sample size in each experiment is the same and the observations in each experiment are drawn from the same underlying population. Since the researcher does not know the true value $\beta$, we treat it as a random variable from his point of view and write $\hat{\beta}_i | \beta \sim i.i.d N(\beta, \text{var}(\hat{\beta}))$. In this setting, the latent $t$-statistics are independent and identically distributed approximately following a normal distribution:

$$X_i^* | \theta \sim i.i.d N(\theta, 1) \quad \text{where} \quad \theta = \frac{\beta - \beta_0}{\text{sd}(\hat{\beta})}$$

is a scaled version of the difference between the true value $\beta$ and its null version hypothesized $\beta_0$ with $\text{sd}(\hat{\beta})$ being the true population standard deviation of $\hat{\beta}$. Note that in this setting, the null and alternative hypotheses (1) are equivalent to

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0.$$ 

Suppose the researcher incurs a constant cost of $c(n + 1) = c \geq 0$ for conducting each
additional study. The researcher has a prior distribution on $\theta$ of $F(\theta)$. This prior distribution may arise, for example, as the result of previous studies conducted by other researchers or as the biases of the researcher. For now we will assume that the researcher does not learn about $\theta$ in the process of conducting his studies so that his subjective distribution on $\theta$ does not change as he conducts more studies. Specializing (3) to the current context, the expected marginal profit from conducting an additional study after conducting $n - 1$ studies already is equal to

$$\mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid X_{n-1}^*) = v \int \mathbb{P}(|X_n|^* > z \mid X_{n-1}, \theta) dF(\theta) \mathbb{I}(X_{n-1} < z) - c$$

(5)

$$= v \int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta) \mathbb{I}(X_{n-1} < z) - c.$$  

At stage $n-1$, the researcher engages in another study if and only if $\mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid X_{n-1}^*) > 0$. That is, he will conduct another study if and only if $X_{n-1} < z$ and

$$\int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta) \geq \frac{c}{v}.$$  

(6)

Now we can make a couple observations:

1. Depending upon the values $c, v, z$ and $\theta$, either the researcher conducts zero studies or continues to conduct studies indefinitely until $X_n^* > z$.

2. Low marginal cost $c$, high payoff $v$, low critical value $z$ or high $\theta$ lead to engagement in studies indefinitely searching for significance.

Given observation 1, it is actually impossible for the journal editor to set the CV $z$ to a value so that the probability that the observed $X_n$ exceeds $z$ is equal to the nominal level $\alpha$ under $H_0$. If $z$ is set low enough for (6) to hold, the researcher will continue to conduct studies until he attains rejection of $H_0$. Conversely, if the editor sets $z$ high enough so that (6) is violated, the researcher will never conduct a single study. In other words, depending upon the editors choice of $z$, the resulting size of the test will be either zero (and so will the power) or one. In both cases, the test becomes completely uninformative.

In the next three sections, we show that this impossibility result is a special case that relies on (i) a constant marginal cost, (iii) a researcher that does not learn and (iii) studies that are independent of each other.

---

5Since $\theta = \frac{\beta - \beta_0}{\text{sd}(\beta)}$, this prior can be elicited directly from a prior on the original parameter of interest $\beta$. 
5. Sampling increasingly costly data

We modify the analysis of section 4 by assuming that the expected cost of conducting an additional study is increasing in the number of studies already run. This modification can incorporate situations in which it becomes increasingly costly to collect new data or run new experiments or could simply capture increasing marginal opportunity costs on the part of the researcher. In terms of the marginal cost function introduced in section 2, we assume that $c : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing and $\lim_{n \to \infty} c(n) = \infty$. These assumptions ensure that new studies eventually become overwhelmingly costly so that a researcher eventually stops conducting new studies pertaining to (1).

We maintain all of the other assumptions of section 4 so that an analogous analysis leads us to the conclusion that at stage $n - 1$, the researcher engages in the $n$th study if and only if $X_{n-1} < z$ and

$$\int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] \, dF(\theta) \geq \frac{c(n)}{\nu}.$$  

This condition can be rewritten as

$$(7) \quad n \leq c^{-1} \left( \nu \int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] \, dF(\theta) \right),$$

where $c^{-1}$ is the inverse of the function $c$, which is well defined and strictly increasing on $\mathbb{R}_+$. The only quantity that the journal editor can choose to influence the number of studies conducted by the researcher is the CV $z$. Taking the point of view of the journal editor, we take $F$, $c$ and $\nu$ as fixed and define $N(z)$ to be the largest (weakly) positive integer that satisfies (7). Thus, $N(\cdot)$ is a function that maps $z \in \mathbb{R}$ into the positive integers. The researcher will stop conducting studies after the $N(z)$th as it will no longer be profitable to do so. If

$$\int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] \, dF(\theta) \leq \frac{c(1)}{\nu},$$

$N(z) = 0$ since it is not profitable to engage in any studies of (1). This occurs when the expected returns to examining the research question are too low because, given the researcher’s beliefs characterized by $F(\theta)$, the significance threshold $z$ is too large relative to $c(1)/\nu$. In fact there is a unique function of $\theta$, $c(1)$, and $\nu$, $Z(\theta, c(1), \nu)$, at which the researcher will decline to study the research question. More specifically, $N(z) \geq 1$ if and only if $z < Z(\theta, c(1), \nu)$. Furthermore, the function $z \mapsto \int [1 - \Phi(z - \theta)] \, dF(\theta)$ is strictly decreasing, so the right-hand side of (7) is strictly decreasing in $z$ (as $\nu > 0$ and $c^{-1}$ is strictly increasing). This implies that $N(z)$ is decreasing in $z$.

Let $\alpha$ be the nominal significance targeted by the editor. The ICCV is the smallest value of $z$
such that

\[(8) \quad P_{H_0}(X_{N(z)} > z) \leq \alpha.\]

Since the function \(z \mapsto \mathbb{P}(X_{N(z)} > z \mid H_0)\) is strictly decreasing from 1 to 0 when \(z\) goes from \(-\infty\) to \(+\infty\), the ICCV is well-defined for any \(\alpha \in (0,1).\) In fact, since this function is continuous, we know that the smallest value of \(z\) such that (8) holds is given by (4). Since \(X_n\) is the maximum of the absolute value of \(n\) mean-zero iid normal variables under \(H_0\), \(P_{H_0}(X_n \leq z) = [\Phi(z) - \Phi(-z)]^n\) for \(z \geq 0\) so that we can rewrite the (implicit) definition of \(z^*\) as

\[1 - [\Phi(z^*) - \Phi(-z^*)]^N(z^*) = \alpha.\]

In general, \(N(z)\) does not have an analytical solution: the computation of \(z^*\) involves an iterative fixed point algorithm. However, this computation can be simplified in at least three special cases for which we can obtain closed-form solutions for \(N(z)\).

The first special case is a mass point prior \(dF(\theta) = \delta(\hat{\theta})d\theta\) for some \(\hat{\theta} \geq 0\), where \(\delta(\cdot)\) denotes the Dirac delta function. In this case,

\[\int [\Phi(z - \theta) - \Phi(-z - \theta)]dF(\theta) = \Phi(z - \hat{\theta}) - \Phi(-z - \hat{\theta})\]

so that

\[N(z) = c^{-1}\left[v[1 - \Phi(z - \hat{\theta}) + \Phi(-z - \hat{\theta})]\right].\]

The second special case is a uniform prior \(dF(\theta) = (b - a)^{-1} 1(a \leq \theta \leq b)d\theta\) for some \(0 \leq a < b\). In this case,

\[\int [\Phi(z - \theta) - \Phi(-z - \theta)]dF(\theta) = (b - a)^{-1} \int_a^b [\Phi(z - \theta) - \Phi(-z - \theta)]d\theta\]

\[= (b - a)^{-1}\left\{\phi(z - b) - \phi(z - a) + (b - z)\Phi(z - b) + (z - a)\Phi(z - a)\right.\]

\[\left.\quad - \phi(-z - b) + \phi(-z - a) - (b + z)\Phi(-z - b) + (a + z)\Phi(-z - a)\right\}\]

so that

\[N(z) = [c^{-1}\left(v[1 - (b - a)^{-1}\left\{\phi(z - b) - \phi(z - a) + (b - z)\Phi(z - b) + (z - a)\Phi(z - a)\right.\right.\]

\[\left.\quad - \phi(-z - b) + \phi(-z - a) - (b + z)\Phi(-z - b) + (a + z)\Phi(-z - a)\right\}]].\]

The third and last special case is a normal prior \(dF(\theta) = \sigma^{-1}\phi((\theta - \mu)/\sigma)d\theta\) for some \(\mu \in \mathbb{R}\) and
\( \sigma > 0 \). In this case,

\[
\int [\Phi(z - \theta) - \Phi(-z - \theta)]dF(\theta) = \sigma^{-1} \int_{-\infty}^{\infty} [\Phi(z - \theta) - \Phi(-z - \theta)]\phi((\theta - \mu) / \sigma)d\theta
\]

\[
= \int_{-\infty}^{\infty} [\Phi(z - \mu - \sigma x) - \Phi(-z - \mu - \sigma x)]\phi(x)dx
\]

\[
= \Phi((z - \mu) / \sqrt{1 + \sigma^2}) - \Phi((-z - \mu) / \sqrt{1 + \sigma^2})
\]

so that

\[
N(z) = \left[ c^{-1}(v[1 - \Phi((z - \mu)/\sqrt{1 + \sigma^2}) + \Phi((-z - \mu)/\sqrt{1 + \sigma^2})]) \right].
\]

For each of these special cases, the editor may estimate the parameters entering the prior distributions using data from previous studies.

We make the following observations:

1. If the expected payoff from rejecting \( H_0 \), \( \nu \), is larger, the ICCV \( z^* \) is larger. This means that editors in high-status journal evaluating research about “important” questions should impose higher CVs. All else fixed, researchers have more incentive to search across many studies when the publication payoff is higher. A larger CV can be used by the editor to counteract the higher rate of false rejections induced by these high payoffs by reducing the incentive to conduct many studies.

2. If the cost of running experiments, \( c(n) \), is larger, the ICCV \( z^* \) is smaller. This means that editors evaluating research with high costs should impose smaller CVs (and vice versa). All else fixed, researchers have less incentive to conduct studies when they are costly. A smaller CV can be used by the editor to account for high research costs by increasing the incentive to conduct studies.

3. The more probability mass a researcher’s prior places on large values of \( \theta \), the larger the ICCV \( z^* \) will be. This means that editors evaluating studies for which researchers are more confident that \( H_0 \) is violated should impose a larger CVs. When they are more confident that \( H_0 \) is false, researchers expect larger payoffs from continuing to conduct studies. All else equal, this gives researchers more incentive to continue conducting studies on \( \theta \). As in case 1. above, a larger CV can be used by the editor to counteract the corresponding higher rate of false rejections.

4. In practice, the journal editor may not know the prior of the researcher. However, it is reasonable to estimate this prior from the outcomes of past studies, either parametrically
or nonparametrically. If there are no past studies to gather data from, it is reasonable to simply use a diffuse prior.

The ICCVs with increasingly costly studies, obtained by Monte-Carlo simulations, are illustrated in figure 1. A first insight from the figure is that the publication bias that we aim to correct—the distortion of the size of hypothesis tests—may be fairly large. The figure displays the size of a two-sided hypothesis test at the standard critical value of 1.96, for various priors (panel A), various cost/benefit ratios (panel C), and various cost elasticities (panel E). If researchers conducted only one study and performed the hypothesis test on it, as is assumed in classical statistics, the size would be 5%. But when researchers sample datasets, the size is much higher than 5%: as high as 30% in panel A, 100% in panel C, and 40% in panel E. So in this context, a significant result is much more likely to be due to luck than advertised by the nominal significance level of 5%.

A second insight from the figure is that the ICCVs required to bring back the actual test size to 5% are above the standard value of 1.96, but not by a tremendous amount. The figure displays the ICCV that delivers a size of 5% for two-sided tests, for various priors (panel B), various cost/benefit ratios (panel D), and various cost elasticities (panel F). Across parameterizations, the ICCV is above 1.96, but never above 3.5.

A third insight from the figure is that prior beliefs about the estimated parameter, and incentives faced by researchers, do not have a very strong impact on the ICCV. Across a broad range of priors and cost/benefit ratios from publication, the ICCV remains between 1.96 and 2.6. The cost elasticity, on the other hand, has an important impact on the ICCV.

6. Learning while sampling data

We now depart from section 4 by allowing the researcher to learn about the true value of $\theta$. To keep the analysis straightforward we assume that as he conducts successive studies, the researcher updates his prior distribution $F(\theta)$ according to Bayes’ rule but maintain all of the other assumptions of section 4. As we will see, learning about the true value of $\theta$ can induce the researcher to stop conducting additional studies, breaking the negative result of section 4. The ICCVs in that case are illustrated in figure 2.

After having conducted $n-1$ studies, instead of using his prior distribution $F(\theta)$ to compute the expected marginal profit from conducting an additional study according to (5), he will incorporate all of the information contained in the $n-1$ studies by using the posterior distribution $F(\theta|X_{n-1})$. Assuming that the researcher’s prior distribution $F(\theta)$ admits a pdf $f(\theta)$, for $n \geq 1$, the posterior
FIGURE 1. ICCVs at 5% significance level when researchers sample increasingly costly data

The figure considers researchers who attempt to obtain significance in two-sided hypothesis tests by sampling iid datasets that are increasingly costly to collect: collecting the nth dataset costs \( c(n) = c_0 \times n^\epsilon \), where \( \epsilon \) is the cost elasticity. All results are obtained from 1,000 Monte-Carlo simulations with the parameter values in table 1, except for those parameter values directly specified on the graph axes. Panels A, C, and E display the size of the hypothesis test at the standard critical value of 1.96, for various priors (panel A), various cost/benefit ratios (panel C), and various cost elasticities (panel E). If researchers collected only one dataset and conducted the test on it, as is assumed in classical statistics, the size would be 5%. But when researchers sample datasets, the size is much higher than 5%, sometimes as high as 100%. Panels B, D, and F display the ICCV that delivers a size of 5%, for various priors (panel B), various cost/benefit ratios (panel D), and various cost elasticities (panel F). Across parameterizations, the ICCV falls between 1.96 and 3.5.
pdf for \( \theta \) takes the form

\[ f(\theta|X_n^*) = \frac{f(X_n^*|\theta)f(\theta)}{\int f(X_n^*|\theta)f(\theta)d\theta}, \tag{9} \]

where the likelihood \( f(X_n^*|\theta) \) is the pdf of a \( N(\theta \iota_n, I_n) \) distribution evaluated at \( X_n^* \) with \( \iota_n \) denoting an \( n \)-vector of one's. This is the joint pdf corresponding to the random vector \( X_n^* \) when the value of \( \theta \) is known. As a notational convention, for \( n = 0 \), \( F(\theta|X_n^*) \) is equal to \( F(\theta_{n+1}) \).

Maintaining all of the other assumptions in section 4, we arrive at the conclusion that at stage \( n - 1 \), the researcher engages in the \( n \)th study if and only if \( X_{n-1} < z \) and

\[ \int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta|X_{n-1}^*) \geq \frac{c}{\theta}. \tag{10} \]

In contrast to the previous section, even if we know \( X_{n-1} < z \), whether or not the researcher chooses to engage in another study depends upon the realization of the random vector \( X_{n-1}^* \). For example, if the researcher obtains a large draw for \( |X_{n-1}^*| \) that is not quite large enough to cross the threshold \( z \), his posterior for \( \theta \) will be updated to shift probability mass toward larger values and (10) can hold. On the other hand, a very small draw for \( |X_{n-1}^*| \) can induce his posterior to shift probability mass toward very small values so that (10) can be violated. These facts break the impossibility result of section 4 so that we can define \( N(z) \) to be the largest \( n \) that satisfies (10) when the data are generated under \( H_0 \), noting that it is a random variable in this context.

The computation of \( z^* \) in this context is complicated by the fact that \( F(\theta|X_{n-1}^*) \) depends upon the realizations of prior studies, which are not observed by the journal editor. Nevertheless, since we know the distribution of \( X_{n-1}^* \) under \( H_0 \), namely \( N(0, I_{n-1}) \), \( z^* \) remains feasible to compute in the presence of learning by the researcher. Though \( N(z) \) is random here, a slight generalization of the argument given in the previous section provides that the ICCV given in (4) is well-defined. To gain intuition, suppose the researcher’s prior on \( \theta \) follows a \( N(\mu, \sigma^2) \) distribution. Using our knowledge of the distribution of \( X_n^* \), we have

\[
 f(X_n^*|\theta) = \frac{1}{\sqrt{(2\pi)^n}} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (X_i^* - \theta)^2 \right)
\]

\[
 f(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (\theta - \mu)^2 \right)
\]
so that

\[
\int f(X_n^*|\theta) f(\theta) d\theta = \frac{1}{\sqrt{(2\pi)^n \sigma^2}} \int \exp \left( -\frac{1}{2} \left[ \sum_{i=1}^{n} (\theta - X_i^*)^2 + \frac{1}{\sigma^2} (\theta - \mu)^2 \right] \right) d\theta
\]

\[
= \frac{1}{\sqrt{(2\pi)^n \sigma^2}} \int \exp \left( -\frac{1}{2} \left[ n + \frac{1}{\sigma^2} \right] (\theta - h)^2 \right) d\theta \exp(k)
\]

\[
= \frac{1}{\sqrt{(2\pi)^n (n\sigma^2 + 1)}} \exp(k),
\]

where

\[
h = \frac{\sigma^2 \sum_{i=1}^{n} X_i^* + \mu}{n\sigma^2 + 1},
\]

\[
k = -\frac{1}{2} \left( \sum_{i=1}^{n} X_i^* + \frac{\mu^2}{\sigma^2} \right) + \frac{\left( \sum_{i=1}^{n} X_i^* + \frac{\mu}{\sigma^2} \right)^2}{2 \left( n + \frac{1}{\sigma^2} \right)}.
\]

Thus,

\[
f(\theta|X_n^*) = \frac{f(X_n^*|\theta) f(\theta)}{\int f(X_n^*|\theta) f(\theta) d\theta}
\]

\[
= \frac{\sqrt{(2\pi)^n (n\sigma^2 + 1)}}{\sqrt{(2\pi)^n + \sigma^2}} \exp \left( -\frac{1}{2} \left[ \sum_{i=1}^{n} (\theta - X_i^*)^2 + \frac{1}{\sigma^2} (\theta - \mu)^2 + \frac{\left( \sum_{i=1}^{n} X_i^* + \frac{\mu}{\sigma^2} \right)^2}{n + \frac{1}{\sigma^2}} - \left( \sum_{i=1}^{n} X_i^* + \frac{\mu^2}{\sigma^2} \right) \right] \right)
\]

\[
= \frac{1}{\sqrt{2\pi \sigma^2_{n+1|n}}} \exp \left( -\frac{(\theta - \mu_{n+1|n})^2}{2\sigma^2_{n+1|n}} \right),
\]

where

\[
\mu_{n+1|n} = \frac{\sigma^2 \sum_{i=1}^{n} X_i^* + \mu}{n\sigma^2 + 1} \quad \text{and} \quad \sigma^2_{n+1|n} = \frac{\sigma^2}{n\sigma^2 + 1}.
\]

This means that $\theta|X_n^*$ has a normal distribution with mean $\mu_{n+1|n}$ and variance $\sigma^2_{n+1|n}$. These ingredients enable the computation of $z^*$ via straightforward Monte Carlo simulation.

This example also allows us to make a few observations:

- The mean of the posterior distribution $\mu_{n+1|n} = \frac{n\sigma^2}{n\sigma^2 + 1} \bar{X}^* + \frac{1}{n\sigma^2 + 1} \mu$ is a weighted average of the
sample mean of previous latent studies $\bar{X}^* \equiv n^{-1} \sum_{i=1}^n X_i^*$ and the prior mean $\mu$, where the former receives relatively more weight as the number of previous studies $n$ increases. For example, in the simplest case of only one previous study and a prior variance equal to the variance of that previous study ($\sigma^2 = 1$), the posterior mean is equal to the arithmetic average of the realization of the previous study and the prior mean.

- As the number of previous studies $n$ grows large, the posterior mean $\mu_{n+1|n}$ converges to the sample mean of the previous studies.

- As the number of previous studies $n$ grows large, the posterior variance $\sigma^2_{n+1|n} = \sigma^2/(n\sigma^2 + 1)$ shrinks toward zero. In conjunction with 2, this implies that for very large $n$ the posterior distribution concentrates heavily around the sample mean of the previous studies, which by the law of large numbers converges to the true value $\theta$. Letting the true value of $\theta$ be denoted by $\theta^*$ this implies that the researcher must eventually stop conducting studies if $1 - \Phi(z - \theta^*) + \Phi(-z - \theta^*) < c/v$. Since $\Phi(\cdot)$ is strictly increasing and $\lim_{x \to \infty} \Phi(x) = 1$ and $c/v > 0$, this implies that for any $\theta^*$, there is a $z$ large enough that induces the researcher to eventually stop conducting studies.

- The analysis of 5.(c) in the previous section for finding $N(z)$ applies here with proper modification. In particular, $N(z)$ is the largest $n$ such that

$$1 - \Phi((z - \mu_{n|n-1})/\sqrt{1 + \sigma^2_{n|n-1}}) + \Phi((-z - \mu_{n|n-1})/\sqrt{1 + \sigma^2_{n|n-1}}) \geq c/v$$

in the context of this section so that $N(z)$ depends upon $X^*_{n-1}$ through $\mu_{n|n-1}$. Given that $X^*_{n-1} \sim \mathcal{N}(0, I_{n-1})$, this allows for the computation of $z^*$ via Monte Carlo simulation.

The ICCVs when researchers learn while sampling data, obtained by Monte-Carlo simulations, are illustrated in figure 2. As in figure 1, the publication bias that we aim to correct—the distortion of the size of hypothesis tests—may be fairly large. The figure displays the size of a two-sided hypothesis test at the standard critical value of 1.96, for various priors (panel A) and various cost/benefit ratios (panel C). When researchers learn while sampling datasets, the size is much higher than 5%: as high as 100% in panel A and 95% in panel C. So in this context, for a broad range of parameter values, a null hypothesis is more likely to be rejected than not even when it is true (as size is above 50% for a broad range of parameter values).

And as in figure 1, the ICCVs required to bring back the actual test size to 5% are above the standard value of 1.96, but not by a tremendous amount. The figure displays the ICCV that delivers a size of 5% for two-sided tests, for various priors (panel B) and various cost/benefit ratios (panel D). Across parameterizations, the ICCV is above 1.96, but never above 3.5.
A. Size distortions for various priors

B. ICCVs for various priors

C. Size distortions for various incentives

D. ICCVs for various incentives

**FIGURE 2. ICCVs at 5% significance level when researchers learn while sampling data**

The figure considers researchers who attempt to obtain significance in two-sided hypothesis tests by sampling iid datasets, and who update their beliefs about the distribution of the \( t \)-statistic after observing the value of the \( t \)-statistic in each new dataset. All results are obtained from 1,000 Monte-Carlo simulations with the parameter values in table 1, except for those parameter values directly specified on the graph axes. Panels A and C display the size of the hypothesis test at the standard critical value of 1.96, for various priors (panel A) and various cost/benefit ratios (panel C). If researchers collected only one dataset and conducted the test on it, as is assumed in classical statistics, the size would be 5%. But when researchers sample datasets, the size is much higher than 5%, sometimes as high as 100%. Panels B and D display the ICCV that delivers a size of 5%, for various priors (panel B) and various cost/benefit ratios (panel D). Across parameterizations, the ICCV falls between 1.96 and 3.4.
A third insight from the figure is that the incentives faced by researchers do not have a very strong impact on the ICCV. Across a broad range of cost/benefit ratios from publication, the ICCV remains between 2.6 and 2.9. Here, unlike in figure 1, prior beliefs about the estimated parameter (especially the prior mean) have a strong impact on the ICCV.

7. Pooling data

Instead of allowing the researcher to learn about the true value of \( \theta \), we take a different departure from section 4 here and have the researcher accumulate data as he moves from one study to the next. The ICCVs in that case are illustrated in figure 3.

To keep the analysis simple, suppose that the \( i^{th} \) study involves collecting an additional \( T \) random variables with associated cost \( c(i) = c \) and the researcher’s estimate of \( \beta \) in the \( i^{th} \) study is equal to the sample mean of the \( n \times T \) independent and identically distributed random variables collected thus far. That is,

\[
\hat{\beta}_i = \frac{1}{iT} \sum_{j=1}^{iT} Y_j
\]

for some sequence of independent and identically distributed random variables \( Y_1, \ldots, Y_{iT} \). In this setting, a standard large sample approximation yields \( \hat{\beta}_i | \beta \sim N(\beta, \text{var}(Y_j)/(iT)) \). However, since they use some of the same underlying random variables in their construction, the different \( \hat{\beta}_i \)'s are no longer independent of one another. Instead, they are jointly normally distributed with

\[
\text{cov}(\hat{\beta}_i, \hat{\beta}_k) = \frac{1}{kT} \text{var}(Y_j)
\]

for \( i \leq k \). In turn, the latent \( t \)-statistics are jointly normally distributed according to

\[
X_i^* | \theta \sim N(\sqrt{i} \theta, 1) \quad \text{where} \quad \theta = \frac{\sqrt{T}(\beta - \beta_0)}{\text{sd}(Y)}
\]

and \( \text{cov}(X_i^*, X_k^*) = \sqrt{i/k} \) for \( i \leq k \).

Due to the correlation between the latent \( t \)-statistics as well as their differing means, specializing (3) to the current context changes the expression for the expected marginal profit from conducting an additional study from (5). Note that in this setting,

\[
X_n^* | X_{n-1}^* = x_{n-1}^*, \theta \sim N \left( \sqrt{n} \theta + \sqrt{\frac{n-1}{n}} \left( x_{n-1}^* - \sqrt{n} \theta \right), \frac{1}{n} \right)
\]
so that
\[
\mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid X^*_{n-1}) = v \int \mathbb{P}(|X^*_n| > z \mid X^*_{n-1}, \theta) dF(\theta) \mathbb{I}(X_{n-1} < z) - c
\]
and the researcher engages in the additional study if and only if \(X_{n-1} < z\) and
\[
\int \left[ 1 - \Phi\left( \sqrt{n}z - \theta - \sqrt{n-1}X^*_{n-1} \right) + \Phi\left( -\sqrt{n}z - \theta - \sqrt{n-1}X^*_{n-1} \right) \right] dF(\theta) \mathbb{I}(X_{n-1} < z) - c \geq \frac{c}{v}.
\]

Similarly to the case with learning, even if we know \(X_{n-1} < z\), the researcher’s choice to engage in another study depends upon the realization of the previous latent \(t\)-statistic \(X^*_{n-1}\). Though the mechanism through which the previous latent \(t\)-statistic determines the researcher’s choice here is different from the case with learning, the qualitative effect is similar: large enough draws of \(X^*_{n-1}\) cause (11) to hold while small enough draws can cause it to be violated. This effect is due to the fact that the researcher is accumulating data, rather than updating his prior, so that the previous latent \(t\)-statistic contains a lot of information about the subsequent \(t\)-statistic. Since the impossibility result of section 4 is also broken in this context, we similarly define (the random variable) \(N(z)\) to be the largest \(n\) satisfying (11) when the data are generated under \(H_0\). In analogy with the previous section, although \(N(z)\) is random here, a slight generalization of the argument given in section 5 provides that the ICCV given in (4) is well-defined.

We make a few additional observations:

1. The analogous comment to Comment 4. in the previous section on learning also applies here. Certain specifications of the researcher’s prior, such as those in Comment 5. of section 5, allow us to analytically evaluate the integral on the left hand side of (11) and simplify the expression for \(N(z)\).

2. Since \(\sqrt{n-1}X^*_{n-1} = (n-1)\theta^* + \sqrt{n-1}Z\), where \(\theta^*\) denotes the true value of \(\theta\) and \(Z \sim N(0, 1)\), the terms \(\sqrt{n}z\) and/or \(\sqrt{n-1}X^*_{n-1}\) dominate any finite value of \(\theta\) in order of magnitude if the researcher has accumulated enough data (for \(n\) large enough). This implies that the effect of any prior distribution \(F(\theta)\) that only places positive probability mass over finite values of \(\theta\) (as is standard) on the decision of the researcher eventually disappears as the researcher continues to accumulate data. This feature of the prior distribution being eventually “washed out” from the decision problem is also apparent in the learning context of the previous section (see Comments 2. and 3. in section 6). However, this occurs in the current context because the researcher accumulates data, rather than
throwing it away in each successive latent study. As more data accumulates, the researcher has an increasingly more information on the value of future t-statistics.

3. If \( H_1 \) holds so that \( \theta^* > 0 \),

\[
\sqrt{n}z - \theta - \sqrt{n-1}X_{n-1}^* = -(n-1)\theta^* + O_p(\sqrt{n})
\]

diverges to negative infinity (in probability) for any finite \( \theta \) as the researcher accumulates more data. This implies that for any standard prior distribution \( F(\theta) \), the left hand side of (11) converges to one (in probability) so that if \( c \leq v \) and the number of studies already conducted is large, there is a very high probability that the researcher will continue to conduct studies and accumulate data until he attains rejection of \( H_0 \).

4. On the other hand, if \( H_0 \) holds so that \( \theta^* = 0 \),

\[
P_{H_0} \left( 1 - \Phi \left( \sqrt{n}z - \theta - \sqrt{n-1}X_{n-1}^* \right) + \Phi \left( -\sqrt{n}z - \theta - \sqrt{n-1}X_{n-1}^* \right) < \frac{c}{\sigma} \right)
\]

is positive for all \( n \) and any finite \( \theta \) and \( c \leq v \). This implies that for any standard prior distribution \( F(\theta) \), if \( H_0 \) is true, the probability that the researcher stops conducting additional studies is positive at every stage. Moreover,

\[
P_{H_0} \left( 1 - \Phi \left( \sqrt{n}z - \theta - \sqrt{n-1}X_{n-1}^* \right) + \Phi \left( -\sqrt{n}z - \theta - \sqrt{n-1}X_{n-1}^* \right) < \frac{c}{\sigma} \right) \\
\geq P_{H_0} \left( 1 - \Phi \left( \sqrt{n}z - \theta - \sqrt{n-1}X_{n-1}^* \right) < \frac{c}{2\sigma}, \Phi \left( -\sqrt{n}z - \theta - \sqrt{n-1}X_{n-1}^* \right) < \frac{c}{2\sigma} \right) \\
= P_{H_0} \left( \frac{-\sqrt{n}z - \theta - \Phi^{-1}(c/2\sigma)}{\sqrt{n-1}} < X_{n-1}^* < \frac{\sqrt{n}z - \theta - \Phi^{-1}(1-c/2\sigma)}{\sqrt{n-1}} \right) \\
= \Phi \left( \frac{\sqrt{n}z - \theta - \Phi^{-1}(1-c/2\sigma)}{\sqrt{n-1}} \right) - \Phi \left( \frac{-\sqrt{n}z - \theta - \Phi^{-1}(c/2\sigma)}{\sqrt{n-1}} \right) \\
\to \Phi(z) - \Phi(-z) > 0
\]

as \( n \to \infty \) so that in contrast to the case above for which \( H_1 \) holds, even as the researcher accumulates more data, the probability that the researcher continues to conduct more studies remains bounded away from one. Thus if \( H_0 \) holds, the researcher must eventually stop conducting studies.

The ICCVs when researchers pool data, obtained by Monte-Carlo simulations, are illustrated in figure 3. Here the publication bias that we aim to correct is moderate. Figure 3 displays the size of a two-sided hypothesis test at the standard critical value of 1.96, for various priors (panel A) and various cost/benefit ratios (panel C). When researchers pool datasets, the size is higher than 5%, indicating size distortion: as high as 25% in panel A and 18% in panel C. Hence, the
The figure considers researchers who attempt to obtain significance in two-sided hypothesis tests by pooling iid datasets. All results are obtained from 1,000 Monte-Carlo simulations with the parameter values in table 1, except for those parameter values directly specified on the graph axes. Panels A and C display the size of the hypothesis test at the standard critical value of 1.96, for various priors (panel A) and various cost/benefit ratios (panel C). If researchers collected only one dataset and conducted the test on it, as is assumed in classical statistics, the size would be 5%. But when researchers sample datasets, the size is higher than 5%, sometimes as high as 25%. Panels B and D display the ICCV that delivers a size of 5%, for various priors (panel B) and various cost/benefit ratios (panel D). Across parameterizations, the ICCV falls between 1.96 and 2.5.
size distortions when researchers pool data are less than when researchers sample data, as in figures 1 and figure 2.

Since size distortions are moderate, the ICCVs required to bring back the actual test size to 5% are above the standard value of 1.96, but not by much. The figure displays the ICCV that delivers a size of 5% for two-sided tests, for various priors (panel B) and various cost/benefit ratios (panel D). Across parameterizations, the ICCV is above 1.96, but never above 2.5.

Figure 3 also shows that the incentives faced by researchers barely affect the ICCV. Across a broad range of cost/benefit ratios from publication, the ICCV remains around 2.3 (panel D). Prior beliefs about the estimated parameter (especially the prior mean) have a stronger impact on the ICCV. However, if researchers were learning while pooling data, it is probable that the effect of prior beliefs would be dampened.

8. General case

In the general case, we would like to allow for dependence across latent studies and potentially different means of the underlying sequence of latent t-statistics to incorporate observational studies with dependent data, cases for which the researcher adds or subtracts new data to a study, instrument selection when using two-stage least squares, regression specification for ordinary least squares, looking through studies of different precision (resulting in different standard errors), etc. In this more general formulation of the problem, the latent outcome of empirical study \( i \geq 1 \) is \( X_i^* \), which is approximately distributed according to the distribution of a \( N(\theta_i, 1) \) distribution with \( \text{cov}(X_i^*, X_j^*) = \omega_{ij} \) so that for any \( n \geq 1 \), we may write

\[
X_n^* \sim N(\theta_n, \Omega_n) \quad \text{with} \quad \theta_n = [\theta_1, \ldots, \theta_n]' \quad \text{and} \quad \Omega_n = [\omega_{ij}]_{i,j=1:n}.
\]

In particular, \( \theta_n = 0 \) under \( H_0 \). We will assume that \( \theta_n \) is unknown to the researcher for all \( n \geq 1 \) but that \( \Omega_n \) is known after \( n \) studies have been conducted. These assumptions approximate the large sample joint distribution of a set of latent t-statistics for testing (1) in a general framework under which standard errors and correlations between estimators can be consistently estimated. We provide concrete examples of these quantities in different research settings below. With a slight abuse of notation, we denote the researcher’s prior distribution on \( \theta_n \) as \( F(\theta_n) \) and also discuss below how a prior distribution on \( \beta \) translates to \( F(\theta_n) \).

**Sampling data.** In the case that each successive study uses an independent sample from the same population to estimate \( \beta \), from the point of view of the researcher we have the following large-sample distributional approximation: \( \hat{\beta}_i | \beta \sim N(\beta, \text{var}(\hat{\beta}_i)) \), where \( \text{var}(\hat{\beta}_i) = \xi^2 / T_i \) for some
\( \zeta^2 \) with \( T_i \) denoting the sample size of the \( i \)th study. In this case, \( \text{sd}(\hat{\beta}_i) \approx \zeta / \sqrt{T_i} \) so that using the form of the latent \( t \)-statistic (2), \( \theta_i = \sqrt{T_i}(\beta - \beta_0) / \zeta \geq 0 \). Since the studies are independent, \( \Omega_n = I_n \). Finally, suppose the researcher has a prior distribution on \( \beta \). His prior on each individual \( \theta_i \) is thus the same distribution shifted by \( \beta_0 \) and scaled by \( \zeta / \sqrt{T_i} \). From the point of view of the researcher, each \( \theta_i \) is a shifted and scaled version of the same underlying random variable \( \beta \), making \( F(\theta_n) \) degenerate.

**Pooling data.** Suppose that each successive study simply adds additional data to the previous study to estimate \( \beta \), where these additional data are independent and collected from the same underlying population. All of the analysis of the previous section applies to this case with the exception of \( \Omega_n = I_n \). Instead, \( \omega_{ij} \) is a decreasing function of \( |i - j| \). In particular, assume that we can approximate \( \hat{\beta}_i \) by a sample mean:

\[
\hat{\beta}_i \approx \frac{1}{T_i} \sum_{t=1}^{T_i} Y_i
\]

for some sequence of independent random variables \( Y_1, \ldots, Y_{T_i} \). This approximation can be used for instance for standard linear regression estimators. Since the researcher accumulates data when forming each estimate, further assume that \( T_i > T_{i-1} \). In this case, we have

\[
\text{cov}(\hat{\beta}_i, \hat{\beta}_j) \approx \frac{1}{T_j} \text{var}(Y_i)
\]

for \( i \leq j \) so that \( \omega_{ij} = \sqrt{T_i/T_j} \) for \( i \leq j \).

**Regression specification.** For this example, we assume that the researcher uses ordinary least squares in the standard linear regression model to estimate the effect of interest. In practice, a typical effect of interest corresponds to the population value of a regression coefficient. However, when the researcher uses different regression specifications across different latent studies, the effect of interest is no longer fixed across studies so that (1)–(2) do not apply in general. Nevertheless, a simple generalization of (1)–(2) can be used for this example. More specifically, suppose that in the \( i \)th study the researcher uses ordinary least squares to estimate a regression coefficient in a regression of \( y_i \) on \( w_i \) from a set of \( T \) independent data points \((y_{i1}, \ldots, y_{iT})\) and \((w_{i1}, \ldots, w_{iT})\) so that

\[
\hat{\beta}_i = \frac{\sum_{t=1}^{T} w_{it} y_{it}}{\sum_{t=1}^{T} w_{it}^2}.
\]
Here, $w_i$ represents the regressor of interest after it has been projected off of the space spanned by the covariates included in the $i^{th}$ regression model, allowing for both different specifications of the regressor of interest and covariates across studies. This framework also allows for different specifications of the dependent variable $y_i$ across latent studies. When using different regression specifications across different studies, the researcher implicitly sets the object of interest in the $i^{th}$ study equal to the population regression coefficient

$$\beta_i = \frac{E(w_{it}y_{it})}{E(w_{it}^2)},$$

where $E$ denotes the expectation operator with respect to the true objective probability measure.

From the point of view of the researcher, standard assumptions yield the following large-sample distributional approximation: $\hat{\beta}_i | \beta_i \sim N(\beta_i, \text{var}(\hat{\beta}_i))$, where $\text{var}(\hat{\beta}_i) = \text{var}(w_{it}y_{it})/TE(w_{it}^2)^2$. When allowing for different specifications across studies (1) must be modified to

$$H_0 : \beta_i = \beta_{i0} \quad \text{versus} \quad H_1 : \beta_i \neq \beta_{i0}$$

and (2) must be correspondingly modified to

$$X_i^* = \frac{\hat{\beta}_i - \beta_{i0}}{\text{se}(\hat{\beta}_i)}$$

so that in large samples

$$X_i^* | \theta_i \overset{iid}{\sim} N(\theta_i, 1), \quad \text{where} \quad \theta_i = \frac{\sqrt{T}(\beta_i - \beta_{i0})}{\text{sd}(w_{it}y_{it})/E(w_{it}^2)} \geq 0.$$ 

In addition,

$$\text{cov} (\hat{\beta}_i, \hat{\beta}_j) \approx \frac{\text{cov}(w_{it}y_{it}, w_{jt}y_{jt})}{TE(w_{it}^2)E(w_{jt}^2)}$$

so that $\omega_{ij} = \text{cov}(w_{it}y_{it}, w_{jt}y_{jt})/\sqrt{\text{sd}(w_{it}y_{it}) \text{sd}(w_{jt}y_{jt})}$. In this example, the researcher’s prior distribution on $(\beta_1, \ldots, \beta_n)$ induces a prior distribution on $\theta_n$ via simple shifting and scaling. It may be reasonable in some examples to assume that the researcher has an identical prior distribution for all $\beta_i$’s, leading to a degenerate $F(\theta_n)$.

**Instrument selection.** By modifying some of the definitions in the previous example, we can also cover the case for which the researcher uses two-stage least squares in a standard linear regression model to estimate the effect of interest. Assuming that the instruments are both strong...
where a partially sophisticated researcher who only partially updates his beliefs.

\[ F(\theta_n; X^*_{n-1}) = \alpha F(\theta_n) + (1 - \alpha)F(\theta_n|X^*_{n-1}), \]

where \( F(\theta_n|X^*_{n-1}) \) denotes the researcher's posterior distribution on \( \theta_n \) and \( \alpha \in [0, 1] \) is a parameter that indexes the sophistication of the researcher with \( \alpha = 1 \) corresponding to an unsophisticated researcher who does not update his beliefs about the parameter of interest, \( \alpha = 0 \) corresponding to a sophisticated researcher who completely updates his beliefs and \( \alpha \in (0, 1) \) corresponding to a partially sophisticated researcher who only partially updates his beliefs.

For the general case, we allow the researcher to learn about the true value of the parameter of interest in a potentially limited fashion. When evaluating the expected marginal profit of conducting an additional study, the researcher uses a convex combination of his prior and the posterior distribution that is fully updated according to Bayes’ rule:

\[ F(\theta_n; X^*_{n-1}) \]

After \( n - 1 \) studies have been conducted, we may then express the expected marginal profit of the \( n^\text{th} \) study (3) as

\[ \mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid X^*_{n-1}) = v \int \mathbb{P}(X^*_{n-1} > z|\theta_n) dF(\theta_n; X^*_{n-1}) \mathbb{I}(X_{n-1} < z) - c(n) \]

where \( X^*_{n|n-1} \sim X^*_n \mid X^*_{n-1} \) such that \( X^*_{n|n-1} \sim N(\theta_{n|n-1}, \omega^2_{n|n-1}) \) with

\[ \theta_{n|n-1} = \theta_n + \Omega_{n,21} \Omega_{n-1}^{-1} [X^*_{n-1} - \theta_{n-1}], \]

\[ \omega^2_{n|n-1} = 1 - \Omega_{n,21} \Omega_{n-1}^{-1} \Omega_{n,12}, \]

\[ \Omega_n = \begin{bmatrix} \Omega_{n-1} & \Omega_{n,12} \\ \Omega_{n,21} & 1 \end{bmatrix}. \]

Finally, the ICCV \( z^* \) is defined in (4) with \( N(z) \) being equal to the largest value of \( n \) such that

\[ \int [1 - \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) + \Phi((-z - \theta_{n|n-1})/\omega_{n|n-1})] dF(\theta_n; X^*_{n-1}) \geq \frac{c(n)}{v} \]
or if the cost function is strictly increasing, the largest $n$ such that

$$n \leq c^{-1} \left( v \int [1 - \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) + \Phi((z - \theta_{n|n-1})/\omega_{n|n-1})]d\alpha(\theta_n; \mathbf{X}_{n-1}) \right).$$

The analogous argument about the existence of such a $N(z)$ to that made in section 5 applies here as well. However, the computation of $z^*$ is complicated by the fact that $F_a(\theta_n; \mathbf{X}_{n-1}^\ast)$ depends upon the realizations of prior studies, which are unobserved by the journal editor or reader. Nevertheless, since we know the distribution of $\mathbf{X}_{n-1}^\ast$ under $H_0$, namely $N(0, \Omega_{n-1})$, $z^*$ remains feasible to compute in the presence of (partial) learning by the researcher: since $\Omega_n$ is known for all $n \geq 1$, it is straightforward to compute $P_{H_0}(X_n > z)$ for any given $n, z$ by Monte Carlo Simulation so that the fixed-point solution to (4), $z^*$, is computationally tractable.\(^6\)

**General results with a normal prior.** If the researcher’s prior distribution is conjugate with the normal distribution, we can provide an analytical expression for his posterior distribution and therefore $F_a(\theta_n; \mathbf{X}_{n-1}^\ast)$. In this subsection, we focus on the case when the researcher’s prior over $\theta_n$ follows a normal distribution. This special case is particularly interesting because it allows us to characterize $N(z)$, and therefore the ICCV defined in (4) more explicitly. Moreover, a normal prior may be natural in many contexts. For example, if the researcher’s prior is based upon averaging estimation results from previous studies, a central limit theorem argument leads to an approximate normal distribution for such an average.

If the researcher’s prior distribution $F(\theta_n)$ admits a pdf $f(\theta_n)$, for $v \geq 2$, the posterior pdf for $\theta_n$ takes the form

$$f(\theta_n|\mathbf{X}_{n-1}^\ast) = \frac{f(\mathbf{X}_{n-1}^\ast|\theta_n)f(\theta_n)}{\int f(\mathbf{X}_{n-1}^\ast|\theta_n)f(\theta_n)d\theta_n},$$

where the likelihood $f(\mathbf{X}_{n-1}^\ast|\theta_n)$ is the pdf of a $N(\theta_{n-1}, \Omega_{n-1})$ distribution evaluated at $\mathbf{X}_{n-1}^\ast$. As a notational convention, for $n = 1$, $f(\theta_n|\mathbf{X}_{n-1}^\ast)$ is equal to $f(\theta_n)$. Suppose in particular that the researcher’s prior on $\theta_n$ follows a $N(\mu_n, \Sigma_n)$ distribution for all $n \geq 1$. In this case,

$$f(\mathbf{X}_{n-1}^\ast|\theta_n) = \frac{1}{\sqrt{(2\pi)^{n-1}\Omega_{n-1}}} \exp\left(-\frac{1}{2}(\mathbf{X}_{n-1}^\ast - \theta_n)^\prime \Omega_{n-1}^{-1}(\mathbf{X}_{n-1}^\ast - \theta_n)\right)$$

$$f(\theta_n) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_n|}} \exp\left(-\frac{1}{2}(\theta_n - \mu_n)^\prime \Sigma_n^{-1}(\theta_n - \mu_n)\right)$$

\(^6\)Extreme value theory can also be used to approximate $P_{H_0}(X_n > z)$ for large $n$ in the case that a simulation-based approach indicates that $N(z)$ is very large.
so that for

\[
\bar{\Omega}^{-1} = \begin{bmatrix}
\Omega_{n-1}^{-1} & 0 \\
0 & 0
\end{bmatrix},
\]

\[
A = -\frac{1}{2} \left( \bar{\Omega}^{-1} + \Sigma_n^{-1} \right),
\]

\[
b = \bar{\Omega}^{-1} X^*_n + \Sigma_n^{-1} \mu_n,
\]

\[
c = -\frac{1}{2} \left( X'^* \bar{\Omega}^{-1} X_n + \mu'_n \Sigma_n^{-1} \mu_n \right),
\]

\[h = -A^{-1} b / 2\] and \[k = c - b' A^{-1} b / 4,\] we have

\[
\int f(X_{n-1}^* | \theta_n) f(\theta_n) d\theta_n
\]

\[
= \frac{1}{\sqrt{(2\pi)^{2n-1} |\Omega_{n-1}| |\Sigma_n|}}
\]

\[
\times \int \exp \left( -\frac{1}{2} \left[ (\theta_n - X^*_n)' \bar{\Omega}^{-1} (\theta_n - X^*_n) + (\theta_n - \mu_n)' \Sigma_n^{-1} (\theta_n - \mu_n) \right] \right) d\theta_n
\]

\[
= \frac{1}{\sqrt{(2\pi)^{n-1} A |\Omega_{n-1}| |\Sigma_n|}} \int \frac{1}{\sqrt{(2\pi)^{n+1/2} |A^{-1}|}} \exp([\theta_n - h]' A [\theta_n - h]) d\theta_n \exp(k)
\]

\[
= \frac{1}{\sqrt{(2\pi)^{n-2} A |\Omega_{n-1}| |\Sigma_n|}} \exp(k)
\]

\[
= \frac{1}{\sqrt{(2\pi)^{n-1} \bar{\Omega}^{-1} \Sigma_n + I_n |\Omega_{n-1}|}} \exp(k).
\]

Thus, (13)–(15) imply

\[
f(\theta_n | X_{n-1}^*) = \frac{\sqrt{(2\pi)^{n-1} |\bar{\Omega}^{-1} \Sigma_n + I_n| |\Omega_{n-1}|}}{\sqrt{(2\pi)^{n-1} |\Omega_{n-1}| |\Sigma_n|}} \exp(-(1/2)(X_{n-1}^* - \theta_{n-1})' \Omega_{n-1}^{-1} (X_{n-1}^* - \theta_{n-1})
\]

\[- (1/2) (\theta_n - \mu_n)' \Sigma_n^{-1} (\theta_n - \mu_n) - k)
\]

\[
= \frac{\sqrt{|\Omega_n^{-1} + \Sigma_n^{-1}|}}{(2\pi)^n} \exp(-(1/2) \theta_n' [\bar{\Omega}^{-1} + \Sigma_n^{-1}] \theta_n + \theta_n' [\bar{\Omega}^{-1} X_n + \Sigma_n^{-1} \mu_n]
\]

\[+ (1/2) [X_n'^* \bar{\Omega}^{-1} + \mu'_n \Sigma_n^{-1}] [\bar{\Omega}^{-1} + \Sigma_n^{-1}]^{-1} [\bar{\Omega}^{-1} X_n + \Sigma_n^{-1} \mu_n].\]
\[
\frac{1}{\sqrt{2\pi^n \left| \Omega_n^{-1} + \Sigma_n^{-1} \right|}} \exp \left( -\frac{1}{2} (\theta_n - \mu_{n|n-1})' \left( \Omega_n^{-1} + \Sigma_n^{-1} \right) (\theta_n - \mu_{n|n-1}) \right),
\]

where
\[
\mu_{n|n-1} = \left( \Omega_n^{-1} + \Sigma_n^{-1} \right)^{-1} \left( \Omega_n^{-1} X_n^* + \Sigma_n^{-1} \mu_n \right).
\]

This in turn implies that \( \theta_n | X_{n-1}^* \) has a multivariate normal distribution with mean \( \mu_{n|n-1} \) and covariance matrix \( \Sigma_{n|n-1} \equiv \left( \Omega_n^{-1} + \Sigma_n^{-1} \right)^{-1} \).

so that this distribution only depends upon the realization of the first \( n - 1 \) studies.

For a given \( X_{n-1}^* \), the left hand side of the inequality (12) defining \( N(z) \) is therefore easy to compute via Monte Carlo simulation since
\[
\int \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) dF_\alpha(\theta_n; X_{n-1}^*) = \alpha \int \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) f(\theta_n) d\theta_n
\]
\[+ (1 - \alpha) \int \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) f(\theta_n | X_{n-1}^*) d\theta_n.\]

The integral inside the first term is the expected value of \( \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) \) when \( \theta_n \) follows a \( N(\mu_n, \Sigma_n) \) distribution, and the integral inside the second term is the expected value of \( \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) \) when \( \theta_n \) follows a \( N(\mu_{n|n-1}, \Sigma_{n|n-1}) \) distribution.

9. Conclusion

Statistical hypothesis testing is a key tool for scientific investigation and discovery. It is used to evaluate existing paradigms, and to assess the effectiveness of new medical treatments, public policies, and other potential remedies to real-world problems. In order to be informative, hypothesis testing relies crucially on properly controlling size—the probability of rejecting a true null hypothesis. A major issue in modern science, however, is that test sizes in published test results are systematically distorted: because scientists are given the incentive to continue to conduct studies until they are able to reject the null hypothesis they are investigating. This is because rejecting a null hypothesis is often considered more interesting from a scientific perspective, and therefore is often required for publication. As a result, a true null hypothesis is much more likely to be rejected than what is advertised in scientific publications.

To correct this size distortion, we construct critical values that are compatible with the
The figure displays the average number of experiments conducted by researchers for different critical values (and in particular at the ICCV). The results for the case “sampling increasingly costly data” come from the simulations in figure 1. The results for the case “sampling data with learning” come from the simulations in figure 2. The results for the case “pooling data” come from the simulations in figure 3. In all the simulations, we set parameter values as in table 1.

Our approach to correct size distortions in hypothesis tests is to construct critical values that take into account researchers’ behavior. Another approach is to constrain researchers’ behavior by asking researchers to register their experiments and analysis plans in advance (Christensen, Freese, and Miguel 2019, part 3). These two strategies result in a very different research process. In figure 4, we show the number of studies conducted by researchers under ICCVs. Researchers are expected to pool about 3 studies on average, or to sample between 4 and 11 studies, depending
on the learning process and cost structure. On the other end, with preregistration, researchers would only conduct one unique study and report those results. Each approach may be more appropriate in different settings. Ours could be more appropriate in the early stage of a research question, when scientific exploration plays a key role. The preregistration approach could be more adapted to later stages, when the research question is well understood and delineated, and it is important to obtain very precise estimates of the parameters of interest.

References


