Precision-based designs for sequential randomized experiments

Mattias Nordin*and Mårten Schultzberg†

May 7, 2024

Abstract

In this paper, we consider an experimental setting where units enter the experiment sequentially. Our goal is to form stopping rules which lead to estimators of treatment effects with a given precision. We propose a fixed-width confidence interval design (FWCID) where the experiment terminates once a pre-specified confidence interval width is achieved. We show that under this design, the difference-in-means estimator is a consistent estimator of the average treatment effect and standard confidence intervals have asymptotic guarantees of coverage and efficiency for several versions of the design. In addition, we propose a version of the design that we call fixed power design (FPD) where a given power is asymptotically guaranteed for a given treatment effect, without the need to specify the variances of the outcomes under treatment or control. In addition, this design also gives a consistent difference-in-means estimator with correct coverage of the corresponding standard confidence interval. We complement our theoretical findings with Monte Carlo simulations where we compare our proposed designs with standard designs in the sequential experiments literature, showing that our designs outperform these designs in several important aspects. We believe our results to be relevant for many experimental settings where units enter sequentially, such as in clinical trials, as well as in online A/B tests used by the tech and e-commerce industry.

*Department of Statistics, Uppsala University. Email: mattias.nordin@statistics.uu.se
†Experimentation Platform team, Spotify. Email: mschultzberg@spotify.com

1 Introduction

A key component in the traditional design of randomized experiments is to decide on the sample size. The experimenter may consider many different aspects such as costs, time, ethics and urgency when deciding on the size of the experiment, but it is generally agreed upon that a well-designed experiment should include a thorough power analysis. In an underpowered study, statistically significant effects are generally overestimates of the true effect size (Gelman and Carlin, 2014a), whereas in a overpowered study, an unnecessary amount of resources are spent in the experiment.

In many experimental settings, units enter into the experiments sequentially. This is, for instance, the case in many clinical trials but also in online A/B testing in the tech and e-commerce industry. In such cases, instead of relying on a fixed-sample design (where the number of observations are decided in advance) the experimenter can ensure that resources are not wasted by continuously monitoring the results and terminate the experiment as soon as results are sufficiently clear to allow a conclusion to be drawn. To do so in a statistically rigorous way, a decision rule which take into account that results are monitored continuously need to be put in place. Sequential tests, originating with Wald (1945), have been developed to do just that. For instance, with the error-spending approach of Lan and DeMets (1983), the experimenter decides on how much type I error to spend at any given point such that the overall type I error is set at a given value, for instance at five percent. If, at any given point, the observed statistic exceeds the implied threshold value, the experiment is terminated. Another sequential testing framework that has gained a lot of attention, especially in the online experimentation literature, is the so-called "always-valid inference" (Johari et al., 2015, 2022; Howard et al., 2021; Waudby-Smith et al., 2024), which bounds the false positive rate by any desired α under continuous monitoring, where the sample size is allowed to approach infinity.

The sequential testing approach is useful when the relevant question is of the form "is the treatment better than the control". However, for many research questions, it is also of interest how large the difference is. In such cases sequential tests will in general work poorly as the usual difference-in-means estimator is a biased estimator of the average treatment effect when early stopping is allowed (Fan et al., 2004). Furthermore, the cost of using sequential tests is that—compared to fixed-sample designs—the power of the test will be lower, because it is necessary to adjust for multiple testing. The

lower power is especially acute for always-valid tests who tend to be very conservative. Finally, if the null hypothesis is true, sequential designs will have to run until a final, pre-specified, maximum sample size (or time) is reached, resulting in potentially inefficient use of data (although, sequential designs may also include futility bounds to mitigate this problem)

In this paper we propose an alternative stopping rule to significance-based stopping, where the goal of the design is to achieve a confidence interval of a given pre-specified width. This width is achieved by continuously forming confidence intervals for the treatment effect and stopping once a targeted width is reached. We call this design a "fixed-width confidence interval design" (FWCID) and introduce both a naive and conservative version of the design. We show that both versions results in consistent estimators of the average treatment effect with asymptotically correct coverage of the confidence interval. In addition, the use of data is asymptotically efficient in the sense that the expected sample size is the same as the sample size that would have been chosen in a fixed-sample design to achieve the target confidence interval, had the variances of the outcome under treatment and control been known. Moreover, the design makes it straightforward to—at any time during the experiment—make a forecast of at what sample size the experiment will terminate, something that is useful from a practical perspective.

Finally, we also show how the idea of forming a stopping rule based on the confidence interval width, i.e., a function of the variance of the treatment effect estimator, can be translated into a design for hypothesis testing, which we call a fixed-power design (FPD). The perhaps most difficult aspect of a fixed-sample design when performing a power analysis is to form beliefs about the outcome distribution under both treatment and control. Even though historical outcome data is sometimes available, it is not trivial to estimate the variance of the treatment effect estimator. For instance, in online experiments, the outcome distribution often changes over time due to seasonality and growing and/or changing user bases, and the fact that treatment effects are often heterogeneous. With heterogeneous treatment effects, the variance under no treatment may not be a good proxy for the variance under treatment. With the FPD, there is no need to guess what the outcome distribution will look like in the design stage. Instead, power is assured by letting the final sample size be stochastic. This feature could be undesirable if the experimenter do not want the experiment to run for "too long". However, we also show that the FPD admits continuous monitoring of expected sample size at which the experiment will terminate. Therefore, the

experimenter can already from an early stage check when the experiment will terminate and take appropriate action (such as terminating the experiment) if the expected sample size is too high. In addition, the fact that sample size is stochastic is a guard against overly optimistic power analyses in fixed-sample designs where it is easy for the experimenter to fall in to the trap of assuming that the variances of the outcome distributions are too small, leading to underpowered studies.

Recent discussions both in the statistics literature and the scientific community at large has criticized the over-reliance on null hypothesis significance testing (e.g., Wasserstein and Lazar 2016; Wasserstein et al. 2019; Gelman and Carlin 2014b; Greenland 2017; Amrhein et al. 2019). Instead, many of these authors have argued that more emphasis should be placed on point estimation and estimates of uncertainty. In this paper, we take exactly such an approach where the two designs we propose, in contrast to much of the sequential testing literature, gives consistent estimators of the average treatment effect together with valid confidence intervals.

The rest of the paper is structured as follows. In Section 2, we introduce the theoretical framework we work in and make some preliminary definitions. In Section 3.1, we introduce the FWCID and prove both consistency and asymptotic efficiency of the naive version of that design. We also introduce the FPD and show how we can use the lessons from the FWCID to construct valid tests with correct power and size. In Section 3.2, we introduce a conservative version of the FWCID and prove consistency and asymptotic efficiency for that version as well. In Section 4, we study small sample behavior in a series of simulation studies where the FWCID and FPD are contrasted with some of the most common sequential designs used today. Finally, Section 5 concludes.

2 Setup

We consider a randomized experiment with treatment indicator $W \in \{0, 1\}$. Let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes of unit i if untreated and treated, respectively. The potential outcomes are independently and identically distributed across units with $\mathbb{E}(Y(0)) = \mu_0$, $\mathbb{E}(Y(1)) = \mu_1$, $0 < \mathbb{V}(Y(0)) = \sigma_0^2 < \infty$ and $0 < \mathbb{V}(Y(1)) = \sigma_1^2 < \infty$.

The target parameter in the experiment is the population average treatment effect (ATE), $\tau = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0))$. We assume that the stable unit

treatment value assumption (SUTVA) holds, which means that we get the observed outcome, Y_i, as

$$Y_i = W_i Y_i(1) + (1 - W_i) W_i(0). \tag{1}$$

We also assume exchangeability, which follows from randomization of W:

$$Y(w) \perp\!\!\!\perp W. \tag{2}$$

For a sample of size n, the difference-in-means estimator can be written as

$$\hat{\tau}_n = \frac{1}{n_1} \sum_{i=1}^{n} Y_i(1) W_i - \frac{1}{n_0} \sum_{i=1}^{n} Y_i(0)(1 - W_i) = \hat{\mu}_{1n} - \hat{\mu}_{0n}, \tag{3}$$

where n_0, n_1 are the number of control and treated units, while $\hat{\mu}_{0n}, \hat{\mu}_{1n}$ are the sample averages of Y for the two groups. Considering the share treated as fixed, the variance of the difference-in-means estimator can be estimated as

$$\hat{\sigma}^2_{\hat{\tau}_n} = \left(\frac{1}{n_1} \hat{\sigma}^2_{1n} + \frac{1}{n_0} \hat{\sigma}^2_{0n} \right), \tag{4}$$

where $\hat{\sigma}^2_{wn}$ is the naive variance estimator for treated and non-treated respectively,

$$\hat{\sigma}^2_{wn} = \frac{1}{n_w} \sum_{i:W_i = w} \left(Y_i(w) - \hat{\mu}_{wn} \right)^2. \tag{5}$$

We focus on the naive variance estimator instead of the standard unbiased one, since our focus in on asymptotics and it will simplify the math later on. For completeness, we also set $\hat{\sigma}^2_{\hat{\tau}_n} = \infty$ if $n_0 \leq 1$ and/or $n_1 \leq 1$. It will be convenient to rewrite the variance in the following way:

$$\begin{aligned} \hat{\sigma}^2_{\hat{\tau}_n} &= \left(\frac{n_0}{n} \hat{\sigma}^2_{1n} + \frac{n_1}{n} \hat{\sigma}^2_{0n} \right) \kappa_n \\ &= \hat{\sigma}^2_{\mathcal{P}_n} \kappa_n, \end{aligned} \tag{6}$$

where $\kappa_n := 1/n_0 + 1/n_1 = n/(n_0 n_1)$. In this paper, we focus on so-called *sequential sampling designs*, i.e., designs where the sampling continues until a *stopping rule* criterion is met.

Definition 1. A sequential sampling design is a design where units enter sequentially such that $\lim_{n \to \infty} n_1/n = p$ a.s. , for $0 < p < 1$, and exchangeability and SUTVA holds for any $n \in \mathbb{N}^+$.

As such, the treatment assignment does not have to be random, but could, for instance, be a design where every second unit goes to treatment and the others to control, as long as exchageablity and SUTVA holds. For any sequential design, the stopping rule, i.e., a rule for when to stop sampling, can be any function of data that satisfies the following rule:

Definition 2. A stopping rule $g(f(X_n), c_n) = N$ is a rule for a function f, a sequence of random variables X_n and a sequence of constants c_n such that sampling is stopped at sample size $n = N$ iff

$$\begin{aligned} &f(X_n) > c_n, \quad \forall n < N, \\ &f(X_N) \leq c_n. \end{aligned} \tag{7}$$

Note that N is not the stopping rule, but the stochastic sample size at which sampling is stopped. Rather, the rule is to stop the first time $f(X_n) \leq c_n$.

3 Fixed confidence interval width designs for ATE estimation

In this section we define a sequential design that uses the observed confidence interval width as a stopping criterion. Formally, we consider a fixed-width confidence interval as in Definition 3.

Definition 3. For a target parameter θ and an estimator $\hat{\theta}_N$, a fixed-width confidence interval with half-width d and confidence level $1 - \alpha$, I_N, is an interval

$$I_N = [\hat{\theta}_N - d, \hat{\theta}_N + d], \tag{8}$$

such that

$$\mathbb{P}(\theta \in I_N) = 1 - \alpha. \tag{9}$$

The difference from fixed-sample confidence intervals is that instead of the estimated variance being stochastic, it is now the sample size that is stochastic.

However, as we show in the following section, asymptotically, we can construct a sequence of fixed-sample confidence intervals and stop once such an interval has half-width smaller than d to construct a valid fixed-width confidence interval. We call such a design a *fixed-width confidence interval design* (FWCID).

3.1 Asymptotically valid inference under FWCID

Theorem 1. *For the target parameter τ and the stopping rule $g(\hat{\sigma}^2_{\hat{\tau}_n}, d^2/z^2_{\alpha/2}) = N$, with $0 < \alpha < 1$, and the estimator of the treatment effect $\hat{\tau}_N$, I_N is an asymptotic fixed-width confidence interval. I.e., it is the case that*

$$\lim_{d\to 0} \mathbb{P}(\tau \in I_N) = 1 - \alpha \tag{10}$$

Proof. The proof follows almost directly from Chow and Robbins (1965) with some minor changes described below. Using Definition 2, the sample size N is given by the following stopping rule

$$\begin{aligned} &\hat{\sigma}^2_{\hat{\tau}_n} > d^2/z^2_{\alpha/2}, \quad \forall n < N, \\ &\hat{\sigma}^2_{\hat{\tau}_N} \leq d^2/z^2_{\alpha/2}. \end{aligned} \tag{11}$$

For ease of notation, let $\sigma^2_{\mathcal{P}} := (1-p)\sigma^2_1 + p\sigma^2_0$ and defining $\hat{p}_n = n_1/n$, we define the following variables:

$$s_n := \frac{\hat{\sigma}^2_{\mathcal{P}_n}}{\sigma^2_{\mathcal{P}}} \frac{p}{\hat{p}_n} \frac{1-p}{(1-\hat{p}_n)} \tag{12}$$

$$t := \frac{z^2_{\alpha/2}\sigma^2_{\mathcal{P}}}{d^2 p(1-p)}. \tag{13}$$

The stopping rule in equation (11) can be rewritten as

$$N = \inf\{n \in \mathbb{N}^+ : s_n \leq n/t\}. \tag{14}$$

Note that $s_n > 0$ a.s. and $\lim_{n\to\infty} s_n = 1$ a.s. It follows that N is a non-decreasing function of t with $\lim_{t\to\infty} N = \infty$ a.s. Moreover, $\lim_{t\to\infty}$ is equivalent to $\lim_{d\to 0}$.

By equation (14), $s_N \leq N/t$ and $s_{N-1} > (N-1)/t$. The second inequality can be rewritten as $s_{N-1}N/(N-1) > N/t$. Hence, we have

$$s_N \leq \frac{N}{t} < \frac{N}{N-1} s_{N-1}. \tag{15}$$

Because $\lim_{t\to\infty} N = \infty$ a.s., it is the case that $\lim_{t\to\infty} N/(N-1) = 1$ a.s.. Finally, because $s_n > 0$ a.s. and $\lim_{n\to\infty} s_n = 1$, from equation (15) it follows that

$$1 = \lim_{t\to\infty} \frac{N}{t} \text{ a.s.} = \lim_{d\to 0} \frac{d^2 N}{z^2_{\alpha/2}\sigma^2_{\mathcal{P}}} p(1-p) \text{ a.s.} \tag{16}$$

We can write the probability that the interval covers τ as

$$\begin{aligned}\mathbb{P}(\tau \in I_N) &= \mathbb{P}(|\hat{\tau}_N - \tau| < d) \\ &= \mathbb{P}\left(\frac{\sqrt{N}\sqrt{p(1-p)}}{\sigma_{\mathcal{P}}}|\hat{\tau}_N - \tau| < \frac{\sqrt{N}\sqrt{p(1-p)}}{\sigma_{\mathcal{P}}}d\right) \qquad (17)\end{aligned}$$

By equation (16), $\sqrt{N}\sqrt{p(1-p)}d/\sigma_{\mathcal{P}}$ converges to $z_{\alpha/2}$ in probability and N/t converges to 1 in probability as $t \to \infty$. By Anscombe (1952), it follows that as $t \to \infty$, $\sqrt{N}(\hat{\tau}_N - \tau)\sqrt{p(1-p)}/\sigma_{\mathcal{P}} \sim \mathcal{N}(0,1)$ Hence, it is the case that, as $t \to \infty$, $\mathbb{P}(\tau \in I_N) = F(-z_{\alpha/2}) - F(z_{\alpha/2})$, which concludes the proof. □

Theorem 1 says that as we let the target interval width go to zero, the coverage for the CI at the stopping point is the intended $1-\alpha$. Moreover, the difference-in-means estimator is consistent under FWCID, formalized in Corollary 1.

Corollary 1. *Consider a sequential experiment with the same stopping rule as in Theorem 1. The difference-in-means estimator is a consistent estimator of the average treatment effect, i.e.,* $\hat{\tau}_N \xrightarrow{p} \tau$ *as* $d \to 0$.

Proof. By Theorem 1, $\hat{\tau}_N$ is asymptotically normal as $d \to 0$. The corollary follows immediately, see e.g. Casella and Berger (2021). □

For readers familiar with the sequential testing literature, both Theorem 1 and Corollary 1 might be somewhat surprising. In the sequential testing literature, the goal is to mitigate inflated false positive rates due to "peeking" at the outcome data during the sampling. Moreover, using sequential tests also gives biased point estimators (Fan et al., 2004). However, in traditional sequential testing, the treatment effect estimator is directly used in the stopping rule, e.g, stopping when the treatment effect is significantly different from zero. Theorem 1 and Corollary 1 instead says that using information about the variance of the estimator as basis for a stopping rule, as opposed to the ATE estimator itself, does not affect the coverage and the estimator remains consistent. The result directly implies that the false positive rate of the corresponding test (i.e., whether the interval covers the null or not) is not inflated, something that we discuss further in Section 3.1.1.

Remark 1. FWCID is asymptotically efficient in the sense that the asymptotic sample size is the same for FWCID as for a fixed-sample design. To see this, let d_f be the half-width of a fixed-sample confidence interval. It is the case that

$$d_f = z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{\mathcal{P}_n}}{n\hat{p}_n(1-\hat{p}_n)}}, \tag{18}$$

which means that

$$\lim_{n\to\infty} \frac{d_f^2 n}{z_{\alpha/2}\sigma^2_{\mathcal{P}}} p(1-p) = 1 \text{ a.s.} \tag{19}$$

Hence, by equation (16), it is the case that

$$\lim_{n\to\infty} d_f^2 n = \lim_{d\to 0} d^2 N \text{ a.s..} \tag{20}$$

This result says that, asymptotically, the number of observations needed for a given confidence interval width is the same for FWCID as for a standard fixed-sample design.[1].

Remark 2. The stopping rule $g(\hat{\sigma}^2_{\hat{\tau}_n}, d^2/z^2_{\alpha/2}) = N$ in Theorem 1 implies that sampling is stopped when the variance of the treatment effect is sufficiently small. By noting that $\hat{\sigma}^2_{\hat{\tau}_n} = \hat{\sigma}^2_{\mathcal{P}_n}/(n\hat{p}_n(1-\hat{p}_n))$, the stopping rule can be rewritten as

$$g\left(\frac{\hat{\sigma}^2_{\mathcal{P}_n} z^2_{\alpha/2}}{d^2\hat{p}_n(1-\hat{p}_n)}, n\right) = N. \tag{21}$$

The advantage of writing the stopping rule this way is that the first argument provides an estimate of when sampling will be stopped. I.e., let $\hat{N}_n$ be the estimated required sample size to get a confidence interval with half-width d, we have

$$\hat{N}_n = \frac{\hat{\sigma}^2_{\mathcal{P}_n} z^2_{\alpha/2}}{d^2\hat{p}_n(1-\hat{p}_n)}. \tag{22}$$

For most designs, p would be known, in which case, $\hat{p}$ could be replaced with p (something that would mainly be useful early on in the experiment when $\hat{p}$ is likely to be volatile).

[1]In fact, it is also possible to show asymptotic efficiency in the sense of Chow and Robbins 1965, which is that $\lim_{d\to 0} \frac{d^2\mathbb{E}(N)}{a^2\sigma^2_{\mathcal{P}}} p(1-p) = 1$. We omit a proof of this result here, but it follows directly from Chow and Robbins 1965

The two equivalent stopping rules are illustrated in Figure 1 for some randomly generated data in a Bernoulli trial. In the top diagram, the stopping rule in Theorem 1 is illustrated where $\hat{\sigma}^2_{\hat{\tau}_n}$ is generally decreasing, with the decreases becoming smoother over time as each individual observation becomes relatively less important. In the bottom diagram, the alternative stopping rule in equation (21) is instead illustrated. Here the expected sample size at which the experiment is stopped does not trend downwards, but is converging towards a finite value.

Remark 3. In the special case where $\mathbb{E}(\hat{p}_n) = 1/2$ and constant treatment effects (i.e., $Y_i(1) - Y_i(0) = \tau$ for all i) and $\hat{p}_n \perp\!\!\!\perp Y_i(w)$, it is the case that $\mathbb{E}(\hat{\tau}_N) = \tau$, i.e., the treatment effect estimator is unbiased.

A sufficient condition for $\mathbb{E}(\hat{\tau}_N) = \tau$ is that $\mathbb{E}(\hat{\tau}_n|\hat{\sigma}^2_{\hat{\tau}_n}) = \mathbb{E}(\hat{\tau}_n) = \tau$ for all n. We have

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_n|\hat{\sigma}^2_{\hat{\tau}_n}) &= \mathbb{E}\left(\hat{\tau}_n|\hat{\sigma}^2_{\hat{\tau}_n}\right) \\
&= \mathbb{E}\left(\hat{\mu}_{1n} - \hat{\mu}_{0n}|\hat{\sigma}^2_{\hat{\tau}_n}\right) \\
&= \mathbb{E}\left(\hat{\mu}_{1n}|\hat{\sigma}^2_{\hat{\tau}_n}\right) - \mathbb{E}\left(\hat{\mu}_{0n}|\hat{\sigma}^2_{\hat{\tau}_n}\right) \\
&= \mathbb{E}\left(\hat{\mu}_{1n}\left|\frac{1}{\hat{p}_n n}\hat{\sigma}^2_{1n} + \frac{1}{(1-\hat{p}_n)n}\hat{\sigma}^2_{0n}\right.\right) - \mathbb{E}\left(\hat{\mu}_{0n}\left|\frac{1}{\hat{p}_n n}\hat{\sigma}^2_{1n} + \frac{1}{(1-\hat{p}_n)n}\hat{\sigma}^2_{0n}\right.\right) \\
&= \mathbb{E}\left(\hat{\mu}_{1n}\left|\frac{1}{\hat{p}_n n}\hat{\sigma}^2_{1n}\right.\right) - \mathbb{E}\left(\hat{\mu}_{0n}\left|\frac{1}{(1-\hat{p}_n)n}\hat{\sigma}^2_{0n}\right.\right) \\
&= \mathbb{E}\left(\hat{\mu}_{1n}\left|\frac{1}{\hat{p}_n n}\hat{\sigma}^2_{1n}\right.\right) - \mathbb{E}\left(\hat{\mu}_{0n} + \tau\left|\frac{1}{(1-\hat{p}_n)n}\hat{\sigma}^2_{0n}\right.\right) + \tau. \qquad (23)
\end{aligned}$$

Because of the constant treatment effect and the fact that $\mathbb{E}(\hat{p}_n) = 1/2$, the conditional distribution of $\hat{\mu}_{1n}|(1/(\hat{p}_n n)\hat{\sigma}^2_{1n})$ equals the conditional distribution of $\hat{\mu}_{0n}|(1/((1-\hat{p}_n)n))\hat{\sigma}^2_{0n}$, and we have $\mathbb{E}(\hat{\tau}_n|\hat{\sigma}^2_{\hat{\tau}_n}) = \tau$ for all n.

3.1.1 Implications for hypothesis testing

Theorem 1 implies that the experimenter decides on a confidence half-width, d, together with a confidence level, α, and let the experiment run until a the stopping rule is satisfied. Alternatively, instead of deciding on a confidence width, the experimenter may consider a null hypothesis, $\tau = \tau_{H_0}$ and a hypothesized alternative, τ_{H_1} for which a given power, $1-\beta$, should be achieved. For a one-sided hypothesis test, we want to find a stopping rule such that the null is rejected with probability α if $\tau = \tau_{H_0}$ and with probability $1 - \beta$

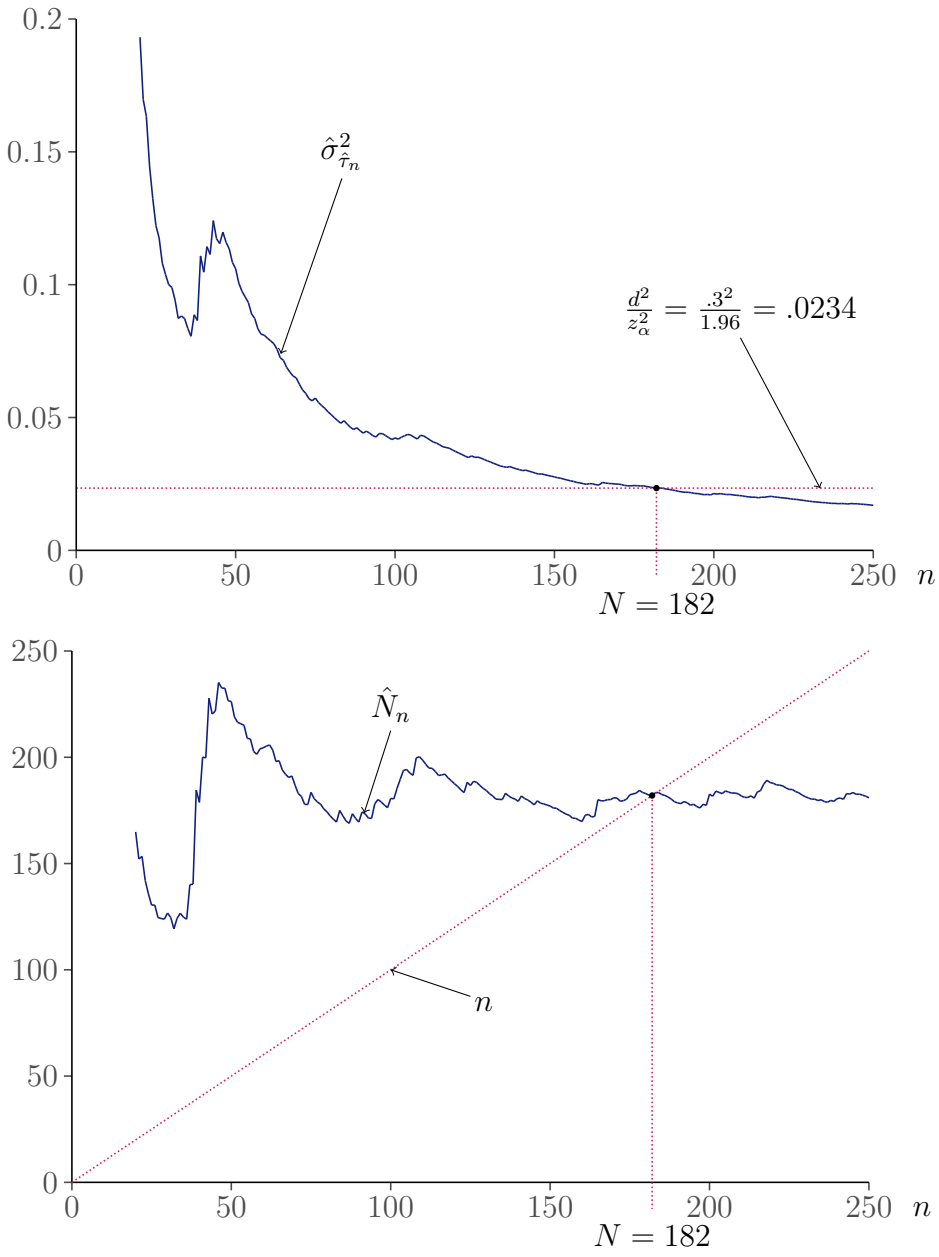

Figure 1: Illustration of stopping rules

Note: The figure illustrates the two equivalent stopping rules $g(\hat{\sigma}^2_{\hat{\tau}_n}, d^2/z^2_{\alpha/2}) = N$ (top diagram) and $g((\hat{\sigma}^2_{\mathcal{P}_n} z^2_{\alpha/2})/(d^2\hat{p}_n(1-\hat{p}_n)), n) = N$ (bottom diagram). Both $Y(0)$ and $Y(1)$ have been drawn from a standard normal distribution, and treatment assignment is determined in a Bernoulli trial with $p = .5$. The half-width of the confidence interval is set at $d = .3$ and the confidence level is 95%.

if $\tau = \tau_{H_1}$. We call such a design a *fixed-power design* (FPD). We formalize this design in Corollary 2.

Corollary 2. *For a null hypothesis of τ_{H_0} and an alternative τ_{H_1}, it is the case that for $\tau_d := \tau_{H_1} - \tau_{H_0}$, the stopping rule $g(\hat{\sigma}^2_{\hat{\tau}_n}, \tau_d^2/(z_\alpha + z_\beta)^2) = N$ implies that*

$$\lim_{\tau_d \to 0} \mathbb{P}(\hat{\tau}_N - z_\alpha \hat{\sigma}^2_{\hat{\tau}_N} > \tau_{H_0} | \tau = \tau_{H_0}) = \alpha, \tag{24}$$

$$\lim_{\tau_d \to 0} \mathbb{P}(\hat{\tau}_N - z_\alpha \hat{\sigma}^2_{\hat{\tau}_N} > \tau_{H_0} | \tau = \tau_{H_1}) = 1 - \beta, \tag{25}$$

for $\tau_{H_1} > \tau_{H_0}$ and

$$\lim_{\tau_d \to 0} \mathbb{P}(\hat{\tau}_N + z_\alpha \hat{\sigma}^2_{\hat{\tau}_N} < \tau_{H_0} | \tau = \tau_{H_0}) = \alpha, \tag{26}$$

$$\lim_{\tau_d \to 0} \mathbb{P}(\hat{\tau}_N + z_\alpha \hat{\sigma}^2_{\hat{\tau}_N} < \tau_{H_0} | \tau = \tau_{H_1}) = 1 - \beta, \tag{27}$$

for $\tau_{H_1} < \tau_{H_0}$.

Proof. We prove this corollary for $\tau_{H_1} > \tau_{H_0}$, but where the exact same steps can be taken to prove the result for $\tau_{H_1} < \tau_{H_0}$. We begin by noting that equation (24) follows directly from Theorem 1, so what remains to be proven is equation (25). Theorem 1 implies that for a given half-width of a confidence interval, d, as $d \to 0$, it is the case that $\hat{\tau} \sim \mathcal{N}(\tau, d^2/z^2)$, for the stopping rule $g(\hat{\sigma}^2_{\hat{\tau}_n}, d^2/z^2) = N$ with $z > 0$, which means that

$$\lim_{d \to 0} \mathbb{P}((\hat{\tau}_N - \tau_{H_1})/\hat{\sigma}_{\hat{\tau}_N} < z | \tau = \tau_{H_1}) = F(z). \tag{28}$$

For finite z, together with the stopping rule which implies that $\tau_d = d$, we get

$$\lim_{\tau_d \to 0} \mathbb{P}((\hat{\tau}_N - \tau_{H_1})/\hat{\sigma}_{\hat{\tau}_N} < z | \tau = \tau_{H_1}) = F(z). \tag{29}$$

We can re-write this equation as

$$\lim_{\tau_d \to 0} \mathbb{P}(\hat{\tau}_N - z\hat{\sigma}_{\hat{\tau}_N} < \tau_{H_1} | \tau = \tau_{H_1}) = F(z). \tag{30}$$

Using the fact that $\tau_{H_0} = \tau_{H_1} - (z_\alpha + z_\beta)\hat{\sigma}_{\hat{\tau}_N}$, we get

$$\lim_{\tau_d \to 0} \mathbb{P}(\hat{\tau}_N - (z + z_\alpha + z_\beta)\hat{\sigma}_{\hat{\tau}_N} < \tau_{H_0} | \tau = \tau_{H_1}) = F(z). \tag{31}$$

Let $z = -z_\beta$ and switching the inequality sign, we get

$$\lim_{\tau_d \to 0} \mathbb{P}(\hat{\tau} - z_\alpha \hat{\sigma}_{\hat{\tau}_N} > \tau_{H_0} | \tau = \tau_{H_1}) = F(z_\beta). \tag{32}$$

By definition, $F(z_\beta) = 1 - \beta$, which proves the result. □

Corollary 2 is a powerful result. It says that for a FPD to have size α and power $1-\beta$, it is sufficient to specify τ_{H_0}, τ_{H_1} together with α and β. This is in contrast to traditional fixed-sample designs where it is also necessary to specify σ_0^2 and σ_1^2 (i.e., the variance of the outcome under control and treatment), something that is often the most difficult part of a power analysis. In some settings, estimates of σ_0^2 may be available with pre-experimental data. However, σ_1^2 may be more difficult to assess if no one historically has received the treatment, something that would be the case in many clinical trial settings, as well as in A/B testing in the tech industry.

Still, the FPD is not a "free lunch" because it comes with one drawback compared to fixed-sample designs: the sample size, N, is stochastic. This fact could be a problem in settings where sampling is costly, and the experimenter does not want to run the risk of sampling for "too long". As noted in Remark 2 for the FWCID, the stopping rule can be rewritten such that—at any point during the experiment—a forecast for how long the experiment will run for can be made as

$$\hat{N}_n = \frac{\hat{\sigma}^2_{\mathcal{P}_n}(z_\alpha + z_\beta)^2}{(\tau_{H_1} - \tau_{H_0})^2 \hat{p}_n(1-\hat{p}_n)}, \tag{33}$$

where $\hat{N}_n$ is interpreted as the estimated required sample size to achieve power of $1-\beta$ if $\tau = \tau_{H_1}$. Hence, if the forecast made suggest that the experiment would need to run for longer than what is defensible for ethical, monetary or other reason, the experiment can be aborted and the resulting fixed-sample confidence interval at that point can be used as confidence interval. Note, of course, that this only works if the decision to terminate is based on $\hat{N}_n$ (thereby implicitly on $\hat{\sigma}^2_{\mathcal{P}_n}$) and not based on $\hat{\tau}_n$.

Finally, we also note that we can get a double-sided test with correct size α and approximate power $1-\beta$ by replacing z_α with $z_{\alpha/2}$.[2]

3.2 Asymptotically conservative FWCID

Theorem 1 and Corollary 2 gives asymptotic guarantees that FWCID gives correct confidence interval coverage and that FPD gives correct size and power. For finite samples, the coverage of the confidence interval under

[2]The power is only approximate because is excludes the very small probability of rejecting the null in the "wrong direction" (i.e., rejecting the null because $\hat{\tau}_N$ is sufficiently smaller than τ_{H_0} when $\tau_{H_1} > \tau_{H_0}$), something that is not relevant in practice. It is possible to get the correct power by noting that $F(z'_\beta) + F(-(2z_{\alpha/2} + z'_\beta)) = 1-\beta$ for some z'_β; a value that is straightforward to solve for numerically.

FWCID will generally be too small since the stopping rule implies that the estimated variance of the treatment effect, $\hat{\sigma}^2_{\hat{\tau}_N}$ is (slightly) smaller, in expectation, than the true variance, $\sigma^2_{\hat{\tau}_N}$. This is intuitive, as we will stop too early only when we underestimate the variance, but never when we overestimate it. Another way of framing this is to say that since $\hat{\sigma}^2_{\hat{\tau}_N} = \hat{\sigma}^2_{\mathcal{P}_N}\kappa_N$, for $d > 0$, it is generally the case that $\mathbb{E}(\hat{\sigma}^2_{\mathcal{P}_N}) < \sigma^2_{\mathcal{P}}$.

To deal with this issue, we would like to find a conservative estimator for $\sigma^2_{\mathcal{P}}$. The way we propose to achieve this is to form a upper confidence sequence for $\sigma^2_{\mathcal{P}}$ that is "always-valid" (also called "anytime valid"). This sequence, U_n, for a target parameter θ is an always-valid $(1-\alpha)$ upper confidence sequence if, given any stopping time n, it holds that:

$$\mathbb{P}(\theta \leq U_n) \geq 1 - \alpha, \forall n \in \mathbb{N}^+, \tag{34}$$

see, e.g., Howard et al. (2021). Here we rely on the asymptotic version of such a confidence sequence as defined in Waudby-Smith et al. (2024). The upper bound of the always-valid confidence sequence forms a conservative sequence of estimates for $\sigma^2_{\mathcal{P}}$ which can be used in the stopping rules in Theorem 1 and Corollary 2. Lemma 1 shows how such a sequence can be formed.

Lemma 1. *$\hat{\sigma}^2_{\mathcal{P}_n} + \sqrt{\mathbb{V}(Z)}\cdot\phi_n(\rho,\alpha_c)$, forms an asymptotic $1-\alpha_c$ always-valid upper confidence sequence for $\sigma^2_{\mathcal{P}}$, where*

$$\phi_n(\rho,\alpha_c) := \sqrt{\frac{2(n\rho_c^2+1)}{n^2\rho^2}\log\left(1+\frac{\sqrt{n\rho^2+1}}{2\alpha_c}\right)}, \tag{35}$$

for the uncentered influence function $Z_i := \frac{1-p}{p}W_i(Y_i-\mu_1)^2 + \frac{p}{1-p}(1-W_i)(Y_i-\mu_0)^2$ and any constant $\rho > 0$.

Proof. By Corollary 3.4 and Proposition B.1 in Waudby-Smith et al. (2024), using the fact that $\mathbb{V}(Z-\sigma^2_{\mathcal{P}}) = \mathbb{V}(Z)$, Lemma (1) holds if $\mathbb{V}(Z)$ is finite and

$$\hat{\sigma}^2_{\mathcal{P}_n} - \sigma^2_{\mathcal{P}} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \sigma^2_{\mathcal{P}}) + o\left(\sqrt{\frac{\log(n)}{n}}\right), \text{ a.s.} \tag{36}$$

Define $\tilde{\sigma}^2_{wn} := \frac{1}{n_w}\sum_{i:W_i=w}^{n}(Y_i(w)-\mu_w)^2$, for $w = 0, 1$ and $\tilde{\sigma}^2_{\mathcal{P}_n} := \frac{1}{n}\sum_{i=1}^{n} Z_i$ as natural variance estimators for σ^2_w and $\sigma^2_{\mathcal{P}}$ if μ_w had been known. Equation

(36) simplifies to

$$\begin{aligned}\hat{\sigma}^2_{\mathcal{P}_n} - \tilde{\sigma}^2_{\mathcal{P}_n} &= \frac{n_0}{n}\hat{\sigma}^2_{1n} + \frac{n_1}{n}\hat{\sigma}^2_{0n} - \frac{1}{n}\sum_{i=1}^{n} Z_i + o\left(\sqrt{\frac{\log(n)}{n}}\right) \\ &= (1-\hat{p}_n)\hat{\sigma}^2_{1n} + \hat{p}_n\hat{\sigma}^2_{0n} - (1-p)\frac{\hat{p}_n}{p}\tilde{\sigma}^2_{1n} - p\frac{1-\hat{p}_n}{1-p}\tilde{\sigma}^2_{0n} + o\left(\sqrt{\frac{\log(n)}{n}}\right) \\ &= (1-\hat{p}_n)\left(\hat{\sigma}^2_{1n} - \frac{1-p}{1-\hat{p}_n}\frac{\hat{p}_n}{p}\tilde{\sigma}^2_{1n}\right) + \hat{p}_n\left(\hat{\sigma}^2_{0n} - \frac{p}{\hat{p}_n}\frac{1-\hat{p}_n}{1-p}\tilde{\sigma}^2_{0n}\right) + o\left(\sqrt{\frac{\log(n)}{n}}\right). \end{aligned} \tag{37}$$

Because of symmetry of $\hat{p}_n\left(\hat{\sigma}^2_{1n} - \tilde{\sigma}^2_{1n}\right)$ and $(1-\hat{p}_n)\left(\hat{\sigma}^2_{0n} - \tilde{\sigma}^2_{0n}\right)$, it is sufficient to show that one of the first two terms is $o(\sqrt{\log(n)/n})$ a.s. for equation (36) to hold. Define $v_n := (1-p)\hat{p}_n/((1-\hat{p}_n)p)$, we have

$$\begin{aligned}\hat{p}_n(\hat{\sigma}^2_{1n} - v_n\tilde{\sigma}^2_{1n}) &= \hat{p}_n\frac{1}{n_1}\sum_{i:W_i=1}(Y_i(1)-\hat{\mu}_{1n})^2 - \hat{p}_n v_n\frac{1}{n_1}\sum_{i:W_i=1}(Y_i(1)-\mu_1)^2 \\ &= \hat{p}_n\left(\frac{1}{n_1}\sum_{i:W_i=1}Y_i(1)^2 - \hat{\mu}^2_{1n}\right) - \hat{p}_n v_n\left(\frac{1}{n_1}\sum_{i:W_i=1}Y_i(1)^2 + \mu_1^2 - 2\hat{\mu}_{1n}\mu_1\right) \\ &= -\hat{p}_n(\hat{\mu}^2_{1n} + v_n\mu_1^2 - v_n 2\hat{\mu}_{1n}\mu_1) + (1-v_n)\hat{p}_n\frac{1}{n_1}\sum_{i:W_i=1}Y_i(1)^2 \\ &= -\hat{p}_n(\hat{\mu}^2_{1n} + \mu_1^2 - 2\hat{\mu}_{1n}\mu_1) + \hat{p}_n(1-v_n)\left(\mu_1^2 - 2\hat{\mu}_{1n}\mu_1 + \frac{1}{n_1}\sum_{i:W_i=1}Y_i(1)^2)\right) \\ &= -\hat{p}_n(\hat{\mu}_{1n}-\mu_1)^2 + \hat{p}_n(1-v_n)\left(\mu_1^2 - 2\hat{\mu}_{1n}\mu_1 + \frac{1}{n_1}\sum_{i:W_i=1}Y_i(1)^2)\right). \end{aligned} \tag{38}$$

The second term is $O(1/\sqrt{n})\cdot o(1)$ a.s., and so we only need to verify that the first term is $o(\sqrt{\log(n)/n})$. Studying the absolute difference, $\hat{p}_n|(\hat{\mu}_{1n}-\mu_1)|$ and using the triangle inequality, we have

$$\begin{aligned}\hat{p}_n|\hat{\mu}_{1n}-\mu_1| &= |\hat{p}_n\hat{\mu}_{1n} - \hat{p}_n\mu_1| \\ &= |\hat{p}_n\hat{\mu}_{1n} - p\mu_1 - (\hat{p}_n\mu_1 - p\mu_1)| \\ &\leq |\hat{p}_n\hat{\mu}_{1n} - p\mu_1| - \mu_1\,|\hat{p}_n - p|\,. \end{aligned}$$

Each of these two terms are mean-centered i.i.d random variables which, by LIL (law of the iterated logarithm), means that

$$\hat{p}_n|\hat{\mu}_{1n}-\mu_1| = O\left(\sqrt{\frac{\log(\log(n))}{n}}\right), \text{ a.s.} \iff \hat{p}_n(\hat{\mu}_{1n}-\mu_1)^2 = O\left(\frac{\log(\log(n))}{n}\right), \text{ a.s.} \tag{39}$$

Hence,

$$|\hat{\sigma}^2_{\mathcal{P}_n} - \tilde{\sigma}^2_{\mathcal{P}_n}| = O\left(\frac{\log(\log(n))}{n}\right) = o\left(\sqrt{\frac{\log(n)}{n}}\right), \text{ a.s.}, \tag{40}$$

which concludes the proof. □

Lemma 1 includes the unknown variance of Z. By replacing μ_w with $\hat{\mu}_w$ we can define

$$\hat{Z}_i := W_i(Y_i(1) - \hat{\mu}_1)^2 + (1 - W_i)(Y_i(0) - \hat{\mu}_0)^2, \tag{41}$$

and estimate the variance as

$$\widehat{\mathbb{V}(Z)} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Z}_i - \frac{1}{n}\sum_{j=1}^{n}\hat{Z}_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Z}_i - \hat{\sigma}^2_{\mathcal{P}_n}\right)^2. \tag{42}$$

Using Lemma 1 we can construct an asymptotic always-valid upper confidence sequence for $\sigma^2_{\mathcal{P}}$ for a given confidence level α_c. The upper bound of this confidence sequence is

$$\hat{\sigma}^2_{\mathcal{P}^{ub}_n} := \hat{\sigma}^2_{\mathcal{P}_n} + \sqrt{\widehat{\mathbb{V}(Z)}} \cdot \phi_n(\rho, \alpha_c). \tag{43}$$

Using this result means that we can now construct a conservative stopping rule as

$$g(\hat{\sigma}^2_{\mathcal{P}^{ub}_n}\kappa_n, d^2/z^2_{\alpha/2}) = N^c. \tag{44}$$

I.e., N^c is the sample size at which sampling is stopped. We can now state the analogous result to Theorem 1 for the stopping rule in equation (44):

Theorem 2. *Assume that* $\mathbb{E}(Y(1)^4) < \infty$ *and* $\mathbb{E}(Y(0)^4) < \infty$. *For the target parameter* τ *and the stopping rule* $g(\hat{\sigma}^2_{\mathcal{P}^{ub}_n}\kappa_n, d^2/z^2_{\alpha/2}) = N^c$, *with* $0 < \alpha < 1, 0 < \alpha_c < 1$, *and the estimator of the treatment effect* $\hat{\tau}_{N^c}$, I_{N^c} *is a fixed-width confidence interval. I.e., it is the case that*

$$\lim_{d\to 0}\mathbb{P}(\tau \in I_{N^c}) = 1 - \alpha \tag{45}$$

Proof. The proof is analogous to the proof of Theorem 1, but where s_n in equation (12) is replaced with:

$$s_n^c := \frac{\hat{\sigma}^2_{\mathcal{P}_n^{ub}}}{\sigma^2_{\mathcal{P}}} \frac{p}{\hat{p}_n} \frac{1-p}{1-\hat{p}_n}. \tag{46}$$

Once again, $s_n^c > 0$ a.s., and so, if we can show that

$$\lim_{n\to\infty} s_n^c = 1, \text{ a.s.}, \tag{47}$$

then the theorem is proven by following the exact same steps as in the proof for Theorem 1. Furthermore, using Lemma 1 (which holds because of finite fourth moments), equation (47) is true if

$$\lim_{n\to\infty} \frac{\hat{\sigma}^2_{\mathcal{P}_n^{ub}}}{\sigma^2_{\mathcal{P}}} = 1, \text{ a.s.} \tag{48}$$

By the definition of $\hat{\sigma}^2_{\mathcal{P}_n^{ub}}$ in equation (43), together with the fact that $\lim_{n\to\infty} \hat{\sigma}^2_{\mathcal{P}_n}/\sigma^2_{\mathcal{P}} = 1$ a.s., we have

$$\lim_{n\to\infty} \frac{\hat{\sigma}^2_{\mathcal{P}_n^{ub}}}{\sigma^2_{\mathcal{P}}} = 1 + \frac{\sqrt{\frac{1}{n}\sum_{i=1}^n \left(\hat{Z}_i - \hat{\sigma}^2_{\mathcal{P}_n}\right)^2} \cdot \phi_n(\rho, \alpha_c)}{\sigma^2_{\mathcal{P}}}. \tag{49}$$

Hence, we need to show that the second term tends to zero almost surely. It is the case that $\phi_n(\rho, \alpha_c) = o(1)$ for $\alpha_c > 0$. We also have

$$\frac{1}{n}\sum_{i=1}^n \left(\hat{Z}_i - \hat{\sigma}^2_{\mathcal{P}_n}\right)^2 = \frac{1}{n}\sum_{i=1}^n \hat{Z}_i^2 - \hat{\sigma}^4_{\mathcal{P}_n} = \\ \frac{1}{n}\left(\sum_{i=1}^n W_i(Y_i(1) - \hat{\mu}_1)^4 + \sum_{i=1}^n (1-W_i)(Y_i(0) - \hat{\mu}_0)^4\right) - \hat{\sigma}^4_{\mathcal{P}_n}. \tag{50}$$

Let μ_{4w} be the fourth central moment for $Y(w)$, it is the case that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \left(\hat{Z}_i - \frac{1}{n}\sum_{j=1}^n \hat{Z}_j\right)^2 = p\mu_{41} + (1-p)\mu_{40} - \sigma^4_{\mathcal{P}} \text{ a.s.} \tag{51}$$

Because we assumed finite fourth moments of the potential outcomes, this term is $O(1)$ a.s. Hence,

$$\lim_{n\to\infty} \frac{\hat{\sigma}^2_{\mathcal{P}^{ub}_n}}{\sigma^2_{\mathcal{P}}} = 1 + O(1) \cdot o(1) \text{ a.s.} = 1 \text{ a.s.}, \tag{52}$$

which proves the theorem. □

Just as the naive FWCID design in Theorem 1 gives the FPD design in Corollary 2, we can use the conservative FWCID design in Theorem 2 to get a conservative version of the FPD design 2 by replacing $\hat{\sigma}^2_{\hat{\tau}_n}$ with $\hat{\sigma}^2_{\mathcal{P}^{ub}_n}\kappa_n$ in Corollary 2.

It is important to distinguish between α and α_c. The former gives the confidence level of the targeted confidence interval for the treatment effect estimator. The latter is the confidence level of the confidence sequence of $\sigma^2_{\mathcal{P}}$. These two values need not be the same and α_c should be chosen based on just how conservative the experimenter wants to be (or, phrased differently, how much to insure against stopping too early), where a smaller α_c means being more conservative (stopping later).

However, regardless of how small a value of α_c is chosen, asymptotically, any $\alpha_c > 0$ leads to asymptotic efficiency in the sense discussed in Remark 1. I.e., the ratio of N and N^c converges to 1 as $d \to 0$. Of course, for finite-sample performance, the choice of α_c matters. One may question why we are not simply using the always-valid sequential confidence sequence directly for the treatment effect estimator, using the results in, e.g., Howard et al. (2021). Doing so would clearly by definition make the CI coverage bounded by $1-\alpha$ at any point, including the stopping point given by a CI-width based rule. The reason is that these always-valid confidence sequences are less efficient. Instead of converging at the rate $O(1/\sqrt{n})$, these sequences instead converge at the rate implied by LIL, i.e., $O(\log(\log(n))/\sqrt{n})$, which means that the ratio of the sample size implied by such a design to N or N^c will tend to infinity asymptotically. We show the poor finite-sample performance (in terms of efficiency) of this design in the simulations below.

4 Simulation Study

To study finite sample behavior of the FWCID, we perform Monte Carlo studies. In addition to naive (Theorem 1) and conservative (Theorem 2)

versions of the FWCID, we also include an always-valid version using the interval in Waudby-Smith et al. (2024). That is, sampling is stopped once the always-valid half-width is smaller than d. The half-width is equal to $v_n \cdot \psi_n(\rho, \alpha)$, where these values are given by (see Appendix A.1 in Maharaj et al. 2023):

$$v_n = \sqrt{\frac{1}{\hat{p}_n}(\hat{\sigma}^2_{1n} + \hat{\mu}^2_{1n}) + \frac{1}{1-\hat{p}_n}(\hat{\sigma}^2_{0n} + \hat{\mu}^2_{0n}) - (\hat{\mu}_{1n} - \hat{\mu}_{0n})^2}, \tag{53}$$

$$\psi_n(\rho, \alpha) = \sqrt{\frac{2(n\rho^2+1)}{n^2\rho^2} \log\left(\frac{\sqrt{n\rho^2+1}}{\alpha}\right)}. \tag{54}$$

Note that $\psi_n(\rho, \alpha)$ differs from $\phi_n(\rho, \alpha)$ because the former is for a symmetric confidence interval, whereas the latter is for a upper confidence interval (i.e., the lower bound is set to $-\infty$). To summarize, the stopping rules for the three designs we consider are:

$$\begin{aligned} \text{FWCID, naive}: &\quad g(\hat{\sigma}^2_{\hat{\tau}_n}, d^2/z^2_\alpha), \\ \text{FWCID, conservative}: &\quad g(\hat{\sigma}^2_{\mathcal{P}^{ub}_n}\kappa_n, d^2/z^2_\alpha), \\ \text{FWCID, always-valid}: &\quad g(v^2_n, d^2/\psi^2_n(\rho, \alpha)). \end{aligned}$$

We study eight different data-generating processes (DGPs), where treatment is decided by a standard Bernoulli trial with probability of treatment equaling $p = .5$. We have the following eight DGPs:

$$\begin{aligned} DGP1: &\quad a_1 Y(0) \sim Normal(0,1), && a_1 Y(1) \sim Normal(0,1) \\ DGP2: &\quad a_2 Y(0) \sim Lognormal(0,1), && a_2 Y(1) \sim Lognormal(0,1) \\ DGP3: &\quad a_3 Y(0) \sim Normal(0,1), && a_3 Y(1) \sim Normal(0.2,1) \\ DGP4: &\quad a_4 Y(0) \sim Lognormal(0,1), && a_4 Y(1) \sim Lognormal(0,1) + .2 \\ DGP5: &\quad a_5 Y(0) \sim Lognormal(0,1), && a_5 Y(1) \sim Lognormal(c_1,1) \\ DGP6: &\quad a_6 Y(0) \sim Gamma(1,1), && a_6 Y(1) \sim Lognormal(c_2, 9/16) \\ DGP7: &\quad a_7 Y(0) \sim Gamma(1,1), && a_7 Y(1) \sim Gamma(1,c_3) \\ DGP8: &\quad a_8 Y(0) \sim Gamma(1,1), && a_8 Y(1) \sim Gamma(c_4,1). \end{aligned}$$

The constants a_j are set such that

$$\mathbb{V}(Y(0))/(1-p) + \mathbb{V}(Y(1))/p = 1, \tag{55}$$

to facilitate easier comparison of results between simulations of different DGPs. The first two DGPs have no treatment effects, whereas the next two have a homogeneous effect of $\tau = .2$. For the last four DGPs, the treatment effect is heterogeneous where $c_1 = 1.266\ldots, c_2 = -.0949\ldots, c_3 = 1.223\ldots, c_4 = 1.210\ldots$ are set such that $\tau = .2$.

We perform simulations for different values of d, where d range from .01 to .5 in increments of .01. For each value of d and DGP, we perform 100,000 replications. Let N^* be the sample size for a fixed-sample confidence interval. The asymptotic value of d for a fixed-sample confidence interval is

$$d^* = \sqrt{\mathbb{V}(Y(0)/((1-p)N^*) + \mathbb{V}(Y(1)/(pN^*)} \cdot z_{\alpha/2}. \tag{56}$$

Using the fact that $p = .5$ together with equation (55), we can solve for N^* as

$$N^* = \frac{4z_{\alpha/2}^2}{d^{*2}}. \tag{57}$$

Let $\overline{N}$ be the value of N^* we get if we replace d^* with d. We set $\alpha = .1$ and so for $d = .01, .02, \ldots, .49, .50$, we get $\overline{N} = 108{,}222; 27{,}055; \ldots; 45; 43$, which gives a rough estimate of how big the sample sizes should be for the different simulations. We let $\alpha_c = .1$, whereas ρ is set such that ϕ_n and ψ_n are minimized for $n = \overline{N}$, respectively (in line with the suggestion of Waudby-Smith et al. 2024).

4.1 Simulation results for FWCID

We show three sets of results: one for bias (i.e., average treatment effect estimate minus population average treatment effect; Figure 2), one for coverage of a 90% confidence interval (Figure 3) and one for relative sample size, $N/\overline{N}$ (Figure 4). The asymptotic results are for when $d \to 0$: hence the convergence goes from right to left in each graph. In general, we have enough replications to make sampling variation negligible; hence, no standard errors or intervals are included in the graphs.

Figure 2 displays the bias of the difference-in-means estimator. In line with Remark 3, we see that when treatment effects are homogeneous, there is no bias at all. With heterogeneous treatment effects, there is bias that depend on the DGP, but, as expected, this bias disappears as $d \to 0$. Generally, the naive design has the largest bias followed by the conservative design with

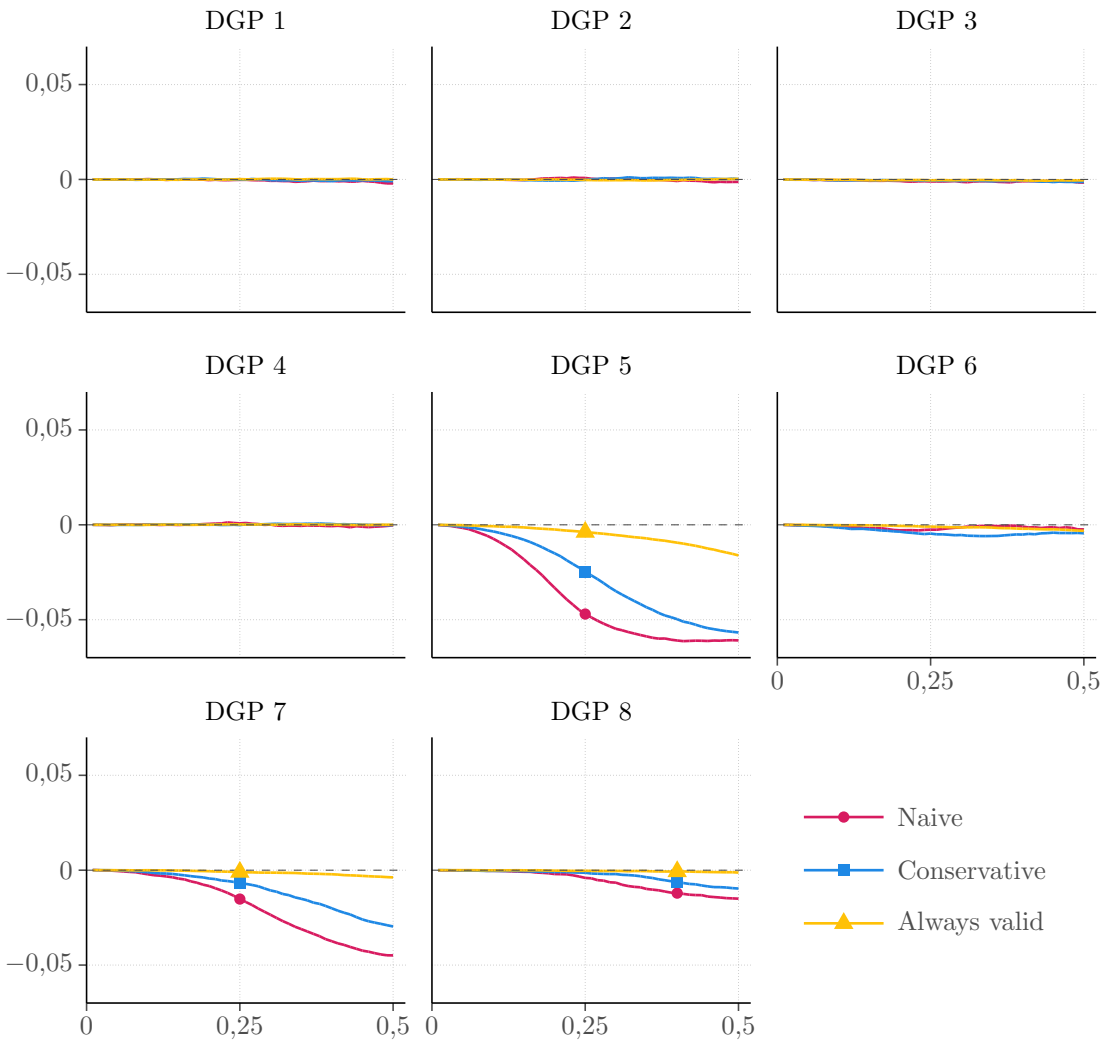

Figure 2: Bias (y-axis) of the difference-in-means estimator as a function of d (x-axis) for the FWCID

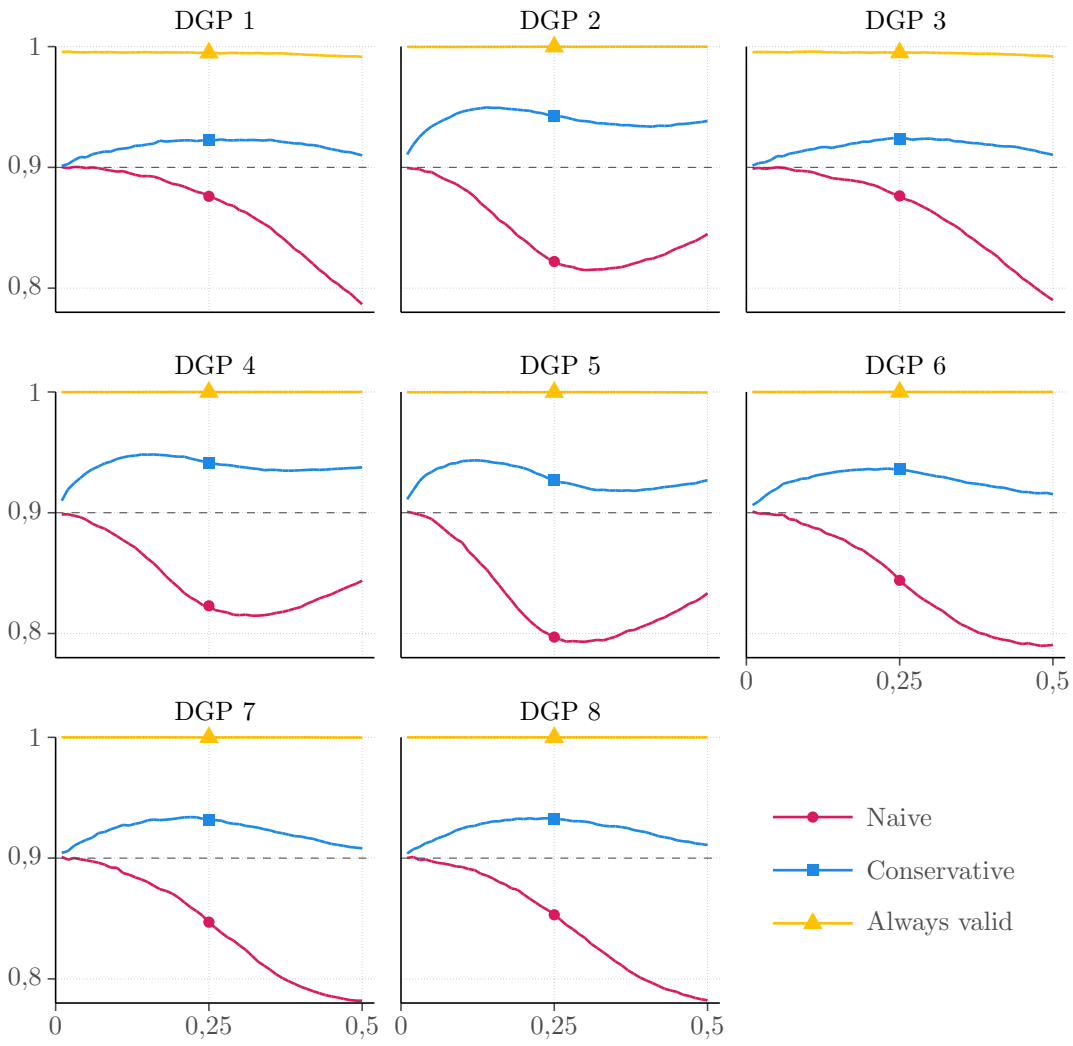

Figure 3: Coverage of a 90% confidence interval (y-axis) of the treatment effect at the stopping point as a function of d (x-axis) for the FWCID

the always-valid having the smallest bias. This is likely due to the fact that sample sizes for a given d is increasing in that order.

Turning to coverage of the 90% confidence interval, we see that the naive interval generally undershoots but converges to 90%, where the opposite is generally the case for the conservative confidence interval. As expected, the always-valid interval has too wide coverage in all settings with the probability that the interval covers the true parameter being almost one for all DGPs.

Finally, the explanation for the pattern in Figure 3 is given in Figure 4. Here the ratio of the average sample size to the sample size of a fixed-sample design is plotted against d. We see that the sample size of the naive design is generally smaller, as expected, but the ratio converges to one as $d \to 0$ (see Remark 1), with the fastest convergence happening when data are normal. For the conservative design, convergence generally comes from above as desired and, again, convergence is fastest for normal data. For the always-valid design, the relative sample size is much larger, especially when data follow a skewed distribution. This conservative behavior is a known property of the always-valid CI sequences due to the LIL.

4.2 Simulation results for hypothesis testing

To evaluate the FPD, we perform simulations for the six DGPs which have an average treatment effect of .2. We set $\alpha = .05$, $\alpha_c = .1$ and let $\tau_{H_0} = 0$, while τ_{H_1} (and, hence, τ_d) is varied from .1 to .4 in increments of .01. We also need to set a maximum sample size, N^{max}, which we set to either $\overline{N}$, $1.5\overline{N}$ or $2\overline{N}$, with $\overline{N}$ being defined as

$$\overline{N} := \frac{(z_\alpha + z_\beta)^2}{\tau_d^2 p(1-p)}. \tag{58}$$

We let the null be $\tau \leq t_{H0}$ and we have $p = .5$. Aside from the two FPD designs (naive and conservative) we include two other designs. The first is the always-valid approach, where a sequence of lower confidence intervals are formed (i.e., the upper bound of the interval is $+\infty$) and the experiment is stopped and the null is rejected as soon as the interval does not cover τ_{H_0}. The second is the commonly used group sequential test (GST, Lan and DeMets 1983) with alpha spending function $\alpha_n = \alpha \cdot (n/N^{max})^2$, which gives a sequence of critical z-values z_n where the experiment is stopped and the null is rejected as soon as the standard z-statistic exceeds z_n. In summary,

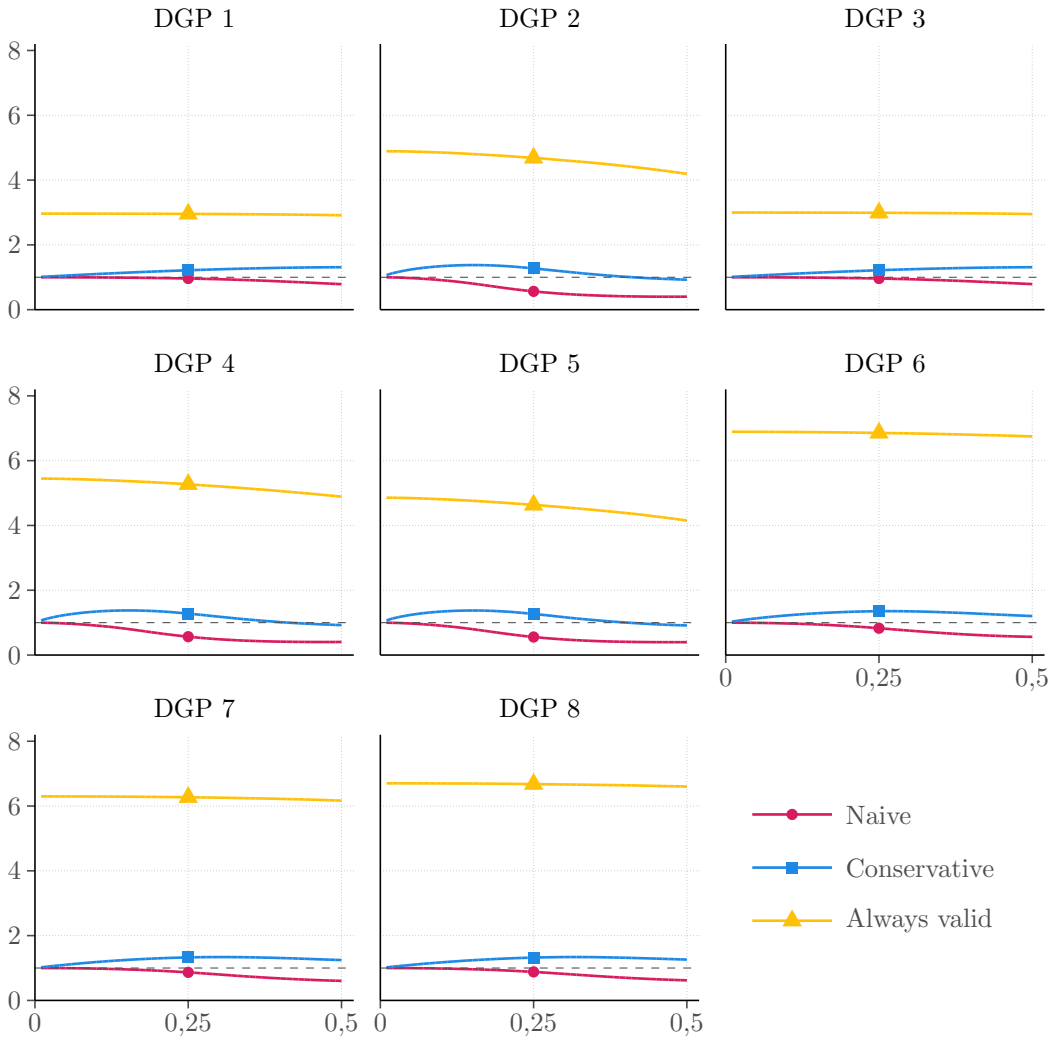

Figure 4: Relative sample size, $E(N)/\overline{N}$ (y-axis) as a function of d (x-axis) for the FWCID

the four stopping rules in the simulations are:

$$\begin{aligned}
\text{FPD, naive}: &\quad g(\hat{\sigma}^2_{\hat{\tau}_n}, \tau_d^2/(z_\alpha + z_\beta)^2), \\
\text{FPD, conservative}: &\quad g(\hat{\sigma}^2_{\mathcal{P}_n^{ub}}\kappa_n, \tau_d^2/(z_\alpha + z_\beta)^2), \\
\text{Always-valid test}: &\quad g(-(\hat{\tau}_n - \tau_{H_0})/v_n, -\phi_n(\rho, \alpha)), \\
\text{GST}: &\quad g(-(\hat{\tau}_n - \tau_{H_0})/\sigma_{\hat{\tau}_n}, -z_n).
\end{aligned}$$

The minus signs for the last two approaches come from the fact that the stopping rule as defined in Definition 2 is to stop as soon as the function of data is *smaller* than the sequence of constants. Once again, ρ for the conservative FPD and the always-valid test are set to minimize ϕ_n, for $n = \overline{N}$ (ϕ_n is used instead of ψ_n here for the always-valid approach since we are only considering one-sided tests). Note that the always-valid test is not a FPD because we stop on significance and not on confidence width.[3]

We begin by studying bias in Figure 5. As expected, both the naive and conservative estimators are unbiased with homogeneous treatment effects (DGP 3 and 4) and approximately unbiased once the sample size gets larger (which is the case when τ_d, and hence τ_{H_1}, gets smaller) for the other DGPs. Also as expected, the GST is severely biased, whereas the always-valid test is somewhere in between.

Turning to power (Figure 6), we first note that the results depend heavily on which N^{max} that is chosen. For $N^{max} = \overline{N}$, the conservative estimator has low power because it is unlikely that the desired confidence width is achieved at the time of $n = N^{max}$. The always-valid approach performs better with a low τ_{H_1} (large sample size) because with some probability, the null will be rejected also with a relatively large confidence width because the treatment effect estimator is unexpectedly large (which also means the estimator is biased, see Figure 5). As expected the GST performs the best.

Note that, according to the theoretical result, the power of both the naive and conservative tests should be $1 - \beta = .8$ when $\tau_d = \tau_{H_1} = \tau = .2$ as long as $N < N^{max}$. For $N^{max} = 2\overline{N}$, we see that that indeed seems to be the case. The DGP with power furthest from .8 is DGP 5, where the power is around .74 for both the naive and conservative test which occurs because there are some simulations where $N = 2\overline{N}$, because the distributions of $Y(0)$ and $Y(1)$ are very skewed.

[3]One could use the always-valid confidence interval to construct a FPD, but, as shown in the simulations for the FWCID, it will have very poor properties with power essentially zero unless N^{max} is set very high.

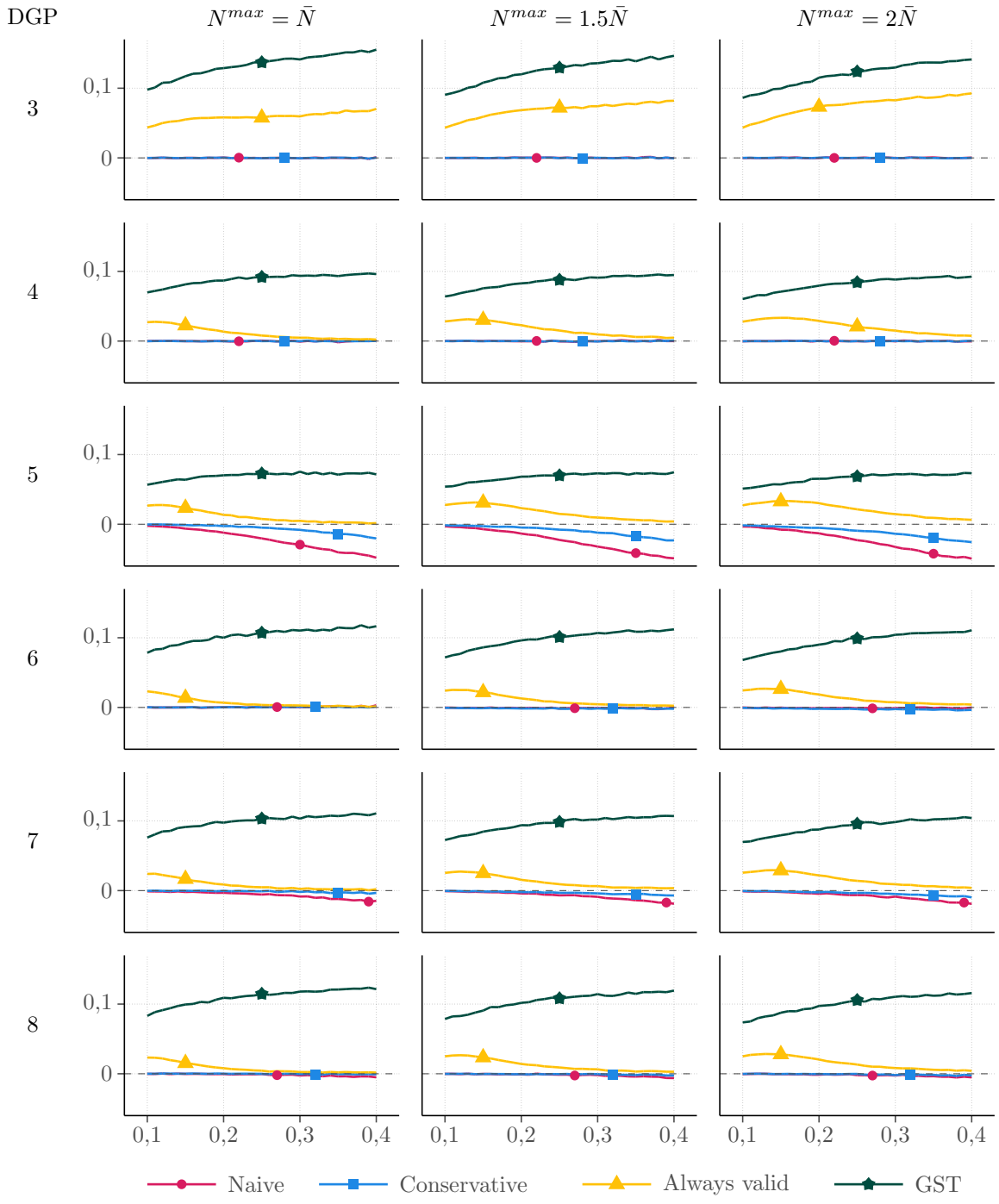

Figure 5: Bias (y-axis) for hypothesis test as a function of τ_{H_1} (x-axis)

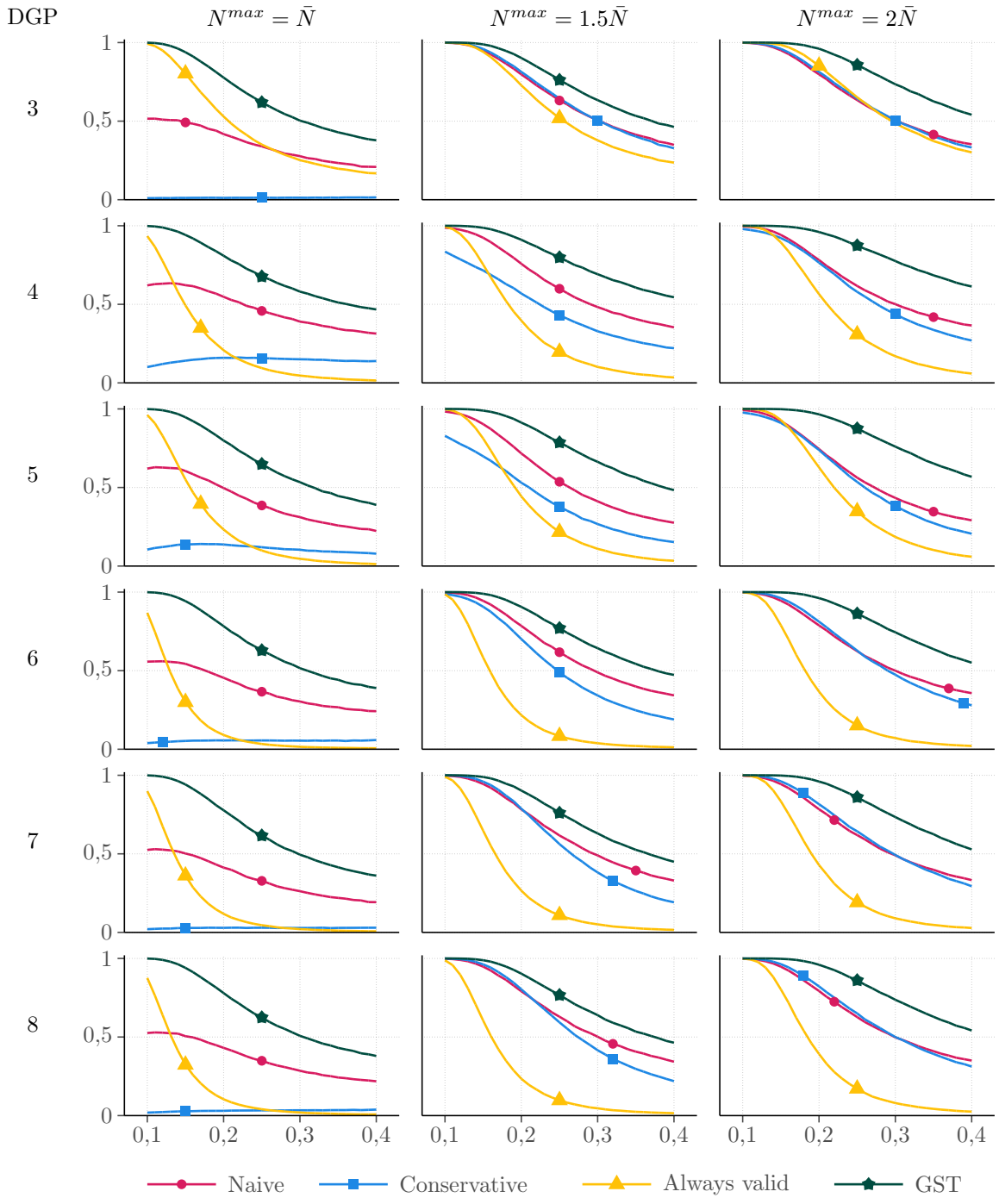

Figure 6: Power (y-axis) for hypothesis tests as a function of τ_{H_1} (x-axis)

Turning to the cases where $N^{max} > \overline{N}$, we see that the two FPDs perform better than the always-valid approach. The reason is that with these stopping rules, it is highly likely that the desired confidence width is achieved before N^{max}, and so the power will depend almost exclusively on whether the treatment effect estimate is large enough at that point. As expected, the GST continues to have the highest power throughout.

Finally, looking at the average sample size at which sampling is stopped (Figure 7), we see that for $N^{max} = \overline{N}$, the conservative stopping rule essentially never binds (i.e., sampling almost never stops before $n = N^{max}$). The same is true for the always-valid approach for large τ_{H_1}. GST generally stops earliest, although it depends on the DGP. For $N^{max} > \overline{N}$, the picture is a bit different. Here the naive stopping rule generally leads to the earliest stopping for large τ_{H_1} and even the conservative approach can stop relatively early. For small τ_{H_1}, the GST once again stops earliest.

4.3 Conclusions from the simulations

For forming a confidence interval of a fixed width—the FWCID—we compare three different methods: the naive, conservative and always-valid approaches. All three are asymptotically valid in the sense that coverage is guaranteed to be at least the desired level (and for the first two, it is asymptotically guaranteed to be exactly at that level). The finite-sample simulations indicate that the naive interval will generally have a coverage smaller than the desired level, whereas the conservative and always-valid will have a coverage larger than the desired level. In most cases, the bias is of minor importance, especially for relatively larger sample sizes (smaller confidence widths).

The big difference between the conservative and always-valid approaches comes when comparing the required sample sizes. The always-valid approach requires between two to seven times as much data as the conservative approach. Our conclusion is therefore that the conservative fixed-width confidence interval is the desirable approach since it gives asymptotic guarantees of both coverage and efficiency, while generally being conservative and at the same time not using too much data for small sample sizes. The naive approach will generally work well for large sample sizes, whereas it is hard to motivate the use of the always-valid confidence interval for FWCID.

For hypothesis testing, we compare the FPD for the naive and conservative versions with the group-sequential test and a always-valid test. The FPD has the advantage over the other two in that the point estimator will gen-

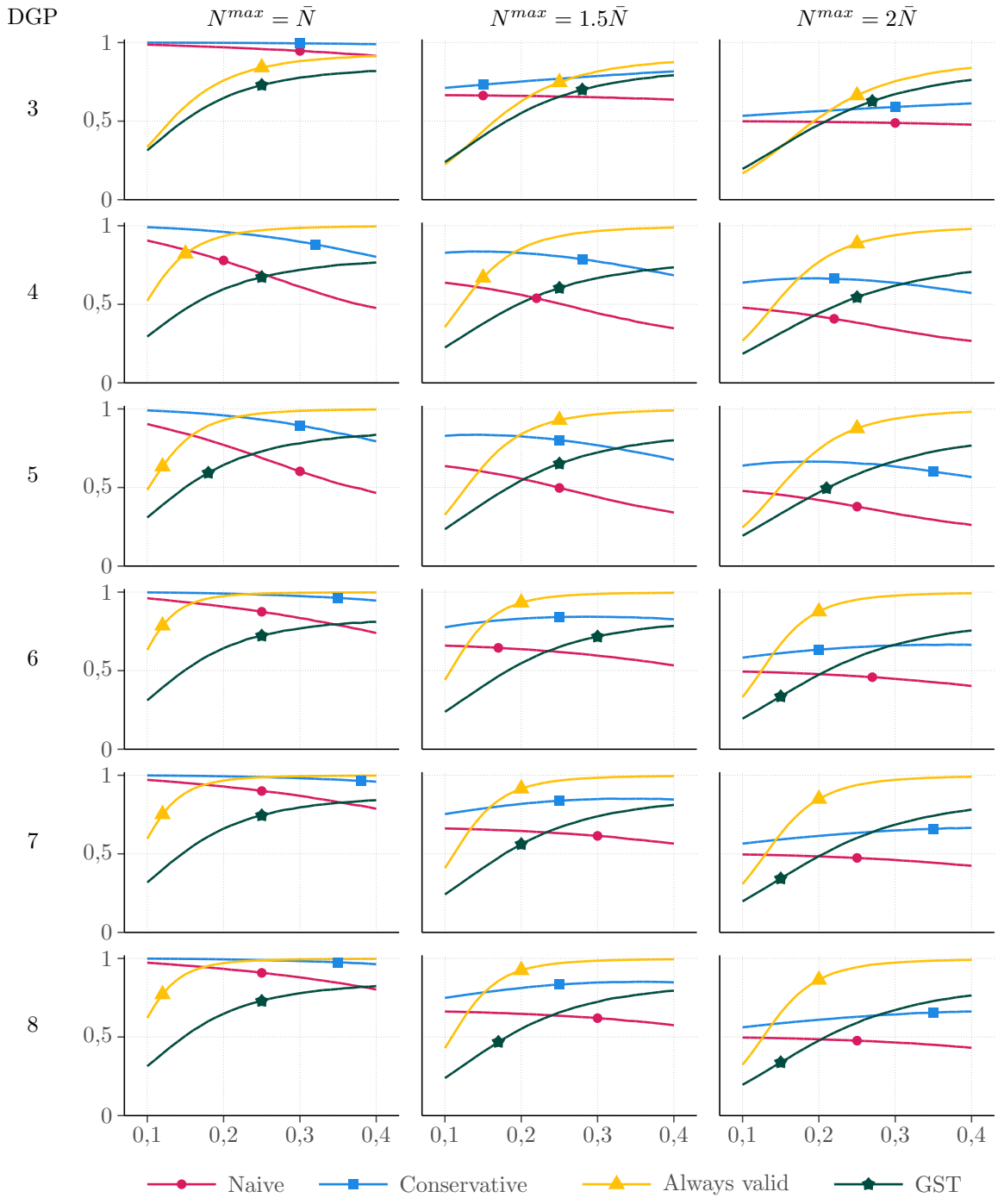

Figure 7: $E(N)/N^{max}$ (y-axis) for hypothesis tests as a function of τ_{H_1} (x-axis)

erally be less biased. However, they will generally be less powerful than the GST, whereas how efficiently they make use of data depends on the specific DGP and parameters used.

5 Discussion

In this paper we propose two sequential designs for randomized experiments where the experiment is stopped once a pre-specified precision of the estimate is reached. In the first design—the fixed-width confidence interval design (FWCID)—the experimenter specifies a confidence interval width and the experiment runs until the stopping rule is met, where the stopping rule is set such that the resulting confidence interval gives a coverage equal to the desired level. We propose two versions of this design, a naive and a conservative, where we prove that both versions give consistent estimators of the average treatment effect and that, asymptotically, the resulting interval has correct coverage and makes efficient use of the data.

The second design—the fixed-power design (FPD)—builds on the same idea of stopping once a given precision is reached, but the stopping rule is set such that a given power is reached while maintaining the size of the test. Just as for the FWCID, the treatment effect estimator is consistent and gives a valid confidence interval, and also comes in a naive and a conservative version. Both the FWCID and FPD allow for any randomization scheme (and even some deterministic ones) as long as the share that will be treated asymptotically is fixed. I.e., all standard randomization schemes, such as Bernoulli trials, biased coin designs and randomized block designs are allowed.

We contrast our proposed designs with traditional fixed-sample designs and other sequential designs. In fixed-sample designs (i.e., the sample size is fixed and decided in advance), the key consideration in the design stage is often to decide on a desired power. However, in our experience working with practitioners, power is often a difficult concept to grasp since it can be difficult to come up with a hypothetical treatment effect. Instead, we believe it can be pedagogically easier to focus on the desired precision of the estimator, such as with the FWCID (of course, this pedagogical point can be achieved also in traditional fixed-sample designs, but is perhaps made more explicit in a design explicitly targeting precision).

In addition, for a power analysis in a fixed-sample design to be performed, it is necessary to come up with estimates for the variances of the outcome

under treatment and control. At companies like e.g. Spotify, for some experiments it is possible to approximate the variance under control using historical data on the outcome. However, in many settings the historical data is simply not similar enough to the outcome data in the experiment due to growing user bases and changing user behaviors. In addition, it can be difficult to come up with any good estimate for the variance under treatment if the treatment has never been seen historically.

With the FPD, the need to provide these variances are alleviated. However, this comes at a cost: The sample size is no longer fixed which means that there is a risk the experiment runs longer than what was originally planned. On the other hand, this is not necessarily a bug, but rather a feature, of the design. It is not uncommon that experimenters come up with optimistic values in their power calculations to show that it is worth running the experiment, leading to underpowered studies. With the FPD, this possibility is partly restricted (it is still possible to come up with an optimistic hypothetical treatment effect), forcing the experimenter to be more honest. Nevertheless, there is usually a limit to how long an experiment can run, and with the FPD (and FWCID) it is possible to, at any point during the experiment, get an estimate of the required sample size before the experiment concludes. If the decision to terminate an experiment early is based solely on the estimated required sample size, such a decision will not compromise the validity of results for experiments going the distance. In addition, for an experiment that is terminated early, traditional point estimates and confidence intervals will be valid. We stress that this strategy is only valid if early termination is based solely on determining that the estimated required sample size is too large. If the experimenter also base the decision on the treatment effect estimate, all bets are off.

Compared to the most commonly used sequential designs, the advantage of both the FWCID and FPD comes from the fact that the treatment effect estimator is consistent and that valid confidence intervals can be constructed without need for adjustments. The FWCID can be implemented using already established "always-valid" approaches for confidence sequences, although the drawback with such approaches is that they are typically less efficient compared to the designs we propose.

In this paper, we have only considered the case with one treatment and one control condition, as well as only studying the difference-in-means estimator. We believe that the main results of this paper could be extended to cover more general settings with more treatment conditions and many dif-

ferent estimators and estimands. We leave the study of such situations for future research.

Acknowledgement

The authors thank Steve R. Howard for feedback and input to this paper.

Replication code

Replication code for the Monte Carlo simulations are available at `https://github.com/mattiasnordin/precision-based-designs`

References

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567(7748):305.

Anscombe, F. J. (1952). Large-sample theory of sequential estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(4):600–607.

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Chow, Y. S. and Robbins, H. (1965). On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean. *The Annals of Mathematical Statistics*, 36(2):457–462.

Fan, X., DeMets, D. L., and Lan, K. G. (2004). Conditional bias of point estimates following a group sequential test. *Journal of Biopharmaceutical Statistics*, 14(2):505–530.

Gelman, A. and Carlin, J. (2014a). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

Gelman, A. and Carlin, J. (2014b). Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6):641–651.

Greenland, S. (2017). For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *European Journal of Epidemiology*, 32(1):3–20.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055 – 1080.

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2022). Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 70(3):1806–1821.

Johari, R., Pekelis, L., and Walsh, D. J. (2015). Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Maharaj, A., Sinha, R., Arbour, D., Waudby-Smith, I., Liu, S. Z., Sinha, M., Addanki, R., Ramdas, A., Garg, M., and Swaminathan, V. (2023). Anytime-valid confidence sequences in an enterprise a/b testing platform. In *Companion Proceedings of the ACM Web Conference 2023*, pages 396–400.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05". *The American Statistician*, 73(sup1):1–19.

Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H., and Ramdas, A. (2024). Time-uniform central limit theory and asymptotic confidence sequences. *arXiv preprint arXiv:2103.06476v9*.