Stopping rule based on effect size stabilization

Benjamin Kowialiewski

December 20, 2023

1 Abstract

XXX

2

2.1 Stopping rule and side effects

Sample size is a critical parameter to consider when running experiments in psychology. This parameter partially determines the probability of detecting a true effect when sampling from a target population. In Null-Significance Hypothesis Testing (NHST), defining sample size using a stopping rule based on p-values leads to side effects. For instance, one can sample from the target population until the p-value reaches significance. This method of sampling from a population inflates type-1 error probabilities and effect sizes. In other words, implementing this method increases the probability of finding an effect when there is none and leads to larger effect sizes compared to what should theoretically be observed if such a stopping rule weren't applied. Therefore, when applying this stopping rule, one ends up with a biased sample that is not representative of the target population.

2.2 Effect size stabilization and type-1 error probability

Recently, Anderson et al. (2022) proposed a method for implementing a stopping rule without inflating type-1 error probability. In this approach, the researcher samples from the population until the effect size stabilizes. Stabilization here refers to the absence of variation in the effect size throughout the sampling process, set against specific arbitrary thresholds. Consider an experiment where a researcher samples from the target population using a within-subject design. With each new participant added to the sample, the effect size (Cohen's d) is calculated. The difference between the obtained effect size and the previously observed effect size before adding the new participant is then assessed. If this difference doesn't exceed 0.05 for 5 consecutive iterations, the sampling process stops. Other-

wise, the sampling process continues until reaching stabilization.

Anderson and colleagues tested this effect size stabilization method in a simulation work. In this, two independent researchers conduct the same experiment concurrently. Researcher A follows the effect size stabilization method described above, while Researcher B, in contrast, doesn't use any stoping rule but terminates the sampling process upon Researcher A's completion. Therefore, both researchers end up with the same sample size. Their sole difference lies in Researcher A's sample being influenced by the stopping rule, while Researcher B's is not. Hence, the sample collected by Researcher B serves as a control against which Researcher A's sample can be compared. This hypothetical scenario can be simulated by generating random values from a normal distribution, each value representing a data point (i.e., one participant) in the sample. Once Researcher A reaches the stopping rule's criteria, the process is repeated to derive distributions of effect sizes and/or p-values for both researchers as needed. This simulation work reveals no difference between the samples collected by both researchers. That is, both researchers reach on average equivalent effect sizes, and this persists when considering a varying number of true effect sizes. Therefore, the method proposed by Anderson and colleagues does not lead to inflated effect sizes, and by extension, does not inflate type-1 error probability.

2.3 The question of power

The stopping rule based on effect size stabilization provides an interesting option for researcher seeking a stopping rule without inflating type-1 error probability. However, Anderson and colleagues' simulation work does not address the issue of power. That is, if a true effect size exists in the population, what would be the probability to find such an effect when applying the stopping rule? This

question has implications for the way researchers determine their sample size. If a stopping rule based on effect size stabilization ensures the ability to find a true effect inherent in the target population, it might potentially serve as an alternative to power analyses.

2.4 The present study

This study addresses the question of power in the context of the stopping rule based on effect size stabilization, as proposed by Anderson and colleagues. A series of simulations is reported wherein a researcher samples from a target population until the sample's effect size stabilizes. The properties of this stopping rule were explored by parametrically modulating two parameters: (1) The true effect size in the population and (2) the number of iterations needed to reach stabilization. The consequences of modulating these parameters were computed for different metrics: (a) The average reached power, (b) the average reached sample size, (c) the average reached effect size, and (d) effect size variability.

3 Methods

Simulations reported in this study involve an imaginary scenario where a researcher conducts an experiment by sampling from a target population. When sampling from the target population, the researcher uses the effect size stabiliation method. The target population presents a true effect size, which can be revealed using a one-sample t-test. A one-sample t-test was chosen for obvious practical reasons. First, one-sample t-tests generalize to within-subject designs, as paired-samples t-tests are one-sample t-tests over the difference between the repeated measures. Second, one-sample t-tests are computationally more efficient to simulate, as they involve the generation of only one value, as opposed to two in the context of independent samples.

Simulations involved an orthogonal manipulation of two parameters: The stopping criterion and effect size in the population. The stopping criterion is the number of times the difference between successive effect sizes does not exceed 0.05. Stopping criteria ranged from 5 to 100. In the context of one-sample t-tests, the effect size is merely the mean of the target population, ranging from 0.0 to 1.0 with a step of 0.01. Hence, there was a total of 96 * 101 = 9696 sets of parameters.

For each set of parameters, samples were generated by drawing random values from a normal distribution using the `rand_distr` package implemented in the Rust programming language. Each time a new random value was added in the sample, the effect size was computed. If the difference between successive effect sizes did not exceed 0.05 for a pre-defined number of iterations (i.e., stopping criterion), the sampling process stopped. This was repeated 10,000 times, resulting in a population of samples. Four metrics were computed across all samples, resulting in four metrics per set of parameters:

- The proportion of samples which differed significantly from 0.0, considering an alpha value of 0.05 using a one-sided one-sample t-test. - The average sample size reached at the end of the sampling process. - The average effect size. - Effect sizes' standard deviation.

Each simulation started by generating a base sample size equivalent to the stopping criterion (e.g., 5) plus one.

$$f(x) = x^{2.0} \tag{1}$$

This is an example of a math equation within a line: $f(x) = x^{2.0}$. This is now properly done.

4 Results

4.1 Checking the stability assumption

The effect size stabilization method hinges on an implicit assumption that effect sizes stabilize over time. Essentially, if the same experiment is repeated many times, the distribution of effect sizes should show more variability for small than large sample sizes. We conducted a test to verify whether this assumption holds true. Figure **??** displays results from 500 simulated experiments, where a researcher samples from a target population assuming a true effect size of 0.5. Each line in the figure represents the evolution of the effect size throughout the sampling process. As can be seen, this assumption is met: There is an important variability among effect sizes at the beginning of the sampling process, and this variability decreases as more participants are added in the samples. This phenomenon merely reflects the fact that, in small samples, extreme variations have a more significant impact than in larger samples.

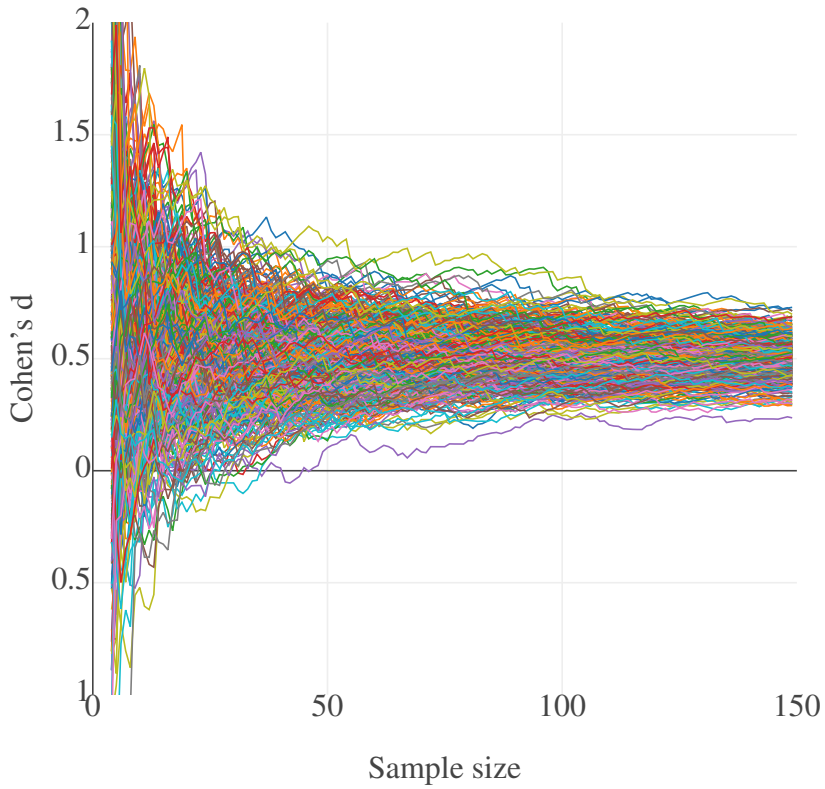

Figure 1: Note