Higher significance with smaller samples: A modified Sequential Probability Ratio Test

Sandipan Pramanik, Valen E. Johnson and Anirban Bhattacharya
Department of Statistics
Texas A&M University

Abstract

We describe a modified sequential probability ratio test that can be used to reduce the average sample size required to perform statistical hypothesis tests at specified levels of significance and power. Examples are provided for Z-tests, T-Tests, and tests of binomial success probabilities. A description of a software package to implement the tests is provided. We also compare the sample sizes required in fixed design tests conducted at 5% significance levels to the average sample sizes required in sequential tests conducted at 0.5% significance levels, and find that the two sample sizes are approximately the same. This illustrates that the proposed sequential tests can provide higher levels of significance using smaller sample sizes.

Introduction

Experimental science relies on the conduct of controlled experiments that test whether effects predicted by a scientific theory can be produced and measured in laboratory settings. Observational science is based on measuring outcomes as they occur naturally, without experimental intervention. In practice, measured outcomes from both observational studies and experiments are subject to random variation and measurement error, and for this
reason hypothesis testing procedures must be employed to determine whether a hypothesized effect has or has not been observed in data. In the classical hypothesis testing paradigm, two types of errors are considered when making this assessment. Type 1 errors occur when the null hypothesis of “no effect” is rejected when the hypothesized effect does not exist. Type 2 errors occur when the null hypothesis is not rejected when the hypothesized effect does exist. To limit claims of false discovery, it is a common practice to design hypothesis testing procedures so that the probability of a Type 1 error (i.e., the significance level) is limited to be less than a pre-specified value, often taken to be $\alpha = 0.05$.

Recent concerns over the replicability of scientific studies have led to calls to reduce the significance thresholds required for declaring a discovery from $\alpha = 0.05$ to $\alpha = 0.005 \ [7, 2]$. Such a policy change would have important ramifications. While it would improve the replicability of scientific claims of discoveries, it would also increase the costs of conducting studies and possibly reduce the number of true discoveries that are made.

The goal of this article is to propose new experimental designs that reduce the sample sizes required to achieve specified Type 1 and Type 2 errors. The new designs represent a modification of the Sequential Probability Ratio Test (SPRT) \[21\] and can be applied to many experiments conducted in the social and natural sciences. On average, the proposed design requires sample sizes that are approximately $1/2$ of the sample size used in corresponding fixed designs when the null hypothesis of no effect is true, and requires sample
sizes that are approximately 20% smaller when alternative hypotheses are true.

In contrast to fixed experimental designs, sequential testing procedures provide a rule for stopping an experiment after observing the outcome of individual experimental items or groups of items. After measuring a group of items, a sequential testing procedure specifies a rule that decides whether to (i) continue an experiment and collect additional observations, (ii) stop an experiment and reject the null hypothesis, or (ii) stop an experiment and accept the null hypothesis.

Perhaps the earliest and most widely known sequential testing procedure is the SPRT proposed by Wald [21]. The SPRT is based on comparing the likelihood ratio between a simple (i.e., point or precise) null hypothesis and a simple alternative hypothesis, and stopping an experiment as soon as the likelihood ratio strongly supports one of the two. More specifically, let $x_1, x_2, \ldots$ represent independent, identically distributed realizations from a distribution with density function $f(x; \theta)$ under both the hypotheses. Suppose the null hypothesis stipulates that $\theta = \theta_0$ and the alternative hypothesis that $\theta = \theta_1$. Then the likelihood ratio statistic in favor of the alternative hypothesis based on $n$ observations can be defined as

$$L(\theta_1, \theta_0; n) = \frac{\prod_{i=1}^{n} f(x_i; \theta_1)}{\prod_{i=1}^{n} f(x_i; \theta_0)}.$$  

To simplify notation, we denote $L(\theta_1, \theta_0; n)$ by $L_n$ in what follows.

The SPRT proceeds by comparing $L_n, n = 1, 2, \ldots$, to constants $A$ and
$B, A > B > 0$, as data from individual items are collected. The procedure stops when

$$L_n \geq A \quad \text{or} \quad L_n \leq B,$$

or equivalently when $L_n$ exits the interval $(B, A)$ for the first time. If $L_n \geq A$, the null hypothesis is rejected; if $L_n \leq B$, the null hypothesis is accepted. The values of $A$ and $B$ determine the Type 1 and Type 2 errors of the test. An important property of the SPRT is that it requires, on average, fewer items to achieve its specified Type 1 and Type 2 errors than any other test having the same or smaller Type 1 and 2 errors.

As an aside, it is worth noting that Wald proposed two modifications of the SPRT to handle composite hypotheses. The “generalized SPRT” considers the ratio of the maximized likelihoods under the respective hypothesis akin to the generalized likelihood ratio test, whereas the “weighted SPRT” replaces the maximized likelihood with an integrated likelihood with respect to certain weight functions for the respective hypothesis. The choice of the weight functions is a tricky issue and there are no known upper bounds on the maximal error probabilities for either method. There is also a substantial literature on a Bayesian decision theoretic treatment of the composite hypothesis testing problem aimed at minimizing the Bayes risk of a sequential test [17]. The Bayes risk is defined in terms of a prior distribution on the parameter space, the cost of collecting each observation, and a loss function on the parameter space associated with making an incorrect decision. A common focus is on testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_1$ ($\theta_1 \geq \theta_0$) in
exponential families, with a strict inequality implying an indifference zone between the two hypothesis. Using Laplace’s integral approximation, the rejection regions for such tests can be connected to the generalized SPRT criterion \[17, 3, 13, 10\].

A key limitation of the SPRT is that it requires the specification of both a null and alternative hypothesis. Specifying an alternative hypothesis is not required in classical hypothesis tests when only Type 1 error constraints have been imposed.

Another limitation of the SPRT is that the sample size required to complete an experiment cannot be bounded prior to the start of an experiment. Regardless of the value of \(N\), there is generally a non-zero probability that more than \(N\) units will be required before \(L_n\) exits the interval \((B, A)\). This feature of the SPRT complicates the practical design of experiments and is addressed in the modification proposed below. An early modification of the SPRT to address this difficulty was proposed in \([1]\), but that modification generally provides less statistical power than the modification that we propose.

From a practical perspective, there are many hypothesis testing contexts where it is not feasible to implement a SPRT. The SPRT cannot be applied in observational studies when subjects are not measured sequentially. Similarly, it cannot be applied when it is not possible to evaluate subjects as soon as they are treated. Such is the case in clinical trials of new disease therapies, which are often conducted at multiple treatment centers. Collation of
data across centers can be time consuming, and it can be difficult to convene review boards. In addition, patient outcomes are often not known for months or even years after a treatment has been administered. To address these challenges, group sequential designs have been developed to allow for the evaluation of patient outcomes only after either a group of patients have been observed, or at scheduled interim analysis times. Readers interested in group sequential designs may refer to [14, 16, 5, 15, 11]. Seigmund [18, 19] provides detailed discussion on the termination of repeated significance tests for group sequential studies with a maximum sample size.

**Method**

We propose a modified SPRT (MSPRT) in which the maximum sample size of an experiment is fixed prior to the start of an experiment, and the alternative hypothesis used to define the rejection region of the test is derived from the size of the test \( \alpha \) (Type I error), the maximum available sample size \( N \), and the targeted Type 2 error \( \beta \) (equal to 1 minus the power). Given these values, the MSPRT is defined in a manner very similar to Wald’s initial proposal.

To set the alternative hypothesis, we find the uniformly most powerful Bayesian test (UMPBT) [8] or approximate UMPBT that matches the rejection region of a classical test of size \( \alpha \) with a sample size of \( N \). For sampling densities that belong to the class of one parameter exponential family models (including Z-tests, tests for proportions, and tests of means of Poisson
counts), UMPBTs exist. For other sampling densities, and in particular for $t$ tests, approximate UMPBTs exist. In general, the values of the parameter that define the alternative hypotheses in these tests are approximately equal to the maximum likelihood estimate of the parameter obtained from data that lie on the boundary of the rejection region of the test. For example, in a one-sided test of a normal mean $\mu$ of size $\alpha$, the UMPBT alternative hypothesis is obtained by taking $\theta_1 = \theta_0 \pm z_\alpha \sigma / \sqrt{n}$. Table 1 provides UMPBT alternatives that can be used in common, one-sided Null Hypothesis Significance Tests (NHST’s). In this table, definitions of alternative hypotheses are based on the maximum sample size $N$.

Given the point alternative hypothesis obtained from the UMPBT or approximate UMPBT, Wald’s SPRT is then conducted until either the likelihood ratio $L_n$ exits the interval $(B, A)$, or a maximum of $N$ items have been tested. The values of $A$ and $B$ for the MSPRT are similar to those used in Wald’s test and are given by

$$A = \frac{1 - \beta}{\alpha} \quad \text{and} \quad B = \frac{\beta}{1 - \alpha}.$$ 

If no decision has been reached after $N$ items have been tested, the null hypothesis is rejected if the likelihood ratio exceeds a threshold, say $\gamma$, that has been determined numerically so as to control the specified type I error of the test. Given $\gamma$, the null hypothesis is accepted and the experiment is terminated if $L_N < \gamma$. If $L_N \geq \gamma$, the alternative hypothesis is accepted. The design parameter $\gamma$ is chosen to be as small as possible while still maintaining
the specified size of the test, $\alpha$. We refer to $\gamma$ as the “Termination Threshold.”

| Test          | $H_0$               | UMPBT alternative          |
|---------------|---------------------|----------------------------|
| Z-test        | $\theta = \theta_0$| $\theta = \theta_0 \pm z_{\alpha} \frac{\sigma}{\sqrt{N}}$ |
| T-test        | $\theta = \theta_0$| $\theta = \theta_0 \pm t_{\alpha;N-1} \frac{s_n}{\sqrt{N}}$ |
| Test for proportion | $\theta = \theta_0$ | $\theta \sim$ a mixture distribution |

Table 1: **UMPBT alternatives for one-sided tests.** For Z and T tests, UMPBT alternative hypotheses have closed form expressions. For tests of proportions, randomized tests must be used to achieve exact Type 1 error control. A (non-randomized) MSPRT can be used to more accurately achieve Type 1 error control, but a mixture distribution is required as the alternative in this setting. This is a slight modification of the UMPBT point alternative as originally defined in [8]. Details and R code for obtaining explicit values for the alternative are described in the Supplementary Information and the R-package MSPRT. The $100(1 - \alpha)^{th}$ quantile of a standard normal distribution and central $t$-distribution with $(N - 1)$ degrees of freedom are denoted by $z_{\alpha}$ and $t_{\alpha;N-1}$, respectively. $\sigma$ denotes the known population standard deviation (sd) in a Z-test, whereas $s_n$ refers to the sample sd (with divisor $(n-1)$) based on $n$ observations.

Figure 1 summarizes the process for conducting a MSPRT for a one-sided test of a normal mean or a population success probability.
In its original form, Wald’s SPRT can be applied only to one-sided tests. In contrast, the MSPRT can be extended for two-sided testing by simultaneously running two one-sided tests of size $\alpha/2$. The test is terminated as soon as the stopping criteria for one of the one-sided tests is satisfied. If both tests continue to the maximum sample size $N$, then a common termination threshold $\gamma$ is again determined so as to maintain the desired Type 1 error of...
the test. If $L_N < \gamma$ for both tests, the null hypothesis of no effect is accepted. Otherwise the null hypothesis is rejected.

**Code availability**

Software required to implement the MSPRT is available from the CRAN R software depository at https://cran.r-project.org/web/packages/MSPRT/index.html.

**Simulation Studies**

This section illustrates the performance of the MSPRT through simulation studies. We examine tests for a binomial proportion, Z-tests and T-tests of size $\alpha = 0.05$ and 0.005. Without loss of generality, we examine one-sided tests with alternative hypotheses of the form $H_1 : \theta > \theta_0$. We also assume that the targeted power of the test is 80% (i.e., $\beta = 0.2$). We compare the MSPRT to standard fixed-design tests having the same size $\alpha$, sample size $N$, and Type 2 error $\beta = 0.2$. For the fixed-design tests, we define $\theta_a$, the fixed-design alternative, as the alternative parameter value that provides the specified $\beta$. 
Figure 2: **Z-test that a population mean equals 0.** Hypothesis test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. The curves in the left plot represent the average proportion of the maximum sample size ($N$) used before the MSPRT terminates in favor of the null or alternative hypothesis. The plot on the right displays the average power of the test against its targeted value of 0.8. In both the plots, the operating characteristics under the alternative are evaluated at the corresponding fixed-design alternatives.
Figure 3: **T-test that a population mean is 0.** Hypothesis test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. In contrast to Figure 2, the population standard deviation is assumed to be unknown. The curves in the left plot represent the average proportion of the maximum sample size ($N$) used before the MSPRT terminates in favor of the null or alternative hypothesis. The plot on the right displays the average power of the test against its targeted value of 0.8. In both the plots, the operating characteristics under the alternative are evaluated at the corresponding fixed-design point alternatives.
Figure 4: Test that a binomial proportion equals 0.2. Hypothesis test of $H_0 : \theta = 0.2$ vs. $H_1 : \theta > 0.2$. The curves in the left plot represent the average proportion of the maximum sample size ($N$) used before the MSPRT terminates in favor of the null or alternative hypothesis. The plot on the right displays the average power of the test against its targeted value of 0.8. In both the plots, the operating characteristics under the alternative are evaluated at the corresponding fixed-design point alternatives.

Figures 2-4 display the average proportion of the fixed design sample size $N$ needed in a MSPRT to achieve nearly equivalent Type 1 and Type 2 errors. In all plots, Type 1 errors are maintained. The subplots on the right depict that average power achieved at the corresponding fixed-design point alternatives.

Figure 2 provides results for the Z-test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ based on samples of independent and identically distributed random variables with known variance. Figure 3 provides results for the corresponding T-test when the sample variance is not known. Figure 4 reports values for the test
of a binomial proportion, in which \( H_0 : \theta = 0.2 \) versus the alternative that \( H_1 : \theta > 0.2 \).

In Figure 4 we see that the discreteness of binomial data causes some non-monotonicity in the proportion of the maximum sample size that is required to reach a decision. This feature of the plot corresponds to the non-monotonicity of power curves for fixed-design tests when sample sizes are increased. For a given a choice of \( N \), the R-package \texttt{MSPRT} finds an “ideal” maximum sample size that accounts for this non-monotonicity. We refer to these values as the “effective sample sizes.” Fig 4 is drawn using only those values as the maximum sample sizes. Further details regarding this issue are provided in the Supplemental Information.

Two features of these plots are noteworthy. First, for the lower Type 1 error of \( \alpha = 0.005 \), the average sample size required by the MSPRT is less than 50% of the sample size required by the fixed-design test when the null hypothesis is true. This finding holds for all three tests. Second, under the alternative hypothesis, the average sample size required for the MSPRT is typically only 80% as large as the sample size required for a fixed-design test.

We next examine the potential benefit that the MSPRT could offer in offsetting the increase of sample size that would be incurred if the bar for declaring a “statistically significant” result was moved from \( p < 0.05 \) to \( p < 0.005 \). Specifically, we compare the sample sizes needed to achieve statistical significance at the 5% level in standard fixed-design tests to the average sample size needed to achieve statistical significance at the 0.5% level.
using the MSPRT.

If the null hypothesis is true, this comparison is straightforward. If not, care must be taken to make sure that the same alternative hypotheses are compared at both levels of significance under the fixed and MSPRT designs. To make this comparison, we determine the $\theta^*$ that achieves the targeted Type 2 error in a fixed-design test of size $0.05$. For that $\theta^*$, we next determine the $N^*$ needed to achieve the same Type 2 error in a fixed-design test of size $\alpha = 0.005$. We then define that $N^*$ to be the maximum sample size for the MSPRT.

Because the average sample size used in the MSPRT depends on whether the null or alternative hypothesis is true, and we are interested in the long run effect of implementing MSPRT over many experiments, it is useful to examine the effect on the total sample size as the proportion of true null hypotheses is varied. Recent research suggests that this proportion is likely to be in the range 0.80-0.95 [4, 9].

Figures 5-7 display the average multiple of the fixed 5% test’s sample size $N$ that is required to perform a MSPRT of size 0.5% as the proportion of tested null hypotheses $\pi_0$ is decreased from 1 (the red line at the bottom) to 0.6 (the blue line). Also displayed is the multiple of $N$ that is required to achieve a Type 1 error of size 0.005 in a fixed-design test (the solid black line at the top). The latter multiple tends to fall between 1.89 and 2.14.

The key finding from Figures 5-7 is that MSPRT tests of size 0.005 require, on average, essentially the same sample sizes that are required to conduct
standard fixed-design tests of size 0.05. This implies that “raising the significance” bar to 0.005 from 0.05 could be accomplished without significantly increasing sample sizes.

Figure 5: **Z-test that a population mean equals 0.** Curves in the left plot represent the average multiple of the sample size in a fixed-design test of size 0.05 required to obtain a MSPRT of size 0.005 of approximately the same power. Average sample sizes are dependent on the proportion of tested null hypotheses that are true. This proportion ($\pi_0$) is coded by color, as indicated. The MSPRT maintains a Type 1 error of 0.005, and its power at $\theta^*$ approximately equals 0.8 for the indicated proportion of $N^*$ (the sample size of the corresponding fixed-design test). The power of the MSPRT is depicted in the plot on the right.
Figure 6: **T-test that a population mean is 0.** Curves in the left plot represent the average multiple of the sample size in a fixed-design test of size 0.05 required to obtain a MSPRT of size 0.005 of approximately the same power. Average sample sizes are dependent on the proportion of tested null hypotheses that are true. This proportion ($\pi_0$) is coded by color, as indicated. The MSPRT maintains a Type 1 error of 0.005, and its power at $\theta^*$ approximately equals 0.8 for the indicated proportion of $N^*$ (the sample size of the corresponding fixed-design test). The power of the MSPRT is depicted in the plot on the right.
Figure 7: Test that a binomial proportion equals 0.2. Curves in the left plot represent the average multiple of the sample size in a fixed-design test of size 0.05 required to obtain a MSPRT of size 0.005 of approximately the same power. Average sample sizes are dependent on the proportion of tested null hypotheses that are true. This proportion ($\pi_0$) is coded by color, as indicated. The MSPRT maintains a Type 1 error of 0.005, and its power at $\theta^*$ approximately equals 0.8 for the indicated proportion of $N^*$ (the sample size of the corresponding fixed-design test). The power of the MSPRT is depicted in the plot on the right.

Discussion

The costs of conducting experiments to test hypothesized effects is often related directly to the number of items or subjects that are tested. When subjects can be tested sequentially, then the use of sequential testing procedures can dramatically reduce these costs. When experiments are designed to test effects that don’t exist (i.e., the null hypothesis is true), then the use of the MSPRT can reduce sample sizes by 20-30% in 5% tests, and by over...
50% in 0.5% tests.

In the context of improving the replicability of scientific studies, the use of MSPRT’s can make the average costs of conducting tests at the 0.5% level approximately equal to the costs of conducting fixed-design experiments at the 5% level of significance.

Potential drawbacks for the implementation of MSPRT’s are the firm requirement to specify the outcome variable and test statistic prior to the start of the experiment. Failure to insure that these quantities are clearly identified a priori would lead to additional opportunities for p-hacking; unethical researchers might apply MSPRT’s to several outcome variables simultaneously, which would negatively impact the control of Type 1 errors. In addition, the conduct of MSPRT’s requires that investigators perform statistical analyses after the acquisition of each subject’s data. In some settings, this requirement might represent an unacceptable cost. From a purely statistical perspective, however, the statistical analyses required for the implementation of an MSPRT in standard settings is quite manageable and can be handled using the R-package MSPRT described in the Supplemental Information.

Acknowledgments

Support for this research was provided by the National Cancer Institutes, R01 CA158113.
References

[1] Anderson, T. W., A Modification of the Sequential Probability Ratio Test to Reduce the Sample Size. *The Ann. of Math. Stat.*, vol.31, no. 1, 1960, pp. 165-197.

[2] Benjamin, Daniel J., Redefine statistical significance. *Nature Human Behaviour*, vol.1, 2017.

[3] Chernoff, H., Sequential Test for the Mean of a Normal Distribution III (Small t). *Ann. of Math. Statist.*, vol.36, no. 1, 1965, 28-54.

[4] Dreber, A. et al., Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, USA vol.112, 2015, pp. 1534315347.

[5] Haybittle, J. L., Repeated assessments of results in clinical trials of cancer treatment. *Brit. J. Radiol.*, vol.44, 1971, pp. 793-797.

[6] Jennison, Christopher and Turnbull, Bruce W., Group Sequential Methods with Applications to Clinical Trials, Chapman and Hall/CRC.

[7] Johnson, Valen E., Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, vol.16, 2013.

[8] Johnson, Valen E., Uniformly most powerful Bayesian tests. *Ann. of Stat.*, vol.41 (4), 2013, pp. 1716-1741.
[9] Johnson, Valen E. et al., On the Reproducibility of Psychological Science. *J. Am. Stat. Assoc.*, vol.112, 2017, pp. 110.

[10] Lai, T., Nearly Optimal Sequential Tests of Composite Hypotheses. *Ann. of Stat.*, vol.16, no. 2, 1988, pp. 856-886.

[11] Lan, K. K. G. and DeMets, D. L., Interim Analysis: The alpha spending function approach. *Statistics in medicine*, vol.13, 1994, pp. 1341-1352.

[12] Lehmann, E. L. and Casella, G., Theory of Point Estimation, 1998, pp. 2nd ed., sec. 1.5.

[13] Lorden, G., Likelihood Ratio Tests for Sequential k-Decision Problems. *Ann. of Math. Statist.*, vol.43, no. 5 1972, 1412-1427.

[14] O'Brien PC, Fleming TR., A Multiple Testing Procedure for Clinical Trials. *Biometrics*, vol.35, 1979, pp. 549-556.

[15] Peto, R., Discussion of papers by J. A. Bather and P. Armitage. *Internat. Statist. Rev.*, vol.53, 1985, pp. 31-34.

[16] Pocock SJ., Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, vol.64, 1977, pp. 191-199.

[17] Schwarz, G., Asymptotic Shapes of Bayes Sequential Testing Regions. *Ann. of Math. Statist.*, vol.33, no. 1 1962, 224-236.

[18] Siegmund, D., Sequential Analysis: Test and Confidence Intervals. New York-Heidelberg-Berlin: Springer-Verlag., 1985
[19] Siegmund, D., Boundary crossing probabilities and statistical applications. *Ann. of Stat.*, vol.14, 1986, pp. 361-404.

[20] T.D. Sterling, Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. *Journal of the American Statistical Association*, vol.54, No. 285 (Mar., 1959), pp. 30-34.

[21] Wald, A., Sequential Tests of Statistical Hypotheses. *Ann. of Math. Statist.*, vol.16, no. 2, 1945, 117-186.