Published in final edited form as:

Stat Pap (Berl). 2019 April ; 60(2): 373–394. doi:10.1007/s00362-019-01095-x.

Asymptotic properties of maximum likelihood estimators with sample size recalculation

Sergey Tarima,

Institute for Health and Equity, Medical College of Wisconsin

Nancy Flournoy

Department of Statistics, University of Missouri-Columbia

Nancy Flournoy: flournoyn@missouri.edu

Abstract

Consider an experiment in which the primary objective is to determine the significance of a treatment effect at a predetermined type I error and statistical power. Assume that the sample size required to maintain these type I error and power will be re-estimated at an interim analysis. A secondary objective is to estimate the treatment effect. Our main finding is that the asymptotic distributions of standardized statistics are random mixtures of distributions, which are non-normal except under certain model choices for sample size re-estimation (SSR). Monte-Carlo simulation studies and an illustrative example highlight the fact that asymptotic distributions of estimators with SSR may differ from the asymptotic distribution of the same estimators without SSR.

Keywords

adaptive designs; asymptotic distribution theory; interim analysis; local alternatives; maximum likelihood estimation; mixture distributions

Mathematics Subject Classification (2010)

62K99; 62L05; 62F05; 62E20

1 Introduction

Recent developments in adaptive designs open additional opportunities for optimizing resource allocation and increasing rewards through study planning. Along with new opportunities, previously disregarded statistical challenges are becoming more acute and require attention from statisticians. For example, the use of a sample-size recalculation (SSR) may lead to biased maximum likelihood estimators (MLEs) and their distributions may be changed.

The main focus of research on SSR has been on controlling error rates in hypothesis testing applications with fixed alternatives. Chapter 11 in Proschan et al (2006) focuses on observed

effect sizes that are differences between two sample means following SSR using methods that are based on a parameter that is ancillary to the effect size, such as the variance under normality. Their Section 11.4 (on SSR methods that use estimates of treatment effects) focuses on controlling error rates. Li et al (2002) derived the joint density of stage I and stage II standardized test statistics, and found it not to be a bivariate normal. This indicates that, in general, a linear combination of these two stage-specific pivots may not be normal.

When SSR is a function of observed treatment effects, classical central limit theorems cannot be casually applied (e.g., to MLEs). In Theorem 1, we show that distributions of pivot statistics are changed by truncation induced by SSR. Authors have begun characterizing MLE behavior in the presence of interim adaptations. Brannath et al (2006) and Graf et al (2016) investigated the bias and mean squared error of MLEs following SSR. Broberg and Miller (2017) explored estimation following two-stage designs with a possible sample size increase planned after the first stage, conditional on stage I data. Measure theoretic challenges related to SSR were discussed in Brannath et al (2012). Mutze et al (2017) explored the use of prior information to improve estimation of a nuisance parameter during SSR.

We study asymptotic MLE behavior following SSR in the context of hypotheses tested against local alternatives. The first paper we found on local alternatives is by Sugiura et al (1973) for likelihood-ratio tests involving multivariate normal data. Their use of local alternatives was motivated by the fact that, under fixed alternatives, limiting distributions of these test statistics degenerated as they approached the null. A test is *consistent* against a fixed alternative if the power function approaches 1 at that alternative while asymptotically maintaining the size of the test. Van der Vaart (1998) argues for the use of local alternatives by pointing out that while consistency is an optimality criterion for tests, most sequences of tests are consistent, and therefore, consistency alone is “too weak to be really useful. To make an informative comparison between sequences of (consistent) tests, (one needs to study) the performance of the tests in problems that become harder as more observations become available. One way of making a testing problem harder is to choose null and alternative hypotheses closer to each other.” In addition, the use of local alternatives reflects the common practice of using large samples to detect small difference and smaller samples to detect large ones. Thus, once the local alternative is fixed, the series of stage I sample sizes we study are inversely proportional to the series of squared hypothesized alternative treatment effects.

The MLE’s asymptotic distributions that we obtain under SSR differ from those of Le Gam (1960) and Jeganathan (1982). Le Gam (1960) introduced *local asymptotic normality* to describe parametric families of densities whose log-likelihood ratios admit approximations of a quadratic nature. Jeganathan (1982) coined the term *local asymptotic mixed normality* to describe limiting distributions in which both the linear and quadratic terms in log-likelihood ratio approximations were random. Popular examples arise from explosive autoregressive processes of first order and Galton-Watson branching processes, a detailed description of which can be found, for example, in a comprehensive review by Roussas and Bhattacharya (2011). MLEs having limiting distributions that are random mixtures of normal variables also have been observed previously in adaptive design settings [Ivanova et

al (2000), Ivanova and Flournoy (2001), May and Flournoy (2009), Lane and Flournoy (2012)].

In contrast to Le Gam (1960) and Jeganathan (1982), the standardized test statistic obtained in our Theorem 1 converge to a *mixture of truncated normal distributions*.

We study MLEs and other estimators under conditions which would permit classical central limit theorems to hold if observations were independent. We show that, using first-stage MLEs to calculate the total sample size, the asymptotic distribution of a full-data MLE may be a mixture, that is not asymptotically normal. In spite of these distributional complexities, various simulation studies have shown that MLEs following SSR can be very competitive (in terms of root mean squared errors and coverage probabilities); see Brannath et al (2006), Wang et al (2010) and Bretz et al (2009).

Section 2 defines several SSR formulae and introduces a simple Z-test example. Large sample properties of MLEs following SSR are explored in Section 3, and then illustrated in Section 4 in several Monte-Carlo experiments. Section 5 shows an illustrative application of the theory to a semi-parametric model; in particular, the large sample distribution of an MLE simulated from a published Cox regression is shown in Figure 4. Section 6 concludes this article with a short summary.

2 Methods

Let $X \sim f_X(x|\theta, \eta)$, where $f_X(x|\theta, \eta)$ is twice differentiable with respect to the parameters and admits the existence of unique and consistent MLEs; η is a nuisance parameter. Additional assumptions are given in Section 3 where we derive asymptotic properties of MLEs after SSR. Primary interest is in testing the null hypothesis $H_0: \theta = 0$ with α -level type I error and $1 - \beta$ power at an alternative $H_1: \theta = h/\sqrt{n} \neq 0$, where n is the stage I sample size and h is a pre-determined local alternative. To reduce complexity, we take θ and η to be scalars. For notational simplicity, we do not consider null hypotheses more general than zero; and the dependence of θ on n is suppressed. We emphasize, however, that the alternative θ depends inversely on the square root of the stage I sample size; hence, the alternatives studied are a sequence of parameters that approaches zero as $n \rightarrow \infty$.

Following the n th observation, the first stage MLE $(\hat{\theta}_1, \hat{\eta}_1)$ of (θ, η) is calculated using $\{X_1, \dots, X_n\}$, and the original estimate of the total sample size required to obtain $1 - \beta$ power at a fixed α level is updated. Then the data on $N - n$ additional subjects are collected if needed. The complete data are $\{X_1, \dots, X_n, X_{n+1}, \dots, X_N\}$. Henceforth, we use subscripts in estimators of θ and η to denote the data collection stage, before (the first stage) or after SSR (the second).

2.1 Sample Size Formula: N_0

Define $\sigma_\theta^2 := \lim_{n \rightarrow \infty} n \text{Var}(\hat{\theta}_1)$, and let $CV := \sigma_\theta h$ denote the *local coefficient of variation*.

We rely on this local definition of the coefficient of variation because the global version, $\sigma_\theta/\theta = \sqrt{n}CV$, diverges to infinity.

We begin with an SSR formula for the total sample size that is suitable with large samples when $\text{Var}(\hat{\theta}_1)$ is known and $Z := \sqrt{n}(\hat{\theta}_1 - \theta)/\sqrt{\text{Var}(\hat{\theta}_1)}$ is asymptotically normal by a Lindeberg-Lévy, Lyapunov or Lindeberg type central limit theorem (CLT):

$$N_0 := \frac{\sigma_\theta^2 c^2}{\theta^2} = n \frac{\sigma_\theta^2 c^2}{h^2} = nc^2 \text{CV}^2, \quad (1)$$

where $c^2 = (z_{1-\alpha/2} + z_{1-\beta})^2$ and z is a quantile of the standard normal distribution. Note N_0 is fully determined by CV^2 and SSR is not formally needed as N_0 is a non-random value.

The validity of the rejection region $|Z| > z_{1-\alpha/2}$ for testing H_0 versus H_1 , and hence of formula (1), relies on applicability of a central limit theorem, unless $\hat{\theta}_1$ itself is normally distributed. If n is small, other sample size formulae may be more appropriate. Finite sample size inference is not studied in this work.

2.2 Sample Size Formulae: N_1 and N_2

One of the following formulae are commonly used for SSR depending on whether only the variance σ_θ^2 is unknown, or whether h is unknown as well:

$$N_1 := n \widehat{\text{Var}}(\hat{\theta}_1) c^2 \theta^{-2} = n^2 \widehat{\text{Var}}(\hat{\theta}_1) c^2 h^{-2} \quad (2)$$

and

$$N_2 := n \widehat{\text{Var}}(\hat{\theta}_1) c^2 \hat{\theta}_1^{-2} = n^2 \widehat{\text{Var}}(\hat{\theta}_1) c^2 \hat{h}_1^{-2}, \quad (3)$$

where $\hat{h}_1 = \hat{\theta}_1 \sqrt{n}$ is the first stage MLE of the local alternative h . Equation (2) is an approximation of N_0 that is designed to detect a minimally clinically relevant effect, $h/\sqrt{n}$, at the predetermined type I error rate α and power $1 - \beta$. Replacing h with $\hat{h}_1$ in N_1 creates N_2 and allows the experiment to stop earlier if the true effect size is higher than expected.

Sample-size recalculation may occasionally lead to sample sizes smaller than n , as well as to overly large, budgetary prohibitive sizes. These are avoided by imposing lower (n) and upper ($n_{\max}$) boundaries on the total sample size. Thus, henceforth, we re-define N_j by the following algorithm:

$$N_j \leftarrow \min \left[\max \left(n, N_j \right), n_{\max} \right], \text{ where } j \in (1, 2).$$

It is not unusual to estimate parameters at boundaries imposed by adaptive designs [e.g., Lane and Flournoy (2012), Tarima et al (2016)]. A performance evaluation of N_j must involve the probabilities that the re-calculated sample size will be on a boundary. Further, we explore the effects of SSR on MLEs in the presence of such lower and upper boundaries.

2.2.1 Example: Z-test. with $X \sim N(\theta, 1)$ —Table 1 characterizes 8 experimental scenarios following SSR when simulated data were $X \sim N(\theta, 1)$. The first 4 scenarios were generated under the local alternative $H_1: \theta = h/\sqrt{n}$ to assess local power; the remainder were under $H_0: \theta = 0$ to assess type I error. In all Monte-Carlo simulations, an upper bound for the sample size was set to $n_{max} = 1000n$; 100, 000 Monte-Carlo simulations were used for each scenario.

The use of the Z-test formula (N_1) led to a slight deflation from nominal power for small stage I samples; but otherwise power was well controlled. Some inflation from nominal type I error observed for smaller sample sizes decreased for larger sizes. From the Shapiro-Wilk P-values, we conject that the distributions of the sample mean of X (i.e., the full sample MLE of θ) are asymptotically normal under both H_0 and H_1 .

With N_2 , Table 1 shows type I error was highly inflated (6% to 8% instead of 5%), but power was well controlled. This is consistent with what has been noted by others [see, for example, Proschan and Hunsberger (1995)]. What is interesting here is that, in addition to presence of asymptotic bias, Shapiro-Wilk tests show asymptotic normality does not hold even for large sample sizes. What are the drivers of this non-normality?

In Section 3, we use $\tilde{\theta} \approx (n/N_2)\hat{\theta}_1 + (1 - n/N_2)\hat{\theta}_2$ to approximate the full-data MLE, where $\hat{\theta}_2$ is the second stage MLE of θ ; so, knowing the distribution of n/N_2 is an important part of understanding the distribution of $\tilde{\theta}$. We present the asymptotic distribution of N_2/n under local alternatives, which is non-degenerate, in contrast to $N_1/n \xrightarrow{P} N_0/n$.

Figure 1 graphs typical behavior of $\log(N_2/n)$ and $V = \sqrt{N_2}(\tilde{\theta} - 0)/\sigma_\theta$. Simulated data were $X \sim N(\theta, 1)$ with $n = 100$ and $n_{max} = 2000n$. A standard normal density is graphed over the density estimates of V . Note, the large frequency with which $\log(N_2/n)$ hits the upper boundary does not change the normality of V whereas frequent hits of the lower boundary makes V non-Gaussian.

2.3 Sample Size Formulae: N_3 , N_4 , N_5 , and N_6

We now define SSR formulae $N_3 - N_6$ to depend on $\hat{\theta}_1$.

A more stable estimate $\widehat{CV}_{boot}^2$ of CV^2 can be obtained by bootstrap resampling, leading to

$$N_3 := nc^2\widehat{CV}_{boot}^2. \quad (4)$$

SSR formulae N_1 , N_2 , and N_3 use different estimates of CV in order to estimate N_0 . But Formula N_0 is justifiable when MLEs have asymptotically normal distributions. We show in Section 4 that the asymptotic distribution of the MLE with SSR is not always normal, even when data are normal.

When the limiting distribution of the full sample MLE may not be normal, it no longer makes sense to automatically calculate c^2 from percentiles of the normal distribution.

Seeking to generalize the fact that $c^2 = (z_{1-\alpha/2} + z_{1-\beta})^2$ secures $1 - \beta$ power against $CV = h/\sigma_\theta$ when the Z statistic is normally distributed, we find c^2 to secure $1 - \beta$ power at a given CV after SSR. In Section 4's Monte-Carlo experiments, we provide c^2 for SSR formulae N_1 – N_6 , securing 80% power at $CV = 1$ for each. Because the true CV is most likely unknown, SSR formulae are found at hypothesized CV values. We call the process of finding c^2 at a hypothesized value of CV *calibration*.

Using calibration to free SSR's dependence on the normal distribution, one can consider additional sample size formulae less directly related to N_0 . For example,

$$N_4 := nc^2 \log \left(1 + \frac{n \widehat{\text{Var}}(\hat{\theta}_1)}{\hat{h}_1^2} \right) = nc^2 \log(1 + \widehat{CV}^2) \quad (5)$$

applies a logarithmic-type transformation to the first stage MLE of CV , $\widehat{CV}$. To require the experiment not stop at the end of stage I, we also consider

$$N_5 := n + N_4 = n + nc^2 \log(1 + \widehat{CV}^2). \quad (6)$$

Another option is to incorporate a prior guess θ_g of θ by adding g fake observations that are identically equal to θ_g to the stage I data:

$$N_6 := c^2 n \frac{(n+g)n^2 \widehat{\text{Var}}(\hat{\theta}_1)}{(n\hat{\theta}_1 + g\theta_g)^2}. \quad (7)$$

N_5 's logarithmic transformation suppresses the effect of small $\hat{h}_1$ which leads to large N_2 and N_3 . N_6 can be extended with various Bayesian [Mutze et al (2017)] and frequentist [Tarima and Pavlov (2006), Dmitriev et al (2014)] approaches for using additional information when θ is estimated as part of the interim analysis. By analogy with N_1 and N_2 we apply upper and lower boundaries, $N_j \leftarrow \min[\max(n, N_j), n_{\max}]$, where $j \in (3, \dots, 6)$.

Figure 2 is similar to Figure 1, but for N_4 and N_5 . The density estimates of V often appear Gaussian, but not always. For example, N_5 SSR is often non-Gaussian, which is a result of the relatively high values of c^2 required to secure 80% power at $CV = 1$. The higher values of c^2 force the total sample sizes to be high for all formulae considered, decreasing the effect of stage I data on the MLE. The normality of V is less likely to hold when SSR formulae originally powered at $\sigma_\theta = h = 1$ are applied at $CV = 0.5$ and $CV = 3$; see Figure 2. At $h = 3$, non normality is observed for N_2 and N_4 .

2.4 Summary of Simulation Results

Distributions of standardized MLEs with SSR can be closely approximated by normal distributions when the major probability mass of N_j is located close to n or is much larger than n (associated with large values of c^2). Normality can be problematic if there is a good chance for N_j to dip below n and substantial probability that N_j is greater than n .

Other SSR formulae can be created that incorporate various aspects of the study design, such as stopping for futility and having boundaries on h and/or σ_θ^2 . For all these and other sample-size formulae, we generally expect the ratio of the total sample size to the stage I sample size, N_j/n , to converge to a stationary distribution as $n \rightarrow \infty$, or to a constant as does N_1/n .

We now address what happens with the distribution of the MLE with SSR as $n \rightarrow \infty$. Theoretical results presented in this manuscript mainly focus on N_1 and N_2 , whereas other N_j are introduced to demonstrate the breadth of the considered approach. More complex SSR procedures without explicit formulae, for example, based on conditional power calculations or other intensive computations are not considered in this manuscript, but the asymptotic results apply to them as well.

3 Large Sample Properties

To describe large sample results, in addition to “convergence in distribution” ($\xrightarrow{d}$) and “convergence in probability” ($\xrightarrow{P}$), we will also use “equal in distribution” ($\stackrel{d}{=}$) and “asymptotically equal in distribution” ($\stackrel{d}{\approx}$).

We call estimates $(\hat{\theta}_1, \hat{\eta}_1)$ of (θ, η) and their statistical models *regular* if, without SSR,

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_1 - \theta \\ \hat{\eta}_1 - \eta \end{pmatrix} \xrightarrow{d} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\theta\theta} & \sigma_{\theta\eta} \\ \sigma_{\theta\eta} & \sigma_{\eta\eta} \end{pmatrix} \right]. \quad (8)$$

where the limiting covariance matrix is positive definite with finite elements. Different sets of assumptions on $f(x|\theta, \eta)$ that cause (8) to be satisfied include sets for independent, identically distributed observations by Cramér [see, for example, Ferguson (1996)], for independent not identically distributed observations [e.g., Philippou et al (1973)], for dependent observations [e.g., Crowder (1976)], and for densities whose support depends on parameters [e.g., Wang et al (2014)]. All sets of conditions include the existence and consistency of the MLE.

Standardized, the first stage MLE of θ is standard normally distributed:

$$\xi_1 := \sqrt{n}(\hat{\theta}_1 - \theta)\sigma_\theta^{-1} \xrightarrow{d} N(0, 1), \quad (9)$$

where $\sigma_\theta^2 := \sigma_{\theta\theta}$.

We consider estimators that are regular without SSR, and we call them MLEs, recognizing that MLEs are only a special case of estimators satisfying (8). Since the total recalculated sample size N_j goes to infinity as $n \rightarrow \infty$, it is informative to consider the asymptotic behavior of N_j/n ratios under different sample size formulae. As will be apparent later, the asymptotic distribution of this ratio is an important component of the asymptotic distribution of the MLE. In Section 3.1, we derive asymptotic distributions of sample size ratios N_1/n and N_2/n SSR, respectively. In Section 3.2, we derive the asymptotic distribution of $\tilde{\theta}$.

3.1 The N_1/n and N_2/n Ratios

By the continuous mapping theorem $\widehat{\text{Var}}(\hat{\theta}_1)$ is the MLE of $\text{Var}(\hat{\theta}_1)$, and since the MLE is consistent by assumption (8), so is $\widehat{\text{Var}}(\hat{\theta}_1)^{1/2}$. Therefore, the ratio N_1/n degenerates to a point-mass:

$$\frac{N_1}{n} \xrightarrow{p} \frac{\sigma_\theta^2}{h^2} c^2. \quad (10)$$

N_2 SSR depends on $\hat{\theta}_1$ and

$$\frac{N_2}{n} \stackrel{d}{=} \frac{\widehat{\text{Var}}(\hat{\theta}_1)/\sigma_\theta^2}{\hat{\theta}_1^2/\sigma_\theta^2} c^2 \stackrel{d}{\approx} \frac{c^2}{\chi_1^2(n\theta^2/\sigma_\theta^2)},$$

where $\chi_1^2(n\theta^2/\sigma_\theta^2)$ is a χ^2 random variable with 1 degree of freedom and non-centrality parameter $n\theta^2/\sigma_\theta^2 = h^2/\sigma_\theta^2$.

Under $H_0: \theta = 0$, $N_2/n \xrightarrow{d} \chi_1^{-2} c^2$, where χ_1^{-2} denotes the inverse chi-squared distribution with 1 degree of freedom. Under $\theta = h/\sqrt{n}$, $N_2/n \xrightarrow{d} c^2 \chi_1^{-2}(h^2/\sigma_\theta^2)$. Now we can write the distribution of N_2/n as

$$\Pr\left(\frac{N_2}{n} \leq r \mid \theta = \frac{h}{\sqrt{n}}\right) \rightarrow \Pr\left(\frac{c^2}{\chi_1^2(h^2/\sigma_\theta^2)} \leq r\right) = 1 - \Pr\left[\chi_1^2\left(\frac{h^2}{\sigma_\theta^2}\right) \leq \frac{c^2}{r}\right].$$

3.2 The Distribution of the MLE of θ after SSR

Henceforth, omitting subscripts on N will refer to a sample size formula in general, with the assumption that N/n converges to a random variable independent of n with support on the positive real numbers. Recall that N is a function of $\hat{\theta}_1$ and $\hat{\eta}_1$.

For large n and $N - n$, ignoring sample-size boundaries, MLEs based on independent groups of subjects allow the following approximation: $\tilde{\theta} \approx \frac{n}{N} \hat{\theta}_1 + \frac{N-n}{N} \hat{\theta}_2$. This large sample approximation holds for regular MLEs calculated without SSR. In the presence of SSR, N depends on $\hat{\theta}_1$ and $\hat{\eta}_1$ which changes an MLE's asymptotic properties. For one dimensional θ and η ,

$$P_{\tilde{\theta}}(x) = \Pr(\tilde{\theta} < x \mid n, n_{\max}) = E_{\hat{\theta}_1, \hat{\eta}_1} \Pr(\tilde{\theta} < x \mid \hat{\theta}_1, \hat{\eta}_1, n, n_{\max}), \quad (11)$$

and using the approximation for $\tilde{\theta}$,

¹Other options for consistent variance estimation include bootstrap and cross-validation.

$$P_{\tilde{\theta}}(x) \approx \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr\left(\hat{\theta}_2 < \frac{xN - ny}{N - n} \mid \hat{\theta}_1 = y, \hat{\eta}_1 = z, n, n_{\max}\right) f_{\hat{\theta}_1, \hat{\eta}_1}(y, z) dy dz,$$

where $f_{\hat{\theta}_1, \hat{\eta}_1}(y, z)$ is the joint density or probability measure function of $(\hat{\theta}_1, \hat{\eta}_1)$.

The distribution of $\tilde{\theta}$ given by (11), as well as power and type I error, can be estimated using Monte-Carlo-type resampling as in Tarima et al (2016). Tarima's resampling procedure is computationally intensive, but not unmanageable. Alternatively, one can consider asymptotic approximations.

The asymptotic variance $\sigma_{\tilde{\theta}}^2$ may or may not depend on the parameter of interest (θ) and/or on nuisance parameters (η). For example, with normal random variables and $\theta = E(\bar{X})$, $\eta = \sigma_{\theta}$ is a nuisance parameter defined in a space orthogonal to θ . In contrast, for Bernoulli random variables, the dependence of θ (a probability of success) on σ_{θ} is clear since $\sigma_{\theta}^2 = \theta(1 - \theta)$.

The ratio N_1/n depends only on $\hat{\eta}_1$ and degenerates to a constant as shown in Equation (10).

We are interested in more general SSR, like N_2 , which may depend on both $\hat{\theta}_1$ and $\hat{\eta}_1$.

Accounting for sample-size boundaries, $\tilde{\theta}$ becomes

$$\begin{aligned} \tilde{\theta} \stackrel{d}{\approx} & I(N \leq n)\hat{\theta}_1 + I(n < N < n_{\max})\left[\frac{n}{N}\hat{\theta}_1 + \frac{N-n}{N}\hat{\theta}_2\right] \\ & + I(n_{\max} \leq N)\left[\frac{n}{n_{\max}}\hat{\theta}_1 + \frac{n_{\max}-n}{n_{\max}}\hat{\theta}_2\right]; \end{aligned} \quad (12)$$

and $\frac{n}{N}\hat{\theta}_1 + \frac{N-n}{N}\hat{\theta}_2$ is, itself, a random mixture. From the asymptotic normality of $\hat{\theta}_1$, the distribution of $\tilde{\theta}$ conditional on $N - n$ converges to that of a truncated normal random variable, where the truncation region is defined by $N - n$. This knowledge allows estimation of

$$p_I = \lim_{n \rightarrow \infty} \Pr(N \leq n) = \lim_{n \rightarrow \infty} \Pr\left(\frac{N}{n} \leq 1\right),$$

which is the asymptotic probability that no additional data need be collected.

Theorem 1 below asserts that distributions of standardized MLEs with SSR may not be normal as they are mixtures of truncated Gaussian random variables. Let $\tau_n := N/(N - n) \xrightarrow{d} \tau$ as $n \rightarrow \infty$, where τ is a random variable defined on $(-\infty, \infty)$. Note τ_n (through N) may be a function of both $\hat{\theta}_1$ and $\hat{\eta}_1$; but dependence on $\hat{\eta}_1$ disappears as $n \rightarrow \infty$ as $\hat{\eta}_1 \xrightarrow{P} \eta$. Then with N_2 SSR, the large sample behavior of τ_n is fully determined by the random variable ξ_1 :

$$\tau = \tau(\xi_1) = \frac{\sigma_\theta^2 c^2}{\sigma_\theta^2 c^2 - (h + \sigma_\theta \xi_1)^2} = \frac{c^2}{c^2 - (h\sigma_\theta^{-1} + \xi_1)^2}.$$

Theorem 1 If $\tau_n \xrightarrow{d} \tau$, then under $\theta = h/\sqrt{n}$, the asymptotic distribution of the standardized MLE,

$$V := \sqrt{N} \left(\frac{\tilde{\theta} - \theta}{\sigma_\theta} \right) = \sqrt{N} \left(\frac{\tilde{\theta} - h/\sqrt{n}}{\sigma_\theta} \right),$$

does not depend on n ; it is

$$\begin{aligned} \Pr(V < v) &\rightarrow p_l \Phi(v \mid \tau < 1) \\ &+ (1 - p_l) \int_{\tau \geq 1} \Phi(\sqrt{\tau}v - \sqrt{\tau - 1}y \mid \tau \geq 1) \phi(y) dy, \end{aligned} \quad (13)$$

where $\tau = \tau(y)$ and

$$\Phi(v \mid A) = \frac{\int_{-\infty}^v I(y \in A) \phi(y) dy}{\int_A \phi(y) dy}.$$

Corollary 1 If $N = n g(\widehat{CV}^2)$ with a nonrandom function $g(\cdot)$, then V does not depend on model parameters under H_0 : $\theta = h = 0$.

Corollary 1 behaves similarly to Wilks's likelihood-ratio test statistic in not depending on either the sample size or the model parameters under H_0 . For example, under H_0 : $\theta = h = 0$, N_2 leads to $\tau = c^2(c^2 - \xi_1^2)^{-1}$ making τ and V independent of h and σ_θ .

3.3 When MLEs without SSR are Not Asymptotically Normal

MLEs are not always asymptotically normal, but may converge to another stationary distribution. For example, given the uniform distribution $Unif(0, \theta)$, where θ is the only model parameter, $X_{(n)} = \max(X_1, \dots, X_n)$ is the MLE of θ and its asymptotic properties are described by $n(\theta - X_{(n)}) \xrightarrow{d} Exp(\theta)$, where $Exp(\theta)$ is the exponential distribution.

More generally, consider an estimator $\tilde{\theta}$ such that $b_n(\tilde{\theta} - \theta) \xrightarrow{d} G$ without SSR, where b_n is a nonrandom sequence diverging to $\pm\infty$ and G is a stationary distribution defined on the real line. Let $V = b_N(\tilde{\theta} - \theta)$, $v_1 = b_n(\hat{\theta}_1 - \theta)$ and assume that the conditional distribution

$$G(v \mid v_1 = y) = \lim_{n \rightarrow \infty} \Pr(V < v \mid v_1 = y)$$

exist for every v_1 .

Theorem 2 If $\tau_n \xrightarrow{d} \tau$, then the asymptotic distribution of V does not depend on n .

$$\Pr(V < v) \rightarrow \int G(v \mid v_1 = y) dG(y). \quad (14)$$

$\Pr(V < v)$ may require Monte-Carlo resampling to calculate the integral, because the functional form of the conditional distribution $G(v \mid v_1 = y)$ may be unknown.

When N/n converges to a random variable, the asymptotic distribution of V can be found using Theorem 2. When N/n converges to a constant, like N_1/n does, then SSR does not change the asymptotic distribution.

An example of SSR under the Uniform distribution, focusing on width of a confidence interval: Given data follow the uniform distribution $Unif(0, \theta)$, the $(1 - \alpha)$ -level confidence interval for $X_{(n)}$ is $(X_{(n)}, X_{(n)}/\alpha^{1/n})$ with its width equal to $X_{(n)}(1 - \alpha^{-1/n})$. To find the sample size required to have an expected confidence interval width of some predetermined size w/n , solve

$$w/n = E(X_{(N)}) (\alpha^{-1/N} - 1) \approx \theta (\alpha^{-1/N} - 1)$$

for N to yield the new sample size:

$$N = - \frac{\log \alpha}{\log \left(1 + \frac{w}{n\theta}\right)}.$$

Note that $[1 + w/(n\theta)]^n \rightarrow e^{w/\theta}$. So

$$\frac{N}{n} = - \frac{\log \alpha}{\log \left(1 + \frac{w}{n\theta}\right)^n} \rightarrow - \frac{\theta}{w} \log \alpha.$$

Thus, by analogy with N_1 , the ratio N/n also converges to a point mass, which does not change the asymptotic distribution, and the sample size can be re-estimated by

$$N_7 = - \frac{nX_{(n)}}{w} \log \alpha.$$

Note that, for every given θ , formula N_7 has an (asymptotic) upper boundary

$N_7 \leq - \frac{n\theta}{w} \log \alpha$ forcing studies with N_7 to stop with stage I data.

3.4 Practical Implications

3.4.1 Normality in Theorem 1—Distributions of Z-scores based on MLEs can be closely approximated by normal distributions when the major probability mass of N is

located close to n , or is much larger than n (associated with large values of c^2) or when N/n converges in probability to a constant. These scenarios follow directly from Theorem 1 and are supported by Figures 1 and 2.

First consider the scenario in which N is much larger than n . Asymptotically, this means that the probability mass of τ_n is approaching 1. Mathematically, we describe this as $\forall \epsilon > 0, \exists B: \forall n > B, \Pr(\tau_n > 1 + \epsilon) < \epsilon$. This situation makes the first term in (13) disappear as $p_I = 0$. The influence of the first stage data is asymptotically fully overwhelmed by the second stage data and the second term becomes the standard normal distribution:

$$(1 - p_I) \int_{y: \tau \geq 1} \Phi(\sqrt{\tau}v - \sqrt{\tau - 1}y) \phi(y) dy = \int \Phi(v) \phi(y) dy = \Phi(v) \quad (15)$$

since $p_I = 0$ and $\tau = 1$ do not depend on y .

The second scenario occurs when N is close to n . Asymptotically this corresponds to the probability mass of τ_n moving toward infinity; mathematically we say that $\forall \epsilon > 0, \exists B: \forall n > B, \Pr(\tau_n > 1/\epsilon) < \epsilon$. This makes the second term in (13) disappear. The influence of the second stage data is asymptotically fully overwhelmed by the first stage data, so $p_I \Phi(v | \tau = 1) = \Phi(v)$ and $p_I = 1$.

The third scenario leading to normal V is convergence of τ_n to a point mass (τ is a constant). SSR N_1 (known h) leads N_1/n , and consequently τ_n , to converge in probability to constant values.

3.4.2 Non-Normality in Theorem 1—Normality can be problematic if both $\Pr(N < n) > 0$ and $\Pr(N > n) > 0$. These are the scenarios N_2 to N_6 . To emphasize the possibility that V can be non-normally distributed we plotted Figure 3, where we considered extremely large $n = 10,000$ and $n_{max} = 2,000,000$ and plotted the distribution of V with SSR N_5 under $h = 0$. In this example, the average total sample size is 29% higher than n ensuring that both stages contributed to V .

3.4.3 Independence of Model Parameters under H_0 —As follows from Corollary 1, SSR N_1 to N_5 lead to model independent distributions of V under H_0 . This directly follows from the fact these sample sizes depend only on CV, whose asymptotic distribution is known under H_0 . This property allows researchers to estimate asymptotic critical values under known SSR.

3.4.4 Non-Normality in Theorem 2—Theorem 2 shows that SSR can change the distribution of V , where V can be any statistic with a stationary non-degenerate asymptotic distribution.

4 Monte-Carlo Experiments

4.1 Hypothesis Testing: Type I Error, Power, and Expected Sample Sizes

As was shown earlier, the asymptotic distributions depend on model parameters θ and η and, when these parameters are assumed, Monte-Carlo resampling can be used to estimate

conditional and marginal critical values for hypotheses tests, as well as their power properties. Table 2 reports the results of 36 Monte-Carlo simulation studies (18 for type I error and 18 for power evaluation) with different choices of SSR and h/σ_θ . The value of c^2 was selected for each sample size formula so that power is at 80% while type I error is controlled at 5% when $\theta = h/\sqrt{n} = 1/\sqrt{30}$. To estimate type I error, data were generated from a $N(0, 1)$ distribution; power was assessed under $N(h/\sqrt{n}, 1)$, $h \in \{1, 0.5, 3\}$ with $n = 30$ and $n_{max} = 600$. Each Monte-Carlo experiment was based on 100,000 repetitions.

4.2 Critical Values

Decisions on acceptance or rejection of null hypotheses were based on four different rules which we call *naive unadjusted critical values*, *conditional critical values estimated by quantile regression*, *conditional critical values estimated by normal approximation* and *marginal critical values*. These are described now.

4.2.1 Naive Unadjusted Critical Values—First, we used an unadjusted critical value assuming asymptotic normality of V . That is, we compared P-values against the traditional 0.05 level to make a decision.

The unadjusted critical value does not provide good control of type I error except in the case of N_1 SSR, where the dependence on stage I data quickly disappears as sample size becomes larger. At smaller sample sizes, there is some minor inflation of type I error which is corrected with adjusted (conditional or marginal) critical values. Power properties and sample sizes do not change much under N_1 SSR. Unfortunately, the researcher is rarely in the ideal situation.

4.2.2 Conditional Critical Values Estimated by Quantile Regression—The second approach conditions critical values on the final sample size. This ensures that, at any sample final size, the type I error will be at the 5% level. These critical values were determined by 100,000 draws under H_0 . The 0.05-level quantile regression [Koenker (2005)] was fit to these P-values transformed to the logit scale. The predictors in this quantile regression model were (1) a dichotomous variable that N hit the lower boundary, (2) a dichotomous variable that N hit the upper boundary, (3) the natural logarithm of the total sample size if the sample size was within the boundaries, and (4) the square of this logarithm. Then for each N , we estimated the predicted 0.05-level quantile. This approach is similar to conditioning on stage I data but not exactly the same because there exists a range of stage I values of $\hat{\theta}_1$ and $\hat{\eta}_1$ associated with $N = n$ and with $N = n_{max}$.

Using conditional critical values approximates type I error control at both conditional and marginal levels. Across all SSR formulae with some very minor deviations, the type I error was controlled at 5%. At the same time, power properties suffer as compared to decisions made with marginal critical values.

4.2.3 Conditional Critical Values Estimated by Normal Approximation—Third, conditional on N , type I error can be estimated using the asymptotic result

$$\Phi(\sqrt{\tau}v - \sqrt{\tau-1}y | H_0, \xi_1 = y) = \alpha/2.$$

This approach can be applied to all SSR formula using $\tau_n = N/(N-n)$ to approximate τ .

Example (N_2 SSR): If $\alpha = 0.05$ and $\beta = 0.2$, under H_0 and $\xi_1 = y$, $\tau(y) = c^2 (c^2 - y^2)^{-1}$. Then, $\sqrt{\tau(y)} = c/\sqrt{c^2 - y^2}$ and $\sqrt{\tau(y)-1} = |y|/\sqrt{c^2 - y^2}$ with $|y| > c$ (a condition for $N > n$) leading to

$$\Phi\left(\frac{CV-y^2}{\sqrt{c^2-y^2}} \middle| H_0, \xi_1 = y\right) = 0.025,$$

and the conditional rejection region

$$\left(-\infty, \frac{y^2}{c} - 1.96\sqrt{1 - \frac{y^2}{c^2}}\right) \cup \left(\frac{y^2}{c} + 1.96\sqrt{1 - \frac{y^2}{c^2}}, \infty\right).$$

This adjustment generally worked better than using conditional critical values calculated using quantile regression: power was approximately 2% higher but at the expense of an occasional slight inflation of type I error by 0.1%–0.2%.

4.2.4 Marginal Critical Values—The fourth approach calculated a single (marginal) critical value ignoring the possibility that type I error may differ at different total sample sizes. This approach does not control the conditional type I error, but shows consistently higher power than conditional control for type I error.

4.2.5 Expected Sample Sizes—Among the SSR methods, N_1 showed the smallest averaged sample size under both H_0 and H_1 . This is an expected finding since this is the only method using a known treatment effect. N_2 and N_3 show similar power with conditional and marginal critical values. N_3 has lower averaged values than N_2 showing benefit of bootstrap estimation of CV^2 . A surprisingly good performance was observed for N_4 : sample size was substantially lower than N_2 and N_3 while controlling the type I error and power.

4.3 Choosing an SSR Procedure

Table 2 shows how researchers can choose a sample size re-estimating procedure. A researcher has to come up with a stage I sample size (n), define an upper boundary for the total sample size n_{max} and assume a CV value. Then, the researcher can choose an SSR for this CV controlling type I error either conditionally or marginally. Table 2 calibrates SSR formulae at $CV = 1$ ($h = 1$, $\sigma_\theta = 1$) so that at $CV = 1$ power is secured at 80%. If the CV is not as assumed (e.g., 0.5 or 3), then power departs from the desired level.

Researchers need to be careful with interpretation of conditional quantities, like conditional type I error. These quantities are random and have no direct interpretation as population

parameters do. As it is shown in Table 2, tight control of conditional type I error may lead to suboptimal designs with higher power and higher expected sample sizes.

5 An Analysis of a Published Cox Proportional Hazards Model

To illustrate practical use of SSR, we reanalyzed a dataset of 7,874 persons available in Terry Therneau's "Survival" R package. This dataset contains serum free light chain (FLC) information, age, sex, creatinine when available, survival status and a follow-up period. This dataset is an extract from a larger dataset of 15,759 persons where the association between FLC and overall survival was established by a Cox proportional hazards model [Dispenzieri et al (2012)].

To determine the effect of FLC (which is a continuous variable defined as the sum of kappa and lambda portions of serum free light chain) on hazards of survival controlling for age, sex and creatinine, we fit the following Cox regression model:

$$\log h(t|FLC, A, M, C) = \log h_b(t) + \beta_1 FLC + \beta_2 A + \beta_3 M + \beta_4 C,$$

where $h_b(t)$ is the baseline survival hazard; A is "Age"; M is an indicator of male sex, and C is a creatinine value. The regression coefficient of FLC (β_1) was estimated as $\hat{\beta}_1 = 0.1368$ ($SE = 0.0093$) with $p < 0.0001$.

This large sample size ensures that very small effects of FLC are detectable with high power. Can we show the effect of FLC exists at a smaller sample size? What happens if an SSR is applied after $n = 100$ with a maximum possible sample size of $n_{max} = 1000$? The same Cox regression applied to the first 100 records gives the regression coefficient of FLC equal to 0.1568 ($SE = 0.0512$). We estimate the standard deviation to be $SD \approx SE/\sqrt{100} = 0.5123$. The application of SSR N_2 ($c^2 = 15.7$) leads to $N = 167.62$ which we round up to 168. From the Cox regression model on 168 persons, the regression coefficient is 0.1257 ($SE = 0.0309$), with $p < 0.0001$ and a Wald 95% confidence interval of 0.0652 to 0.1862. Application of Theorem 1 leads to a wider 95% confidence interval of 0.0415 to 0.2503 and a highly significant P-value < 0.0001 .

The use of N_4 ($c^2 = 19.0$) leads to $N = 193$. Both N_2 and N_4 were calibrated to detect $CV = 1$ with 80% power.

To convert this illustrative example into a simulation study, we repeated drawing stage I data, performing N_2 SSR, drawing stage II data and performing final analysis 100,000 times on the randomly reshuffled FLC dataset. The distribution of V and the total sample MLE on these 100,000 iterations are presented on Figure 4. Using CLT critical values for statistical significance, the null hypothesis of no FLC effect was rejected in 96.2% cases, whereas the rejection rate was 95.6% with the use of local alternatives. This comes at an expense of type I error, which was 5.6% using local alternatives and 6.6% using CLT. Resampling under the null hypothesis was conducted by a random permutation of the FLC variable under the assumption that FLC is independent of other model covariates. Coverage probability for nominal 95% confidence intervals was 77.5% using CLT and 89.7% using local alternatives.

Limitations of this resampling are that the full sample is considered to be the whole population and the underlying data generating model is not known, as it is in traditional Monte-Carlo simulation studies reported in Section 4. Then, the parameter of interest, say, $\beta_1 = E(\hat{\beta}_1)$, is defined as an application of the Cox model to the whole dataset of 7,874 persons. But at smaller sample sizes, Cox regression estimates are known to have bias, which decreases with increased sample size. One million random samples of a predetermined sample size (e.g., 100, or 200, or 1000) allowed us to accurately estimate the value of $E(\hat{\beta}_1|N)$ at different sample sizes. For example, $E(\hat{\beta}_1|N = 100) = 0.1812$, $E(\hat{\beta}_1|N = 200) = 0.1719$, and $E(\hat{\beta}_1|N = 1000) = 0.1515$. When we consider coverage probability of the nominal 95% CI for $E(\hat{\beta}_1|N)$, we get 94.2% for local alternatives and only 81.7% for the CLT based CI.

This illustrative example showed the advantage of Theorem 1 over the CLT.

6 Summary

It has been previously shown that MLEs continue to be competitive estimators despite SSR. However, the use of SSR in the middle of data collection calls MLEs' asymptotic properties into question. Our Theorems 1 and 2 establish when, after SSR, the asymptotic distribution under the null and local alternatives may be a random mixture of distributions that should be used in place of asymptotic distributions that apply without SSR. Given MLEs are asymptotically normal without SSR, mixtures arises from boundaries imposed on the re-estimated sample size and from SSR algorithms that depend on estimates of the effect size.

In Theorem 1, these non-degenerate asymptotic mixture distributions are obtained at the usual $\sqrt{n}$ -convergence rate of the standardized test statistic. By analogy with Wilks's phenomenon under the null hypothesis, the limiting distribution does not depend on model parameters, except for the probabilities that the total (random) sample size falls at lower and upper boundaries.

Theoretical properties are illustrated with a few simulation studies and an application to a real life dataset. We encourage researchers to use SSR in their practice, but to be aware that if a treatment effect is used in an SSR then the asymptotic properties of MLEs may be different than they are without SSR.

Note, that in addition to Theorem 1's application to SSR, it also applies to group sequential analyses in which $N = n_1$ (early stopping) or $N = n_1 + n_2$ (full sample). Also, to elicit the asymptotic distribution for multiple interim analyses, simply apply Theorem 2 recursively. These results will be studied further in a subsequent manuscript.

Acknowledgments:

We thank Dr. Assaf P. Oron, an unknown referee and the editor for their very valuable feedback on our manuscript.

7: Appendix

Proof (Theorem 1): The joint asymptotic density (8) is fully determined by $(\hat{\theta}_1, \hat{\eta}_1, \hat{\theta}_2, \hat{\eta}_2)$, and marginally under $\theta = h/\sqrt{n}$,

$$\xi_{n1} = \sqrt{n}\sigma_\theta^{-1}(\hat{\theta}_1 - h/\sqrt{n}) \rightarrow \xi_1 \stackrel{d}{=} N(0, 1). \quad (16)$$

By the continuous mapping theorem and the consistency of MLEs, σ_θ evaluated as a function of $\hat{\theta}_1$ and $\hat{\eta}_1$ is a consistent estimator of σ_θ and using it in place of σ_θ , (16) still holds. Similarly, conditionally on N ,

$$\xi_{n2} = \sqrt{N-n} \cdot \sigma_\theta^{-1}(\hat{\theta}_2 - h/\sqrt{n}) \rightarrow \xi_2 \stackrel{d}{=} N(0, 1). \quad (17)$$

Recalling that, with large n and $N-n$, the full sample MLE is approximately distributed as $(n/N)\hat{\theta}_1 + (1-n/N)\hat{\theta}_2$, a large sample approximation follows:

$$V = \sqrt{N}\sigma_\theta^{-1}(\tilde{\theta} - h/\sqrt{n}) \approx \sqrt{\frac{n}{N}}\xi_1 + \sqrt{1-\frac{n}{N}}\xi_2 \xrightarrow{d} \sqrt{\frac{\tau-1}{\tau}}\xi_1 + \sqrt{\frac{1}{\tau}}\xi_2$$

and

$$\begin{aligned} \Pr(V < v) &\rightarrow \Pr\left(\sqrt{\frac{\tau-1}{\tau}}\xi_1 + \sqrt{\frac{1}{\tau}}\xi_2 < v\right) \\ &= E_{\xi_1} \Pr(\xi_2 < \sqrt{\tau}v - \sqrt{\tau-1}y | \xi_1 = y). \end{aligned} \quad (18)$$

It is possible that, for some values of $\hat{\theta}_1$ and $\hat{\eta}_1$, the total sample size N is not large enough to claim normality; but the probability of this event converges to zero. This follows directly from the equality

$$1 = \Pr(N \leq n) + \Pr(n < N \leq n + \sqrt{n}) + \Pr(N > n + \sqrt{n}).$$

The event $n < N \leq n + \sqrt{n}$ includes all possible values of N where the asymptotic normality of V may not hold. The probability of this event

$$\Pr(n < N \leq n + \sqrt{n}) = \Pr\left(1 < \frac{N}{n} \leq 1 + \frac{1}{\sqrt{n}}\right) \rightarrow 0$$

when N/n converges to a stationary random variable with a right-continuous distribution at 1.

Then, for large n ,

$$\Pr(V < v) \approx \int_{-\infty, \tau \geq 0}^{\infty} \Phi(\sqrt{\tau}v - \sqrt{\tau-1}y)\phi(y)dy. \quad (19)$$

Equations (18) and (19) are not defined for $\tau = 1$ which corresponds to values of $\hat{\theta}_1$ and $\hat{\eta}_1$ associated with $N = n$. When the bound is incorporated the proof is complete. **Q.E.D.**

Proof (Corollary 1): The $\widehat{CV}^2$ is asymptotically distributed as a noncentral chi-squared random variable with 1 d.f. whose distribution is fully determined by CV^2 . Then, asymptotically, under H_0 , $\widehat{CV}^2$ is a central chi-squared random variable with 1 d.f. which completes the proof. **Q.E.D.**

References

- Brannath W, Koenig F, Bauer P (2006) Estimation in flexible two stage designs. *Statistics in Medicine* 25(19):3366–3381 [PubMed: 16220489]
- Brannath W, Gtjzhr G, Bauer P (2012) Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association* 107(498):824–832
- Bretz F, Koenig F, Brannath W, Glimm E, Posch M (2009) Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 28(8): 1181–1217 [PubMed: 19206095]
- Broberg P, Miller F (2017) Conditional estimation in two-stage adaptive designs. *Biometrics* 73(3): 895–904 [PubMed: 28099993]
- Crowder MJ (1976) Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society Series B (Methodological)* pp 45–53
- Dispenzieri A, Katzmnn JA, Kyle RA, Larson DR, Therneau TM, Colby CL, Clark RJ, Mead GP, Kumar S, Melton LJ, et al. (2012) Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population In: *Mayo Clinic Proceedings*, Elsevier, vol 87, pp 517–523 [PubMed: 22677072]
- Dmitriev Y, Tarassenko P, Ustinov Y (2014) On estimation of linear functional by utilizing a prior guess In: Dudin A, Nazarov A, Yakupov R, Gortsev A (eds) *Information Technologies and Mathematical Modelling*, Springer International Publishing, Cham, pp 82–90
- Ferguson T (1996) *A Course in Large Sample Theory*. Routledge, New York
- Graf AC, Gtjzhr G, Brannath W (2016) Precision of maximum likelihood estimation in adaptive designs. *Statistics in Medicine* 35(6):922–941, sim.6761. [PubMed: 26459506]
- Ivanova A, Flournoy N (2001) A birth and death urn for ternary outcomes: stochastic processes applied to urn models. Chapman and Hall/CRC, Boca Raton
- Ivanova A, Rosenberger WF, Durham SD, Flournoy N (2000) A birth and death urn for randomized clinical trials: Asymptotic methods. *Sankhy, Series B* 62(1):104–118
- Jeganathan P (1982) On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. *Sankhya: The Indian Journal of Statistics, Series A* (1961–2002) 44(2): 173–212
- Koenker R (2005) *Quantile Regression*. Econometric Society Monographs, Cambridge University Press
- Lane A, Flournoy N (2012) Two-stage adaptive optimal design with fixed first-stage sample size. *Journal of Probability and Statistics* 2012:n/a–n/a
- Le Cam LM (1960) Locally asymptotically normal families of distributions. *Univ California Publ Statist* 3:37–98
- Li G, Shih WJ, Xie T, Lu J (2002) A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* 3(2):277–287 [PubMed: 12933618]
- May C, Flournoy N (2009) Asymptotics in response-adaptive designs generated by a two-color, randomly reinforced urn. *The Annals of Statistics* 37(2):1058–1078
- Mutze T, Schmidli H, Friede T (2017) Sample size re-estimation incorporating prior information on a nuisance parameter. *Pharmaceutical Statistics* 17(2):126–143, pst.1837. [PubMed: 29181869]
- Philippou AN, Roussas G, et al. (1973) Asymptotic distribution of the likelihood function in the independent not identically distributed case. *The Annals of Statistics* 1(3):454–471

- Proschan MA, Hunsberger SA (1995) Designed extension of studies based on conditional power. *Biometrics* 51(4):1315–1324 [PubMed: 8589224]
- Proschan MA, Lan KKG, Wittes JT (2006) *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer
- Roussas GG, Bhattacharya D (2011) Revisiting local asymptotic normality (LAN) and passing on to local asymptotic mixed normality (LAMN) and local asymptotic quadratic (LAQ) experiments In: *Advances in Directional and Linear Statistics*, Springer, pp 253–280
- Sugiura N, et al. (1973) Asymptotic non-null distributions of the likelihood ratio criteria for covariance matrix under local alternatives. *The Annals of Statistics* 1(4):718–728
- Tarima S, Pavlov D (2006) Using auxiliary information in statistical function estimation. *ESAIM: Probability and Statistics* 10:11–23
- Tarima S, He P, Wang T, Szabo A (2016) Interim sample size recalculation for observational studies. *Observational Studies* 2(2):65–85
- Van der Vaart A (1998) *Asymptotic Statistics*. Cambridge University Press
- Wang H, Flournoy N, Kpamegan E (2014) A new bounded log-linear regression model. *Metrika* pp 695–720
- Wang Y, Li G, Shih WJ (2010) Estimation and confidence intervals for two-stage sample-size-flexible design with LSW likelihood approach. *Statistics in Biosciences* 2(2): 180–190

Fig. 1.

Distributions of $\log(N_2/n)$ and $V = \sqrt{N_2}(\tilde{\theta} - 0)$ with $N_2(c^2 = 15.7)$, SSR under H_0 and local alternative hypotheses $h = 0.5, 1$, and 3 ; $X \sim N(\theta, 1)$, $n = 100$, $n_{max} = 2000n$.

Fig. 2.

Distribution of $\log(N/n)$ and $V = \sqrt{N}(\tilde{\theta} - \theta)$ with $X \sim N(\theta, 1)$ following $N_4(c^2 = 19)$ and $N_5(c^2 = 15)$ SSR under $h = 0.5$ and $h = 3$; $n = 100$, $n_{max} = 2000n$.

Fig. 3. Histogram of the full sample standardized MLE with SSR N_5 ; $X \sim N(0, 1)$; $V = \sqrt{N}\tilde{\theta}$, ($c^2 = 20$), $n = 10,000$, $n_{max} = 2,000,000$, and 30,000 Monte-Carlo resamples. The solid line plots the standard Normal density.

(a) Histogram of $\log(N_2/n)$ (b) Density of V **Fig. 4.**

Distributions of $\log(N_2/n)$ from the free light chain example, 100,000 resamples. The standard normal Gaussian curve is graphed with dashed line. The solid lined Gaussian density uses sample mean and sample standard deviation.

Table 1Simulation of the Z -test with $X \sim N(\theta, 1)$.

SSR	n	Average N	$\theta = h/\sqrt{n}$	Average $\bar{X}$	Power	Shapiro-Wilk P-value
N_1	10	83	0.316	0.317	0.741	<0.001
N_1	100	769	0.100	0.100	0.793	0.672
N_2	10	35	0.316	0.393	0.785	<0.001
N_2	100	414	0.100	0.122	0.809	<0.001
N_1	10	36	0	0.000	0.054	<0.001
N_1	100	1003	0	0.001	0.054	0.251
N_2	10	89	0	0.001	0.092	<0.001
N_2	100	2879	0	0.000	0.085	<0.001

Table 2

Type I Error and Power. “Naive” decisions are based on the critical value of 0.05. “C1” decisions are based on conditional critical values estimated by quantile regression; “Mar” decisions are based on marginal critical values and “C2” decisions are based on conditional critical values estimated by the normal approximation. SSR formulae were calibrated for 80% power at $CV = 1$; scenarios with $CV = 3$ and $CV = 0.5$ reveal effects of model misspecification in SSR.

		Type I error		Power			
			Adjusted		Adjusted	EN	
CV	SSR(c^2)	Naive	C1,Mar,C2	Naive	C1,Mar,C2	H_0	H_1
1	N_1 (8.3)	.051	.050,.050,.052	0.800	0.797,0.795,0.826	249	249
1	N_2 (15.7)	.064	.050,.050,.051	0.895	0.801,0.875,0.818	480	381
1	N_3 (17.8)	.068	.050,.050,.052	0.892	0.803,0.866,0.822	470	379
1	N_4 (19.0)	.080	.050,.050,.051	0.862	0.801,0.847,0.824	355	293
1	N_5 (15.0)	.063	.050,.050,.051	0.892	0.805,0.874,0.822	464	370
1	N_6 (20.0)	.060	.050,.050,.051	0.881	0.802,0.865,0.810	443	353
3	N_1 (8.3)	.051	.051,.050,.052	1.000	1.000,1.000,1.000	250	250
3	N_2 (15.7)	.064	.050,.050,.051	0.999	0.771,0.999,0.811	479	85
3	N_3 (17.8)	.068	.051,.050,.052	0.999	0.780,0.998,0.842	470	95
3	N_4 (19.0)	.063	.051,.050,.051	1.000	0.858,1.000,0.932	355	107
3	N_5 (15.0)	.061	.051,.050,.052	1.000	0.817,0.999,0.833	463	89
3	N_6 (20.0)	.059	.051,.050,.051	1.000	0.983,1.000,0.923	443	108
.5	N_1 (8.3)	.051	.050,.050,.051	0.303	0.301,0.299,0.383	250	250
.5	N_2 (15.7)	.064	.050,.050,.050	0.503	0.471,0.454,0.566	479	452
.5	N_3 (17.8)	.064	.050,.050,.050	0.503	0.471,0.454,0.566	479	452
.5	N_4 (19.0)	.062	.050,.050,.051	0.490	0.460,0.450,0.552	464	438
.5	N_5 (15.0)	.061	.050,.050,.051	0.472	0.446,0.439,0.537	444	419
.5	N_6 (20.0)	.055	.051,.050,.051	0.541	0.513,0.526,0.611	546	500