Most Powerful Test Sequences with Early Stopping Options

Sergey Tarima · Nancy Flournoy

Received: date / Accepted: date

Abstract Sequential likelihood ratio testing is found to be most powerful in sequential studies with early stopping rules when grouped data come from the one-parameter exponential family. First, to obtain this elusive result, the probability measure of a group sequential design is constructed with support for all possible outcome events, as is useful for designing an experiment prior to having data. This construction identifies impossible events that are not part of the support. The overall probability distribution is dissected into stage specific components. These components are sub-densities of interim test statistics first described by Armitage, McPherson and Rowe (1969) that are commonly used to create stopping boundaries given an \( \alpha \)-spending function and a set of interim analysis times. Likelihood expressions conditional on reaching a stage are given to connect pieces of the probability anatomy together.

The reduction of the support caused by the adoption of an early stopping rule induces sequential truncation (not nesting) in the probability distributions of possible events. Multiple testing induces mixtures on the adapted support. Even asymptotic distributions of inferential statistics are mixtures of truncated distributions. In contrast to the classical result on local asymptotic normality (Le Cam 1960), statistics that are asymptotically normal without stopping options have asymptotic distributions that are mixtures of truncated normal distributions under local alternatives with stopping options; under fixed alternatives, asymptotic distributions of test statistics are degenerate.

Keywords Adaptive designs · adapted support · group sequential designs · local asymptotics · interim hypothesis testing · likelihood ratio tests

Institute for Health and Society, Medical College of Wisconsin, 8701 Watertown Plank Rd, Wauwatosa, WI, 53226 · Department of Statistics, University of Missouri, 600 S State St., Apt. 408 Bellingham, WA 98225
1 Introduction

We define a sequential experiment to be one in which the decision to stop collecting data is based on data collected previously in the study. Wetherill and Glazebrook \cite{52} emphasize that "two aspects of a sequential procedure must be clearly distinguished, the stopping rule, and the manner in which inferences are made once observations are stopped. ... in the design problem, it is important to know how probable are various possible results." We distinguish the probability framework underpinning these two activities, and also a third - the probability framework underlying interim hypothesis tests.

Dodge and Romig \cite{9} proposed the first known sequential test procedure in which a decision to stop or continue collecting data was based on prior data, recognizing that decisions to stop or continue a trial made based on prior observations could substantially reduce the expected numbers of required subjects; it was a two-stage design. Bartky \cite{3} devised a multiple sequential testing procedure for binomial data based on Neyman and Pearson \cite{35}'s likelihood ratio test that Wald \cite{50} cites as a "forerunner" to his more general sequential probability ratio test (SPRT) procedure in which the probabilities of type I and II errors are controlled. Extensions with stopping decisions based on groups of subjects [Group Sequential Designs (GSDs)] are given in Jennison and Turnbull \cite{21}.

Neyman and Pearson \cite{35} show that likelihood ratio tests are most powerful for testing a simple null versus a simple alternative hypotheses. In Ferguson \cite{12}, the Karlin-Rubin theorem is viewed as an extension of Neyman-Pearson approach to most powerful testing of composite hypotheses. The Karlin-Rubin theorem applies to the one-dimensional exponential family. With sequential stopping options, some elements of the sample space become impossible, which changes the distributions of statistics. Sufficient statistics become dependent on the random sample size; see Blackwell \cite{5}. If statistics belong to the exponential family without early stopping options, then they belong to a curved exponential family when exposed to early stopping options (Efron \cite{10}, Liu and Hall \cite{29}, Liu et al. \cite{30}). Section 3.2 shows that despite being from a curved exponential family, the sequential tests based on likelihood ratios continue to be most powerful for any $\alpha$-spending function.

1.1 Notation and A Simple Example

This example demonstrates a couple important repercussions of sequential stopping rules:

- The support is reduced,
- Bivariate normal random variables become non-observable; the observable bivariate random variable is a mixture of truncated normal random variables,

Let $X_1$ and $X_2$ be $N(\theta, 1)$ random variables with an unknown location parameter $\theta$. 

1.1.1 Non-sequential experiments

If $X_1 = x_1$ alone is observed, the log-likelihood

$$l(\theta | X_1 = x_1) = -2 \log \sqrt{2\pi} - 0.5(x_1 - \theta)^2$$

is maximized at $\hat{\theta}_1 = x_1$.

If both $X_1 = x_1$ and $X_2 = x_2$ are observed independently, the random variable $(X_1, X_2)$ is defined on the probability space $(\mathbb{R}^2, \mathcal{B}, P)$, where $\mathbb{R}$ is the sample space $[\mathbb{R} = (-\infty, \infty)]$, $\mathcal{B}$ is the Borel $\sigma$-algebra on $\mathbb{R}^2$ and $P$ is a bivariate normal distribution with mean vector $(\theta, \theta)$, units variances and zero correlation.

The log-likelihood function is

$$l(\theta | X_1 = x_1, X_2 = x_2) = -2 \log \sqrt{2\pi} - 0.5(x_1 - \theta)^2 - 0.5(x_2 - \theta)^2$$

and the maximum likelihood estimator (MLE) of $\theta$ is $\hat{\theta}_2 = (x_1 + x_2)/2$.

1.1.2 Sequential experiments: likelihood, support and probability measures

What happens in sequential settings when $X_2$ is only observed if $X_1 < 2.18$? Let $D$ denote the random stopping stage. Then $D = 1 + I(X_1 < 2.18)$, where $I(\cdot)$ is an indicator function. In this simple example, $D$ is also the random sample size.

The joint distribution of $(X, D)$ and the marginal distribution of $X$ are the same:

$$f_X(x|\theta) = f_{X,D}(x,d|\theta) = \phi(x_1 - \theta) [\phi(x_2 - \theta)]^{I(d=2)}$$

$$= [\phi(x_1 - \theta)]^{I(d=1)} [\phi(x_1 - \theta)\phi(x_2 - \theta)]^{I(d=2)},$$

where $\phi(\cdot)$ is the standard normal density. The representation of $f_X(x|\theta)$ in the first line of (1) partitions the density according to data collection stages, while the representation in the second line partitions the density according to stopping stages. In canonical form,

$$f_{X,D}(x,d|\theta) = h(x) \exp \left( [x_1 + I(d=2)x_2] \theta - [1 + I(d=2)] \frac{\theta^2}{2} \right),$$

where $h(x) = \exp \left( -\frac{1}{2} [x_1^2 + I(d=2)x_2^2 + I(d=2) \log (\sqrt{\pi}) + \log (\sqrt{\pi})] \right)$. Thus, the density (2) belongs to the curved exponential family with a sufficient statistic

$$\left( \sum_{k=1}^d x_k, d \right) = [x_1 + I(d=2)x_2, 1 + I(d=2)].$$

Curved exponential families were defined by Efron [10] and the sufficient statistic with a random number of summands $\sum_{k=1}^D x_k$ was derived by Blackwell.
Probability distribution (2) is a special case of the exponential family derived in Liu et al. [30] [see their formula (2.6)]. A more general probability distribution of $X$ is presented in Section 2.

The log-likelihood function

$$l(\theta | X_1 = x_1, X_2 = x_2, D = d) = \begin{cases} [\phi(x_1 - \theta)]^{I(d=1)} & \text{if } D = 1, \\ [\phi(x_1 - \theta)\phi(x_2 - \theta)]^{I(d=2)} & \text{if } D = 2 \end{cases}$$

is maximized at $\hat{\theta} = I(d = 1)x_1 + I(d = 2)(x_1 + x_2)/2$. Consequently, the score function, the MLE and the observed information are the same as for the non-sequential experiment. What changes?

The joint support of $X_1$ and $X_2$ changes because $X_2$ becomes impossible (not just missing) when $X_1 \geq 2.18$. The random variable $(X_1, X_2)$ is non-observable when $X_1 \geq 2.18$ and the joint distribution of $X_1$ and $X_2$ is therefore truncated and not normal. Formally, the support for joint density can be decomposed into support for the experiment stopping with $X_1$ and support for the experiment continuing to observe $X_2$:

$$T = \{x_1 \geq 2.18\} \cup \{\{x_1 < 2.18\} \cap \{x_2 \in \mathbb{R}\}\} \subset \mathbb{R}^2.$$ 

This $T$ is a special case of the support formalized in Liu et al. [30]. If $A \in T$, then

$$P_{(D)} = \Pr((X_1, X_2) \in A) = \Pr(X_1 \geq 2.18)\Pr(X_1 \in A | X_1 \geq 2.18) + \Pr(X_1 < 2.18)\Pr((X_1, X_2) \in A | X_1 < 2.18)$$

is a probability measure on the $\sigma$-algebra $\sigma(T)$. Thus, the observable random variable

$$X_T = \begin{cases} X_1 & \text{if } D = 1, \\ (X_1, X_2) & \text{if } D = 2. \end{cases}$$

is defined on the probability space $(\mathcal{T}, \sigma(\mathcal{T}), P_{(D)})$.

In contrast to $(X_1, X_2)$, $X_T$ is observable for this sequential experiment. The MLE, is a random variable defined on $(\mathcal{T}, \sigma(\mathcal{T}), P_{(D)})$ and its probability distribution is a mixture of the left truncated normal random variable $\{X_1 | D = 1\}$ and an average of the right truncated normal random variable $\{X_1 | D = 2\}$ and the normal random variable $X_2$.

Thus, in this sequential experiment, MLEs are not normal random variables as illustrated in Figure 1. The first column shows the distribution of the MLE if the experiment stopped at stage 1, which is left truncated normal. The seconds of histograms shows the distribution of the MLE if the experiment proceeded to stage 2. This distribution is a mixture of right truncated data from stage 1 and untruncated normal data from stage 2. The final column shows unconditional distribution of the MLE.
1.2 The Scope of this Paper

We consider sequential experiments having a small finite number of interim decision points, that is, experimental set-ups for which Martingale central limit theorems and Brownian theory are not suitable. Our interest is in experiments that aim primarily on a hypothesis test of effect size. We focus on characterizing the effect of sequential stopping rules on probability distributions of test statistics.

In this manuscript, Section 2 introduces notation for GSDs with stopping rules dependent on a parameter of interest (through the distributions of the test statistics). This section also presents distributions of cumulative test statistics conditional on reaching a stage, conditional on stopping at a stage, and unconditional defined on the probability space with truncation-adapted support. All of these probability distributions are truncated or truncated-mixtures. Section 3 presents likelihood-based inference on the truncation-adapted probability space. In section 3.1 a local asymptotic distribution of the MLEs is found to be non-degenerate; it is a mixture of truncated normal distributions. Section 3.2 shows that the possibility of early stopping does not change the monotonicity of likelihood ratios in the one-parameter exponential family. Thus, stage-specific tests continue to be uniformly most powerful by the Karlin-
Rubin theorem, which makes sequential tests based on monotone likelihood ratio uniformly most powerful. Throughout Sections 2 and 3 theoretical results are illustrated by a two stage example, Pocock [41]. Finally, Section 4 concludes this article with a summary and a discussion of impact within the contemporary research environment.

2 Probability Distributions with Early Stopping

Let $X$ denote subjects' outcome variable and assume that, when observed in isolation, it has a probability distribution function or a probability mass function $f_X = f_X(x|\theta)$. To simplify the material, the term density is used to refer to probability measures without formally distinguishing between them.

Let a sequence of $X$s be observed with the primary objective of testing the null hypothesis $H_0 : \theta = 0$ with overall $\alpha$-level type 1 error and $1 - \beta$ power at an alternative $H_1 : \theta = \theta_1$.

It is convenient to group the random sample into stages separated by the interim analysis times: $X_1, X_2, \ldots$, where $X_k = (X_{n(k-1)+1}, \ldots, X_{n(k)})$; $n(k) = \sum_{j=1}^{k} n_j$; here for simplicity $n_j$ is a pre-specified number of observations in stage $j$, $1 \leq k \leq K$; $n(0) = 0$ and $K - 1$ is the maximum number of interim analyses permitted. Every stage is assumed to be “reachable”, that is, there is a positive probability of reaching each stage. Data collection at each stage is followed by a hypothesis test that results in a decision to stop the study or to enroll a new group of patients; except that if stage $K$ is reached, the experiment stops after $n(K)$ observations regardless of the last $n_K$ observations' values.

2.1 Stopping Decisions

Let $T(k)$ be a function of observations $X(k) = (X_1, \ldots, X_{n(k)})$ that is compared against a cutoff value $c_k$ to determine whether to stop at stage $k$ or continue through stage $k + 1$, $1 \leq k < K - 1$. These decisions are defined by the events $\{\cap_{j=1}^{k-1} \{T(j) \leq c_j\} \cap \{T(k) > c_k\}\}$ and $\cap_{j=1}^{K-1} \{T(j) \leq c_k\}$, respectively, and are conveniently summarized by a random variable denoting the stopping stage:

$$D = K \cdot I (\cap_{j=1}^{K-1} \{T(j) \leq c_j\}) + \sum_{k=1}^{K-1} k \cdot I (\cap_{j=1}^{k} \{T(j) \leq c_j\} \cap \{T(k) > c_k\});$$

$D \in \{1, \ldots, K\}$ will appear as random index such as in $X_{(D)}$ to emphasize that the stopping stage is unknown and is described probabilistically though the random variable $D$. $X_{(d)}$ is the random variable $X_{(D)}$ conditioned on stopping with stage $D = d, d = 1, \ldots, K$.

It is important to account for $D$ in probability statements about $T_{(d)}$ because $D$ determines the observations’ probability support as illustrated in Section 1.1.
2.2 After deciding to stop

Fig. 2 Support associated with different stopping decisions when $K = 3$; $c_k$ is the critical value for stopping at stage $k$.

If study stops at stage $D = d < K$, $H_0$ is rejected. If $d = K$, the final hypothesis test determines acceptance or rejection of $H_0$. At the time of the final analysis, the density of observations conditional on stopping at stage $d$
(i.e., the density of $X_{(d)} = X_{(D)}|\{D = d\}$) is

$$f^C_{X_{(d)}} = f_{X_{(D)}}(x_{(D)}|D = d, \theta) = \frac{f_{X_{(D)}}(x_{(D)})}{Pr_{\theta}(D = d)} Pr_{\theta}(D = d),$$

where $f^s_{X_{(d)}} = [f_{X_{(d)}}(x_{(d)}|\theta)]^{I(D=d)}$ denotes the sub-density with support defined by $D = d$ (see the exemplary sketches in Figure 2). Similarly, for a statistic $T_{(d)} = T(X_{(d)})$, the conditional on $D = d$ density is $f^C_{T_{(d)}}$.

In contrast, if the stopping rule is not random and the experiment stops with $n_{(d)}$ observations, $Pr_{\theta}(D = d) \equiv 1$ and the observations have density $f_{X_{(d)}}$.

2.3 At the Time of Experimental Design

Prior to data collection, both $D$ and $X_{(D)}$ are unknown and the joint density of $X_{(D)}$ can be written in several ways:

$$f_{X_{(D)}}(x_{(D)}|\theta) = \sum_{d=1}^{K} [f_{X_{(d)}}(x_{(d)}|\theta)]^{I(D=d)}$$

$$= \sum_{d=1}^{K} f^s_{X_{(d)}} = \sum_{d=1}^{K} Pr_{\theta}(D = d) f^C_{X_{(d)}},$$

(5)

The joint density is a mixture of densities corresponding to possible outcome vectors; i.e., these densities are defined on non-overlapping regions of the density’s support.

When a test statistic $T_{(d)}$ summarizes $d$th stage data, the density of $T_{(D)}$ can be written analogous to Equation (5) as

$$f_{T_{(D)}}(t_{(D)}|\theta) = \sum_{d=1}^{K} [f_{T_{(d)}}(t_{(d)}|\theta)]^{I(D=d)}$$

$$= \sum_{d=1}^{K} f^s_{T_{(d)}} = \sum_{d=1}^{K} Pr_{\theta}(D = d) f^C_{T_{(d)}},$$

(6)

Even if every stage-specific test statistic $T_{d}$ is normally distributed, the distribution of $T_{(D)}$ is not.

2.4 At Interim Hypothesis Testing

Suppose at stage $d - 1$, the decision was made to continue sampling, the support for the $d$th test statistic is characterized by $D \geq d$. The density of
\( T_{(D)} \) conditional on \( D \geq d \) is

\[
f_{T_{(D)}}(t_{(D)}|D \geq d, \theta) = \frac{\sum_{k=d}^{K} f_{T_{(k)}}^{sub}}{\Pr_{\theta}(D \geq d)} = \sum_{k=d}^{K} \frac{\Pr_{\theta}(D = k)}{\Pr_{\theta}(D \geq d)} f_{T_{(k)}}^{C.} \tag{7}
\]

Again, even if each \( T_k \) is normal, the distributions of \( T_{(D)}|D \geq d \) are not.

### 2.4.1 Connection with Armitage’s algorithm

SAS’s popular SEQDESIGN procedure, R’s gsDesign package, Cytel’s EAST and others assess type I and power properties using a recursive sub-density formula (Armitage et al. [1]) to evaluate the distribution of \( T_{(D)}|D \geq d \). Armitage’s subdensity is \( f_{T_{(D)}}^{sub}|D \geq d = \sum_{k=d}^{K} f_{T_{(k)}}^{sub} \) and

\[
f_{T_{(D)}}(t_{(D)}|D \geq d, \theta) = \frac{f_{T_{(D)}}^{sub}|D \geq d}{\Pr_{\theta}(D \geq d)} \tag{8}
\]

For example, at \( K = 2 \), the sub-density of \( T_{(D)}|D \geq d \) is

\[
f_{T_{(D)}}^{sub}|D \geq 2(t|\theta) = \int_{-\infty}^{c_1} f_{T_{(D)}|D \geq 2}(t_1|\theta) f_{T_{(1)}}(t_1|\theta) dt_1 \tag{9}
\]

and its density is \( f_{T_{(D)}}|D \geq 2(t|\theta) = f_{T_{(D)}}^{sub}|D \geq d(t|\theta) \left( \int_{c_1}^{\infty} f_{T_{(1)}}(t_1|\theta) dt_1 \right)^{-1} \). Recursively, the density conditional on reaching the \( d \)th interim analysis is

\[
f_{T_{(D)}}(t|D \geq d, \theta) = \frac{\int_{-\infty}^{c_{d-1}} f_{T_{(D)}|D \geq d-1}(t|d_{d-1}, \theta) f_{T_{(d-1)}}(t_{d-1}|\theta) dt_{d-1} dt_{d-1}}{\int_{-\infty}^{c_{d-1}} f_{T_{(d-1)}}(t_{d-1}|\theta) dt_{d-1}}. \tag{10}
\]

### 2.5 \( \sigma \)-fields and support defined by a set of critical values

At the time of \( d \)th hypothesis test, given \( D \geq d \), the values \( \{x_j : j > n_{(d)}\} \) are not observable and hence do not contribute to the density; indeed, they do not belong to the adaptation-rule driven sample space, and consequently, they do not belong to a \( \sigma \)-field of the random process being monitored; hence, they do not belong to the sequential experiment as a whole. In this paper, by analogy with structural zeroes in contingency tables, these values are excluded from the sample space.

The \( d \)th stage-specific test statistic \( T_d = T(X_d) \) is defined on a probability space \((T_d, \sigma(T_d), P_d)\), where \( T_d \) is typically a real line \((\mathcal{R})\), \( \sigma(T_d) \) is Borel \( \sigma \)-field and \( P_d \) is a probability measure on the measurable space \((T_d, \sigma(T_d))\). A sequence of nested \( \sigma \)-fields \( \mathcal{F}_{(d)} := \sigma(T_1) \times \cdots \times \sigma(T_d) \), \( d = 1, \ldots, K \), creates a filtration \( F = (\mathcal{F}_{(d)})_{d \leq K} \) on the product probability space \((T, \mathcal{F}_{(K)}, P)\), where \( T = T_1 \times \cdots \times T_K \) and \( P = P_1 \times \cdots \times P_K \). But in the presence of possible stopping, not all combinations of \((T_1, \ldots, T_K)\) are possible. The cumulative test statistics \( T_{(d)} = T_{(d)}(T_1, \ldots, T_d) \) are defined only on a subspace of the sample space \( T_1 \times \cdots \times T_d \). Thus, probability environment substantially changes.
2.5.1 **Interim hypothesis testing**

The statistic \( T_{(1)} = T_1 \) conditional on reaching stage 1 \((D \geq 1)\) is defined on the sample space \( \mathcal{T}_{(1)} = \mathcal{T}_1 \); so \( \sigma (\mathcal{T}_{(1)}) = \mathcal{F}_{(1)} \). The statistic \( T_{(2)} \) conditional on \( D \geq 2 \) is defined on

\[
\mathcal{T}_{(2)} = \{ (t_1, t_2) : \{ t_1 \leq c_1 \} \},
\]
and \( \sigma (\mathcal{T}_{(2)}) \subset \mathcal{F}_{(2)} \). Further, for \( d \in \{3, \ldots, K\} \), \( T_{(d)} \) conditional on \( D \geq d \) is defined on

\[
\mathcal{T}_{(d)} = \{ (t_1, \ldots, t_d) : \{ t_{(j)} \leq c_j \}, j = 1, \ldots, d-1 \}.
\]

For all \( d \), the support \( \mathcal{T}_{(d)} \) and the \( \sigma \)-field \( \sigma (\mathcal{T}_{(d)}) \) is reduced by the possibility of early stopping: \( \sigma (\mathcal{T}_{(d)}) \subset \mathcal{F}_{(d)} \). This creates new measurable spaces \((\mathcal{T}_{(d)}, \sigma (\mathcal{T}_{(d)}))\) for interim tests at every possible stage \(1, \ldots, K\). Armitage’s recursive sub-density formula \cite{armitage2010} is defined on this measurable space; see Equation \((9)\) for \( K = 2 \). Re-scaling yields a density function \cite[see Equation \((10)\)]{armitage2010} which defines a probability measure to complete the probability space \((\mathcal{T}_{(d)}, \sigma (\mathcal{T}_{(d)}), P_{(d)})\), where the probability measure \( P_{(d)} \) is determined by density \((10)\):

\[
P_{(d)}(A) = \int_A f_{\mathcal{T}_{(d)}}(t|D \geq d, \theta) \text{d}t, \quad A \in \sigma (\mathcal{T}_{(d)}).
\]

2.5.2 **After the stop decision is made**

At each interim stage \( d = 1, \ldots, K-1 \), the decision to reject or accept \( H_0 \) splits \( \mathcal{T}_{(d)} \) into two non-overlapping regions denoted \( \mathcal{T}^{\text{stop}}_{(d)} \) and \( \mathcal{T}^{\text{cont}}_{(d)} \), respectively. Since \( \mathcal{T}^{\text{stop}}_{(d)} \cap \mathcal{T}^{\text{stop}}_{(d+1)} = \emptyset \) and \( \mathcal{T}^{\text{cont}}_{(d)} \subset \mathcal{T}_{(d+1)} \), then the sets \( \mathcal{T}^{\text{stop}}_{(1)}, \ldots, \mathcal{T}^{\text{stop}}_{(K-1)} \), and \( \mathcal{T}_{(K)} \) make a partition of the sample space of \( T_{(D)} \):

\[
\mathcal{T}_{(K)} + \sum_{k=1}^{K-1} \mathcal{T}^{\text{stop}}_{(k)} = \left( \bigcup_{d=1}^{K-1} \mathcal{T}^{\text{stop}}_{(d)} \right) \cup \mathcal{T}_{(K)} = \bigcup_{d=1}^{K} \mathcal{T}_{(d)} \subset \mathcal{T}.
\]

We define a probability space for the observable random variable \( T_{(d)} \) using the probability measures \( P_{(d)} \) defined by \( f_{\mathcal{T}_{(d)}}^C \) on the measurable space \( (\mathcal{T}^{\text{stop}}_{(d)}, \sigma (\mathcal{T}^{\text{stop}}_{(d)})) \). At \( D = K \), the probability measure \( P_{(K)} \) defined by \( f_{\mathcal{T}_{(K)}}^C \) completes the probability space for a measurable space \((\mathcal{T}_{(K)}, \sigma (\mathcal{T}_{(K)}))\).

Thus, at the end of the study, at \( D = d \), the researcher operates with an observable random variable \( T_{(d)} \).

2.5.3 **At the design stage**

The conditional random variables, \( T_{(d)} \), defined on non-overlapping \( \sigma \)-fields are combined together into the unconditional random variable \( T_{(D)} \sim f_{\mathcal{T}_{(D)}} \), defined on the sample space \( \left( \bigcup_{d=1}^{K} \mathcal{T}_{(d)} \right) \cup \mathcal{T}_{(K)} \). The probability distribution defined on the \( \sigma \)-field on this sample space is \( P_{(D)} = \sum_{d=1}^{K} \Pr (D = d) P_{(d)} \).
2.5.4 Impossible events

The set \( T_0 = T \setminus \cup_{d=1}^{K} \mathcal{T}(d) \) contains all impossible combinations of \((t_1, \ldots, t_K)\) under a chosen stopping rule. If \( K = 2 \), for example, then \( T \setminus \cup_{d=1}^{2} \mathcal{T}(d) = \{(t_1, t_2) : t_1 > c_1\} \). Without the possibility of early stopping, all combinations of \((t_1, \ldots, t_K) \in T\) would be possible.

2.6 Pocock’s Example: One-Sided Two-Group Sequential Z-test

Pocock [41] proposed a simple two-stage procedure for testing \( H_0 : \theta = 0 \) with a pre-determined power at \( H_1 : \theta = \theta_1 \) on normal data. With \( n_1 = n_2 \), \( c_1 = c_2 = 2.18 \) is used to secure the overall type 1 error rate \( \alpha = 0.025 \) with a one-sided \( z \)-test. Let \( n_0 = 0 \) and

\[
Z_k = \frac{1}{\sqrt{n_k}} \sum_{i=n_{(k-1)}+1}^{n_{(k)}} X_i = \sqrt{n_k} \cdot \bar{X}_k \sim \mathcal{N}(\theta, 1), \quad k = 1, 2.
\]

The study is stopped for efficacy at stage 1 if \( Z_1 \geq 2.18 \) and proceeds to stage 2 when \( Z_1 < 2.18 \). If the study is stopped at stage 1, the support for \( \bar{X}_1 \mid Z_1 \geq 2.18 \) starts at \( 2.18 / \sqrt{n_1} \) and stretches to \( +\infty \). If the study continues through stage 2, support for \( \bar{X}_1 \mid Z_1 < 2.18 \) ranges from \(-\infty\) to \( 2.18 / \sqrt{n_1} \).

The test statistic for the second stage, under \( Z_1 < 2.18 \),

\[
Z_{(2)} = \frac{\sqrt{n_1}}{\sqrt{n_1 + n_2}} Z_1 + \frac{\sqrt{n_2}}{\sqrt{n_1 + n_2}} Z_2 = \frac{1}{\sqrt{n_1 + n_2}} \sum_{i=1}^{n_1+n_2} X_i.
\]

The distribution of \( Z_{(2)} \) is a mixture of a right-truncated normal \( Z_1 \mid Z_1 < 2.18 \) and the normal \( Z_2 \).

Figure 3 shows histograms of \( Z_1 \mid Z_1 \geq 2.18 \), \( Z_{(2)} \mid Z_1 < 2.18 \), and \( Z_{(D)} \) estimated from 100,000 Monte-Carlo samples assuming \( \theta = 2.18 \). These histograms based on \( n_1 = n_2 = 100 \) are almost identical to the histograms in Figure 4 based on \( n_1 = n_2 = 1 \). This highlights on the important message that non-normality continues to be present even asymptotically [47].

3 Likelihood-based Inference with EarlyStopping

If the stopping rule is not random and the experiment stopped with \( n_{(d)} \) observations, \( \Pr_{\theta}(D = d) \equiv 1 \), the likelihood is

\[
\mathcal{L}^{fix}(\theta | k, \mathbf{x}_{(d)}) = f_{\mathbf{x}_{(d)}}.
\]

Considering the joint density \( f \) conditional on the data \( (d, \mathbf{x}_{(d)}) \) observed at the end of a sequential experiment, the likelihood is

\[
\mathcal{L}(\theta | d, \mathbf{x}_{(d)}) = f^{sub}_{\mathbf{x}_{(d)}}.
\]
The indicator in (4) emphasizes that support for the random variables is reduced by the conditioning; see this illustrated in Figure 2. Note also in Figure 2 that the support conditional on stopping at one stage is disjoint from the support conditional on stopping at another stage.

$L$ is a function of $\theta$ and the observed data $(d, x_{(d)})$, but $L$ is a continuous function only of $\theta$, as discontinuities in $(d, x_{(d)})$ arise from the mixture distribution of $X_{(D)}$. These discontinuities are inherited by MLEs and other statistics derived from them. Conditional on $(d, x_{(d)})$, MLEs maximizing $L$ are

$$\hat{\theta} = \arg \max_{\theta} f_{X_{(d)}}.$$

For every $(d, x_{(d)})$, $f_{X_{(d)}}^{ub} = f_{X_{(d)}}$ for all $\theta$; consequently,

$$\hat{\theta}^{fix} := \arg \max_{\theta} f_{X_{(d)}} = \arg \max_{\theta} f_{X_{(d)}}^{ub} = \hat{\theta},$$

that is, as is well known, making stopping decisions does not alter maximum likelihood point estimates; they are the same whether obtained by maximizing $L$ or $L^{fix}$ and other observed statistics derived from the likelihood (e.g., the score function and the observed information) are unaffected as well.

The following example extends the simple one in Section 1.1 with $n_1 = n_2 = 1$ to arbitrary $n_1$ and $n_2$ to illustrate (as is proven later) that, although
maximum likelihood point estimates and test statistics are unaffected by early stopping decisions, their probability distribution does not tend to normality even with larger sample sizes.

3.1 Large Sample Properties

The $d$th stage-specific MLEs $\hat{\theta}_d$ of $\theta$ and their statistical models are called regular if, without the possibility of early stopping,

$$\xi_d = \sqrt{n_d} \left( \hat{\theta}_d - \theta \right) \xrightarrow{d} N(0, \sigma^2),$$

where $0 < \sigma < \infty$. Assumption (13) was described in Tarima and Flournoy [47] to include the more specific assumptions for

- independent, identically distributed observations by Cramér [e.g., for example, Ferguson [11]],
- independent not identically distributed observations [e.g., Philippou et al. [39]],
- dependent observations [e.g., Crowder [7]],
- and densities whose support depends on parameters [e.g., Wang et al. [51]].

All these specific sets of assumptions include assumptions of the existence and consistency of the MLE.

With large samples, the MLE at the stage $d$ analysis can be approximated recursively by

$$\hat{\theta}(d) \approx \frac{n(d-1)}{n(d)} \hat{\theta}_{(d-1)} + \frac{n_d}{n(d)} \hat{\theta}_d \approx \sum_{j=1}^{d} \frac{n_j}{n(d)} \hat{\theta}_d,$$

where $\hat{\theta}_{(d-1)}$ is an MLE based on cumulative data from stages 1 to $d-1$ and $\hat{\theta}_d$ is the $d$th stage-specific MLE. The standardized $d$th stage-specific MLE is so,

$$T_d = \sqrt{n(d)} \sum_{j=1}^{d} \frac{n_j}{n(d)} \left( \frac{\hat{\theta}_j - \theta}{\sigma} \right) = \sum_{j=1}^{d} \sqrt{\frac{n_j}{n(d)}} \xi_j \approx \sqrt{n(d)} \left( \frac{\hat{\theta}_d - \theta}{\sigma} \right),$$

that is, $T_d$ is (approximately) the standardized $d$th stage-specific MLE. The asymptotic properties of $T_d(D)$ depend on the existence and distribution of a limiting random variable $r(D)$ defined by

$$\sum_{d=1}^{K} I(D = d) \frac{n(d)}{n_d} \xrightarrow{d} \sum_{d=1}^{K} I(D = d) r(d) = r(D),$$

where $r(d) = \lim_{n_d \to \infty} n(d)/n_d$ is the asymptotic ratio of the $d$th cumulative-stage and stage-specific sample sizes [Theorem 1 in Tarima and Flournoy [47]]; $r(D)$ is a multinomial random variable with support on $r(d)$. 

Assume the limits \( n_d / n_j \to r(d) \in (0, \infty), j \leq d \) exist with \( r(d) = r(d) \). Then given \( D = d \),

\[
T(d) \to \sum_{j=1}^{d} \frac{\xi_j}{\sqrt{r(d)j}}.
\]

While \( \xi_j \xrightarrow{d} \mathcal{N}(0, 1) \), the distributions used for the final analysis, the interim analysis and experimental design, respectively, are mixtures of truncated distributions:

\[
\Pr_{\theta} (T(D) < v | D = d) \to \Pr_{\theta} \left( \sum_{j=1}^{d} \frac{\xi_j}{\sqrt{r(d)j}} < v \big| D = d \right),
\]

\[
\Pr_{\theta} (T(D) < v | D \geq d) \to \sum_{k=d}^{K} \Pr_{\theta} (D = k) \Pr_{\theta} \left( \sum_{j=1}^{k} \frac{\xi_j}{\sqrt{r(k)j}} < v \big| D = k \right),
\]

\[
\Pr_{\theta} (T(D) < v) \to \sum_{k=1}^{K} \Pr_{\theta} (D = k) \Pr_{\theta} \left( \sum_{j=1}^{k} \frac{\xi_j}{\sqrt{r(k)j}} < v \big| D = k \right).
\]

Pocock’s Example: Large Sample Properties

Under assumption (13) with large sample sizes, \( \hat{\bar{\theta}}(D) = I(D = 1)\hat{\bar{\theta}}(1) + I(D = 2)\hat{\bar{\theta}}(2) \) are standardized as

\[
T(D) = \sqrt{n(D)} \left( \hat{\bar{\theta}}(D) - \bar{\theta} \right) / \sigma
\]

\[
= I(D = 1)\sqrt{n_1} \left( \hat{\bar{\theta}}(1) - \bar{\theta} \right) / \sigma + I(D = 2)\sqrt{n_1 + n_2} \left( \hat{\bar{\theta}}(2) - \bar{\theta} \right) / \sigma. \quad (16)
\]

Assume the limit \( I(D = 1) + I(D = 2) \frac{n_1 + n_2}{n_1 \sqrt{n_2}} \xrightarrow{d} r(D) \) exists as \( n_1 \to \infty \). Then, adopting the local alternative hypothesis \( \bar{\theta} = h / \sqrt{\pi} \) yields a non-degenerate limiting distribution of \( T(D) \) that models both stages of the experiment:

\[
\Pr_{\theta} (T(D) < v) \to p_1 \Phi (v | D = 1)
\]

\[
+ (1 - p_1) \int_{-\infty}^{c_1} \Phi (\sqrt{r(D)}v - \sqrt{r(D)} - 1y) \phi (y | D = 2) \, dy \quad (17)
\]

where \( p_1 = \lim_{n_1 \to \infty} \Pr_{\theta} (D = 1) \) is the limiting stage 1 stopping probability; \( \phi \) and \( \Phi \) denote the standard normal density and cumulative distribution function, respectively. With a fixed alternative \( \theta \), \( \Pr_{\theta} (T(D) < v) \to \Phi (v) \) because \( p_1 \to 1 \) with probability 1, and modeling related to stage 2 data is lost.

In Pocock’s example, \( T(D) = I(D = 1)Z_1 + I(D = 2) (Z_1 + Z_2) / \sqrt{2} \) and

\[
\Pr_{\theta} (T(D) < v) \to \Pr_{\theta} (Z_1 > c_1) \Phi (v | Z_1 > c_1)
\]

\[
+ \Pr_{\theta} (Z_1 \leq c_1) \int_{-\infty}^{c_1} \Phi (\sqrt{2}v - y) \frac{\phi (y)}{\Pr (y \leq c_1)} \, dy \quad (18)
\]
Note if $Z_1 \leq c_1$, then
\[
\Pr_{\theta} \left( \frac{Z_1 + Z_2}{\sqrt{2}} < v | Z_1 \leq c_1 \right) = \int_{-\infty}^{c_1} \phi \left( \sqrt{2}v - y \right) \frac{\phi(y)}{\Pr(y \leq c_1)} dy,
\]
which is a continuous mixture of distributions.

3.2 Most Powerful Group Sequential Tests

Let $X \sim f_X(\theta)$, where $f$ belongs to the one-parameter exponential family. Without a possibility of early stopping, the likelihood for a realization $x = (x_1, \ldots, x_n)$ of a random sample $X = (X_1, \ldots, X_n)$ is
\[
L(\theta|x) = \prod_{i=1}^{n} f_X(x) = h(x) g(T(x)|\theta) = h(x) e^{\eta(\theta) T(x) + A(\theta)}
\]
where all relevant information about $\theta$ is absorbed by a sufficient statistic $T(x)$. Assume the test statistic $Z$ is a one-to-one transformation of $T$.

If $LR(t) = g(t|\theta)/g(t|\theta_0)$ has monotone likelihood ratio (MLR) in $t$, then the Karlin-Rubin theorem provides uniformly most powerful (UMP) tests.

In sequential testing settings, when $d$th stage is reached $(D \geq d)$, the interim likelihood
\[
L(\theta|D \geq d, x_{(d)}) = I(D \geq d) h(x_{(d)}) e^{\eta(\theta) T(x_{(d)}) + A_d(\theta)}
\]
has the associated interim likelihood ratio
\[
LR(t|D \geq d) = \frac{L(\theta|D \geq d, x_{(d)})}{L(\theta_0|D \geq d, x_{(d)})}.
\]
For every $d$th stage hypothesis test, $D \geq d$ and
\[
LR(t|D \geq d) = \exp \left[ (\eta(\theta_1) - \eta(\theta_0)) T_{(d)} + (A_d(\theta_1) - A_d(\theta_0)) \right]
\]
which means that the MLR property is preserved with early stopping.

**Definition:** A sequence of $\alpha_d$-level interim tests $\{T_{(1)} > c_1\}, \ldots, \{T_{(D)} > c_D\}$ will be called the sequential test.

**Theorem 1** For any fixed $\{\alpha_1, \ldots, \alpha_K\}$ and $(n_1, \ldots, n_K)$ (1) the interim LR test $\{T_{(d)} > c_d\}$ is a UMP $\alpha_d$-level test and (2) no sequential test is more powerful than the sequential test based interim LRs.

**Proof.** (1) As shown above, conditional on reaching stage $d$, $(D \geq d)$, $T_{(d)}$ continues to be sufficient in exponential families and the MLR property is preserved. By the Karlin-Rubin theorem, the test based on $T_{(d)}$ is uniformly most powerful at $d^{th}$ stage.

(2) Using Table I notation,
\[
1 - \beta(\theta) = \sum_{k=1}^{K} \left( \prod_{j=1}^{k-1} [\beta_j(\theta)] \right) [1 - \beta_k(\theta)] = 1 - \prod_{k=1}^{K} \beta_k(\theta).
\]
At $K = 1$, $\alpha_1$ and $n_1$ uniquely define $c_1$ and $\{T(1) > c_1\}$ is a UMP by part 1 with a stage-specific power curve $1 - \beta_1(\theta)$. At an arbitrary stage $d$, $\alpha_d$ and $n_d$ uniquely define $c_d$ and, by part 1 of the theorem, the stage-specific power $1 - \beta_d(\theta)$, is the highest. Thus, for any choice of $(\alpha_1, \ldots, \alpha_K)$ and $(n_1, \ldots, n_K)$, and consequently $\{T(d) > c_d\}$, the power of the sequential test based on LR is $1 - \beta(\theta) = 1 - \sum_{d=1}^{K} \beta_d(\theta)$. This power is the highest for any given $\theta$, because the stage specific type 2 errors $\beta_d(\theta)$ are the lowest for each $d$ at any $\theta$. Q.E.D.

Pocock’s Example: Likelihood Ratio

For $X_i \sim \mathcal{N}(\theta, 1)$, $i = 1, \ldots, n$, the likelihood ratio for testing $H_0 : \theta = \theta_0$ vs $H_A : \theta = \theta_1 > \theta_0$ is

$$\log LR(t) = -n (\log \theta_0 - \log \theta_1) - t/\theta_1,$$

which has rejection region $\{t > c\}$, where $t = \sum_{i=1}^{n} X_i$ is a sufficient statistic. Using $Z = t/\sqrt{n} \sim f_Z = \mathcal{N}(\theta_1, 1)$, the likelihood ratio test is $\{Z > c/\sqrt{n}\}$.

With Pocock’s example, under $D \geq 1$, $Z_1$ is a normal random variable used for stage 1 hypothesis testing, and

$$\log LR(Z_1|D \geq 1) = -n_1 (\log \theta_0 - \log \theta_1) - \sqrt{n_1} Z_1/\theta_1.$$

Then,

$$\log LR(Z_{(2)}|D \geq 2) = -(n_1 + n_2) (\log \theta_0 - \log \theta_1) - \sqrt{n_1 + n_2} Z_{(2)}/\theta_1 + C,$$

where $C = \log \Pr(Z_1 < c_1|\theta_0) - \log \Pr(Z_1 < c_1|\theta_1)$ is independent of data given $D \geq 2$. By Theorem, $\{Z_{(d)} > c_d\}$ is the UMP test at stage $d$ given $n_d$ and $\alpha_d$, and no other sequential test is more powerful overall.

### Table 1

| Feature             | Mathematical Definition |
|---------------------|------------------------|
| overall type 1 error| $\alpha = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{d-1} \Pr_0(Z_{(j)} \leq u_j)$ |
| stage-specific type 1 error | $\alpha_d = \Pr_0\left(Z_{(d)} > u_d \cap \bigcap_{j=1}^{d-1} \{Z_{(j)} \leq u_j\}\right)$ |
| $\alpha$-spending function | $\alpha(d) = \sum_{d=1}^{K} \alpha_j \prod_{j=1}^{d-1} \Pr_0(Z_{(j)} \leq u_j)$ |
| overall type 2 error | $\beta = \sum_{j=1}^{n} \beta_j(\theta) \prod_{j=1}^{d-1} \Pr_0(Z_{(j)} \leq u_j)$ |
| stage-specific type 2 error | $\beta_d(\theta) = \Pr_0\left(Z_{(d)} > u_d \cap \bigcap_{j=1}^{d-1} \{Z_{(j)} \leq u_j\}\right)$ |
| $\beta$-spending function | $\beta(d)(\theta) = \sum_{d=1}^{K} \beta_j(\theta) \prod_{j=1}^{d-1} \Pr_0(Z_{(j)} \leq u_j)$ |
| overall power       | $1 - \beta(\theta)$ |
| stage-specific power | $1 - \beta_d(\theta)$ |
| cumulative power    | $1 - \beta(d)(\theta)$ |
4 Impact and Summary

To establish that a sequence of likelihood ratio tests are most powerful, we began by constructing the joint probability distribution over the set of possible events. The use of an early stopping criterion eliminates the possibility of some realizations of cumulative test statistics \( Z = (Z_1, \ldots, Z_K) \). This makes an otherwise normal random process \( Z \) unobservable. On its true stopping rule adapted support, the distribution of \( Z \) is a mixture of truncated distributions at each stage. Thus, unadapted distributional assumptions should be modified to take into account the planned adaptation scheme. Liu and Hall [29] and Liu et al. [30] recognized change in support in the one-parameter exponential family and investigated bias estimation. Schou and Marschner [44] recognized presence of truncation in the joint distribution of stage-specific test statistics, but they mostly focused on bias in meta-analytic studies. The adapted support is critical for derivation MLEs’ and test statistics’ distributions in Section 2.

In Section 2 distinct probability measures are derived for design [unconditional], interim hypothesis testing [conditional on collecting data up to the time of interim testing] and when the study is completed [conditional on deciding to stop at a particular stage]. These probability distributions formalized a new probabilistic framework (Section 2) which is used in Section 3.2 to show no testing sequence is more powerful than sequential likelihood ratio tests.

Likelihood ratio tests are most powerful for testing simple hypotheses and, under monotone likelihood ratio, for testing composite hypotheses; see Neyman and Pearson [35] and Karlin-Rubin theorem in Ferguson [12]. The Karlin-Rubin theorem works with the one-parameter exponential family and Section 3.2 shows that most powerful testing continues to hold for sequential tests. Even though distributions of sample-size-dependent sufficient statistics (Blackwell [5]) belong to a curved exponential family (Efron [10], Liu and Hall [29], Liu et al. [30]), the distributions conditional on the sample size still belong to one-parameter exponential family. This fact is used in Section 3.2 to show that likelihood ratio sequential tests are most powerful tests with any predetermined \( \alpha \)-spending function. This result is applicable to many common group sequential designs.

Treatment-effect-dependent (non-ancillary) stopping rules are part of many common GSDs including Pocock [41], O’Brien & Fleming [37], and Haybittle-Peto designs [18, 38]. If the data follow a normal distribution with known variance (without possibility of early stopping), then these designs are most powerful for their \( \alpha \)-spending functions. Similarly, application of Simes test (see Simes [45]) and its recently proposed modification for sequential testing (see Tamhane et al. [46]) cannot have higher power than group sequential tests from Jennison and Turnbull [21] with the same \( \alpha \)-spending function.

Historically, many researchers have relied, and currently rely, on joint normality (e.g., Jennison and Turnbull [21], Proschan et al. [42], Kunz et al. [24]). Critical adjustments are often done with recursive sub-density estimation; see Armitage et al. [1]. Section 2.4.1 places Armitage’s sub-density formula in the broader probability framework of possible events and confirms,
from the prospective of this $\sigma$-algebra, that the common current practice of using Armitage’s formula for calculating distributions of interim test statistics is appropriate.

There is plenty of evidence that adaptive designs make statistics non-normally distributed. Demets and Lan [8] point out that the distribution of stage-specific test statistics is not normal and should be estimated recursively. Jennison and Turnbull [21] plot the density of a normal test statistic used in GSD settings [pages 174-177], where discontinuity points clearly show non-normality. Li et al. [27] find the joint density of stage 1 and stage 2 standardized test statistics not to be bivariate normal. Local asymptotic non-normality was established following sample size recalculations (SSRs) that depend on an interim observed treatment effect (Tarima and Flournoy [47, 48]); and a GSD with a single interim analysis can be viewed as a special case of an SSR. MLEs converging to random mixtures of normal variables have been found in other adaptive designs (Ivanova et al. [20], Ivanova and Flournoy [19], May and Flournoy [32], Lane and Flournoy [25], Flournoy et al. [14]).

Milanzi et al. [33] developed a likelihood approach that applies when the early stopping rule does not depend on the parameter of interest. In this case, sample size adaptation is ancillary to the treatment effect and asymptotic normality of MLEs holds. Gnedenko and Korolev [16], Bening et al. [4], Christoph et al. [6], Korolev and Zeifman [23] assume that the distribution of the random sample size does not depend on previously collected data; Molenberghs et al. [34] assumes this asymptotically.

Nevertheless, convergence of sample means to non-normal random variables was shown even for ancillary random sample sizes. In [16], Gnedenko and Korolev show convergence of standardized sums with random number of summands of infinitely divisible random variables to mixtures of stationary distributions. They give conditions for convergence to a mixture of normal distributions. Bening et al. [4] and Christoph et al. [6] explore convergence to mixtures of normal distributions and to Student’s limit distribution. When convergence is mixed (see, for example Häusler and Luschgy [17]), Lin et al. [28] shows how norming with the observed information can result in a normal limit. However, the requirement for mixed convergence appears strong, and it does not cover the limiting mixtures obtained in this paper.

The impact of early stopping is pervasive. It affects the probabilistic characterisation of the tests (e.g., type I error and Fisher Information) as well as the distributions of MLEs and test statistics. Its effect on Fisher information, when stopped at different stages, is not widely recognized.

During the design phase, before observations are taken, the full form of the joint density [5], accounting for all possible events, is appropriate. This contrasts with the current practice of using the density assuming the experiment will continue through stage $K$. More details on the differences resulting from these two design approaches will be the subject of another paper.

However, normality and asymptotic normality assumptions continue to be directly used with non-normally distributed statistics. We identify two main reasons for this.
1. Many researchers consider large sample properties against a fixed treatment effect independent of the sample size. From one point of view, a treatment effect is a population quantity which does not change with sample size. But if one develops an asymptotic approximation to the testing environment using a fixed treatment effect, the statistical experiment stops at the first interim analysis with probability one for any consistent test; test statistics degenerate to a point mass; see Section 7.4 in Fleming and Harrington [13]. Under a fixed treatment effect, the power converges to one and cannot be used to compare different testing procedures. This issue triggered development of various descriptions of asymptotic relative efficiency. The most popular approach is Pitman asymptotic relative efficiency [40], where asymptotic power is evaluated under local alternatives; see Nikitin [36]. In addition, local alternatives clearly reflect actual practice for experiment planning. Experiments are never planned for a statistical power = 1. Small sample size studies (pre-clinical, animal studies) are planned to detect large effect sizes, moderate sample sizes (typical phase 3 studies) are used to detect moderate effect sizes, and large sample sizes (epidemiological studies, like vaccine studies) are used to detect small differences. Koopmeiners et al. [22] explored MLEs conditional on stopping, but assumed asymptotic normality to evaluate their uncertainty. Martens and Logan [31] relied on asymptotic normality for evaluating regression coefficients under the Fine–Gray model in GSD settings. Asendorf et al. [2] evaluated asymptotic properties with SSR under a fixed alternative for negative binomial random variables.

2. Some researchers investigate local asymptotic properties when early stopping is not possible or is ancillary to the treatment effect. This also leads to asymptotic normality. Scharfstein et al. [43] show that without possibility of early stopping, “time-sequential joint distributions of many statistics … are multivariate normal with an independent increments covariance structure” under local alternatives. These results are generally consistent with the classical results on local asymptotic normality of Le Cam [26], where both mean and variance of the limiting normal distribution depend on the parameter of interest; see Chapter 7 of van der Vaart [49]. However, Section 3.1 shows that the possibility of early stopping destroys local asymptotic normality; the limiting distribution of standardized test statistics is a mixture of truncated normal distributions. Similar findings were previously proved for non-ancillary sample size recalculations; see Tarima and Flournoy [47].

Gao et al. [15] is a rare exception in not making a normality assumption; these authors mostly deal with set operations and probabilities and, using stage-wise ordering of events, they calculate P-values, confidence intervals, and a median unbiased estimate of the parameter of interest.

It is recommended that the full form of the joint density [5], accounting for all possible events, be used for study design before observations are taken.
This contrasts with the current practice of using the density assuming the experiment will continue through stage $K$.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Armitage, P., C. McPherson, and B. Rowe (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A* 132, 235–244.
2. Asendorf, T., R. Henderson, H. Schmidli, and T. Friede (2019). Sample size re-estimation for clinical trials with longitudinal negative binomial counts including time trends: Sample size re-estimation with longitudinal negative binomial counts. *Statistics in Medicine* 38(9), 1503–1528.
3. Bartky, W. (1943). Multiple sampling with constant probability. *The Annals of Mathematical Statistics* 14, 363–377.
4. Bening, V., N. Galieva, and R. Korolev (2012). On rate of convergence in distribution of asymptotically normal statistics based on samples of random size. In *Annales Mathematicae et Informaticae*, pp. 17–28.
5. Blackwell, D. (1947, 03). Conditional expectation and unbiased sequential estimation. *Ann. Math. Statist.* 18(1), 105–110.
6. Christoph, G., M. M. Monakhov, and V. V. Ulyanov (2020). Second-order chebyshev–edgeworth and cornish–fisher expansions for distributions of statistics constructed from samples with random sizes. *Journal of Mathematical Sciences*, 1–29.
7. Crowder, M. J. (1976). Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society. Series B (Methodological)* 38(1), 45–53.
8. Demets, D. L. and K. K. G. Lan (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine* 13(13–14), 1341–1352.
9. Dodge, H. and H. Romig (1929). A method of sampling inspection. *The Bell System Technical Journal* 8, 613–631.
10. Efron, B. (1975, 11). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* 3(6), 1189–1242.
11. Ferguson, T. (1996). *A Course in Large Sample Theory*. New York: Routledge.
12. Ferguson, T. S. (2014). *Mathematical statistics: A decision theoretic approach*, Volume 1. Academic press.
13. Fleming, T. R. and D. P. Harrington (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons.
14. Flournoy, N., C. May, and C. Tommasi (2018). The effects of adaptation on inference for non-linear regression models with normal errors. *arXiv preprint arXiv:1812.03970*. 
15. Gao, P., L. Liu, and C. Mehta (2013). Exact inference for adaptive group sequential designs. *Statistics in medicine* 32(23), 3991–4005.
16. Gnedenko, B. V. and V. Y. Korolev (1996). *Random summation: limit theorems and applications*. CRC press.
17. Häusler, E. and H. Luschgy (2015). *Stable convergence and stable limit theorems*, Volume 74. Springer.
18. Haybittle, J. (1971). Repeated assessment of results in clinical trials of cancer treatment. *The British journal of radiology* 44(526), 793–797.
19. Ivanova, A. and N. Flournoy (2001). *A birth and death urn for ternary outcomes: stochastic processes applied to urn models*. Chapman and Hall/CRC, Boca Raton.
20. Ivanova, A., W. F. Rosenberger, S. D. Durham, and N. Flournoy (2000). A birth and death urn for randomized clinical trials: Asymptotic methods. *Sankhy, Series B* 62(1), 104–118.
21. Jennison, C. and B. Turnbull (1999). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press.
22. Koopmeiners, J. S., Z. Feng, and M. S. Pepe (2012). Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. *Statistics in Medicine* 31(5), 420–435.
23. Korolev, V. and A. Zeifman (2019). From asymptotic normality to heavy-tailedness via limit theorems for random sums and statistics with random sample sizes. In *Probability, Combinatorics and Control*, pp. XX–XX. IntechOpen.
24. Kunz, C. U., S. Jörgens, F. Bretz, N. Stallard, K. V. Lancker, D. Xi, S. Zohar, C. Gerlinger, and T. Friede (2020). Clinical trials impacted by the covid-19 pandemic: Adaptive designs to the rescue? *Statistics in Biopharmaceutical Research* 0(0), 1–17.
25. Lane, A. and N. Flournoy (2012). Two-stage adaptive optimal design with fixed first-stage sample size. *Journal of Probability and Statistics* 2012, n/a–n/a.
26. Le Cam, L. M. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* 3, 37–98.
27. Li, G., W. J. Shih, T. Xie, and J. Lu (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* 3(2), 277–287.
28. Lin, Z., N. Flournoy, and W. F. Rosenberger (2020). Random norming aids analysis of non-linear regression models with sequential informative dose selection. *Journal of Statistical Planning and Inference* 206, 29–42.
29. Liu, A. and W. Hall (1999). Unbiased estimation following a group sequential test. *Biometrika* 86(1), 71–78.
30. Liu, A., W. Hall, K. F. Yu, and C. Wu (2006). Estimation following a group sequential test for distributions in the one-parameter exponential family. *Statistica Sinica* 16(1), 165–181.
31. Martens, M. J. and B. R. Logan (2018). A group sequential test for treatment effect based on the Fine–Gray model. *Biometrics* 74(3), 1006–
1013.

32. May, C. and N. Flournoy (2009). Asymptotics in response-adaptive designs generated by a two-color, randomly reinforced urn. *The Annals of Statistics* 37(2), 1058–1078.

33. Milanzi, E., G. Molenberghs, A. Alonso, M. G. Kenward, A. A. Tsiatis, M. Davidian, and G. Verbeke (2015). Estimation after a group sequential trial. *Statistics in Biosciences* 7(2), 187–205.

34. Molenberghs, G., M. G. Kenward, M. Aerts, G. Verbeke, A. A. Tsiatis, M. Davidian, and D. Rizopoulos (2012). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy; sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research* 21, 11–41.

35. Neyman, J. and E. S. Pearson (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231(694-706), 289–337.

36. Nikitin, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press.

37. O’Brien, P. C. and T. R. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics*, 549–556.

38. Peto, R., M. Pike, P. Armitage, N. Breslow, D. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto, and P. Smith (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* 34(6), 585–612.

39. Philippou, A. N., G. Roussas, et al. (1973). Asymptotic distribution of the likelihood function in the independent not identically distributed case. *The Annals of Statistics* 1(3), 454–471.

40. Pitman, E. (1948). *Lecture Notes on Nonparametric Statistics*. New York: Columbia University.

41. Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64(2), 191–199.

42. Proschan, M. A., K. K. G. Lan, and J. T. Wittes (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer.

43. Scharfstein, D. O., A. A. Tsiatis, and J. M. Robins (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* 92(440), 1342–1350.

44. Schou, I. M. and I. C. Marschner (2013). Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Statistics in Medicine* 32(28), 4859–4874.

45. Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika* 73(3), 751–754.

46. Tamhane, A. C., J. Gou, and A. Dmitrienko (2020). Some drawbacks of the simes test in the group sequential setting. *Statistics in Biopharmaceutical Research* 12(3), 390–393.

47. Tarima, S. and N. Flournoy (2019a). Asymptotic properties of maximum likelihood estimators with sample size recalculation. *Statistical Papers* 60,
23–44.
48. Tarima, S. and N. Flournoy (2019b). Distribution theory following blinded and unblinded sample size re-estimation under parametric models. *Communications in Statistics - Simulation and Computation* 0(0), 1–12.
49. van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.
50. Wald, A. (1947). *Sequential Analysis*. New York: Wiley.
51. Wang, H., N. Flournoy, and E. Kpamegan (2014). A new bounded log-linear regression model. *Metrika* 77(5), 695–720.
52. Wetherill, G. and K. Glazebrook (1986). *Sequential Methods in Statistics* (3 ed.). New York: Chapman and Hall.