Screening for Dispersion Effects by Sequential Bifurcation

BRUCE E. ANKENMAN, Northwestern University
RUSSELL C.H. CHENG, University of Southampton
SUSAN M. LEWIS, University of Southampton

The mean of the output of interest obtained from a run of a computer simulation model of a system or process often depends on many factors; many times however only a few of these factors are important. Sequential bifurcation is a method that has been considered by several authors for identifying these important factors using as few runs of the simulation model as possible. In this paper, we propose a new sequential bifurcation procedure whose steps use a key stopping rule that can be calculated explicitly, something not available in the best methods previously considered. Moreover we show how this stopping rule can also be easily modified to efficiently identify those factors that are important in influencing the variability rather than the mean of the output. In empirical studies, the new method performs better than previously published fully sequential bifurcation methods in terms of achieving the prescribed Type I error. It also achieves higher power for detecting moderately large effects using fewer replications than earlier methods. To achieve this control for midrange effects, the new method sometimes requires more replications than other methods in the case where there are many very large effects.

Additional Key Words and Phrases: Stopping rules; Controlled sequential bifurcation; Simulation, Multiplicative variance model, Fixed width confidence intervals

1. INTRODUCTION

We consider use of the linear statistical model in the exploratory study of system performance when this performance depends on a large number of explanatory factors but only a few of these factors are likely to have an important effect. The system could be anything, but recent literature has shown that the linear model can be particularly helpful in early stage screening for important factors in simulation models of physical or organizational systems such as manufacturing processes or logistical networks.

We assume performance to be measured as a univariate output response Y, obtained from runs of a simulation model, and the problem is to identify those factors which have a substantial effect using as few runs as possible. We shall call this the *location effects problem*.

For many systems the direction of influence of each factor is known. For example, in a manufacturing system, two factors affecting the output rate of the system are the speed of operation of a machine and the number of machines available. Increasing these factors can be expected to increase output rate. Thus the direction of influence of their effects on the output of interest are known.

When the direction of influence of each factor is known, the method of sequential bifurcation (SB) is a simple but effective way of making simulation runs step by step in a way that enables elimination of non-influential factors not just one at a time but

Author's addresses: B.E. Ankenman, Department of Industrial Engineering, Northwestern University, Evanston, IL 60044; R.C.H. Cheng and S.M. Lewis, Mathematics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

often in large groups simultaneously. The full set of factors is systematically divided into smaller and smaller subsets, each of which either does not contain any factors whose effect is important or still contains one or more factors whose effect is important. In the first case, all the factors in the subset can be eliminated as being unimportant. In the second case, unless the subset contains only one effect, it is further subdivided.

The original SB method was proposed and studied in Bettonvil and Kleijnen [1997] for the deterministic case. Cheng [1997] gave a version for the stochastic case. More general versions for the stochastic case, notably the controlled sequential bifurcation (CSB) method given by Wan et al. [2006] and an improvement of this called CSB-X given in Wan et al. [2010] both of which control the power of detecting important effects at each step and the overall Type I error, all under heterogeneous variance conditions. In subsequent work, Wan and Ankenman [2007], Shen and Wan [2009], Sanchez et al. [2009], Shen et al. [2010] incorporate fuller exploratory refinements such as the inclusion of fractional factorial designs (FFDs). Although such augmented approaches are more comprehensive and are a useful addition, they arguably lose some of the essential simplicity of SB which remains attractive for initial exploratory work.

In summary, CSB-X appears to be the most attractive published pure SB method to date for handling the location effects problem. This paper focuses on the pure SB approach and makes the following three contributions to the methodology:

(i) We propose an alternative method which we shall call the *Anscombe Fully Sequential Bifurcation Method* (AFSB). This works in a similar way to CSB-X in that both methods use a *sequential stopping rule* to decide how many additional simulation runs to add at each step of the algorithm. The main difference is that the rule used in AFSB is based on a result by Anscombe [1953], which is *explicit* and so is more easily calculated than the one used in CSB-X which requires the simultaneous solution of a pair of double-integral equations. Anscombe's result is an asymptotic one which is however second order accurate compared with the well known stopping rule discussed by Chow and Robbins [1965], which is only first order accurate. Both the Chow and Robbins' result and Anscombe's result can only be directly applied to construct *symmetric* confidence intervals (CIs). We include theoretical results showing how the results of both Chow and Robbins and Anscombe can be extended to *non-symmetric* CIs. We believe our results are new, and they underpin our proposed procedures at least asymptotically.

(ii) We shall call the problem of identifying those factors that are important in influencing, not the mean, but the *variability* of the output, the *dispersion effects problem*. It is easy to modify the AFSB method to handle this problem. Interestingly the resulting algorithm takes a more explicit and simpler form than that for the location effects problem. We again need to know the direction of influence of each factor, only now on the variability. For instance, in the previously mentioned manufacturing example, we might be able to assume that an improved output rate will also lead to increased variability in the output, so that an increase in the speed of operation of a machine, or in the number of machines, will both lead to an increase in output variability. If the experimenter is uncomfortable with this assumption, the methods proposed here could be set in a more general procedure like the one discussed in Shen et al. [2010].

(iii) We shall show that the CSB-X algorithm, as set out in Wan et al. [2010], contains some redundancy in that the runs specified to be added at each step may not actually all be required in applying the stopping rule. The power or Type I error properties of the method are not altered by the removal of this redundancy, which simply concerns the way additional runs are added, and is independent of the stopping rule used. This accelerated use of the stopping rule is implemented in the AFSB algorithm.

The "X" in CSB-X refers to the fact that the method incorporates a foldover method for eliminating bias in the estimation of main effects when interaction effects are present. To simplify exposition, we have not included foldover in our formulation of AFSB, though it could be incorporated in just the same way. In comparisons, to avoid confusion, we shall refer to CSB-X when implemented without foldover as the CFSB (short for Controlled Fully Sequential Bifurcation) method.

We should point out here that the statistical model being investigated is a very basic one involving a linear model with normal errors and only two levels for each independent variable. This model however is the one used by previous authors with whose work we wish to compare the methods presented in this paper. Moreover the methods that we discuss are intended for and are suited to exploratory work. Once factors deemed important have been identified, additional work using more complex models may then be needed. For example if a factor needs to be explored at additional levels, this could be handled by treating it as multiple factors each having just two levels.

In Section 2 we set out the SB process in detail and introduce the linear model to be used in the statistical analysis, We also describe the AFSB method as applied to the location effects problem. In Section 3 we show how the AFSB method can be modified to tackle the dispersion effects problem.

Some simple examples are given in Section 4 showing the AFSB method provides reasonable control, for both the location and dispersion effects problem, of the two types of statistical error and to give readily interpretable results. A realistic example with 92 factors is also discussed. Some final comments are set out in Section 5.

2. THE SB METHOD

In this section we set out the SB algorithm in detail. The algorithm operates by carrying out a step by step partition of the set of all factors into successively smaller subsets. This process is recursive. At each step a subset of the current partition is selected at random and examined. The result of the examination determines the importance or otherwise of the factors in the subset as a whole and also how to carry out the next step of the partition. We begin with a more detailed description of the overall algorithm.

2.1. SB Algorithm

The algorithm partitions the factors into subsets. The order of factors is fixed throughout all the steps of the partitioning process. For simplicity a factor will be labeled by its index position j in the ordered list $\{j = 1, ..., k\}$. This order is assumed to be arbitrary, but as noted by previous authors, if prior information enables the factors to be labeled in either increasing or decreasing order of importance, the efficiency of the SB method will be improved. The increased efficiency comes when small effects are grouped together in the subsets. In numerical investigations not shown in detail in this paper, we found that when factors are put in decreasing order, the number of design points and observations can be reduced by 20-30%. Since proper ordering requires prior knowledge of the effect sizes, factors will be put in arbitrary order for the numerical examples in this paper.

As with all sequential bifurcation methods that involve random error, the order of the factors and the method of partitioning will affect the specific factors that are declared important or not important. The properties of SB that relate to the probability of Type I and Type II error hold in general but do not guarantee that the procedure is invariant to factor order or partitioning strategy.

Each subset formed in the partitioning process is made up of a number of contiguous factors in the ordered list. We denote the subset of factors $j = k_0, k_0 + 1, ..., k_1$ by $\{k_0, k_1\}$ where $k_0 \leq k_1$. At each step of the partitioning process a subset is selected at random, $\{k_0, k_1\}$ say, and examined using a routine $E(k_0, k_1)$, to be specified later.

The following specification of the SB algorithm focuses on how the subset partitioning is carried out, showing the way $E(k_0, k_1)$ appears in this process.

SB Algorithm (Partitioning)
Step 0. Let S be the set of subsets into which the k factors are partitioned. Initialize S to contain the one subset $\{1, k\}$.
Step 1. Remove a subset from S and denote it by $\{k_0, k_1\}$. (Thus when this step is carried out for the first time, $\{k_0 = 1, k_1 = k\}$.) Examine this subset using $E(k_0, k_1)$, a routine that includes making additional simulation runs and which returns a value when it is finished. There are three possible outcomes:

If $E(k_0, k_1) = 0$ all the factors $j = k_0, k_0 + 1, ..., k_1$ are classified as unimportant and eliminated from all further examination.

If $E(k_0, k_1) = 1$ and $k_0 = k_1$, then there is only one factor k_0, in the subset and this is classified as important and thus requires no additional testing.

If $E(k_0, k_1) = 1$ and $k_0 < k_1$, then $\{k_0, k_1\}$ is split into two subsets $\{k_0, k'\}$ and $\{k' + 1, k_1\}$. Different authors use different rules for setting k'. For example, Kleijnen et al. [2006] always choose a k' to make the number of factors in the set $\{k_0, k'\}$ to be equal to a power of two. We will use $k' = \lfloor (k_0 + k_1)/2 \rfloor$ so that the two subsets are of the same size if $k_0 + k_1$ is odd. These two subsets are then placed into the set of subsets, S.
Step 2. The process is repeated from Step 1 until there are no subsets left in S.
At the end of the algorithm each factor will have been classified as either important or unimportant.

Set out in this way it can be seen that the partitioning of subsets is distinct from the precise form that the routine $E(\cdot)$ takes. We can therefore discuss the routine quite separately but before doing so we need to describe our assumptions about the form of the observations obtained from simulation runs in order to understand how the routine makes additional simulation runs and classifies factors.

2.2. Statistical Model of Simulation Output

The observations used in the routine $E(\cdot)$ are assumed to be of the form

$$\begin{aligned} Y_i &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i & (1) \\ &= \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i(\mathbf{x}_i) \quad i = 1, 2, ..., n & (2) \end{aligned}$$

when n simulation runs have been made. Here Y_i is the observed response in the ith run; there are k factors, and x_{ij} is the level of factor j in the ith run; β_0 is a general mean; and β_j, $j = 1, 2, ..., k$, are the unknown factor location effects. In the vector form we have written $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)^T$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})^T$.

The $\varepsilon_i(\mathbf{x}_i)$ are observational errors. We assume that they are normally distributed, $N(0, \sigma^2(\mathbf{x}_i))$, and that the standard deviation takes the form

$$\log \sigma(\mathbf{x}_i) = \gamma_0 + \sum_{j=1}^{k} \gamma_j x_{ij} = \gamma_0 + \boldsymbol{\gamma}^T \mathbf{x}_i \tag{3}$$

with γ_j, $j = 1, 2, ..., k$ unknown factor dispersion effects.

For the location effects problem our key assumption is that we know the direction of each factor effect. We can therefore, if necessary with a change of sign to x_j, assume with no loss of generality that

$$\beta_j \geq 0, \quad j = 1, 2, ..., k. \tag{4}$$

Similarly, when considering the dispersion effects problem, our key assumption will be, if necessary with a change of sign to x_j, that

$$\gamma_j \geq 0, \quad j = 1, 2, ..., k. \tag{5}$$

To be clear, in this paper we will address the dispersion effect problem and the location effects problem separately and not attempt to simultaneously identify both location and dispersion effects with the same set of tests. This allows x_j to have a different sign for location effects testing than it has for dispersion effects testing and thus both (4) and (5) can hold. Although there might be substantial computational efficiency in attempting to tackle both of these problems at the same time, we leave this for future research.

We shall refer to the combination of factor levels used in a given run as a *design point*. We assume, without loss of generality, that each factor is applied at one of two levels scaled to be 0 and $+1$.

As in most SB procedures in the recent literature, the only design points we use take the form $\mathbf{x}(l)$, $0 \leq l \leq k$, with components

$$\begin{aligned} x_j &= 1 \text{ if } 1 \leq j \leq l \\ &= 0 \text{ if } l < j \leq k \end{aligned}$$

with the understanding that $\mathbf{x}(0)$ has all components $x_j = 0$, $j = 1, 2, ..., k$ and $\mathbf{x}(k)$ has all components $x_j = 1$, $j = 1, 2, ..., k$. It will be convenient to refer to such a design point as *design point* l. In Ankenman et al. [2006], design points of other forms are considered for the case of deterministic responses.

The use of design points of the type $\mathbf{x}(l)$ enables a simple estimate of the sum

$$\mu(k_0, k_1) = \sum_{j=k_0}^{k_1} \beta_j$$

to be constructed by taking observations in pairs, one at $\mathbf{x}(k_1)$

$$\begin{aligned} Y(k_1) &= \beta_0 + \boldsymbol{\beta}^T \mathbf{x}(k_1) + \varepsilon[\mathbf{x}(k_1)] \\ &= \beta_0 + \sum_{j=1}^{k_1} \beta_j + \varepsilon[\mathbf{x}(k_1)] \end{aligned}$$

and one at $\mathbf{x}(k_0 - 1)$

$$\begin{aligned} Y(k_0 - 1) &= \beta_0 + \boldsymbol{\beta}^T \mathbf{x}(k_0 - 1) + \varepsilon[\mathbf{x}(k_0 - 1)] \\ &= \beta_0 + \sum_{j=1}^{k_0 - 1} \beta_j + \varepsilon[\mathbf{x}(k_0 - 1)]. \end{aligned}$$

The difference is

$$\begin{aligned} d(k_0, k_1) &= Y(k_1) - Y(k_0 - 1) = \sum_{j=k_0}^{k_1} \beta_j + \varepsilon(k_0, k_1) \\ &= \mu(k_0, k_1) + \varepsilon(k_0, k_1) \end{aligned} \tag{6}$$

where we have written

$$\varepsilon(k_0, k_1) = \varepsilon[\mathbf{x}(k_1)] - \varepsilon[\mathbf{x}(k_0 - 1)],$$

which is clearly normal as it is a linear combination of normally distributed random variables. The variance of $\varepsilon(k_0, k_1)$ depends only on $k_0, k_1,$ and so can be written as $\sigma^2(k_0, k_1)$.

Writing $d_i(k_0, k_1), i = 1, 2, ..., n,$ for the differences, each of the form in (6), of n pairs of observations, and their average difference

$$D_n(k_0, k_1) = n^{-1}\sum_{i=1}^{n} d_i(k_0, k_1) = \mu(k_0, k_1) + n^{-1}\sum_{i=1}^{n} \varepsilon_i(k_0, k_1) \tag{7}$$

estimates $\mu(k_0, k_1)$ with increasing accuracy as n increases. All the terms β_j contributing to $\mu(k_0, k_1)$ are non-negative so that if $D_n(k_0, k_1)$ is small we can say that none of the factors $k_0,\ k_0 + 1, ..., k_1$ are important. If $D_n(k_0, k_1)$ is large we can say that at least one is important, but the important factor can only be identified if $k_0 = k_1$. From (7), $D_n(k_0, k_1)$ is also normally distributed since its error is an average of independent normally distributed errors. Its variance is a function of σ^2 and is treated as known or estimated depending on whether σ^2 is known or estimated.

We can now describe the routine $E(k_0, k_1)$.

2.3. The Routine $E(k_0, k_1)$

In our proposed method, AFSB, the routine $E(k_0, k_1)$ applied to a selected subset $\{k_0, k_1\}$ operates by making additional simulation runs at either or both the design points $k_0 - 1$ and k_1. After an additional run is made, a call is made to the function $StopTest(\cdot)$. This has two parts. First, it uses a *stopping rule*, to be described later, to decide whether to stop making additional runs. If the decision is to add a further run, the $StopTest$ function returns a value of -1.

If no further runs are to be made, the second part of $StopTest(\cdot)$ is applied, which carries out an (statistical) *importance test*, also described later, to declare if the subset of factors under consideration is unimportant or (possibly) important, returning respectively the value 0 or 1. The routine $E(\cdot)$ is exited at this point and the function $E(\cdot)$ takes on the value of $StopTest$.

The following formal description of $E(\cdot)$ focuses on how runs are added in $E(\cdot)$, and where $StopTest(\cdot)$ is applied.

In just this section alone, to avoid having to use sub-subscripts we write $L = k_0 - 1$ and $U = k_1$. Let n_L and n_U be the number of observations already made at design points L and U on entry into $E(k_0, k_1)$. The value m is the minimum number of observations required at both design points when applying the stop rule to allow a variance estimate to be calculated. Based on a large numerical study to be described we have set $m = 5$ in the numerical examples given later. The precise form of $E(k_0, k_1)$ is as follows.

```
Function $E(k_0, k_1)$
  [$L \equiv k_0 - 1, U \equiv k_1$]
  $n = 0$
  $E = -1$
  Do {
    Increment $n$
    If($n_L < n \leq n_U$){Add a run at $L$; Increment $n_L$}                                   E(1)
    ElseIf($n_U < n \leq n_L$){Add a run at $U$; Increment $n_U$}                               E(2)
    ElseIf($n > n_L$){Add a run at $L$ and at $U$; Increment both $n_L$ and $n_U$}              E(3)
    If($n \geq m$ [and $n_L = n_U = n$]){$E = StopTest(n, k_0, k_1)$}                            E(4)
  }While($E = -1$)
  Return $E$
```

In the Do loop, n is increased one step at a time from $n = 1$. Steps E(1), E(2) and E(3) add observations at whichever design point, L or U, has fewer observations, until the number of observations at both are equal, after which observations are added to both levels. This way of adding observations means that, as n is increased, there are always at least n observations at each of the design points L and U.

Our method, AFSB, is exactly the SB Algorithm of the previous section incorporating $E(\cdot)$ but *excluding* the conditions set out in the square brackets of step E(4). $StopTest$, which appears in E(4), will therefore be called immediately when n reaches the minimum value m, and will then be called for every $n \geq m$. Note that it is possible that n_L and n_U are already both non-zero when a set $\{k_0, k_1\}$ is first examined because previous examination of adjoining sets have already resulted in observations being made at L and U. In such cases, not requiring that $n = n_L = n_U$ means we do not require $n \geq \max(n_L, n_U)$ before the stopping rule is applied. This can be more efficient, as we can stop with $m \leq n < \max(n_L, n_U)$ or even with $m \leq n < \min(n_L, n_U)$. In these latter cases only the first n observations at L and at U are used.

The CFSB method as given in (Wan et al. [2010], Figure 3) adds runs in a similar way to $E(k_0, k_1)$, but with a *different* $StopTest$, and with a step E(4) that *includes* the conditions in square brackets. Thus n has to satisfy $n \geq \max(m, n_L, n_U)$ before the $StopTest$ calculation is called, a less efficient condition than in the AFSB method, as it sometimes will require an n to be reached that is greater than needed before any testing starts. Numerical experimentation suggests that a saving of 15% - 25% in computational effort is typically obtained using the more efficient, but equally effective AFSB condition. Moreover, numerical results fully reported below indicate the satisfactory practical performance of AFSB with regard to size and power.

We are now in a position to discuss the $StopTest$ function used in the AFSB method.

2.4. Stopping Rules using Fixed Width Confidence Intervals

The first function of $StopTest$ is to decide if further runs are needed in the examination of the current subset of factors. This uses a modification of a well known approach to the sequential construction of a CI of given width w and level of confidence. The stopping rule is simply to stop making additional runs once a preselected CI width is obtained. For simplicity of notation we suppress the dependence of quantities on k_0 and k_1 and write D_n for $D_n(k_0, k_1)$, μ for $\mu(k_0, k_1)$, and so on. Since all the effects are positive and we are testing for large effects, the upper limit of the CI will be primarily driven by the specified probability of Type II error, β. This follows because we will only declare effects unimportant if the upper limit of the CI is below a user specified threshold of importance. Whereas the lower limit of the CI will be primarily driven by the specified probability of Type I error, α. This is because we only misidentify an unimportant effect as important when the lower limit of the CI is greater than a user specified threshold of unimportance. This dependence structure often leads to an asymmetric CI. (Note that here we have used β without a subscript to denote the probability of making a Type II error, as is conventional. This should not be confused with the subscripted β_j used to denote coefficients in the model (1) of the observations.)

Let the CI be of form $(D_n - w_1, D_n + w_2)$ where $w_1, w_2 > 0$. Conditioned on n, we can obtain such a CI from the probability statement

$$\Pr(z_\beta \leq \sqrt{n}(D_n - \mu)/\sigma \leq z_{1-\alpha}) = 1 - \alpha - \beta \tag{8}$$

where $\sqrt{n}(D_n - \mu)/\sigma$ is the standardised form of D_n so that it is distributed as a $N(0, 1)$ variable, and where z_β and $z_{1-\alpha}$ are the β and $1-\alpha$ quantiles of the $N(0, 1)$ distribution.

We shall assume that

$$0 < \alpha,\ \beta < 1/2 \tag{9}$$

so that $z_\beta < 0$ and $z_{1-\alpha} > 0$. The probability statement in (8) can be written as

$$\Pr(D_n - w_1 \le \mu \le D_n + w_2) = 1 - \alpha - \beta$$

where we set

$$w_1 = \sigma z_{1-\alpha}/\sqrt{n} \text{ and } w_2 = -\sigma z_\beta/\sqrt{n}. \tag{10}$$

Both are positive under the assumption (9).

If the width of the interval is to be

$$w = w_1 + w_2 \tag{11}$$

we must have

$$w_1 = \frac{z_{1-\alpha}}{z_{1-\alpha} - z_\beta} w \text{ and } w_2 = -\frac{z_\beta}{z_{1-\alpha} - z_\beta} w. \tag{12}$$

Moreover replacing w_1 and w_2 in (11) by their expressions in (10) and then squaring gives

$$n^{-1}\sigma^2 = \frac{w^2}{(z_{1-\alpha} - z_\beta)^2}. \tag{13}$$

The values of α and β are user selected. Thus if σ^2 is known, a CI of given width w is obtained simply by selecting n to satisfy (13). We call this the *Known Sigma Stopping Rule.*

As n is integer, rounding may be needed, but conventionally the smallest n is taken for which the left-hand side of (13) is less than or equal to the right-hand side to ensure an actual width no greater than w is achieved with least effort. We shall do this with all the stopping rules we consider.

Consider now the situation where σ^2 is not known. The sample variance,

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - D_n)^2 \tag{14}$$

can now be used as a first order approximation for σ^2. To obtain a given width w when σ is known, n must satisfy (13). When σ is not known a simple way to determine n is to take an initial sample of m observations, with $m \ge 2$, then:

$$\text{Take } N \text{ as the first } n \ \ge m \text{ for which } n^{-1}s_n^2 \le \frac{w^2}{(z_{1-\alpha} - z_\beta)^2}. \tag{15}$$

We shall call this the *basic stopping rule*. Here N is a random variable. For the case $\alpha = \beta$, Chow and Robbins [1965] show that selecting N in this way is asymptotically consistent in the sense defined in Lemma 1 below. However Lemma 1 shows that (15) is also asymptotically consistent when $\alpha \ne \beta$.

Lemma 1. Let

$$J_n = [D_n - w_1,\ D_n + w_2] \tag{16}$$

where w_1 and w_2 are as defined in (12). Let N be as defined in (15). If $0 < \sigma^2 < \infty$ then

$$\lim_{w\to 0} \Pr(\mu \in J_N) = 1 - \alpha - \beta \quad \text{(asymptotic consistency).}$$

Proof. This is given in the Online Supplement.

The original Chow and Robbins [1965] result and our extension in the Lemma are of course just asymptotic ones becoming precise only as $w \to 0$. We offer them partly as a reassurance that our methods have some theoretical foundation if only in the limit, in the same way that other asymptotic results are used in the literature, for example in evaluating maximum likelihood statistics. However small sample properties of the distribution of N have been discussed quite thoroughly by Starr [1966b] and Starr [1966a] and Woodroofe [1977] for the case $\alpha = \beta$. In particular Starr [1966b] showed that use of N gives reasonably accurate coverage, and that it is more efficient than the well-known two-stage procedure proposed by Stein [1945]. The effect of (15) is to produce somewhat narrow CIs so that the coverage is less than the nominal value of $(1-2\alpha)$. Given the simplicity of (15), its use in SB might be entertained, at least when this use is exploratory, though some caution needs to be exercised. When (15) is used as the stopping rule in the routine $E(\cdot,\cdot)$ of the SB algorithm, we shall call this the *basic fully sequential bifurcation* (BFSB) method.

Anscombe [1953] discusses second order accurate versions of (15) which we now consider. The method is not often mentioned but one might expect it to be more accurate than (15). It is as easy to implement.

Anscombe [1953] discusses only the case $\alpha = \beta$. We extend this to where α and β are not necessarily equal using a stopping rule that is a generalization of that given in his equation (A2.32); here and with other equations we refer to from Anscombe's paper we have added an 'A' to the equation number to indicate its source.

The stopping rules considered by Anscombe require calculation of the sum of squares in the formula for the sample variance (14). We follow Anscombe's method of doing this by using the $d_1, d_2, ...d_n$ of (7) to construct independent (actually $\sigma\chi_1^2$) variates

$$U_i = \frac{1}{i(i+1)}\{id_{i+1} - \sum_{j=1}^{i} d_j\}^2 \quad i = 1, 2, ..., n-1, \tag{17}$$

so that $\sum_{i=1}^{n-1} U_i = \sum_{i=1}^{n}(d_i - D_n)^2$. Thus, if further observations $d_{n+1}, d_{n+2}, ...$ are observed, the corresponding $U_n, U_{n+1}, ...$can be added without changing the previous U_i.

Equation (A2.32), the general form of Anscombe's stopping rule, is to take an initial sample of m observations, with $m \geq 2$, then:

$$\text{Take } N \text{ as the first } n \ \geq m \text{ for which } s_{n-1} = \sum_{i=1}^{n-1} U_i \leq cn^B(1 - \frac{b}{n} + O(n^{-2})) \tag{18}$$

where b and c are quantities depending on α, β and w, to be determined, and $B > 1$ is a constant to be selected. The version of (18) given by Anscombe actually uses β rather than B. In anticipation of the result, we shall set $B = 2$ in all that follows; this simplification avoids confusion with our use of β as a probability used in defining the CI.

Anscombe studies the behavior of the stopping rule (18) in terms of ν where

$$\nu = \frac{\sigma^2}{c} + b - 1. \tag{19}$$

Here ν has the simple interpretation of being σ^2/c, the number of observations needed to achieve a CI of given width when σ^2 is known, to which has been added the constant $b-1$. Anscombe shows that under certain assumptions (set out in the Online Supplement), if

$$b = 2 + g + \frac{1}{2}(z_{1-\alpha}^2) \text{ and } c = \frac{w^2}{4z_{1-\alpha}^2}$$

where g is a certain waiting time constant, then the stopping rule (18) gives a symmetric CI, $I_n = [D_n - w/2, D_n + w/2]$, of width w. Moreover we have $\nu = 4z_{1-\alpha}^2(\sigma/w)^2 + b - 1$, so that $\nu \to \infty$ as $w \to 0$. The CI then satisfies

$$\Pr(\mu \in I_N) = 1 - 2\alpha + O(\nu^{-3/2}) \tag{20}$$

so that it is asymptotically second order accurate.

The value of c ensures the interval width is w with a coverage of $(1 - 2\alpha)$ that is correct to first order, i.e. the error is $O(\nu^{-1})$, whilst the value of b ensures the result is correct to second order. With only a single parameter b available for this fine tuning, it is clear that the result (20) is a special case, achievable because a symmetric CI allows a particular choice of b to yield second order accuracy for both the probability of μ being above the CI and of being below it. When $\alpha \neq \beta$ we cannot achieve second order accuracy for both probabilities individually with a single stopping rule of the form (18) (though it may be possible to estimate the two probabilities separately using a single sequential procedure that stops only when both probabilities are obtained to second order accuracy; but we have not considered this). We can however achieve second order accuracy for just one of the probabilities, or for a combination of both, as the following Proposition shows.

Proposition. Suppose Assumptions 1, 2 and 3 given in Anscombe [1953] p.10 hold (the assumptions are set out in detail in the Proof of the Proposition in the Online supplement), with p in Assumption 3 set equal to $p = 3/2$, and with Anscombe's β, which we call B, set to $B = 2$. For the stopping rule (18) let $b = 2 + g + (\tau/2)$, where τ is a function of α and β that we use to generalize Anscombe's result. Also, let $c = w^2/(z_{1-\alpha} - z_\beta)^2$ and ν be as defined in (19), that is $\nu = (z_{1-\alpha} - z_\beta)^2(\sigma/w)^2 + b - 1$. Let α and β be fixed and let τ be a quantity which we can choose and which can depend on α and β. Let $\sigma/w \to \infty$ so that $\nu \to \infty$:
(i) If $\tau = z_{1-\alpha}^2$, then

$$\Pr(D_N - \mu \leq w_1) = 1 - \alpha + O(\nu^{-3/2}),$$

and

$$\Pr(-w_2 \leq D_N - \mu \leq w_1) = 1 - \alpha - \beta + O(\nu^{-1}).$$

(ii) If $\tau = z_\beta^2$, then

$$\Pr(D_N - \mu \leq -w_2) = \beta + O(\nu^{-3/2})$$

and

$$\Pr(-w_2 \leq D_N - \mu \leq w_1) = 1 - \alpha - \beta + O(\nu^{-1}).$$

(iii) If $\tau = [z_{1-\alpha}\varphi(z_{1-\alpha}) - z_\beta\varphi(z_\beta)]^{-1}[z_{1-\alpha}^3\varphi(z_{1-\alpha}) - z_\beta^3\varphi(z_\beta)] = \tau_0$, say, then

$$\Pr(-w_2 \leq D_N - \mu \leq w_1) = 1 - \alpha - \beta + O(\nu^{-3/2}).$$

(iv) For any other (finite) value of τ, we have

$$\Pr(-w_2 \leq D_N - \mu \leq w_1) = 1 - \alpha - \beta + O(\nu^{-1}).$$

We have moreover (as $\sigma/w \to \infty$)

$$E(N) = \frac{(z_{1-\alpha} - z_\beta)^2\sigma^2}{w^2} + \frac{1+\tau}{2} + O(\nu^{-1/2}). \tag{21}$$

Proof. This is given in the Online Supplement.

In AFSB, we use the Proposition with $\tau = \tau_0$ so that the stopping rule is:

Take N as the first n $(\geq m)$ for which

$$\{n[n - 2.676 - (\tau_0/2)]\}^{-1} \sum_{i=1}^{n-1} U_i \leq \frac{w^2}{(z_{1-\alpha} - z_\beta)^2}. \tag{22}$$

We can make an immediate comparison of the rule (22) with (15) the lefthand side of which can be written as $n(n-1)^{-1} \sum_{i=1}^{n-1} U_i$. Thus, certainly when τ_0 is positive, the lefthand side of (22) is larger than that of (15) by a factor of $(n-1)/[n - 2.676 - (\tau_0/2)]$ for all n. This means that a larger value of N is always needed by (22); i.e. though the CI is of fixed width, (22) always requires μ to be estimated more accurately than (15) in forming the interval.

The rule (22) requires a CI width of w to be specified. We set $w = \Delta_1 - \Delta_0$, where (Δ_0, Δ_1), with $0 < \Delta_0 < \Delta_1$, is a user-specified interval. In AFSB, Δ_0 is a threshold of unimportance (i.e. all effects that are less than Δ_0 are unimportant) and Δ_1 is a threshold of importance (i.e. all effects greater than Δ_1 are critically important and thus should be identified with high probability). Effects that are between Δ_0 and Δ_1 are effects that are not negligible, but are small enough that declaring them insignificant would not be highly detrimental. Clearly both Δ_0 and Δ_1 should be chosen by a user who is familiar with the system under test and can assess the impact of ignoring effects that fall between Δ_0 and Δ_1. The thresholds Δ_0 and Δ_1 are used in the importance test that is applied once the stopping rule (22) is satisfied, and we discuss this next.

2.5. The Importance Test

We turn now to the importance test applied in $StopTest$ once the stopping rule (22) is satisfied. Once we have stopped adding further runs, we can calculate w_1 and w_2 as given in (12) to obtain the upper and lower limits of the CI as

$$C_U = D_n - w(z_{1-\alpha} - z_\beta)^{-1} z_\beta \tag{23}$$

and

$$C_L = D_n - w(z_{1-\alpha} - z_\beta)^{-1} z_{1-\alpha}, \tag{24}$$

so that $C_U - C_L = \Delta_1 - \Delta_0$ as required. We must therefore have:

Either $C_U \leq \Delta_1$ in which case we set $E = 0$ and eliminate all the factors k_0, $k_0 + 1, ..., k_1$ as being unimportant;

Or Else $C_L \geq \Delta_0$ in which case we set $E = 1$ to indicate that some of the factors k_0, $k_0 + 1, ..., k_1$, might be important.

The value E is kept at its entry value of $E = -1$ whilst n is incremented until (22) is satisfied. Summarizing, $StopTest$ operates as follows:

$$StopTest(n, k_0, k_1) = \begin{cases} -1 & \text{if (22) is not satisfied} \\ 0 & \text{if (22) is satisfied and } C_U \leq \Delta_1 \\ 1 & \text{if (22) is satisfied and } C_L \geq \Delta_0 \end{cases}. \tag{25}$$

This completes the description of the AFSB algorithm, the whole being summarised in the specifications of (i) the SB Algorithm (Partitioning), (ii) the routine $E(k_0, k_1)$ and (iii) $StopTest(n, k_0, k_1)$. We note five properties of the full process.

(1) At any given step, the process guards against eliminating a factor, j say, when $\beta_j \geq \Delta_1$. From the form of C_U we find that, if $j \in \{k_0, k_1\}$, then

$$\Pr(C_U \leq \Delta_1 | \beta_j \geq \Delta_1) \leq \beta. \tag{26}$$

Thus at any step of the process we can be asymptotically $(1-\beta)100\%$ confident that β_j will *not* be found unimportant if $\beta_j \geq \Delta_1$. Thus Δ_1 is a threshold of importance, and the importance test ensures there is small probability of elimination of factors if $\beta_j \geq \Delta_1$. When just one coefficient β_j is being considered we see that (26) is just a statement that the *power* of the test is $(1-\beta)$ at $\beta_j = \Delta_1$.
(2) Likewise, at any given step, the test protects against any factor j being classified as important when $\beta_j \leq \Delta_0$. To see this note that we never finally classify a factor as being important until it is being considered individually (i.e. when the examination routine is of the form $E(j,j)$). In this case we find that

$$\Pr(C_L \geq \Delta_0 | \beta_j \leq \Delta_0) \leq \alpha. \tag{27}$$

Thus at any step of the SB process we can be asymptotically $(1-\alpha)100\%$ confident that β_j will not be found important if $\beta_j \leq \Delta_0$. Thus Δ_0 is a threshold of unimportance, with the test ensuring there is small probability of classifying a factor as important if $\beta_j \leq \Delta_0$. When just one coefficient β_j is being considered we see that (27) is just a statement that the *size* of the test is α at $\beta_j = \Delta_0$.
(3) Any factor found important will automatically have been located to within a CI of prescribed width $w = \Delta_1 - \Delta_0$. Thus the width determines the accuracy to which coefficients found important are determined.
(4) If the difference $\Delta_1 - \Delta_0$ is small relative to their magnitudes, we are then close to the limiting case where $\Delta_1 = \Delta_0$ with this common value acting as a dividing point below which a factor is deemed unimportant and above which it is considered important. A small w does of course result in potentially greatly increased computational effort.
(5) Though not shown in (14) for notational simplicity, the calculation of the sample variance (14) is specific to the subset $\{k_0,\, k_1\}$ being examined. The AFSB algorithm thus automatically handles the heterogeneous variance assumed in the model.

2.6. Numerical Properties of the Anscombe Stopping Rule and Test

The accuracy of the stopping rule and importance test relies on the asymptotic accuracy of the Proposition. We have not tried to assess theoretically the magnitude of the $O(\nu^{-r}), r = 1/2$, 1, 3/2 remainder terms appearing in the expressions given in the Proposition. Instead we have studied the behavior of the stopping rule and importance test numerically, estimating the actual size and power of the test as parameters are varied. Our numerical study is sufficiently detailed to allow us, in the same spirit as in the work of Singham and Schruben [2012] who discuss how to obtain symmetric CIs with accurate coverage for the case $\alpha = \beta$, to provide specific guidelines on how to choose w and m in practice so that the target size and power settings of the importance test are more than met. We note that AFSB is invariant to shifts in the values of Δ_0 and Δ_1 when $w = \Delta_1 - \Delta_0$ is held constant, so that we need only consider w and not Δ_0 and Δ_1 separately.

Figure 1 in Singham and Schruben [2012] shows that coverage accuracy improves as $w \to 0$, as expected from Lemma 1, with the limit being approached from below. The figure also shows that if $w \to \infty$ with $\alpha = \beta$ held constant, or that if $\alpha = \beta \to 0$ with w held constant, then the coverage becomes greater than the target value. Our own numerical study corroborates this behavior.

We report results for three test sizes: $\alpha = 0.01, 0.05, 0.1$ and for the power levels $(1-\beta) = 0.8, 0.9, 0.95$. The behavior of (22) depends on w and σ only through the ratio σ/w. We have therefore, with no loss of generality, set $\sigma = 1$ to examine the numerical behavior of both the rule and the test for w in the range $[0.05, 5]$ and for $m = 3, 5, 7, 10, 20$. These values cover the range of values likely to be of practical use. Of special interest are the values of m as this governs the initial effort needed in setting up the sequential analysis. In terms of overall effort needed to carry out a full SB analysis we would wish to set m as small as possible while ensuring that the accuracy of the rule is maintained in terms of its size and power. For each combination of α, β, w and m values we applied the rule to $100,000$ pairs of sequences $\{d_i \sim N(0,1), i = 1, 2, ...\}$ and $\{d'_i \sim N(w,1), i = 1, 2, ...\}$ recording how many times the final test declared the result important. The normal distributions of d_i and d'_i correspond to the case $\Delta_0 = 0$ and $\Delta_1 = w$. There is no loss of generality in setting $\Delta_0 = 0$ as the test, being translationally invariant, depends only on the difference between Δ_0 and Δ_1, and not on their absolute values. When the mean of the d'_i is equal to w, the proportion of times the test returned an important result in the $\{d_i\}$ and the $\{d'_i\}$ sequences respectively estimates α, the size of the test, and $(1 - \beta)$, the power of the test.

The estimated size and power of the test and the average value of N are plotted as functions of w in Figure 1 for the nine combinations of $(\alpha, (1 - \beta))$ values considered, for the case $m = 5$. The cases $m = 3, 7, 10, 20$ are given in the Online Supplement. In all these figures we have, for clarity, restricted the range of w plotted to focus on those observed values of size and power relatively close to their target values.

The results are particularly satisfactory as far as the power is concerned. For all combinations of (α, β, m) examined experimentally: (i) the observed power was almost exactly the target value for the smaller values of w, with the observed power increasing as w increases; (ii) the observed size did take values above the target value but only in a relative small interval of the w range around $w \simeq 0.5$, thereafter decreasing as w increased. For the cases $\alpha = 0.1/0.05/0.01$ the largest observed size over all β and w, and for all $m \geq 5$, was $0.106/0.056/0.014$.

However account has also to be taken of the fact that for small w values the size of $E(N)$ becomes impractically large. Our recommendation is, where possible, to set $m = 5$, together with choice of w restricted to $w \geq 0.7/0.8/1.0$ according as $\alpha = 0.1/0.05/0.01$. Doing this ensures, for all three α values, that the estimated actual size is always less than the target value, and that the estimated actual power is always greater than the target power. $E(N)$ is then between 11 and 19, with in all cases, irrespective of β, $E(N)$ falling to 5 or 6 as w increases to $w = 5$.

Fuller details are reported in the Online Supplement. Thus, for example, if values of w less than $w = 0.7$, are needed, we see from Figures 2 and 3 given in the Supplement that use should not be made of $\alpha = 0.01$. Using either $\alpha = 0.1$ or $\alpha = 0.05$ and setting $m \geq 7$ allows any value of $w \geq 0.1$ to be used whilst keeping the test size less than 10% over its target value for all the three β values we have considered. In allowing this latitude over test size we remark that, in SB, our emphasis is on correctly identifying important factors, so ensuring that the actual size is no larger than the target size of the test is perhaps not quite so important as ensuring that the power is at least as great as the target power. Note that we have not recommended use of $w = 0.05$ as N is impractically large in this case. Plots of how N varies for small N are shown in Figure 4 in the Online Supplement.

The behavior of the stopping rule governs the worst case performance of SB process as a whole in terms of the overall number of observations required. SB is intended for use in situations there are only a small number, k say, of factors that are important, or at least that are not unimportant. It is quite easy to see that the number of runs needed by SB to identify these factors will be $O(k)$, that is, roughly Mk where M is a factor not depending on k. However M will depend on w/σ; in fact M will likely be comparable to N, so that a rough estimate of the likely number of observations needed for the SB process as a whole would be $kE(N)$. Our numerical results, given here and in the Online Supplement, highlight the importance of not choosing w too small relative to σ if $E(N)$ is not to be impractically large. Thus in choosing w, it will be necessary to have some initial idea of what the value of σ should be. It may therefore be necessary to carry out preliminary runs to obtain an estimate of σ before choosing w.

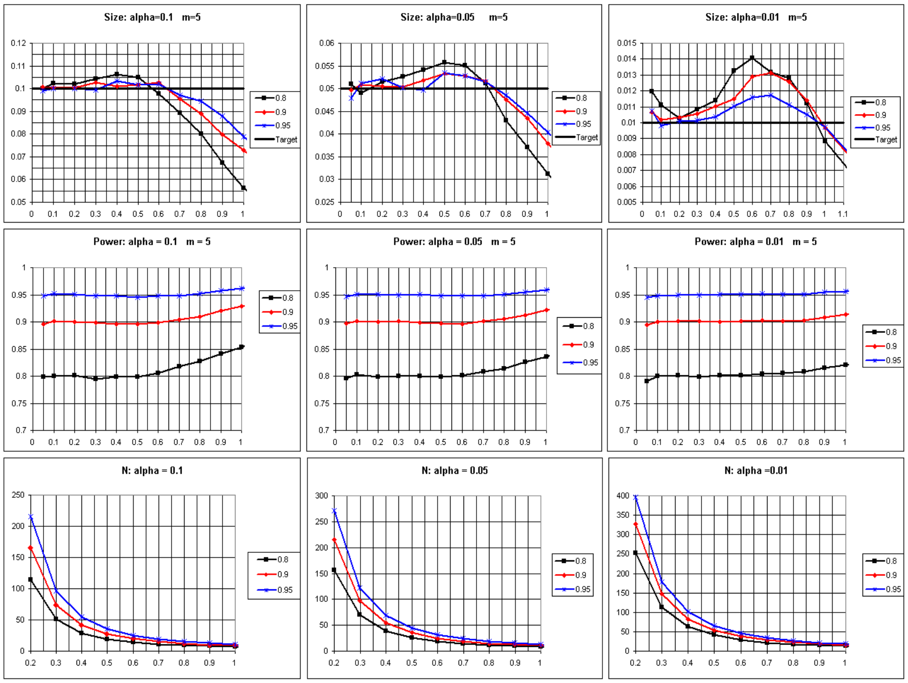

Fig. 1. Observed size, power and N as functions of w for the stopping rule (22) and importance test of equation (25). The three lines in each plot correspond to the three power values. All the size/power graphs respectively decrease/increase monotonically as w increases past 1 off the plotted area.

We mention here a possible alternative using the procedure P1 given by Liu [1997]. Like Anscombe's original procedure, this is applied to a sequence Y_i, $i = 1, 2, ...$ of independent $N(\mu, \sigma^2)$ variables to construct a symmetric two-sided $(1 - \alpha)100\%$ CI of given width $w = 2d$ for μ, but with given power $(1 - \beta)$ (in Liu [1997] this power is denoted by β, but we retain our use of β here). The procedure uses a stopping rule to construct the CI, followed by a two-sided test of whether to reject the hypothesis H_0: $\mu = 0$. The test has power $(1-\beta)$ in the sense that the probability of rejection is at least $(1-\beta)$ if $|\mu| \geq d$. Procedure P1,unmodified, can actually replace our stopping rule (22) and importance test (25), however, being a two-sided procedure it is not efficient in our application, requiring significantly more runs. It should be possible in principle to modify P1 so that it is one sided, but the details do not look straightforward, especially as Theorem 4.1 in Woodroofe [1977] on which P1 depends, is only proved in outline in Woodroofe [1977] with some key details not explained. We have therefore not studied this possibility further as this would require a separate article to handle properly. However we have carried out some preliminary investigation of how P1 might be modified, which is reported in the Online Supplement.

3. THE DISPERSION EFFECTS PROBLEM

The AFSB algorithm may be modified to address the problem of estimating dispersion effects instead of locations effects. The SB Algorithm (Partitioning) and the function subroutine $E(k_0, k_1)$ have exactly the same form as in the location effects problem. The only change is in $StopTest(\cdot)$. This *still* takes the form (25). However the calculation of C_L, C_U and the stopping rule are modified in a way which we now describe.

In function $E(k_0, k_1)$ we again obtain n pairs of observations at points $\mathbf{x}(k_0 - 1)$ and $\mathbf{x}(k_1)$, namely,

$$Y_i[\mathbf{x}(k_0 - 1)] \;\; i = 1, 2, ..., n \quad \text{and} \quad Y_i[\mathbf{x}(k_1)] \;\; i = 1, 2, ..., n$$

separately, to form corresponding sequences

$$V_i[\mathbf{x}(k)] = \frac{1}{i(i+1)}\{iY_{i+1}[\mathbf{x}(k)] - \sum_{j=1}^{i} Y_j[\mathbf{x}(k)]\}^2 \;\; i = 1, 2, ..., n-1, \;\; k = k_0 - 1, \; k_1.$$

These expressions are simply the independent components of the sample variances of the two sequences when decomposed (using Helmert's transformation) into a form convenient for updating as observations are added. Assuming the observations take the form (1), the V_i are distributed as

$$V_i[\mathbf{x}(k)] \sim \sigma_i^2[\mathbf{x}(k)]C_{ik}, \quad k = k_0 - 1, k_1$$

where the C_{i,k_0-1} and C_{ik_1} are mutually independent chi-squared variates, χ_1^2, with one degree of freedom. The sample versions of the logged standard deviations in (3) are thus and

$$0.5 \log V_i[\mathbf{x}(k)] = \gamma_0 + \sum_{j=1}^{k} \gamma_j + \xi_{ik} \quad i = 1, 2, ..., n-1, \;\; k = k_0 - 1, \; k_1 \tag{28}$$

where the ξ_{i,k_0-1}, ξ_{ik_1} $i = 1, 2, ..., n-1$, are all mutually independent variables each distributed as a $\frac{1}{2}\log(\chi_1^2)$ variable.

The sequence

$$\begin{aligned} h_i(k_0, k_1) &= 0.5\{\log V_i[\mathbf{x}(k_1)] - \log V_i[\mathbf{x}(k_0 - 1)]\} \\ &= \sum_{j=k_0}^{k_1} \gamma_j + \delta_i \\ &= \lambda(k_0, k_1) + \delta_i, \quad i = 1, 2, ..., n-1 \end{aligned}$$

is the dispersion effects equivalent of the sequence $d_i(k_0, k_1)$ appearing in (7) that was used to test the importance of the combined factor location effect $\mu(k_0, k_1)$. We can therefore use, in direct analogue to D_n,

$$\begin{aligned} H_n &= \frac{1}{n-1}\sum_{i=1}^{n-1} h_i(k_0, k_1) \\ &= \lambda(k_0, k_1) + \frac{1}{n-1}\sum_{i=1}^{n-1} \delta_i \end{aligned} \tag{29}$$

to test the importance of $\lambda(k_0, k_1)$ in the factor dispersion effects problem.

A remarkable feature of H_n is that the δ_i have distributions which are completely known. As already noted, they are independently and identically distributed, with each being the difference between two independent $\frac{1}{2}\log(\chi_1^2)$ variables. The density of each δ_i is therefore given by the convolution

$$f_\delta(x) = \int_{-\infty}^{\infty} g(x+u)g(u)du$$

where $g(x) = (2/\pi)^{1/2}e^x e^{-e^{2x}/2}$ is the density of a $\frac{1}{2}\log(\chi_1^2)$ variable. This reduces to the (symmetric about 0) form

$$f_\delta(x) = \frac{2e^{-x}}{\pi(1+e^{-2x})},$$

with variance $Var(\delta) = \pi^2/4$. This density and the variance are derived in the Online Supplement.

Convergence of the sum $\sum \delta_i$ to normality is rapid as the δ_i are symmetric with finite first and second moments. We show, in the Online Supplement, that the CDF of the sum of just two δ_i matches the normal CDF with a maximum difference that is less than 0.001. In the calculation of the fixed width CI we can therefore use (13), (with standard normal quantiles z_β and $z_{1-\alpha}$) to fix the number of observations needed as:

$$\text{Stop at } n = \frac{\pi^2(z_{1-\alpha} - z_\beta)^2}{4w^2} + 1. \tag{30}$$

We have added the unity term to account for the fact that W_n is the average of only $(n-1)$ w_i when the number of original Y observations is n. The formula in (30) is the dispersion effects equivalent of the known sigma stopping rule (13) for the location effects case.

The statistical model (1) with errors given by (3) is not the only model that can be handled by the methods considered in this paper. Any multiplicative model can be written in a logarithmic form that is amenable to a similar analysis. For example if instead of the standard deviation, the variance is assumed to be

$$\log \sigma^2(\mathbf{x}_i) = \gamma_0 + \sum_{j=1}^{k} \gamma_j x_{ij} = \gamma_0 + \boldsymbol{\gamma}^T \mathbf{x}_i$$

then the only change is that (30) becomes

$$n = \frac{\pi^2(z_{1-\alpha} - z_\beta)^2}{w^2} + 1.$$

If there is uncertainty about the model it might be safer to assume that the variance of the δ_i is not known. In this case we can form a sample variance sequence from the δ_i in direct analogy to the sample variance sequence U_i given in (17), i.e.

$$V_i = \frac{1}{i(i+1)}\{ih_{i+1} - \sum_{j=1}^{i} h_j\}^2 \quad i = 1, 2, ..., n-2$$

and, as in (22), we can apply the stopping rule:

Take as N the first $n \geq m$ for which

$$\{n[n - 2.676 - (\tau_0/2)]\}^{-1} \sum_{i=1}^{n-2} V_i \leq \frac{w^2}{(z_{1-\alpha} - z_\beta)^2}. \tag{31}$$

In the Online Supplement we present the results of a numerical study showing that the stopping rule (31) and the size and power of the corresponding importance test behave almost identically to those of the stopping rule (22) and its corresponding test, so that our recommendations on how to select w and m can still be applied in the dispersion case. Fuller details are given in the Online Supplement.

Summarising, the only changes needed for AFSB to handle the dispersion effects problem is to replace D_n by H_n as given in (29) in the calculation of the CI limits (23) and (24), and to replace the stopping rule (22) by (30) or possibly (31).

We compare the use of (30) and (31) in the next section.

4. NUMERICAL EXAMPLES

In this section, we will use numerical examples to demonstrate the performance of the various algorithms that have been presented.

For the location effects problem, we will compare four algorithms in order to compare the effects of (1) acceleration (i.e. excluding the condition in the square brackets of step E(4) of $E(\cdot)$); (2) the use of the basic stopping criterion in (15); and (3) the use of Anscombe‘s stopping rule in (22). Algorithm performance is measured by the ability of an algorithm to hold the probability of Type I error to the specified level of 0.05, while producing high power to identify effects that are larger than $\Delta_0 = 2$ and very high power $(1 - \beta)$ for identifying effects that are larger than $\Delta_1 = 4$. We will use three experimental trials, each focusing a different range of factor sizes. The first trial uses effect sizes from 0 to 6 to build a sense of the power across a wide range. The second trial replicates the experimental trials by Wan et al. [2010] with factor levels between 2 and 6 with many of the factors larger than $\Delta_1 = 4$. The third trial uses relatively small but non-negligible factor effects between 2 and 4. We will then discuss the performance of the location effect algorithms on a well-known Ericsson simulation example from the literature.

For the dispersion effects problem, we use a 32 factor example to test the performance of our proposed algorithm and we compare the use of the known sigma stopping rule in (30) with the unknown sigma stopping rule of (31). We also test the performance of our algorithm as the order of factors is changed since the order of factors is known to affect the efficiency of SB procedures. Finally, we will then compare the performance of our dispersion effect algorithms with an example based on the Ericsson simulation.

4.1. The Location Effects Problem

For the location effects problem we shall compare four algorithms:

(i) The Original CFSB (OrigCFSB) of Wan et al. [2010], i.e. CSB-X but without the foldover. The conditions in square brackets in step E(4) of $E(\cdot)$ are therefore included.

(ii) Accelerated CFSB (AccelCFSB). This is CFSB as in (i), but with the conditions in square brackets in step E(4) of $E(\cdot)$ omitted.

(iii) Accelerated basic fully sequential bifurcation (AccelBFSB). This is simply the SB algorithm using the basic stopping rule (15), again with the conditions in square brackets in step E(4) of $E(\cdot)$ omitted.

(iv) Accelerated AFSB (AccelAFSB). This uses the Anscombe stopping rule (22), again with the conditions in square brackets in step E(4) of $E(\cdot)$ omitted.

We report the results of three experimental trials. All involve 10 factors like those given in Table 2 of Wan et al. [2010]. In all three trials, a set of metaexperiments were carried out, where each metaexperiment comprised 1000 identical experiments with observations artificially generated as in (1).

4.1.1. Experimental Trial 1: Identifying a wide range of effects. The same 9 metaexperiments for each of the four FSB methods were carried out. In each metaexperiment the values of all 10 factor coefficients were the same in all the experiments, but with this common value changing between metaexperiments. Specifically in metaexperiment i: $i = 1, 2, ..., 9$, $\beta_j = \beta^{(i)}$, $j = 1, 2, ..., 10$. The nine $\beta^{(i)}$ values were 0.0, 1.0, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 6.0. The errors were heteroscedastic with $\sigma(\boldsymbol{\gamma}, \mathbf{X}) = \sum_{i=1}^{10} \gamma_j X_j$, where $\boldsymbol{\gamma} = (0.0, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0)$ and $\mathbf{X}$ are the factor levels.

Table I. Trial 1: Proportion found important out of 1000 observations for each metaexperiment.

Metaexperiment	Coef. Size	OrigCFSB	AccelCFSB	AccelBFSB	AccelAFSB
1	0	0.00	0.00	0.00	0.00
2	1	0.00	0.00	0.00	0.00
3	2	0.01	0.04	0.05	0.04
4	2.5	0.05	0.11	0.19	0.18
5	3	0.26	0.34	0.43	0.43
6	3.5	0.79	0.73	0.71	0.72
7	4	0.95	0.91	0.90	0.91
8	5	1.00	0.98	1.00	1.00
9	6	1.00	0.99	1.00	1.00

Table II. Trial 1: Average number of replicates required per trial from 1000 observations.

Metaexperiment	Coef. Size	OrigCFSB	AccelCFSB	AccelBFSB	AccelAFSB
1	0	1436	**1414**	3358	3365
2	1	11322	**8244**	9177	9193
3	2	18259	13180	13410	13452
4	2.5	17767	15124	13781	13837
5	3	22876	19055	13803	13842
6	3.5	21511	17913	13802	13842
7	4	15422	13315	13802	13842
8	5	8942	7799	13802	13842
9	6	6311	5548	13802	13842

The same accuracy parameters: $\alpha = 0.05$, $\beta = 0.1$, $\Delta_0 = 2$, $\Delta_1 = 4$ were used in every experiment. The proportion of time each factor is found important should therefore be 0.05 when $\beta^{(i)} = 2.0$ ($= \Delta_0$) and should be 0.90 when $\beta^{(i)} = 4.0$ $(= \Delta_1)$.

The results are shown in Table I. The OrigCFSB method is conservative. The results for AccelCFSB and AccelAFSB are very similar with the proportion of time each factor is found important for the cases $\beta^{(i)} = 2.0$ and 4.0 close to their nominal 0.05 and 0.90 values. AccelBFSB does very well in this example and its results are comparable to those for AccelCFSB and AccelAFSB.

The average number of replicates needed per experiment for all 9 metaexperiments and for the four methods is given in Table II. AccelAFSB and AccelBFSB perform very similarly and are best when all the factor coefficient values lie between Δ_0 and Δ_1. AccelCFSB performs best outside this range.

4.1.2. Experimental Trial 2: Identifying large effects. In this trial just four metaexperiments were carried out, one for each of the four methods. For all the experiments in all four metaexperiments, the factor coefficients were the same as those used in Wan et al. [2010], namely: 2.0, 2.44, 2.88, 3.32, 3.76, 4.2, 4.64, 5.08, 6. The errors were heteroscedastic with $\sigma(\boldsymbol{\gamma}, \mathbf{X}) = \sum_{i=1}^{10} \gamma_j X_j$,with $\gamma_j = \beta_j$ all j.

The same accuracy parameters: $\alpha = 0.05$, $\beta = 0.1$, $\Delta_0 = 2$, $\Delta_1 = 4$ were used in every experiment.

Table III gives the number of times each factor was found important in each of the metaexperiments, and also the average number of replications used in each experiment for that metaexperiment.

In this trial a relatively large number of factor coefficients (the last four β_j) are greater than the importance threshold $\Delta_1 = 4$. This results in fewer replications needed by both CFSB methods compared with AccelAFSB and AccelBFSB. However AccelAFSB is more accurate at the Δ_0 threshold than either CFSB method or AccelBFSB, with the OrigCFSB performing especially poorly. At the Δ_1 threshold and above, OrigCFSB, AccelBFSB and AccelAFSB all display similar quite accurate power

Table III. Trial 2: Proportion found important out of 1000 trials and average number of replicates per experiment, for each metaexperiment.

Metaexperiment	1	2	3	4
Coef. Size	OrigCFSB	AccelCFSB	AccelBFSB	AccelAFSB
2.00	0.000	0.025	0.079	0.047
2.44	0.027	0.079	0.165	0.157
2.88	0.158	0.227	0.319	0.334
3.32	0.505	0.486	0.528	0.551
3.76	0.780	0.774	0.759	0.763
4.20	0.938	0.900	0.894	0.910
4.64	0.979	0.937	0.969	0.972
5.08	0.993	0.967	0.991	0.993
5.52	0.995	0.981	1.000	0.996
6.00	0.999	0.985	0.999	1.000
Ave. number of reps	21386	17818	26690	26731

Table IV. Trial 3: Proportion found important out of 1000 trials and average number of replicates per experiment, for each metaexperiment.

Metaexperiment	1	2	3	4
Coef. Size	OrigCFSB	AccelCFSB	AccelBFSB	AccelAFSB
2.0	0.001	0.034	0.070	0.048
2.0	0.010	0.042	0.069	0.049
2.5	0.044	0.110	0.201	0.186
2.5	0.078	0.114	0.180	0.181
3.0	0.331	0.366	0.394	0.422
3.0	0.332	0.353	0.423	0.416
3.5	0.760	0.702	0.710	0.714
3.5	0.760	0.723	0.697	0.685
4.0	0.947	0.916	0.908	0.907
4.0	0.931	0.909	0.889	0.906
Ave. number of reps	29249	24367	19490	19544

in identifying those $\beta_j > \Delta_1$ as being important, whilst AccelCFSB again did not perform well.

4.1.3. Experimental Trial 3: Identifying small, but important, effects. The layout of this trial is the same as that in the second, comprising four metaexperiments of 1000 identical experiments each using one of the four FSB methods with all experiments using $\alpha = 0.05$, $\beta = 0.1$, $\Delta_0 = 2$, $\Delta_1 = 4$ and again with heteroscedastic errors: $\sigma(\boldsymbol{\gamma}, \mathbf{X}) = \sum_{i=1}^{10} \gamma_j X_j$, with $\gamma_j = \beta_j$, all j.

The same β_j values were used in all experiments namely: 2.0, 2.0, 2.5, 2.5, 3.0, 3.0, 3.5, 3.5, 4.0, 4.0. Thus all the β_j's lie between Δ_0 and Δ_1. The results are given in Table IV.

The conditions favour AccelAFSB as all factor coefficients lie between Δ_0 and Δ_1. This results in fewer replications needed by AccelAFSB compared with AccelCFSB. AccelBFSB requires the fewest replications but AccelAFSB runs it quite close. As might be expected OrigCFSB requires the greatest number of replications.

At the Δ_0 threshold, the two factors with $\beta_i = 2$ should be found important about 5% of the time. This was the case for AccelAFSB. The percentage for AccelBFSB was rather high, whilst for AccelCFSB it was somewhat low and for OrigCFSB it was very low.

At the Δ_1 threshold the two factors with $\beta_i = 4$ should be found important about 90% of the time. Both AccelBFSB and AccelAFSB performed fairly accurately. The AccelCFSB was marginally high whilst for OrigCFSB is was clearly high. Overall AccelAFSB performed the best in this example.

Table V. Ericcson Example: Average number of runs per experiment vs. Experimental standard deviation.

Standard Deviation	OrigCFSB	AccelCFSB	AccelBFSB	AccelAFSB
0.06	134	134	134	134
0.48	226	185	163	214
0.96	762	529	448	551
1.92	2859	2011	1833	1936
2.88	6310	4432	4148	4256
3.84	11172	7935	7346	7432

Summarising all three Trials: If the expectation is that there will be many factor coefficients larger than Δ_1 and calculation of the stopping rule boundaries is not an issue, then AccelCFSB might be preferable. But if the expectation is that most factor coefficient values will fall in the Δ_0 and Δ_1 range or calculation of the stopping rule boundaries is an issue, then AccelAFSB would be preferable.

4.1.4. Ericsson Example with 92 factors. We now consider an example based on a supply chain simulation in the mobile communications industry at the Ericsson company in Sweden, as reported in Kleijnen et al. [2006]. It involves 92 factors.

To simplify the study, we did not make explicit runs of the simulation model in order to generate direct data for input to the SB analysis. Instead we generated data from a *metamodel* of the form (1) with variance structure of the form (3) fitted to data obtained from 128 runs obtained from the actual simulation model. Because the metamodel was used for the dispersion as well as the location effects problems, these actual simulation runs focused on the dispersion characteristics of the simulation model, and were made with different settings of the dispersion coefficients as the design points based on an orthogonal statistical design. We then fit a metamodel of the standard deviation/variance, as given by (28), to this simulation run data (using standard regression techniques). This gave (least squares) estimated values of γ_j: $\hat{\gamma}_j$, $j = 0, 1, ..., k$. In experiments with the SB method, input data was not then generated from the simulation model directly, but instead from the fitted statistical metamodel (1), with added normal errors ε_i of the form (3) whose standard deviations are given using the fitted coefficients :

$$\hat{\sigma}(\mathbf{x}_i) = \sigma \exp(\sum_{j=1}^{k} \hat{\gamma}_j x_{ij}). \tag{32}$$

Thus, as far as the numerical study of the SB analysis was concerned, the estimates $\hat{\gamma}_j$ were treated as the true factor dispersion coefficients. Note however that $\hat{\gamma}_0$ is not used directly. If it had its effect would simply have been to multiply $\hat{\sigma}(\mathbf{x}_i)$ by the factor $\exp(\hat{\gamma}_0)$. We have instead included its effect in the factor σ treating this as a parameter which we can vary between experiments.

The same 128 simulation data set used to obtain the $\hat{\gamma}_j$ also provided estimates, $\hat{\beta}_j$, of the β_j appearing in (1), so that the overall fitted metamodel produced data that could be used in the study of both location and dispersion problems.

For the location effects problem, the above fitted model, with heterogeneous variance structure (32), was used to generate observations in six trials each comprising four metaexperiments carried out for each of the four SB methods compared in the previous subsection. Each metaexperiment in the set comprised 1000 experiments all with the same error SD parameter σ in (32), but with σ varying between the six trials taking the specific values $\sigma = 0.06,\ 0.48,\ 0.96,\ 1.92,\ 2.88,\ 3.84$. The accuracy parameters were set at $\alpha = \beta = 0.1,\ \ \Delta_0 = 0.5$, $\Delta_1 = 2$ throughout.

Table VI. Ericsson Example: Proportion of times the top ten location effects were found important vs experimental standard deviation.

	γ	0.104	**-0.036**	0.033	0.142	0.134	0.224	0.091	0.078	**0.074**	0.161
	Factor ID	#11	**#47**	#89	#85	#86	#87	#43	#45	**#49**	#92
σ	β_i	0.332	**0.570**	0.617	1.155	1.161	1.186	1.290	1.297	**2.080**	4.179
0.06	OrigCFSB	0.000	**0.000**	0.000	0.055	0.094	0.189	0.738	0.788	**1.000**	1.000
	AccelCFSB	0.000	**0.000**	0.000	0.055	0.094	0.189	0.738	0.788	**1.000**	1.000
	AccelBFSB	0.000	**0.000**	0.000	0.055	0.094	0.189	0.738	0.788	**1.000**	1.000
	AccelAFSB	0.000	**0.000**	0.000	0.055	0.094	0.189	0.738	0.788	**1.000**	1.000
0.48	OrigCFSB	0.000	**0.001**	0.012	0.402	0.408	0.437	0.284	0.401	**0.994**	1.000
	AccelCFSB	0.001	**0.002**	0.019	0.383	0.403	0.447	0.280	0.369	**0.990**	1.000
	AccelBFSB	0.002	**0.011**	0.037	0.427	0.425	0.440	0.265	0.392	**0.982**	1.000
	AccelAFSB	0.000	**0.006**	0.029	0.416	0.419	0.449	0.262	0.391	**0.989**	1.000
0.96	OrigCFSB	0.006	**0.006**	0.009	0.413	0.402	0.424	0.269	0.379	**0.984**	1.000
	AccelCFSB	0.007	**0.024**	0.017	0.411	0.408	0.439	0.226	0.385	**0.935**	0.997
	AccelBFSB	0.037	**0.057**	0.052	0.434	0.415	0.450	0.224	0.347	**0.854**	0.997
	AccelAFSB	0.025	**0.038**	0.036	0.397	0.439	0.464	0.220	0.342	**0.914**	1.000
1.92	OrigCFSB	0.012	**0.014**	0.015	0.427	0.397	0.439	0.242	0.383	**0.984**	0.999
	AccelCFSB	0.019	**0.021**	0.024	0.406	0.437	0.451	0.242	0.326	**0.883**	0.994
	AccelBFSB	0.037	**0.053**	0.035	0.411	0.453	0.446	0.231	0.323	**0.830**	0.996
	AccelAFSB	0.036	**0.037**	0.031	0.443	0.450	0.459	0.206	0.351	**0.891**	1.000
2.88	OrigCFSB	0.008	**0.007**	0.013	0.407	0.412	0.440	0.266	0.367	**0.978**	1.000
	AccelCFSB	0.018	**0.027**	0.016	0.382	0.416	0.448	0.217	0.362	**0.868**	0.989
	AccelBFSB	0.042	**0.053**	0.034	0.425	0.447	0.456	0.239	0.351	**0.893**	1.000
	AccelAFSB	0.019	**0.032**	0.051	0.397	0.454	0.459	0.209	0.356	**0.918**	1.000
3.84	OrigCFSB	0.017	**0.021**	0.020	0.394	0.410	0.450	0.259	0.393	**0.969**	1.000
	AccelCFSB	0.019	**0.024**	0.023	0.417	0.391	0.433	0.203	0.338	**0.854**	0.976
	AccelBFSB	0.043	**0.047**	0.033	0.419	0.447	0.454	0.223	0.335	**0.906**	1.000
	AccelAFSB	0.035	**0.046**	0.041	0.421	0.426	0.446	0.236	0.378	**0.908**	1.000

Table V shows the average number of replications required per experiments in each of the 24 metaexperiments.

In the trial where σ was smallest with $\sigma = 0.06$, all four methods required the same number of replications in each experiment. For all the other trials, where $\sigma > 0.06$, Orig CFSB required the most replications whilst Accel BFSB required the fewest. The number of replications required by the other two methods, Accel CFSB and Accel AFSB, was between these, but with Accel AFSB always only just higher than Accel BFSB. Comparing Accel CFSB and Accel AFSB, the former performed slightly better for lower σ, viz when $\sigma = 0.48$ and $\sigma = 0.96$, but this was reversed with Accel AFSB requiring fewer replications than Accel CFSB for larger σ, viz when $\sigma = 1.92$, $\sigma = 2.88$ and $\sigma = 3.84$.

The performance of all four methods is quite similar in terms of identifying important factors. Table VI shows how often the 10 largest factors values were found important for each method in each of the six trials as σ varied. All the other factors had coefficient values well below the lower threshold value of $\Delta_0 = 0.5$. In fact the largest number of times any was found important in any metaexperiment in any trial was 60 compared with the 100 times that would have corresponded to the $\alpha = 0.1$ error threshold of a marginally unimportant factor. In the vast majority of cases the number of times any one of these factors was found important was well below this 100.

In Table VI, two columns are highlighted (bold). One corresponds to Factor 49 whose true coefficient value of 2.080 is closest to the upper threshold of $\Delta_1 = 2.0$, so that, with $\beta = 0.1$, a method that is accurate should be finding this factor important approximately 90% of the time. When $\sigma = 0.06$ all methods found the factor always important. This might be explained by the fact that the coefficient value is greater than the threshold of importance value of $\Delta_1 = 2.0$, and for such a small σ all methods are able to identify the factor as being important. However we would expect the count to fall to somewhere just above 90% as σ increases. However Orig CFSB remains consistently high, Accel CFSB falls away to well below 90% and Accel BFSB displays the greatest variation about 90%. Accel AFSB performance quite consistently with values about 90% once σ has increased past $\sigma = 0.48$.

The other highlighted column corresponds to factor 47 whose true coefficient value of 0.57 is the lowest value just above the $\Delta_0 = 0.5$ threshold. In this case all the methods seem to be biased too low, even in the case where $\sigma = 0.06$ when one might be expecting methods to find the factor greater than 10% of the time. As sigma increases the importance count rises for all the methods, but there is still a distinct bias in all cases. Accel BFSB seems best followed by Accel AFSB, then Accel CFSB, then Orig CFSB.

In conclusion Accel AFSB has performed perhaps the most satisfactorily at the upper threshold level, but none of the methods are very satisfactory at the lower level in this example, though Accel AFSB is the least unsatisfactory.

4.2. Dispersion Effects Problem

We now turn to the performance of AccelAFSB for the dispersion effects problem. Recall that we address this problem separately from the location effects problem. We have simply chosen the case of symmetric confidence levels with $\alpha = \beta = 0.1$ as being typical of what might be used in practice for levels of confidence.

We also have to select the threshold levels Δ_0 and Δ_1 and here we have also used values that might be adopted in practice. Note that, in the case of the dispersion effects model, and in contrast to the location effects model, Δ_0 and Δ_1 are thresholds for the coefficients γ_j appearing in equation (28). Thus $\gamma_j = \Delta_i$ corresponds to the situation where increasing the jth factor level from $x_j = 0$ to $x_j = 1$ would increase the standard deviation by a factor $\exp(\Delta_i)$. Thus $\Delta_i = \log(1+a_i)$ corresponds to a change in standard deviation of $a_i \times 100\%$. As the dispersion effects are positive, we need only consider the case where $0 < \Delta_0 < \Delta_1$ so that $0 < a_0 < a_1$.

We have used three pairs of Δ_0, Δ_1 values:

$$\Delta_0 = \log(1.1) \simeq 0.0953, \quad \Delta_1 = \log(1.25) \simeq 0.2231, \tag{33}$$

$$\Delta_0 = \log(1.5) \simeq 0.4055, \quad \Delta_1 = \log(3.0) \simeq 1.0986 \tag{34}$$

and

$$\Delta_0 = \log(2.0) \simeq 0.6931, \quad \Delta_1 = \log(5.0) \simeq 1.6094. \tag{35}$$

Bearing in mind the inherent variability in the estimation of standard deviations, the first pair represents a very stringent condition where an increase in standard deviation of 25% is already deemed important whilst an increase can be no more than 10% if it is to be deemed unimportant. We might consider the second pair to be what might be used in typical practical situations, where there has to be a 200% increase in the standard deviation for the change to be deemed important and an increase of up to 50% would be considered unimportant. The third pair might be used in exploratory situations where there has to be a fairly extreme increase of 400% to be deemed truly important and an increase of up to 100% would still be considered unimportant.

4.2.1. Example with 32 factors. In this example there are 32 factors and we compare the performance of the dispersion effects algorithm using the stopping rules in (30) and (31). We also compare the performance of our recommended stopping rule, (31), with various orders of the factor effects.

We carried out four metaexperiments with 1000 independent experiments in each, all using the same following settings. Eight of the factors contribute a dispersion effect

Table VII. 32 Factor Experiment

Metaexperiment	1	2	3	4
Stopping Rule	Known Sigma	Anscombe	Anscombe	Anscombe
Factor Order	Ranked	Ranked	Mixed	Very Mixed
Prop.(coef.=log(3.0) found imp.)	0.9075	0.8919	0.8851	0.8718
Prop.(coef.=log(1.5) found imp.)	0.0823	0.0849	0.0625	0.0631
Prop.(coef.=0 found imp.)	0.0000	0.0000	0.0004	0.0007
Mean number of design points	16	16	20	22
Mean observations/design point	36	52	36	36

of $\gamma = \log(3)$; eight with $\gamma = \log(1.5)$; the remaining 16 factors all have dispersion effect $\gamma = 0$. We set $\alpha = \beta = 0.1$, $\Delta_0 = \log(1.5) \simeq 0.4055$, $\Delta_1 = \log(3.0) \simeq 1.0986$. Thus a quarter of the effects are precisely at the lower threshold level and a quarter at the higher threshold level.

First metaexperiment: The factors were rank ordered by size. Thus all 8 factors with $\gamma = \log(3)$ were placed first, then all 8 factors with $\gamma = \log(1.5)$, with the remaining 16 factors where $\gamma = 0$ coming last. In each experiment the factors were classified using the SB method with the known sigma stopping rule in (30). With the given values of α, β, Δ_0 and Δ_1, we would expect roughly 90% of the trials to declare the first 8 factors important, 90% of the trials to declare the second set of 8 factors unimportant, and few of the remaining factors ever to be declared important. Table VII gives the results showing the results to be much in line with what we expect. If anything, the method is somewhat conservative at the lower threshold with less than 10% of factors at the lower threshold incorrectly classified as important.

The average number of design points used in each trial was 16.36 and the mean number of observation per design point was 36.

Second metaexperiment: This was exactly the same as the first experiment, except for comparison we changed the stopping rule to (31), the case where σ is assumed unknown. The results are given in Table VII. It will be seen that the percentage of time each factor is found important is very similar to those in the first metaexperiment. The main change is that the mean number of observations used per design point has increased from 36.0 to 51.9. This result suggests, not surprisingly, that use of the known sigma stopping rule is preferable. It may be that use of the unknown sigma stopping rule is more robust against model misspecification as it is well-known that sequential stopping rules are not too dependent on the precise error structure. However we have not investigated this point in this paper. We shall use the known sigma stopping rule in all the remaining dispersion examples considered in the paper, but in many real experiments the unknown sigma stopping rule might be preferred to guard against model uncertainty.

Third metaexperiment: This was exactly the same as the second metaexperiment, with the same set of γ effects but in a mixed order namely: five at $\gamma = \log(3)$, then 3 at $\gamma = 0$, then 3 at $\gamma = \log(3)$, then 5 at $\gamma = 0$, then 5 at $\gamma = \log(1.5)$, then 6 at $\gamma = 0$, then 3 at $\gamma = \log(1.5)$ and finally 2 at $\gamma = 0$. The results are given in Table VII. Though not as good as in the first experiment, they still seem quite satisfactory. Here the SB process displays a downward bias in declaring certain factors, that are actually important, to be important less often than expected. In particular the factors at 5 and 11, which should be declared important with probability $1 - \beta$, are declared important less frequently than they should be, whilst the factors at 21 and 28 which should be declared unimportant with probability $1 - \alpha$ are declared unimportant more often than this. The reason for this seems to be that all these factors are positioned in such a way that they are likely to become grouped in the SB process with a large number of unimportant factors. Their effect is then dominated by the combined effect of these unimportant factors. Thus, for example, 5 becomes grouped with 6, 7, and 8, whilst 28 becomes grouped with 25, 26 and 27.

It will be seen that apart from specific factors that might be affected in this way, the majority of the factors are found important or not important the right proportion of times indicating that the $E(k_0, k_1)$ test itself operates satisfactorily. The example is perhaps a fairly stringent test in that the non-zero factor values are so clearly grouped at the critical Δ values and are also ranked in a regular pattern.

Fourth metaexperiment: This again uses the same set of γ effects as in the three metaexperiments already discussed, but with the order even more mixed. The first three coefficients are $\gamma_1 = \log(3)$, $\gamma_2 = \log(1.5)$, $\gamma_3 = 0$; this pattern of three is repeated for eight groups, followed by the last eight γ's all equal to zero. The results are given in Table VII and are quite similar to those of the third metaexperiment. Those coefficients where $\gamma = \log(3)$ are found important slightly less than the nominal 90% of the time with some more adversely affected because of grouping effects. The coefficients where $\gamma = \log(1.5)$ display the same pattern in an even more marked way, being found important somewhat less than the nominal 10% level.

The mean number of design points required to be examined has increased from 16.36 in fully ranked case, to 20 in the mixed ranking order case, and to 22.0 in the very mixed ranking order case. The mean number of observations needed per design point remained stable, indeed constant, at 36 in all cases.

Though artificial, this 32 factor example gives a clear indication of the basic properties of the proposed SB method, suggesting it has understandable and reasonably robust characteristics. The example in the next section provides a more realistic instance of what might occur in practice.

4.2.2. Ericsson Example. We consider the dispersion effects problem for the Ericsson example. Recall that the data from this example is constructed after fitting a metamodel to 128 runs of the simulation. Thus, the coefficients in this model are only estimates of the true coefficients of the actual simulation. The fitted dispersion coefficient values are labeled as 'CoeffTrue' in Table VIII. Of the 92 estimated $\hat{\gamma}_j$ values, 49 were positive and 43 were negative. This might appear to seriously break the positivity condition (5) requiring all the $\gamma_j \geq 0$, for SB to be applicable. However, of the 43 estimates for which $\hat{\gamma}_j < 0$, only 4 had p-values less than 0.1 from the fitting of the 128 simulation runs. This means that the true coefficients of the simulation may not be negative at all. Thus condition (5) is largely satisfied, and the few significantly negative $\hat{\gamma}_j$ are what might happen in practice. The observations in the experiments were generated from the fitted metamodel for variability including these negative dispersion effects and thus provide a test of the SB method under realistic conditions.

The most significant positive dispersion coefficients were, in decreasing order of importance, those corresponding to $j = 87$, 92, 85, 86 and 11, all with p-values less than 0.01.There is some overlap of these factors with those factors identified in Kleijnen et al. (2006) as having important location effects. However this overlap is not uniform in that some factors such as $j = 88$ or $j = 90$ identified in Kleijnen et al. (2006) as having important location effects have insignificant fitted dispersion effects.

Table VIII shows the results of three pairs of metaexperiments, each metaexperiment being made up of 1000 independent experiments.

In the first pair we set $\alpha = \beta = 0.1$, $\Delta_0 = \log(1.1)$ and $\Delta_1 = \log(1.25)$. Only factor 87, with $\gamma_{87} = .2236$, is at the $\Delta_1 = \log(1.25) = 0.2231$ threshold level of importance. When the factors are placed in their original order, it can be seen that factor 87 was found important nearly 90% of the time. When the factors are reordered according to their 'true' (i.e. fitted) coefficient values, there is a some degradation in the proportion of time factor 87 was found important, but the degradation is slight.

Table VIII. Ericcson Factor Experiment

Metaexperiment	1	2	3	4	5	6
Factor Order	Original	Ranked	Original	Ranked	Original	Ranked
Δ_0	ln(1.10)	ln(1.10)	ln(1.5)	ln(1.5)	ln(2)	ln(2)
Δ_1	ln(1.25)	ln(1.25)	ln(3.0)	ln(3.0)	ln(5)	ln(5)
Prop.(coef. 87 found imp.)	0.896	0.882	0.001	0.014	0.000	0.002
Prop.(next highest coef found imp.)	0.531	0.379	0.000	0.007	0.000	0.001
Mean number of design points	16.1	24.7	4.1	7.2	2.6	3.7
Mean observations/design point	994	994	36	36	21	78

The main difference between the two cases is in the number of design points required, with respectively 16.1 and 24.7 observations needed per design point. Interestingly it is the case where the coefficients are in ranked order that requires more design points. This seems to be due to the fact that when a large number of coefficients, with values either close to or in between Δ_0 and Δ_1, are grouped together then more observations are needed to determine their status and most importantly, there is a higher probability that they will be found important. Thus, for example, the factors 11, 17, 25 and 43, with dispersion coefficient values between 0.0907 and 0.1044, and which are isolated in their original positions, are rapidly eliminated as being unimportant in the experiments of the first metaexperiment. But when they are bunched together, as in the second metaexperiment, then they are found important a much higher proportion of the time (though still less than 10%) and this requires the extra observations to determine their non-zero values.

The other main feature of this first pair of metaexperiments is the very large number, 994, of observations (i.e. simulation runs) needed per design point. This is a reflection of the stringent Δ threshold levels employed and shows that use of such threshold values may result in an unacceptably high number of simulation runs being needed.

The second pair of metaexperiments uses the same metamodel of dispersion effects, but the threshold levels are relaxed to the more pragmatic values: $\Delta_0 = \log(1.5)$ and $\Delta_1 = \log(3)$. In the metamodel the 'true' dispersion effects are all now definitely less than Δ_0 and so should be found unimportant. With factors in their original order, none is found important more than 0.1% of the time. When ranked order is used, the results are almost identical. The only very slight difference, occurring with one or two of the factors with the largest dispersion coefficients. The one with the largest coefficient, factor 87, is found important the highest proportion of the time, but even this value of 1.4% is still well below the allowed upper limit of 10%. Compared with the first pair of metaexperiments, the main change is that the mean number of design points needed in a trial is now only 4.1. Moreover only 36 observations are required per design point.

The final pair of metaexperiments is where the threshold levels are further relaxed to $\Delta_0 = \log(2)$ and $\Delta_1 = \log(5)$. In these metaexperiments, whether the factors are in ranked order or not, the coefficients are found unimportant practically all the time. Only 2.6 design points are needed in the case where the factors are in their original order, rising slightly to 3.7 design points for the case where the coefficients are in ranked order. In both cases there is a further reduction, to 21, in the number of observations needed per design point.

Our conclusion from this example therefore is that use of threshold levels like (34) or (35) would allow the SB method to be used in a very practical way to provide a rapid check of whether there is any serious heteroscedasticity in a response output of interest.

5. FINAL COMMENTS

In this article, we have introduced a new stopping rule based on a fixed width CI to improve the performance of sequential bifurcation screening. We demonstrated this method for both the location effects problem and the dispersion effects problem. We also developed a way to accelerate sequential bifurcation procedures by modifying the procedure to eliminate redundant observations that were commonly taken in previous descriptions of the SB algorithm.

For the location effects problem, Accel CFSB, our accelerated version of controlled SB originally proposed by Wan et al. [2010], and Accel AFSB, our new method, seem the most efficient. The performance of AFSB is generally more robust particularly in delivering stated Type I error level at the lower threshold Δ_0 and producing the highest power with the fewest replications for identifying midrange effects. Whereas, the Accel CFSB, is somewhat more efficient at finding very large effects. Computationally, Accel AFSB is easier to implement than Accel CFSB as it has a stopping rule with a boundary that can be explicitly calculated.

For the dispersion problem the modified AFSB method is easily implemented and provides reasonable control of the size of the two types of statistical error. It thus provides an efficient and practical method of screening to check if there is any significant heteroscedasticity in response output.

The numerical examples were obtained using algorithms implemented in VBA macros in a spreadsheet. The spreadsheet is available from the authors.

6. ACKNOWLEDGMENT

This work was supported by EPSRC grant GR/S91673 and carried out whilst Bruce Ankenman was visiting Southampton Statistical Sciences Research Institute.

REFERENCES

Ankenman, B., Cheng, R., and Lewis, S. 2006. A polytope method for estimating factor main effects efficiently. In Proceedings of the 2006 Winter Simulation Conference, L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, Eds. IEEE, Piscataway, N.J., 369–375.

Anscombe, F. 1953. Sequential estimation. Journal of the Royal Statistical Society, Ser. B 15, 1–29.

Bettonvil, B. and Kleijnen, J. 1997. Searching for important factors in simulation models with many factors: Sequential bifurcation. European Journal of Operational Research 96, 180–194.

Cheng, R. 1997. Searching for important factors: sequential bifurcation under uncertainty. In Proceedings of the 1997 Winter Simulation Conference, D. W. S. Andradottir, K.J. Healy and B. Nelson, Eds. IEEE, Piscataway, N.J., 275–280.

Chow, Y. and Robbins, H. 1965. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. Annals of Mathematical Statistics 36, 457–462.

Kleijnen, J., Bettonvil, B., and Persson, F. 2006. Screening for the Important Factors in Large Discrete-Event Simulation Models: Sequential Bifurcation and Its Applications. Springer, New York, N.Y.

Liu, W. 1997. On some sample size formulae for controliing both size and power in clinical trials. Journal of the Royal Statistical Society, Ser. D 46, 239–251.

Sanchez, S., Wan, H., and Lucas, T. 2009. A two-phase screening procedure for simulation experiments. ACM Transactions on Modeling and Computer Simulation 19.

Shen, H. and Wan, H. 2009. Controlled sequential factorial design for simulation factor screening. European Journal of Operational Research 198, 511–519.

Shen, H., Wan, H., and Sanchez, S. 2010. The hybrid approach for simulation factor screening. Naval Research Logistics 57, 45–57.

Singham, D. and Schruben, L. 2012. Finite-sample performance of absolute precision stopping rules. INFORMS Journal on Computing 24, 624–635.

Starr, N. 1966a. On the asymptotic efficiency of a sequential procedure for estimating the mean. Annals of Mathematical Statistics 37, 1173–1185.

STARR, N. 1966b. The performance of a sequential procedure for the fixed-width interval estimation of the mean. Annals of Mathematical Statistics 37, 36–50.

STEIN, C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. Annals of Mathematical Statistics 16, 243–258.

WAN, H. AND ANKENMAN, B. 2007. Two-stage controlled fractional factorial screening for simulation experiments. Journal of Quality Technology 39, 126–139.

WAN, H., ANKENMAN, B., AND NELSON, B. 2006. Controlled sequential bifurcation: a new factor-screening method for discrete-event simulation. Operations Research 54, 743–755.

WAN, H., ANKENMAN, B., AND NELSON, B. 2010. Improving the efficiency and the efficacy of controlled sequential bifurcation. INFORMS Journal on Computing 22, 482–492.

WOODROOFE, M. 1977. Second order approximations for sequential point and interval estimation. Annals of Mathematical Statistics 5, 84–995.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.