Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Three-phase optimal design of sensitivity experiments

C.F. Jeff Wu [a,*], Yubin Tian [b]

[a] Georgia Institute of Technology, United States
[b] Beijing Institute of Technology, China

ARTICLE INFO

Article history:
Received 5 December 2012
Received in revised form
7 May 2013
Accepted 16 October 2013

Keywords:
D-optimal designs
Quantile estimation
Robbins–Monro
Sensitivity testing
Stochastic approximation
Up-and-down method

ABSTRACT

In sensitivity testing the test specimens are subjected to a variety of stress levels to generate response or nonresponse. These data are used to estimate the critical stimulus (or threshold) level of the experimental object. Because of its versatile applications, several sensitivity testing procedures have been proposed and used in practice. There remains the outstanding question of finding an efficient procedure, especially when the sample size is small and the interest lies in the extreme percentiles. In the paper we propose a novel three-phase procedure, dubbed 3pod, which can be described as a trilogy of "search-estimate-approximate". A core novel idea is to choose the stress levels to quickly achieve an overlapping data pattern which ensures the estimability of the underlying parameters. Simulation comparisons show that 3pod outperforms existing procedures over a range of scenarios in terms of efficiency and robustness.

1. Introduction

In some experimental investigations each experimental unit has a critical stimulus level that cannot be observed directly. The distribution of these critical levels (or thresholds) over test specimens is of primary interest. In pyrotechnic applications the stimulus (or stress) level, denoted by x, may be the drop height of an explosive, or the pressure on a pile of ammunition, and the response, denoted by y, is binary (e.g., explosion/non-explosion). In industrial applications, x may be the wind speed in a wind-tunnel experiment or the dosage of a compound in a toxicity study. Sensitivity tests are commonly used to estimate some aspects of this distribution. Numerous methods have been proposed for conducting sequential sensitivity testing. But there remains the outstanding question of finding a sequential procedure that works well when the test sample size is small, especially for estimating *extreme* quantiles. For example, there is no clear winner among the competing methods in a comprehensive simulation study reported by Young and Easterling (1994). The median is considered because it is easier to estimate (i.e., more information in the data). And engineers often use it as a benchmark, especially during product development. However, to assess the reliability of a product for field use, the interest and focus is often on the extreme percentiles with $p \geq 0.9$ or even $p \geq 0.99$.

The problem can be formulated as follows. Let $y=1$ or 0 denote the binary outcome, response or nonresponse respectively and $F(x)$ denote the probability Prob $(y=1|x)$ of response at a given stimulus level x. Usually we consider the location-scale model:

$$F(x,\mu,\sigma) = G((x-\mu)/\sigma), \tag{1}$$

* Corresponding author. Tel.: +1 404 352 0885.
E-mail address: jeffwu@isye.gatech.edu (C.F.J. Wu).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.10.007

where μ and $\sigma > 0$ are unknown parameters and G is a known distribution function. Define x_p to be the pth quantile of the distribution, i.e., $F(x_p) = p$. Hence $x_p = \mu + \sigma G^{-1}(p)$. Many methods have been proposed for quantile estimation. A review will be given in Section 2. However, when p is large, e.g., $p = 0.99$ or 0.999, and the sample size is small to moderate, accurate estimation of extreme quantiles is still a challenging problem.

The main purpose of this paper is to develop a new *three-phase sequential procedure* that can find the desired stimulus levels more precisely by quickly and efficiently exploiting the information in the test data and expert knowledge. In Section 2, we will give a brief review of some existing methods that are known to be competitive or used in practice. In Section 3, we describe the proposed procedure. Its phase I is to generate some response and nonresponse values, to identify a reasonable experimental range, and to move the stress levels to achieve an overlapping pattern. Phase II is to facilitate the maximum likelihood estimation of the parameters in the assumed model and to spread the stress levels for optimal parameter estimation. Once the data pattern is in good shape, we move to Phase III which uses an efficient approximation scheme to get the stress levels converge to the unknown quantile quickly. This three-phase procedure can be viewed as a trilogy of "search–estimate–approximate". Section 4 gives one example to illustrate the key steps of the three-phase procedure. Simulation comparisons of the procedure with some existing methods are given in Section 5. Concluding remarks and further questions are given in Section 6.

2. Review of some existing methods

The literature contains many studies that compare a large number of methods. An early one with a good review is Wetherill (1963). Some of the papers mentioned in this section also have good reviews. Therefore we will confine our review in this section to those that will be compared with our method in Section 3. First is the up-and-down method (Dixon and Mood, 1948), a.k.a. the Bruceton test. It increases (and respectively decreases) the x value by a step length d if $y = 0$ (and resp. $y = 1$). Because of its simplicity, it is still widely used, even though it has been known among researchers to be inefficient in most situations. It is only for estimating the median $x_{0.5}$. The Robbins and Monro (1951) stochastic approximation procedure and its adaptive version (Lai and Robbins, 1979) are known to be much more efficient than the up-and-down method. Joseph (2004) recognized that the procedure, originally developed for continuous data, is not well suited for binary data. He then modified the Robbins–Monro (RM) procedure as follows. Let $\theta = x_p$ and assume the prior distribution of θ is $N(x_1, \tau_1^2)$. Consider the following stochastic approximation scheme:

$$x_{i+1} = x_i - a_i(y_i - b_i), \quad i \geq 1, \tag{2}$$

where a_i and b_i are two sequences of constants. Let $Z_i = x_i - \theta$, $i \geq 1$. He proposed to choose a_i and b_i such that $E(Z_{i+1}^2)$ is minimized subject to the condition $E(Z_{i+1}) = 0$. Under the normal approximation of the distribution of Z_i by $N(0, \tau_i^2)$, he showed that the solution is given by

$$b_i = \Phi\left\{\frac{\Phi^{-1}(p)}{(1+\beta^2\tau_i^2)^{1/2}}\right\}, \quad a_i = \frac{1}{b_i(1-b_i)\frac{\beta\tau_i^2}{(1+\beta^2\tau_i^2)^{1/2}}\phi\left\{\frac{\Phi^{-1}(p)}{(1+\beta^2\tau_i^2)^{1/2}}\right\}},$$
$$\tau_{i+1}^2 = \tau_i^2 - b_i(1-b_i)a_i^2, \quad \beta = \frac{G'(G^{-1}(p))}{\phi(\Phi^{-1}(p))}\cdot\frac{1}{\sigma}, \tag{3}$$

where $\Phi(\cdot)$ is the stand normal CDF and $\phi(\cdot)$ is its density function. For the sake of brevity, we shall refer to this binary version of the RM procedure as the Robbins–Monro–Joseph (RMJ) procedure. If the true distribution is normal ($G = \Phi$), then β in (3) reduces to σ^{-1}. Joseph (2004) recommended the choice

$$\tau_1 = \frac{c}{\Phi^{-1}(0.975)}. \tag{4}$$

We choose $c = 5$ in the simulation study for RMJ. As is common in stochastic approximation, the last x value is used as the estimate of the unknown value x_p. That is, for N iterations, use x_{N+1} as the estimate.

Another alternative is the MLE recursive method due to Wu (1985). Approximate the true unknown model by using a parametric model $F(x|\gamma)$ like the logit or probit indexed by a few parameters, e.g., $\gamma = (\mu, \sigma)$. After n runs, let $\hat{\gamma}_n$ be the MLE of γ. The next run is chosen at the level x_{n+1} defined as $F(x_{n+1}|\hat{\gamma}_n) = p$. However the method can only be used if the MLE of (μ, σ) in the model (1) exists. According to Silvapulle (1981), a necessary and sufficient condition for the existence of MLE is to have an *overlapping pattern* in the data. That is, the largest x value among the y_i's with $y = 0$, denoted by M_0, should be larger than the lowest x value among the y_i's with $y = 1$, denoted by m_1, namely

$$M_0 > m_1. \tag{5}$$

Wu (1985) recognized that his method needs to start with an initial design that satisfies this condition. A Bayesian extension of Wu's method was proposed by Joseph et al. (2007). We refer to $[m_1, M_0]$ as the *overlapping interval*. If $M_0 \leq m_1$, we refer to $[M_0, m_1]$ as the *separation interval* because the region for nonresponse ($y = 0$) is separated from the region for response ($y = 1$). If a separation pattern is observed, taking the next stress level x within the separation interval will not change the separation pattern (but the interval will get shorter). We refer to this as "trapped in separation". At a certain point, the next stress level should be taken outside the separation interval in order to break the logjam.

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

Note that the RM-type procedures do not need an initial design because it does not require the use of an MLE estimate. Neyer (1994) was the first to propose, among other things, a systematic method for generating a good initial design. His method consists of three parts. In the first part, a modified binary search is used to generate both response and nonresponse values and to "close in" on the region of interest; the second part is devoted to finding an overlapping pattern quickly, i.e., to get a unique MLE; in the third part the *D*-optimality criterion is used to generate future design points (i.e., stress levels). Neyer's method is the most novel and effective new method in the last 20 years. It was developed with military applications in mind. Other unpublished works in the same context include Langlie (1962) and Rothman et al. (1965). Another recent work is Dror and Steinberg (2008), which used a Bayesian *D*-optimal design approach for optimally estimating the unknown parameters in a generalized linear model. When specialized for the binary data problem considered here, their main interest is the optimal estimation of the parameters μ and σ, not quick approximation or estimation of extreme quantiles.

3. A three-phase sequential procedure of sensitivity testing

We shall call the proposed procedure in this section a *three-phase optimal design* of sensitivity experiments, or 3pod for abbreviation. The term 3pod or "tripod" is appropriate because its empirical performance (see Section 5) is most steady among the competing methods.

3.1. Phase I design

The objective of the Phase I design is to quickly identify a reasonable experimental range by obtaining some response and nonresponse values and to move the stress levels to achieve an overlapping pattern. This phase consists of three stages: (1) obtain some response and nonresponse values; (2) search for overlapping regions; (3) enhance the overlapping regions. Note that the need for reaching an overlapping pattern was discussed in Section 2.

(1) *Obtain response and nonresponse values*:
First, based on past experience or data from similar products, we can guess a reasonable range (μ_{min}, μ_{max}) of the location parameter μ, as well as the value σ_g of the scale parameter σ, where the subscript g stands for "guess". Also, these initial guesses should satisfy

$$\mu_{max} - \mu_{min} \geq 6\sigma_g.$$

There are five scenarios about the initial range (μ_{min}, μ_{max}) relative to the true value μ: (μ_{min}, μ_{max}) is too much to the left, too much to the right, too wide, too narrow, or is symmetrical around μ. This stage of the design should quickly detect which of the five scenarios holds and adjust the design sequence accordingly so that the stress levels can move toward symmetry around μ.
First, run tests at $x_1 = \frac{3}{4}\mu_{min} + \frac{1}{4}\mu_{max}$ and $x_2 = \frac{1}{4}\mu_{min} + \frac{3}{4}\mu_{max}$. Let the observed values be denoted by y_1 and y_2 respectively. There are four cases:

(i) If $(y_1, y_2) = (0, 0)$, it indicates that (μ_{min}, μ_{max}) is to the left of μ. Thus the experimental range should be extended toward the right. Run test at $x_3 = \mu_{max} + 1.5\sigma_g$. If $y_3 = 1$, move to stage I2; if $y_3 = 0$, run test at $x_4 = \mu_{max} + 3\sigma_g$. If $y_4 = 1$, move to stage I2. If $y_4 = 0$, there is a strong indication that the range is not large enough. Increase the x value by $1.5\sigma_g$ each time and run test until $y = 1$ is observed.
(ii) If $(y_1, y_2) = (1, 1)$, it indicates that (μ_{min}, μ_{max}) is to the right of μ. Run test at $x_3 = \mu_{min} - 1.5\sigma_g$. If $y_3 = 0$, move to stage I2; if $y_3 = 1$, run test at $x_4 = \mu_{min} - 3\sigma_g$. If $y_4 = 0$, move to stage I2. If $y_4 = 1$, as argued in case (i), decrease the x value by $1.5\sigma_g$ each time and run test until $y = 0$ is observed.
(iii) If $(y_1, y_2) = (0, 1)$, move to stage I2.
(iv) If $(y_1, y_2) = (1, 0)$, it indicates that the range (μ_{min}, μ_{max}) is too narrow around $x_{0.5}$. To expand the range, run experiments at $x_3 = \mu_{min} - 3\sigma_g$ and $x_4 = \mu_{max} + 3\sigma_g$ respectively. The corresponding observed values are denoted as y_3 and y_4. Then move to stage I2.

(2) *Search for overlapping regions*: This stage has two steps (i) and (ii):
(i) Recall from (2) that $m_1 < M_0$ indicates an overlapping pattern.
(a) If $m_1 < M_0$, move to the next stage I3.
(b) If $m_1 \geq M_0$, compute the MLE $\hat{\mu}$ of μ using the model in Eq. (1) with $\sigma = \sigma_g$ and choose the next level at $\hat{\mu}$. If an overlapping pattern is observed (which implies that $\hat{\mu}$ must lie outside the separation interval), move to the next stage I3. If the separation pattern continues, repeat the procedure of using $\hat{\mu}$ for the next level until

$$m_1 - M_0 < 1.5\sigma_g. \tag{6}$$

Once Eq. (6) is met, we should choose the x levels outside the interval $[M_0, m_1]$ to avoid being "trapped in separation". This is the next step (c) or (d).
(c) If $m_1 - M_0 < 1.5\sigma_g$, $m_1 \geq M_0$, and $k_0 > k_1$, where k_0 and k_1 are the number of $y_i = 0$ and $y_i = 1$ respectively, run test at $m_1 + 0.3\sigma_g$. If $y = 0$, move to stage I3; otherwise, run test at $M_0 - 0.3\sigma_g$. If $y = 1$, move to stage I3; if $y = 0$, it means that an overlapping interval is not yet found and that σ_g may be too large; move to the next step (ii).

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

(d) If $m_1 - M_0 < 1.5\sigma_g$, $m_1 \geq M_0$, and $k_0 \leq k_1$, run test at $M_0 - 0.3\sigma_g$. If $y=1$, move to stage I3; otherwise, run test at $m_1 + 0.3\sigma_g$. If $y=0$, move to stage I3; if $y=1$, move to step (ii).

Reduce the guessed value of σ_g by setting $\sigma_g = \frac{2}{3}\sigma_g$. Update the m_1 and M_0 from (i), and repeat the step (i) until

(ii) $m_1 < M_0$.

(3) *Enhance the overlapping regions*: If $M_0 - m_1 \geq \sigma_g$, run test at $(M_0+m_1)/2$; if $0 < M_0 - m_1 < \sigma_g$, run test at $(M_0+m_1)/2+0.5\sigma_g$ and $(M_0+m_1)/2-0.5\sigma_g$ respectively. Then move to phase II.

3.2. Phase II design

Its objective is to choose stress levels to *optimize the parameter estimation* in the assumed model. Following Neyer (1986), we use the *D*-optimal design criterion to choose stress levels. Compute the MLE $(\hat{\mu}_k, \hat{\sigma}_k)$ of (μ, σ) based on the observed values $(x_1, y_1), \ldots, (x_k, y_k)$. To prevent the estimates from exceeding the design region, truncate the estimates as follows:

$$\tilde{\mu}_k = \max\{\underline{x}, \min(\hat{\mu}_k, \overline{x})\}, \quad \tilde{\sigma_k} = \min\{\hat{\sigma}_k, \overline{x} - \underline{x}\}, \tag{7}$$

where $\underline{x} = \min\{x_1, \ldots, x_k\}$, $\overline{x} = \max\{x_1, \ldots, x_k\}$. In other words, $\tilde{\mu}_k$ should be within $[\underline{x}, \overline{x}]$, and $\tilde{\sigma}_k$ should not exceed $\overline{x} - \underline{x}$. Then, choose the next point x_{k+1}, such that the determinant of the Fisher information matrix evaluated at $(\tilde{\mu}_k, \tilde{\sigma}_k)$ and $\{x_1, \ldots, x_k, x_{k+1}\}$ is maximized. Run test at x_{k+1} and observe y_{k+1}. Suppose n_1 runs have been allocated for Phase I and Phase II designs. Repeat the iteration until all n_1 tests are completed.

3.3. Phase III design

The objective of the Phase III design is to choose stress levels to be placed near the unknown x_p in order to obtain more information about x_p. Here, we use the Robbins–Monro–Joseph method. See the discussion and justification in Section 2. This phase consists of two stages.

(1) *Choice of the initial value*: We use

$$x_{n_1+1} = \tilde{\mu}_{n_1} + G^{-1}(p)\tilde{\sigma}_{n_1}, \tag{8}$$

to estimate x_p, and also as the starting value of Phase III, where $\tilde{\mu}_{n_1}$ and $\tilde{\sigma}_{n_1}$ are defined in Eq. (7), based on the n_1 observations in Phases I and II. Calculate the Fisher information matrix $I_{n_1}(\tilde{\mu}_{n_1}, \tilde{\sigma}_{n_1})$. Denote the inverse matrix of $I_{n_1}(\tilde{\mu}_{n_1}, \tilde{\sigma}_{n_1})$ by $[I^{ij}_{n_1}]$, where $I^{ij}_{n_1}$ is its (i,j)-th element. Then let

$$\tau_1^2 = I^{11}_{n_1} + \{G^{-1}(p)\}^2 I^{22}_{n_1}. \tag{8a}$$

For stability reason (explanation in Section 3.4 under III), we truncate τ_1^2 within [2.3429, 6.5079].

(2) *Optimal approximation to* x_p: Denote by y_{n_1+1} the observation at x_{n_1+1}. For the remaining design sequence, use the following iterative scheme:

$$x_{n_1+i+1} = x_{n_1+i} - a_i(y_{n_1+i} - b_i), \quad i \geq 1, \tag{9}$$

where x_{n_1+i} is the ith stress level of Phase III, y_{n_1+i} is the corresponding observed value, and a_i and b_i $i \geq 1$ are the coefficients in Eq. (3) with the *exception* that β in Eq. (3) is replaced by 0.5β.

Repeat the procedure until n_2 tests are completed, where n_2 is the designated number of tests for phase III. Since phase III uses the RMJ procedure, the last x value is used as the estimate of the unknown value x_p.

For easy reference, a flow chart for phase I of 3pod is given in Fig. 1.

3.4. Further discussions on the 3pod procedure

The most novel feature of 3pod is phase I. First, it generates at least one x with $y=1$ and with $y=0$ and to identify a reasonable experimental range (in stage I1). Then it moves the design sequence to quickly achieve an overlapping data pattern (in stage I2 and I3). In Phase II we choose additional stress levels to optimize parameter estimation. Phase I was inspired by Neyer's parts 1 and 2, whose purpose is to get the data to achieve an overlapping pattern, and by Wu's (1985) earlier observation on such a need. But there are major differences in how this is carried out. Neyer immediately used the *D*-optimal design criterion (evaluated at $\hat{\mu} = \frac{1}{2}(M_0+m_1)$ and σ_g) to generate the next stress levels. Although the *D*-optimal design approach is useful for optimal parameter estimation, it is not well suited for quickly achieving an overlapping data pattern. Our approach is to devote stages 2 and 3 of phase I to reach and enhance the overlapping pattern. The steps involved use direct search based on the information in y, $m_1 - M_0$ and σ_g. This difference is responsible for the superior performance of 3pod over Neyer's method in having a much smaller percentage of simulation runs that fail to achieve an overlapping pattern. See the comparisons in Tables 2–4 of Section 5. Only in Phase II, we use *D*-optimal design to optimize parameter estimation as in Neyer. Typically not many samples are devoted to this stage, while Neyer devotes most of its

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

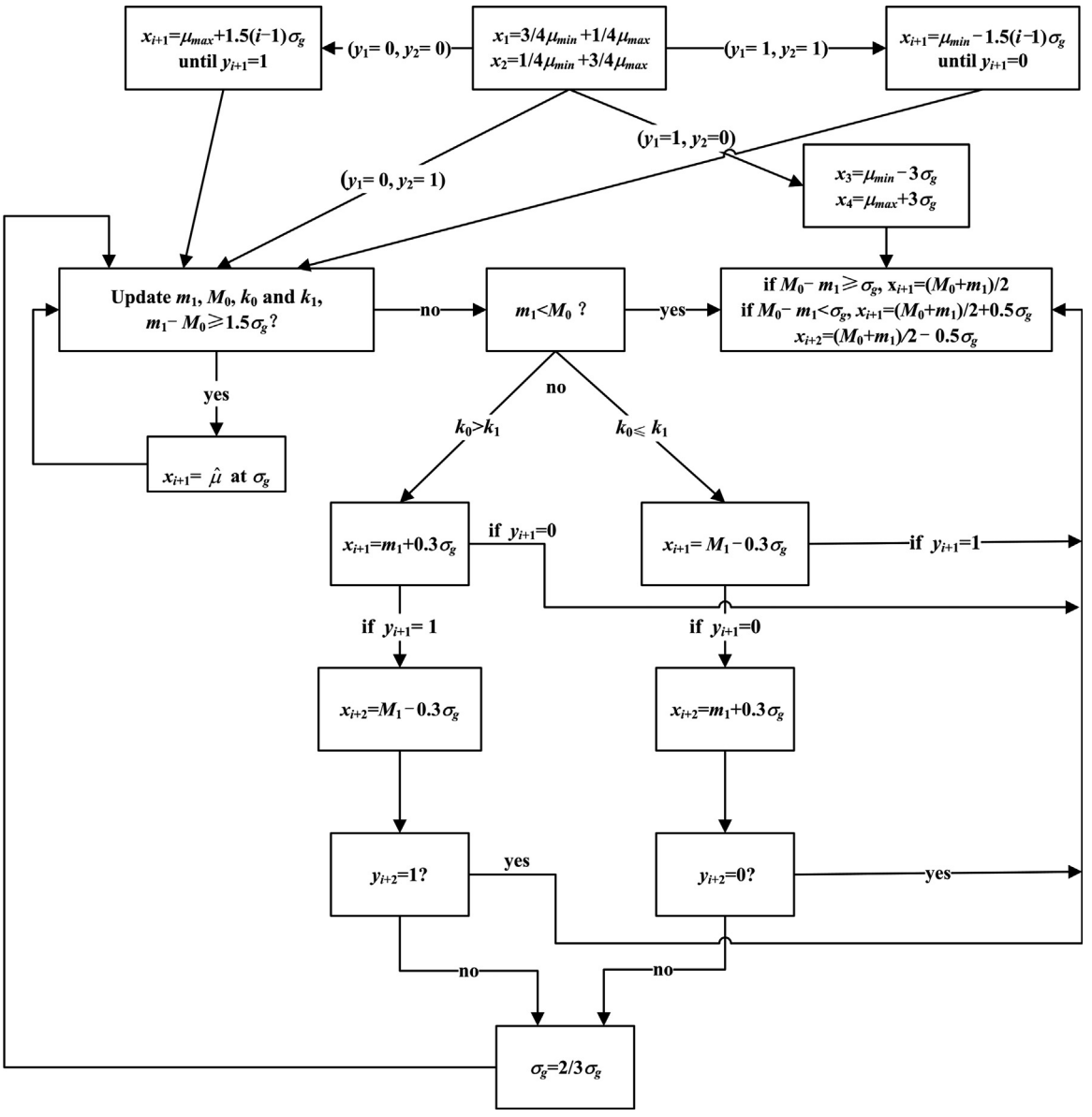

Fig. 1. A flow chart showing Phase I of the 3pod procedure.

samples using the *D*-optimality criterion. Another innovation is the adoption of the RMJ procedure in phase III. Instead of using *D*-optimal design as in part 3 of Neyer, we use the RMJ procedure which is known to converge quickly to extreme quantiles.

We will now give further explanation or justification for some steps in 3pod. The steps are labeled according to the labeling in the procedure. For example, I2ib refers to case b of step i in stage 2 of phase I.

I1(i) and I1(ii): It usually does not require more than two or three runs to complete the step unless the range (μ_{min}, μ_{max}) is off and σ_g is too small. These two conditions are reasonable requirements in most practical work. If in doubt, the investigator should devote some resources to resolve this issue. The purpose of this stage is to widen the range if necessary so that the procedure will not be trapped in a wrong region.

I1(**iv**): Why are only two additional runs required? There are only four possibilities for the four y values arranged from low to high x values: (i) (0 1 0 1) is the best outcome because it already has a good overlapping pattern. (ii) (0 1 0 0) and (1 1 0 1) both have an overlapping pattern. In I3, it may add additional runs to "enhance" the overlapping pattern. (iii) (1 1 0 0) is the least likely scenario. It can only happen if $y=1$ at x_3 $(=\mu_{min}-3\sigma_g)$ and x_1 $(=\mu_{min}+\frac{1}{4}(\mu_{max}-\mu_{min}))$, and $y=0$ at x_2 $(=\mu_{max}-\frac{1}{4}(\mu_{max}-\mu_{min}))$ and at x_4 $(=\mu_{max}+3\sigma_g)$. Because $x_1-x_3=x_4-x_2=\frac{1}{4}(\mu_{max}-\mu_{min})+3\sigma_g$ and $x_2-x_1=\frac{1}{2}(\mu_{max}-\mu_{min})$, its probability of

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

occurrence is very low unless both $\mu_{max}-\mu_{min}$ and σ_g are very small, an unlikely scenario as we argued above.

I2ib: If there is only one $y=1$ at m_1 and one $y=0$ at M_0, it is easy to see that $\hat{\mu}=\frac{1}{2}(M_0+m_1)$ for any symmetric distribution. This amounts to taking the next level at the middle of the range $[M_0, m_1]$ and is called a *binary search* by Neyer. In general $\hat{\mu}\neq\frac{1}{2}(M_0+m_1)$ and can be quite different if there are more 0's than 1's (or vice versa). Therefore the proposed approach is more general than Neyer's binary search method and exploits the data more efficiently. As long as the new level $\hat{\mu}$ is still within the range $[M_0, m_1]$, as argued in Section 2, the separation pattern will remain. And that is why we use the rule in Eq. (6) to get out of the separation interval. The choice of the constant 1.5 in Eq. (6) is based on the empirical studies. It should not be too small so that the design sequence can quickly get out of the separation trap.

I2ic and I2id: When $n_0 > n_1$, it indicates that the design sequence lies at the lower end (i.e., to the left of the median) of the curve. Thus the next level should be above m_1. The choice of the constant 0.3 in $0.3\sigma_g$ is to ensure that the step is not too large so that $y=0$ can be observed to the right of m_1, i.e., to achieve an overlapping pattern. (Note the similarity to part 2 of Neyer's method, in which σ_g is decreased by a multiple of 0.8 each time in order to achieve a "faster overlap of the data".) A similar remark applies to case d. When $n_0=n_1$, we can choose the next level to the left or to the right of the interval $[M_0, m_1]$. For simplicity, we keep this case in d.

I2(ii): When this step is reached, it indicates that the search range is too large and σ_g should be reduced to enhance the chance of observing an overlapping pattern. The constant 2/3 is based on the empirical studies. A smaller value like 3/5 may be considered.

I3: It takes one to two more runs to enhance the overlapping pattern so that the MLE estimation in phase II can be performed more efficiently.

II: It is known that for the probit model the *D*-optimal design with fixed sample size places equal number of points on two symmetrical quantiles x_p with $p=0.128$ and 0.872. See Morgan (1992, p. 342). Its net effect is to spread out the points more widely. When the same criterion is used after some data are collected in phase I, its net effect is harder to characterize. A theoretical result due to Yang and Stufken (2009) may be relevant in this regard. They showed that any optimal approximate (i.e., continuous) design, regardless of one-stage or multi-stage, is based on two points. Since this result assumes that the design weight is continuous, it is not directly applicable to the finite sample situation considered here. In the illustrative example in Section 4, we will see how this step helps to fill in a gap in the design sequence. The truncation step in Eq. (7) is rarely invoked but can actually happen, especially when the overlapping data pattern is barely satisfied.

III: Instead of using an initial guess x_1 as in RMJ, we use an updated estimate (see Eq. (8)) of x_p based on all data in phase I and II as the starting value of phase III. From the simulation studies in Section 5, the RMJ can perform erratically when its x_1 value is wrongly guessed and σ_g is too large. Our procedure does not suffer from this because, by the time we reach phase III, we will have accumulated enough data and information so that the updated guess of σ is not far off and the initial value in Eq. (8) is closer to the true value x_p. Another reason for the need of an updated initial value is that, once we start the RMJ, we will not use the MLE estimation any more. The terminated value of the RMJ iterations will be the estimate of the quantile x_p. Our simulation experience suggests that τ_1^2 in Eq. (8a) based on data in phases I and II can sometimes be rather small. Given the recommendation by Joseph (2004) for choosing $\tau_1=c/\Phi^{-1}(0.975)$ in Eq. (4), we restrict c in the range [3, 5]. The corresponding range for τ_1^2 is [2.3429, 6.5079]. Because we choose $c=5$ in the simulation study for the RMJ procedure, we want the corresponding c value in phase III of 3pod not to exceed 5. This explains why we choose 5 as the upper bound of the range [3, 5]. The lower bound 3 is somewhat arbitrary as long as it prevents the choice of small τ_1^2 values, which will cause the iteration to take small incremental steps. The use of truncation appears to stabilize the performance of 3pod. We also found that, even when τ_1^2 is chosen within the wide range [2, 16], the difference it makes on the RMSE values in the simulation study is insignificant, i.e., in the second decimal place. In Eq. (9), we modified the RMJ procedure by replacing β by 0.5β. This choice gives a more steady performance in the convergence of the RMJ iterations to the true value. One practical reason for halving the β value is to improve the confidence level of the upper confidence limit for the final estimate $\hat{x}_p$, a subject for future work. However, whether we choose 0.5β or β or a value in between, it does not change the overall picture in the conclusions regarding 3pod in Section 5.

4. Illustration

We use the following example to illustrate the key steps of the proposed 3pod method. Suppose that the underlying distribution is normal $F(x)=\Phi((x-\mu)/\sigma)$ with $\mu=10$, $\sigma=1$. Table 1 and Fig. 2 show the sequence of the first 30 points of the 3pod design for estimating $x_{0.9}=(11.2816)$. We choose $(\mu_{min}, \mu_{max})=(0, 22)$, $\sigma_g=3$, $n_1=15$, and $n_2=15$. The guessed σ_g value of 3 is much larger than the true σ value of 1. The first two x values are: $x_1=\frac{3}{4}\cdot 0+\frac{1}{4}\cdot 22=5.5$, $x_2=\frac{1}{4}\cdot 0+\frac{3}{4}\cdot 22=16.5$; $y_1=0$ and $y_2=1$ are observed. This falls in case I1(iii) because $(y_1, y_2)=(0, 1)$; move to stage 2, step 2ib. The MLE $\hat{\mu}=\frac{1}{2}(5.5+16.5)=11$, which is x_3. Observe $y_3=0$. Update $M_0=11$ and $m_1=16.5$. Because $m_1-M_0=5.5>1.5\sigma_g=4.5$, continue step I2ib. Compute the MLE $\hat{\mu}=13.8$ and take $x_4=13.8$. Note that it is slightly to the right of $\frac{1}{2}(M_0+m_1)=13.75$

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

Table 1
The sequence of the first 30 points of the 3pod design for estimating $x_{0.9}$.

No. i	Level x_i	Result y_i	Comment
1	5.5	0	I1(iii)
2	16.5	1	I1(iii)
3	11	0	I2ib
4	13.8	1	I2ib
5	10.1	0	I2id
6	14.7	1	I2id
7	10.4	1	I2(ii)
8	11.7	1	I3
9	9.7	1	I3
10	7.3	0	II
11	7.8	0	II
12	8.1	0	II
13	12.2	1	II
14	8.5	0	II
15	11.8	1	II
16	11.7106	1	III
17	11.4896	1	III
18	11.2980	0	III
19	12.3899	1	III
20	12.2393	1	III
21	12.1033	1	III
22	11.9796	1	III
23	11.8660	1	III
24	11.7612	1	III
25	11.6638	1	III
26	11.5730	1	III
27	11.4878	1	III
28	11.4077	1	III
29	11.3321	1	III
30	11.2605	1	III
31	11.1925		III

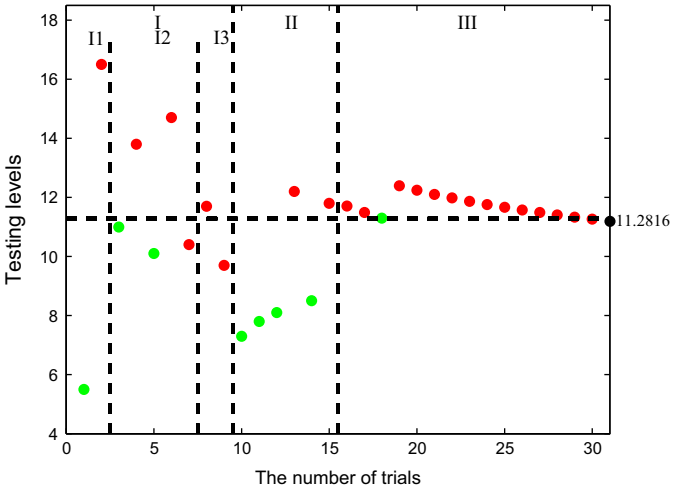

Fig. 2. An illustrative example. (*Note*: A response is marked in red, a nonresponse in green.) (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

because of the lack of symmetry in the data. Observe $y_4=1$. Because $m_1-M_0=13.8-11=2.8<4.5$ satisfies Eq. (6) and $n_0=n_1=2$, move to step I2id and take $x_5=M_0-0.3\sigma_g=11-0.3\times 3=10.1$; observe $y_5=0$. Still no overlapping pattern; stay in step I2id, take $x_6=m_1+0.3\sigma_g=14.7$ and observe $y_6=1$. Still no overlapping pattern; move to step I2(ii) and set $\sigma_g=\frac{2}{3}\times 3=2$. Return to step I2id, choose $x_7=11-0.3\times 2=10.4$ and observe $y_7=1$. An overlapping pattern is achieved! Move to the next stage I3. Because $M_0-m_1=11-10.4=0.6<2$ $(=\sigma_g)$, take x_8 and x_9 at $\frac{1}{2}(M_0+m_1)\pm 0.5\sigma_g$ $(=11.7, 9.7)$. Observe $y_8=y_9=1$. Move to phase II. Recall $k=9$ in this step. Compute the MLE $\hat{\mu}_9=9.9726$, $\hat{\sigma}_9=2.0705$. Since they do not exceed the range $[\underline{x},\overline{x}]=[5.5, 16.5]$ and $\overline{x}-\underline{x}=11$ respectively, we have $\tilde{\mu}_9=\hat{\mu}_9$ and $\tilde{\sigma}_9=\hat{\sigma}_9$. Choose $x_{10}=7.3$ based on the D-optimal criterion evaluated at $\tilde{\mu}_9=9.9726$ and $\tilde{\sigma}_9=2.0705$. Because $n_1=15$, we use the same criterion to choose $x_{11},\ldots,x_{15}$.

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

Then move to Phase III for x_{16} to x_{30}. From these x values we note that the RMJ method in Phase III does its job well. The first three iterations (x_{16}, x_{17}, x_{18}) descend toward the true value. Then it makes an upward correction at x_{19} because $y_{18}=0$. After that, it descends steadily toward the true value.

In Fig. 2, the test runs with $y=1$ (and resp. $y=0$) are marked in red and green respectively. The true quantile 11.2816 is marked in black at the right end. From the color pattern, it is easily seen that the separation pattern persists in the first six runs. Only run 7 breaks the pattern but only barely because the only overlap is for the pair ($x_7=10.4$, $y_7=1$) and ($x_3=11$, $y_3=0$). The x distance between the two points is 0.6. Since an overlap is observed, we move to I3 to enhance the pattern. The next two runs have $y=1$ (in red) and the data overlap is more pronounced as can be seen from the figure. The next phase II is to use D-optimality criterion to add further points. As we discussed in Section 3, a D-optimal design tends to choose points that are more spread out. The first three points (in green) in II are to the left of all the previous points except the first one ($x_1=5.5$). Therefore these three points are chosen very much according to the D-optimal design theory, i.e., to fill in a gap on the left side. But the next three points (red, green, red) in II are harder to explain by this theory. They are all in the middle of the range and thus cannot be explained by this geometric argument. As explained in Section 3, the six points in phase II are chosen to "strengthen" the parameter estimation. This should help the choice of the starting value (run 16) in phase III. The RMJ sequence in phase III behaves as expected. It hovers around the true value (black point) with one big jump from one side to the other (green dot to red dot, low to high).

5. Simulation study

In this section we make simulation comparisons of the proposed method, dubbed as 3pod, with four other methods reviewed in Section 2. Two distributions are used for generating the simulation samples: the normal distribution $\Phi((x-\mu)/\sigma)$ and the logistic distribution $LG((x-\mu)/(\sigma/1.8138))$ with $\mu=10$ and $\sigma=1$, where $LG(z)=(1+e^{-z})^{-1}$. Note that, for the logistic distribution, σ in the normal distribution should be divided by $1.8138(=\pi/\sqrt{3})$ so that the normal and the logistic sensitivity distributions have the *same* standard deviation. This is important for comparing the performances under the two distributions. Because the methods under comparisons assume normality in the procedures, the simulation study based on the normal samples is to study their estimation *efficiency*, while that for the logistic samples is to study their *efficiency/robustness*. Because μ and σ are unknown to the experimenter, we use μ_g and σ_g as their "guessed" values. For the five methods under comparisons, we choose their initial values as follows:

1. Up-and-down: It needs the initial value x_1 and the step length d. We choose $x_1=\mu_g$ and $d=\sigma_g$.
2. Neyer: It needs a range $[\mu_{min}, \mu_{max}]$ for μ and a guessed value of σ. Here we choose σ_g for σ and $\mu_{min}=\mu_g-4\sigma_g$, $\mu_{max}=\mu_g+4\sigma_g$.
3. Wu: We use parts 1 and 2 of Neyer's method to generate overlapping pattern. After that, we use Wu's MLE scheme for generating the design sequence. The choice of initial values is thus the same as in Neyer's.
4. 3pod: It needs a range $[\mu_{min}, \mu_{max}]$ for μ and a guessed value of σ. Choice is the same as in Neyer's.
5. Robbins–Monro–Joseph (RMJ): It needs an initial value x_1, τ_1 and σ (see Eq. (3)). We choose $x_1=\mu_g+\Phi^{-1}(p)\sigma_g$, $\tau_1=2.5$ and σ_g for σ. We choose 2.5 for τ_1 (equivalent to $c=5$ in Eq. (4)) to ensure a 0.95 *a priori* probability (in the normal case) for the parameter $\theta(=x_p)$ to be in the interval $[x_1-5, x_1+5]$.

Recall that, except for the RMJ method, the first four methods all require the estimation of the parameters μ and σ in model (1) and that such estimates only exist if the data satisfy the overlapping pattern defined in Eq. (5). Before embarking on a simulation study of the efficiencies of procedures, we need to know "how often the required overlapping pattern can be achieved for a given sample size n." A sequential experimental run of size n is called *successful* if its data satisfy the overlapping pattern in Eq. (5). If not, it is said to be *wasted*. The first part of the simulation study is to find out how often each of the first four methods encounters wasted runs in order to obtain 1000 successful runs. The *smaller* is the number of wasted runs for achieving 1000 successful runs, the *better* is the procedure because it measures the failure rate using such a method for a given test sample size. Because Wu's method has the same initial parts as Neyer's, its performance will be identical to that of Neyer's. Therefore only three methods are compared in this part of the study.

For each of the three methods (up-and-down, Neyer, 3pod), the entries in Table 2 give the numbers of wasted runs in order to generate 1000 successful runs for $n=40$ ($n_1=25$, $n_2=15$ for 3pod), over a range of values for μ_g and σ_g, i.e.,

Table 2
Comparison of three methods, $n=40$ ($n_1=25, n_2=15$ for 3pod). (Each entry gives the range of the number of wasted runs in order to have 1000 successful runs.)

Method	$\mu_g=9-11$				
	$\sigma_g=0.5$	$\sigma_g=1.0$	$\sigma_g=2.0$	$\sigma_g=3.0$	$\sigma_g=4.0$
Up-and-down	1–2	32–40	106–1775	2158–37 681	45 595–1 530 915
Neyer	23–34	74–84	414–528	498–1103	2142–2411
3pod	0	0–1	0–4	6–16	14–30

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

$\mu_g = 9, 10, 11, \sigma_g = 0.5, 1.0, 2.0, 3.0, 4.0$. The choice of 9 and 11 for μ_g is a reasonable representation of what guessed values of the median (i.e., the μ value) the investigator can make because their corresponding response probability values are 0.159 and 0.841 for the normal case and 0.140 and 0.860 for the logistic case. These cover a broad range of guessed values around 0.5 the investigator can make. Tables 3 and 4 give the corresponding values for $n=60$ ($n_1 = 30, n_2 = 30$ for 3pod) and $n=80$ ($n_1 = 35, n_2 = 45$ for 3pod). To save space, a range of values is given in each entry as μ_g varies from 9, 10 to 11. The first observation from the tables is the deteriorating performance as the σ_g value increases. This is expected as a bigger σ_g value indicates a more fluctuating behavior of the procedure. Among the three methods, the up-and-down method performs very poorly. For $\sigma_g = 3$ and 4, it will take thousands to a million wasted runs. Only for $\sigma_g = 0.5$ and 1, it outperforms Neyer but still lags behind 3pod. Because of its very unstable performance, we will not consider it in the next phase of the simulation study. Furthermore, 3pod consistently outperforms Neyer and particularly so for large σ_g values like 2, 3 and 4. This can be explained by the fact that 3pod devotes its first phase to accelerate the search toward overlapping patterns while Neyer's attempts to do the same by using a less efficient binary search. Another aspect of the study is the dependence on the sample size n. As n increases from 40 to 60 and 80, up-and-down and Neyer have fewer wasted runs. For the 3pod method, as n increases, we also increase the initial sample size n_1 (for phase I) from $n_1 = 25$ to 30 for $n=60$ and to 35 for $n=80$. Note that the number of wasted runs for 3pod depends only on phase I. Therefore, if n_1 remains at 25 for $n=60$ and 80, this number would not change. By increasing the initial size by an increment of 5, the number of wasted runs in Tables 3 and 4 is reduced to nearly zero. It is an encouraging result because it suggests that, even for large σ_g value like 3.0 and 4.0, a mere increase of 5 runs for phase I will eliminate almost all the wasted runs in Table 2.

Table 3
Comparison of three methods, $n=60$ ($n_1 = 30, n_2 = 30$ for 3pod). (Each entry gives the range of the number of wasted runs in order to have 1000 successful runs.)

Method	$\mu_g = 9$–11				
	$\sigma_g = 0.5$	$\sigma_g = 1.0$	$\sigma_g = 2.0$	$\sigma_g = 3.0$	$\sigma_g = 4.0$
Up-and-down	0–1	1–7	22–1052	1202–24934	29157–1020277
Neyer	21–33	68–77	408–626	494–1041	2029–2334
3pod	0	0	0–1	0–2	0–3

Table 4
Comparison of three methods, $n=80$ ($n_1 = 35, n_2 = 45$ for 3pod). (Each entry gives the range of the number of wasted runs in order to have 1000 successful runs.)

Method	$\mu_g = 9$–11				
	$\sigma_g = 0.5$	$\sigma_g = 1.0$	$\sigma_g = 2.0$	$\sigma_g = 3.0$	$\sigma_g = 4.0$
Up-and-down	0–1	1–2	2–662	787–18 519	21396–764958
Neyer	16–32	62–74	393–521	482–993	1971–2289
3pod	0	0	0	0	0–1

Table 5A
RMSE for estimation of $x_{0.9}$, $n=40$ ($n_1 = 25$, $n_2 = 15$ for 3pod), true distribution = normal.

	$\sigma_g = 0.5$	$\sigma_g = 1.0$	$\sigma_g = 2.0$	$\sigma_g = 3.0$	$\sigma_g = 4.0$
$\mu_g = 9$					
Neyer	0.4798	0.4957	0.5095	0.4675	0.5268
3pod	0.4284	0.4534	0.4686	0.4472	0.4606
Wu	0.3787	0.3541	0.3976	0.3467	0.4504
RMJ	0.3109	0.2605	0.3035	0.3529	0.3929
$\mu_g = 10$					
Neyer	0.4596	0.4644	0.4958	0.4817	0.4626
3pod	0.4505	0.4520	0.4897	0.4423	0.4498
Wu	0.3879	0.3565	0.3768	0.4148	0.4178
RMJ	0.2967	0.2632	0.3065	0.3595	0.4046
$\mu_g = 11$					
Neyer	0.5681	0.5001	0.5005	0.6202	0.7446
3pod	0.4436	0.4480	0.4780	0.4583	0.4439
Wu	0.5185	0.3823	0.4137	0.5338	0.5718
RMJ	0.3054	0.2730	0.3147	0.3605	0.5139

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

In the second part of the simulation study, we compare the estimation performance of 3pod with three other methods (Neyer, Wu, RMJ) through simulations. For each method, we simulated 1000 repeated samples with a given sample size n from the normal distribution $\Phi((x-\mu)/\sigma)$ with $\mu=10$ and $\sigma=1$. Only successful runs are retained in the 1000 samples. Then we computed the bias (Bias) and the root mean-squared error (RMSE) of the 1000 estimates for each method. Because in most cases Bias is a much smaller component of RMSE, to save space, we only give the RMSE results in Tables 5A–7A. Separate discussion on Bias will be given later in the few cases where it plays a dominant role. The Bias results can be found in the online supplementary materials. Table 5A shows the results for estimating $x_{0.9}$ with $n=40$ over a range of guesses for μ and σ, i.e., $\mu_g=9,10,11$, $\sigma_g=0.5,\ 1.0,\ 2.0,\ 3.0,\ 4.0$. Tables 6A and 7A show the corresponding results for $x_{0.99}$ and $x_{0.999}$ with $n=60$ and $n=80$ respectively. We include $x_{0.99}$ and $x_{0.999}$ in the study as they represent the high reliability requirement in practical applications. We increase the sample size from 40 to 60 and 80 respectively because more extreme percentiles would require larger samples. For the 3pod method, two versions are considered: $n_1=25$ and 30 for $n=60$, and $n_1=25$ and 35 for $n=80$. Its purpose is to see how sample size allocation to phase I and II and phase III will affect the estimation efficiency.

In terms of the RMSE, the relative rankings of the four methods can be described in three broad scenarios as follows:

A. For $n=40$ and $\mu_g=9$, 10, we have

RMJ>Wu>3pod>Neyer

in descending order of performance. For $\mu_g=11$ and $\sigma_g=0.5,1.0,2.0,3.0$, Wu and 3pod are comparable but 3pod has a more stable performance. For $\mu_g=11$ and $\sigma_g=4.0$, 3pod beats RMJ and the rest.

Table 5B
Values of $x_1-x_{0.9}$, $x_1=$ starting value, $x_{0.9}=$ true value under normal.

	σ_g=0.5	σ_g=1.0	σ_g=2.0	σ_g=3.0	σ_g=4.0
μ_g=9	−1.6408	−1	0.2816	1.5631	?2.8447
μ_g=10	−0.6408	0	1.2816	2.5631	?3.8447
μ_g=11	0.3592	1	2.2816	3.5631	4.8447

Table 6A
RMSE for estimation of $x_{0.99}$, $n=60$, true distribution=normal.

	$\sigma_g=0.5$	$\sigma_g=1.0$	$\sigma_g=2.0$	$\sigma_g=3.0$	$\sigma_g=4.0$
$\mu_g=9$					
Neyer	0.7160	0.6310	0.7445	0.6834	0.5309
3pod ($n_1=25,n_2=35$)	0.5342	0.5434	0.5369	0.5231	0.5515
3pod ($n_1=30,n_2=30$)	0.5574	0.5532	0.5737	0.5542	0.5619
Wu	1.3000	1.2464	2.1362	1.3276	1.2057
RMJ	0.4633	0.4005	0.5064	1.3509	3.9470
$\mu_g=10$					
Neyer	0.6225	0.6270	0.6987	0.6678	0.4409
3pod ($n_1=25,n_2=35$)	0.5719	0.5971	0.5919	0.4489	0.4579
3pod ($n_1=30,n_2=30$)	0.6033	0.5859	0.6078	0.5213	0.5580
Wu	1.3432	1.2803	1.7334	1.2176	1.2840
RMJ	0.4537	0.4128	0.4752	2.3509	4.9470
$\mu_g=11$					
Neyer	0.8675	0.6616	0.7621	0.8921	0.8699
3pod ($n_1=25,n_2=35$)	0.5585	0.5642	0.5611	0.5365	0.5233
3pod ($n_1=30,n_2=30$)	0.5765	0.5789	0.5892	0.5752	0.5525
Wu	1.1976	1.0108	1.2602	1.0297	1.3105
RMJ	0.4629	0.4172	0.7808	3.3509	5.9470

Table 6B
Values of $x_1-x_{0.99}$, $x_1=$ starting value, $x_{0.99}=$ true value under normal.

	σ_g=0.5	σ_g=1.0	σ_g=2.0	σ_g=3.0	σ_g=4.0
μ_g=9	−2.1632	−1	1.3263	3.6527	5.979
μ_g=10	−1.1632	0	2.3263	4.6527	6.979
μ_g=11	−0.1632	1	3.3263	5.6527	7.979

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

Table 7A
RMSE for estimation of $x_{0.999}, n=80$, true distribution = normal.

	$\sigma_g=0.5$	$\sigma_g=1.0$	$\sigma_g=2.0$	$\sigma_g=3.0$	$\sigma_g=4.0$
$\mu_g=9$					
Neyer	0.8149	0.7850	0.8817	0.8398	0.5460
3pod ($n_1=25, n_2=55$)	0.7696	0.7632	0.7776	0.6802	0.7433
3pod ($n_1=35, n_2=45$)	0.8371	0.7959	0.7713	0.7277	0.7679
Wu	1.3693	1.8537	2.7452	2.1485	2.0030
RMJ	0.6970	0.5841	0.5282	4.0104	7.4440
$\mu_g=10$					
Neyer	0.6793	0.7627	0.8758	0.7850	0.4917
3pod ($n_1=25, n_2=55$)	0.7857	0.8580	0.7842	0.6815	0.7286
3pod ($n_1=35, n_2=45$)	0.8168	0.8323	0.7909	0.7220	0.7347
Wu	1.2997	1.9930	2.5588	1.8412	2.3299
RMJ	0.6835	0.5618	1.4748	5.0104	8.4440
$\mu_g=11$					
Neyer	0.9837	0.7928	0.9245	1.0653	0.9861
3pod ($n_1=25, n_2=55$)	0.8035	0.7772	0.7968	0.7081	0.7555
3pod ($n_1=35, n_2=45$)	0.8311	0.7950	0.7800	0.7423	0.7555
Wu	1.3410	1.5844	2.3133	2.0253	2.3918
RMJ	0.7087	0.6622	2.4748	6.0103	9.4440

Table 7B
Values of $x_1 - x_{0.999}$, x_1 = starting value, $x_{0.999}$ = true value under normal.

	$\sigma_g=0.5$	$\sigma_g=1.0$	$\sigma_g=2.0$	$\sigma_g=3.0$	$\sigma_g=4.0$
$\mu_g=9$	−2.5451	−1	2.0902	5.1805	8.2707
$\mu_g=10$	−1.5451	0	3.0902	6.1805	9.2707
$\mu_g=11$	−0.5451	1	4.0902	7.1805	10.2707

B. For $n=60$, $\mu_g=9, 10$ and $\sigma_g=0.5, 1.0, 2.0$, and $\mu_g=11$ and $\sigma_g=0.5, 1.0$; and for $n=80$, $\mu_g=9$ and $\sigma_g=0.5, 1.0, 2.0$, and $\mu_g=10, 11$ and $\sigma_g=0.5, 1.0$, we have

RMJ>3pod>Neyer>Wu.

One exception is $n=80$, $\mu_g=10$ and $\sigma_g=0.5, 1.0$, for which Neyer beats 3pod and is comparable to RMJ. But in these two cases, Neyer would take 30 and 74 wasted runs respectively while 3pod does not have any wasted run (see Table 3).

C. For $n=60$, $\mu_g=9, 10$ and $\sigma_g=3.0, 4.0$, and $\mu_g=11$ and $\sigma_g=2.0, 3.0, 4.0$; and for $n=80$, $\mu_g=9$ and $\sigma_g=3.0, 4.0$, and $\mu_g=10, 11$ and $\sigma_g=2.0, 3.0, 4.0$, we have

3pod>Neyer>RMJ and Wu,

where RMJ and Wu perform equally poor. One exception to the rankings is when $n=80$, $\mu_g=10$ and $\sigma_g=4.0$, for which Neyer does much better than 3pod. But this should be balanced against the fact that from Table 4 Neyer would take 2289 wasted runs in this case while the number of wasted runs for 3pod is much smaller: 22 for $n_1=25$ and 1 for $n_1=35$. Therefore the much smaller RMSE value of 0.4917 for Neyer should not be taken too seriously as the price of taking up a large number of wasted runs is big.

Among the simulation results, the most striking and puzzling is the *sudden deterioration* of the RMJ method when σ_g increases to 3.0 and 4.0 for $n=60$ and to 2.0, 3.0 and 4.0 for $n=80$. When it does very poorly, its RMSE is dominated by its Bias component, and in many situations the two are identical. This suggests that, for all the 1000 simulation samples, the estimates from RMJ cluster tightly around a value which is far from the true quantile. A typical example is given below to demonstrate this problem. Take $n=60$, $\mu_g=10$ and $\sigma_g=4.0$ for estimating $x_{0.99}$ ($=12.3263$). The initial value for RMJ is $x_1=19.3054$, which is much larger than 12.3263. Its y_1 is 1. The next two iterations are x_2 ($=19.228$), $y_2=1$ and x_3 ($=19.1548$), $y_3=1$. Each of the following 58 iterations makes a very small step toward 12.3263 and the corresponding y value is always 1. When it terminates, $x_{61}=17.2733$ is taken as the estimate of $x_{0.99}$ ($=12.3263$), with the bias equal to 4.9470. Because the initial x value is so far above the true value, the final x value after 60 iterations is still far above the true value. Therefore the same (x, y) sequence appears in each of the 1000 simulation samples. This leads to a zero simulation variance and the simulation bias and RMSE are both equal to 4.9470 as shown in Table 6A.

The previous example suggests that RMJ's performance may be affected by how far the starting value x_1 is from the true value x_p. In view of RMJ's excellent performance in many cases, we need to understand what factors affect its performance. Underneath each of Tables 5A–7A, we give Tables 5B–7B which lists the values of the difference $x_1 - x_p$ for the same set of μ_g

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

and σ_g values and for $p=0.9, 0.99, 0.999$, where the starting value $x_1=\mu_g+\Phi^{-1}(p)\sigma_g$ and x_p is computed under the normal distribution (with $\mu=10$ and $\sigma=1$). We use shade to highlight entries in these three tables that correspond to situations in which RMJ performs worse than 3pod. All of them occur when both the σ_g values and the x_1-x_p values are large, which will be illustrated later in Tables 9B–10B. On the other hand, a truly outstanding performance of RMJ may be attributed to a good choice of the starting value. In the ideal case of $\mu_g=10$ and $\sigma_g=1$, $x_1-x_p=0$ in Tables 5B–7B, RMJ performs the best at $\sigma_g=1$ than at other σ_g values. For example, in Table 5A under $\mu_g=10$, RMJ has the smallest RMSE 0.2632 among all σ_g values. Same for Tables 6A–7A. However, no such strong association between RMSE and x_1-x_p can be observed in other parts of Tables 5A–7A. Other observations from these tables are given below:

1. Only 3pod and Neyer perform consistently over the range of σ_g values. And 3pod outperforms Neyer's in most cases.
2. For the estimation efficiency of 3pod, it is better to devote more samples (i.e., $n_2=35$ over 30 and $n_2=55$ over 45) to phase III, which concerns approximation. But a larger n_1 value (i.e., smaller n_2 value) has the advantage of taking fewer wasted runs as shown in Tables 2–4. How to trade-off these two objectives is a question of further interest.
3. Overall Neyer performs better than Wu and is more stable.
4. To save space, we do not report the results of the up-and-down method (available in supplemental materials). Even in cases where it takes a small number of wasted runs, it is consistently worse than RMJ in terms of the RMSE. For $\sigma_g=0.5$, it is also worse than 3pod while for $\sigma_g=1.0$, it beats 3pod.

In the third part of the simulation study, we repeated part two by changing the distribution from normal to logistic (with $\mu=10$ and $\sigma=1$ in $LG((x-\mu)/(\sigma/1.8138))$; see definition in the beginning of the section). The results are summarized in Tables 8A–10A and 8B–10B, using the same format as in Tables 5A–7A and 5B–7B. The RMSE values in Tables 8A–10A are generally larger than those in Tables 5A–7A. This is expected because the procedures should perform less well when the assumed distribution is different from the true distribution. The overall pattern in the comparisons among the four procedures is the same as before. The only new observation is that 3pod outperforms RMJ, but only slightly, in two additional cases: (i) $n=60$, $\mu_g=9$ and $\sigma_g=2$, (ii) $n=80$, $\mu_g=11$ and $\sigma_g=1$. Note that these happen for small σ_g values. Another difference, which has less significance, is where Neyer beats the other procedures. Recall that there are three such cases for the normal distribution (two in scenario B and one in scenario C). For the logistic distribution, there are four cases all with $\sigma_g=4$: (i) $n=60$, $\mu_g=9, 10$ (ii) $n=80$, $\mu_g=9, 10$. As we remarked in the corresponding discussion for the normal distribution, Neyer's better performance should be weighed against the large number of wasted runs it accumulates before reaching 1000 successful runs. From Tables 3–4, the number of wasted runs for $\sigma_g=4$ and $n=60, 80$ ranges from 1971 to 2334. Such large numbers suggest that Neyer's method cannot function properly. In retrospect, Neyer should have been removed in the ensuing simulation studies on efficiency for these cases. There are also minor differences in the relative rankings of the worst two procedures in some cases. Details on these will not be reported here as they matter less than finding out who are the best two. To summarize, other than the cases discussed above, the rankings in scenarios A–C for the normal distribution carry over to the logistic distribution.

Next we turn to the comparisons between Tables 8B–10B versus Tables 5B–7B. The former has 18 shaded entries as opposed to 16 in the latter. That is, for the logistic distribution there are two additional cases in which RMJ performs poorly than for the normal distribution. These two cases were discussed in the previous paragraph. For the 16 shaded cases for normal and the corresponding 16 cases for logistic, the values of x_1-x_p are high. This confirms the previous observation on the poor performance of RMJ when the starting value x_1 is far from the true value x_p (near the end of discussion on part two of simulation). For the two new cases for logistic, the corresponding values of x_1-x_p are not high but should be discounted

Table 8A
RMSE for estimation of $x_{0.9}$, $n=40$ ($n_1=25$, $n_2=15$ for 3pod), true distribution=logistic.

	$\sigma_g=0.5$	$\sigma_g=1.0$	$\sigma_g=2.0$	$\sigma_g=3.0$	$\sigma_g=4.0$
$\mu_g=9$					
Neyer	0.4794	0.4991	0.5250	0.4733	0.5723
3pod	0.4785	0.4594	0.4909	0.4864	0.5239
Wu	0.3889	0.3685	0.4164	0.3738	0.5156
RMJ	0.3020	0.2785	0.3362	0.3794	0.4353
$\mu_g=10$					
Neyer	0.4549	0.4616	0.4973	0.4967	0.5017
3pod	0.4798	0.4559	0.4992	0.5003	0.5289
Wu	0.4001	0.3626	0.4010	0.5047	0.4733
RMJ	0.2916	0.2838	0.3386	0.3923	0.4698
$\mu_g=11$					
Neyer	0.5592	0.4871	0.5026	0.6149	0.7523
3pod	0.4960	0.4705	0.4845	0.4947	0.5416
Wu	0.5223	0.3713	0.4348	0.5524	0.6041
RMJ	0.3010	0.2975	0.3454	0.4122	0.6396

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

Table 8B
Values of $x_1 - x_{0.9}$, x_1 = starting value, $x_{0.9}$ = true value under logistic.

	σ_g=0.5	σ_g=1.0	σ_g=2.0	σ_g=3.0	σ_g=4.0
μ_g=9	−1.5706	−0.9298	0.3517	1.6333	2.9148
μ_g=10	−0.5706	0.0702	1.3517	2.6333	3.9148
μ_g=11	0.4294	1.0702	2.3517	3.6333	4.9148

Table 9A
RMSE for estimation of $x_{0.99}$, n=60, true distribution=logistic.

	$\sigma_g = 0.5$	$\sigma_g = 1.0$	$\sigma_g = 2.0$	$\sigma_g = 3.0$	$\sigma_g = 4.0$
$\mu_g = 9$					
Neyer	0.8520	0.8115	0.9090	0.8293	0.5909
3pod ($n_1 = 25, n_2 = 35$)	0.7036	0.6632	0.7188	0.7652	0.7961
3pod ($n_1 = 30, n_2 = 30$)	0.7434	0.6894	0.7580	0.7283	0.7664
Wu	1.3791	1.2497	1.9453	1.7400	1.5708
RMJ	0.5560	0.5424	0.7656	1.5710	4.1978
$\mu_g = 10$					
Neyer	0.7832	0.7918	0.8381	0.7690	0.5391
3pod ($n_1 = 25, n_2 = 35$)	0.7263	0.6816	0.6987	0.7532	0.8567
3pod ($n_1 = 30, n_2 = 30$)	0.7439	0.7098	0.7113	0.7483	0.7805
Wu	1.3991	1.2828	1.8885	1.8354	1.4246
RMJ	0.5571	0.5245	0.6720	2.5311	5.1978
$\mu_g = 11$					
Neyer	1.0202	0.8506	0.9447	1.0570	0.9349
3pod ($n_1 = 25, n_2 = 35$)	0.7234	0.6615	0.7076	0.7755	0.8167
3pod ($n_1 = 30, n_2 = 30$)	0.7548	0.7154	0.6974	0.7735	0.8101
Wu	1.2768	1.0170	1.4854	1.0816	1.6489
RMJ	0.5500	0.5734	0.9380	3.5311	6.1978

Table 9B
Values of $x_1 - x_{0.99}$, $x_1 =$ starting value, $x_{0.99}$ = true value under logistic.

	σ_g=0.5	σ_g=1.0	σ_g=2.0	σ_g=3.0	σ_g=4.0
μ_g=9	−2.3702	−1.2071	1.1193	3.4456	5.7720
μ_g=10	−1.3702	−0.2071	2.1193	4.4456	6.7720
μ_g=11	−0.3702	0.7929	3.1193	5.4456	7.7720

Table 10A
RMSE for estimation of $x_{0.999}$, n=80, true distribution=logistic.

	$\sigma_g = 0.5$	$\sigma_g = 1.0$	$\sigma_g = 2.0$	$\sigma_g = 3.0$	$\sigma_g = 4.0$
$\mu_g = 9$					
Neyer	1.3811	1.3342	1.4745	1.4232	0.9936
3pod ($n_1 = 25, n_2 = 55$)	1.1383	1.0505	1.0080	1.2161	1.2226
3pod ($n_1 = 35, n_2 = 45$)	1.1673	1.1306	1.0219	1.1463	1.1749
Wu	1.3694	1.7279	2.4104	2.3432	2.2477
RMJ	1.0064	0.7154	0.6144	3.9623	7.2339
$\mu_g = 10$					
Neyer	1.2837	1.3385	1.3866	1.2922	0.8834
3pod ($n_1 = 25, n_2 = 55$)	1.1551	1.0691	1.0868	1.1375	1.0871
3pod ($n_1 = 35, n_2 = 45$)	1.2039	1.1671	1.0460	1.1950	1.1259
Wu	1.3537	1.8082	2.4986	2.4748	2.2794
RMJ	1.0010	0.7640	1.4695	4.9623	8.2339
$\mu_g = 11$					
Neyer	1.5355	1.3289	1.4696	1.6232	1.3817
3pod ($n_1 = 25, n_2 = 55$)	1.1489	1.0672	1.0044	1.1005	1.1815
3pod ($n_1 = 35, n_2 = 45$)	1.2055	1.1605	1.0002	1.1203	1.1487
Wu	1.5319	1.4790	2.2917	1.9754	2.2671
RMJ	1.0108	1.1927	2.4602	5.9623	9.2339

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

Table 10B
Values of $x_1 - x_{0.999}$, x_1 = starting value, $x_{0.999}$ = true value under logistic.

	σ_g=0.5	σ_g=1.0	σ_g=2.0	σ_g=3.0	σ_g=4.0
μ_g=9	−3.2628	−1.7177	1.3726	4.4628	7.5530
μ_g=10	−2.2628	−0.7177	2.3726	5.4628	8.5530
μ_g=11	−1.2628	0.2823	3.3726	6.4628	9.5530

in the discussion on dependency on $x_1 - x_p$ because RMJ lags only slightly behind 3pod in these two cases. Since most of the shaded entries occur for large σ_g, it suggests that another factor for the poor performance of RMJ is having large σ_g. This observation deserves an explanation. First we note that, as the σ value in the RMJ procedure increases, the β value in Eq. (3) decreases and likewise a_i in Eq. (3) decreases as it is proportional to β. A smaller a_i in the RMJ iterations in Eq. (2) means a smaller increment it makes in the iterations. This will slow down the convergence of the RMJ sequence if its starting value is far from the true value. The same argument shows that a small σ value in the RMJ procedure gives a bigger increment in the RMJ iterations. Therefore it is less affected by having a starting value far from the true value.

The final observation concerns the allocation of sample sizes n_1 and n_2 in 3pod. As in the normal case, a larger n_2 is generally preferred. But there are exceptions in the logistic case (i.e., μ_g=9, 10, σ_g=3, 4 in Table 9A; μ_g=9, σ_g=3, 4 and μ_g=11, σ_g=4 in Table 10A). Why does this happen only in the logistic case? We have no good explanation.

Based on the discussions for the normal and logistic distributions, two clear winners have emerged: 3pod and RMJ. Between them, we make the following recommendations:

(i) When σ_g is close to the true σ value or when σ_g is small, the Robbins–Monro–Joseph (RMJ) method is the best.
(ii) When σ_g is larger than the true σ value, the 3pod method is the best in most situations.
(iii) Because of the very unstable performance of RMJ for large guessed values of σ and the difficulty for practitioners to come up with a good guessed value, 3pod is overall a safer (but not necessarily better) choice.

6. Concluding remarks and further discussion

Research on sensitivity testing has a long history since the early work of Dixon and Mood (1948). It has periodically attracted attention among researchers and users. Although the problem has not been a hot topic in recent decade, finding an efficient and robust sensitivity testing procedure remains an outstanding challenge because there has not been a consensus on which is the best procedure. Among the better informed users, Neyer's (1994) method has been viewed as a preferred choice. Our research was partly motivated by the desire to find a better procedure than Neyer's.

The proposed 3pod procedure consists of three phases. The purposes of the three phases can be described as those of *search, estimate and approximate* respectively. The most original and novel is phase I, which tries to quickly achieve an overlapping data pattern and then to enhance the pattern. Even though Wu (1985) recognized the importance of achieving overlapping pattern in sequential sensitivity testing, he did not incorporate it in the design procedure. Neyer was the first to incorporate the achieving of overlapping pattern in the design strategy. This is quite novel because it works on *getting data to ensure parameter estimability* while standard practice in optimal design is to achieve estimation efficiency by assuming parameter estimability. For experiments with large samples, estimability is not an issue. However, for sensitivity testing with small samples due to expensive test items or short duration, parameter estimability is not even guaranteed. Thus a good procedure should address this issue. On the other hand, procedures like stochastic approximation do not face this issue because it does not require estimation of parameters in an underlying model. Turning to phase II of 3pod, it is the same as Neyer's use of *D*-optimal design to spread out stress levels. Here the focus is on getting estimation efficiency. Since the purpose of phase III is to accelerate convergence to the unknown quantile x_p after the region of interest is identified in phases I and II, we use Joseph's (2004) modification of the Robbins–Monro procedure for binary data (dubbed as RMJ in the paper).

From the extensive simulation study, 3pod and RMJ have emerged as the overall winners. RMJ performs excellently in many cases but can suddenly deteriorate when the starting value is far from the true value or when σ_g, the guessed value of σ, is large. In contrast, 3pod performs more steadily and has the best performance when RMJ does poorly. Another conclusion from the simulation study is that 3pod consistently outperforms Neyer. Two reasons account for the better performance of 3pod. First, it employs a more elaborate and efficient search scheme in phase I than the binary search in Neyer's method. Second, it uses the RMJ in its phase III to accelerate convergence. Neyer does not have such a provision in its procedure. Given the excellent performance of RMJ, this may be the more important reason.

One unique feature of the 3pod procedure is that it consists of "modules" that can be reassembled for other purposes. Its I1, stage 1 of phase I, is to obtain a response (R) and a nonresponse (NR); I2 is to achieve overlapping; II is for optimal parameter estimation and III is for quick approximation. These four modules: I1(R-NR), I2 (overlap), II (estimate), III (approximate), may be deployed in a different order to suit the purpose of the experiment. For example, in an ongoing work

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007

on sequential testing with hybrid response (i.e., first part binary and second part continuous only if $y=1$ is observed), we have used these modules in a different order. Another possible application is to use I1 first and then switch to RMJ. Since RMJ can behave erratically if its starting value is far from the true value and σ_g is large, this modification can avoid the problem of having many initial runs with the same y value as demonstrated in the example in Section 5.

The excellent performance of RMJ in many cases came as a surprise since Joseph's (2004) work has received scant attention in the literature. A key and remaining question is how to choose between 3pod and RMJ. RMJ seems to do well for small guessed values of σ. How to quantify this in a practical situation? This will require further investigation. Since 3pod has a good and robust performance throughout the simulation study, it may be tempting to make it an overall choice. But it underperforms RMJ when the latter does well. We think there is room for 3pod to be further improved. One possibility is to find a way to modify and streamline phases I and II so that 3pod can devote more samples to phase III to take advantage of the good properties of RMJ. Since the starting value for RMJ in phase III is based on all data in phases I and II (see Eq. (8)), a slightly shortened phases I and II should still supply a good starting value for III. For example, should I3 (to enhance overlapping) be dropped to give the saved samples for phase III? How to shorten phases I and II is an issue that warrants further investigation.

There is finally the question of estimation after the data collection is completed. In general, the issue of design and that of estimation are decoupled in the present context. The investigator can choose to use a frequentist or Bayesian approach for estimation regardless of the strategies used in sequential design. But the method of estimation also depends on the parameters of interest. For example, if the interest is in some extreme percentile x_p with p close to 0 or 1, both the RMJ and 3pod use the last x value (when the design is terminated) as the estimate of x_p. There are two justifications. First, it is known that the Robbins–Monro type stochastic approximation (which is shared by RMJ and phase III of 3pod) is asymptotically consistent, i.e., its sequence converges to x_p regardless of the assumptions on the underlying distribution F. If one tries to use the whole data to estimate an extreme percentile, it will only work if the parameter assumption on F is valid. Otherwise, the estimate can be severely biased. However, if the parameter of interest is the median or a moderate percentile, such estimate will not be much affected by the parameter assumption. If the interest lies in estimating the location and scale parameters μ and σ in model (1), then the whole data can be used and the validity of estimation depends on the same parametric assumption. Because 3pod collects data over a reasonable range of x values in its phase I, the whole data can also be used to estimate other parameters of interest such as those discussed above.

Acknowledgments

Wu's research is supported by ARO Grant W911NF-08-1-0368 and NSF Grant DMS 1007574. The authors are grateful to Dianpeng Wang for his meticulous assistance in the simulation study, to Paul Roediger for helpful comments and to David Steinberg for an insightful suggestion on rescaling the σ parameter in the logistic distribution in the simulation study.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jspi.2013.10.007.

References

Dixon, W.J., Mood, A.M., 1948. A method for obtaining and analyzing sensitivity data. J. Am. Stat. Assoc. 43, 109–126.
Dror, H.A., Steinberg, D.M., 2008. Sequential experimental designs for generalized linear models. J. Am. Stat. Assoc. 103, 288–297.
Joseph, V.R., 2004. Efficient Robbins–Monro procedure for binary data. Biometrika 91, 461–470.
Joseph, V.R., Tian, Y., Wu, C.F.J., 2007. Adaptive designs for stochastic root-finding. Stat. Sin. 17, 1549–1565.
Lai, T.L., Robbins, H., 1979. Adaptive design and stochastic approximation. Ann. Stat. 7, 1196–1221.
Langlie, H.J., 1962. A Reliability Test Method for One-Shot Items, Publication U-1792. Aeronutronic (Div. of Ford Motor CO.), Newport Beach, CA.
Morgan, B.J.T., 1992. Analysis of Quantal Response Data. Chapman and Hall, London.
Neyer, B.T., 1994. *D*-optimality-based sensitivity test. Technometrics 36, 61–70.
Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. 29, 373–405.
Rothman, D., Alexander, M.J., Zimmerman, J.M., 1965. The Design and Analysis of Sensitivity Experiments, vol. 1. NASA CR-62026.
Silvapulle, M. J., 1981. On the existence of maximum likelihood estimators for the binomial response models. J.R.Stat.Soc., Ser. B 43, 310–313.
Wetherill, G.B., 1963. Sequential estimation of quantal response curves. J. R. Stat. Soc., Ser. B 25, 1–48.
Wu, C.F.J., 1985. Efficient sequential designs with binary data. J. Am. Stat. Assoc. 80, 974–984.
Young, L.J., Easterling, R.G., 1994. estimation of extreme quantiles based on sensitivity tests: a comparative study. Technometrics 36, 48–60.
Yang, M., Stufken, J., 2009. Support points of locally optimal designs for nonlinear models with two parameters. Ann. Stat. 37, 518–541.

Please cite this article as: Wu, C.F.J., Tian, Y., Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.10.007