BIOMETRICS 55, 396–402
June 1999

Two-Stage Design of Quantal Response Studies

R. R. Sitter

Department of Mathematics and Statistics, Simon Fraser University,
Burnaby, British Columbia V5A 1S6, Canada
email: sitter@cs.sfu.ca

and

C. F. J. Wu

Department of Statistics, University of Michigan,
Ann Arbor, Michigan 48109-1027, U.S.A.

SUMMARY. In a quantal response study, there may be insufficient knowledge of the response relationship for the stimulus (or dose) levels to be chosen properly. Information from such a study can be scanty or even unreliable. A two-stage design is proposed for such studies, which can determine whether and how a follow-up (i.e., second-stage) study should be conducted to select additional stimulus levels to compensate for the scarcity of information in the initial study. These levels are determined by using optimal design theory and are based on the fitted model from the data in the initial study. Its advantages are demonstrated using a fishery study.

KEY WORDS: Binary data; c-optimality; D-optimality; F-optimality; Logit; Phase II trials; Probit.

1. Introduction

In a quantal response study, a subject is given a stimulus at level x and exhibits a response or nonresponse with probability $p(x)$ or $1-p(x)$. In medical or animal studies, x may be the dosage of a new formulation or drug and the response a positive reaction to the treatment; in sensitivity testing, x may be the pressure applied to explosives or other devices and the response an explosion; in economic valuation of recreational activities studies, x may be a dollar amount offered and the response an indication of willingness to pay this amount more for access to the activity. This type of study arises in a wide variety of scientific investigations. A central goal in such studies is to estimate some aspects of the quantal response curve $p(x)$, a function of x. Quite commonly, the pth percentile L_p or several percentiles are of interest to the investigator, where L_p satisfies $p(L_p) = p$. Alternatively, if a parametric model for $p(x)$ is assumed, interest may focus on the estimation of the parameters in the model, which leads to various optimality criteria.

Though many of the ideas throughout this paper hold more generally, for the purpose of illustration, we will restrict attention to a simple but commonly used class of models. We assume a parametric model $p(x) = H[\beta(x-\mu)]$ for the quantal response curve, where μ and β are unknown parameters and $H(\cdot)$ a specified distribution function. A number n_i of independent observations of this type are taken at k stimulus levels, $x_1, \ldots, x_k$, and the number of responses at each stimulus level, $r_1, \ldots, r_k$, are observed. Thus, $r_1, \ldots, r_k$ are mutually independent binomial random variables, $r_i \sim \text{bin}(n_i, p(x_i))$, with log-likelihood

$$L(\mu, \beta) \propto \sum_{i=1}^{k} r_i \log p(x_i) + (n_i - r_i) \log q(x_i).$$

The maximum likelihood estimators (MLE) $\hat{\mu}$ and $\hat{\beta}$ can then be found for μ and β.

The Fisher information matrix for $(\hat{\mu}, \hat{\beta})^{\mathrm{T}}$ under this model is

$$I(\mu, \beta) = n \begin{pmatrix} \beta^2 S_0 & -S_1 \\ -S_1 & S_2/\beta^2 \end{pmatrix}, \tag{1}$$

where

$$\begin{aligned} S_0 &= \Sigma_{i=1}^{k} \lambda_i w(z_i), \\ S_1 &= \Sigma_{i=1}^{k} \lambda_i z_i w(z_i), \\ S_2 &= \Sigma_{i=1}^{k} \lambda_i z_i^2 w(z_i), \\ z_i &= \beta(x_i - \mu), \\ \lambda_i &= n_i/n, \\ n &= \Sigma_i n_i, \end{aligned}$$

and $w(t) = \{H'(t)\}^2/[H(t)\{1-H(t)\}]$. For many of the most commonly used models, $H'(\cdot)$ is symmetric about zero and therefore the median stimulus level $L_{0.5} = \mu$.

A commonly used approach to these studies is what we call unistage designs. The basic framework of unistage designs for

quantal response experiments consists of choosing the number of stimulus levels k, the set of stimulus levels $\{x_i\}_{i=1}^k$, and the number of subjects at each level $\{n_i\}_{i=1}^k$. This is done before the experiment, and then the entire experiment is run.

Optimal design focuses entirely on the precision of estimates, which in most applications is measured by the asymptotic variance–covariance matrix of $(\hat{\mu}, \hat{\beta})$,

$$\mathrm{AV}(\hat{\mu}, \hat{\beta}) = I^{-1}(\mu, \beta).$$

Depending on the interest of the experimenter, different functions of this matrix may be used as the basis for comparing the precision of competing designs. Unfortunately, $I(\mu, \beta)$ depends on μ and β, which are unknown at the design stage. It is usually assumed that the experimenter has some *a priori* knowledge of the response curve, and good initial values μ_0 and β_0 are chosen based on this knowledge. These might be available from some previous related experiment(s), some preliminary dose-ranging study, or pretesting of some kind. Since μ_0 and β_0 are obtained from previous experimentation, one might argue correctly that unistage designs are in fact the second stage of a less formal multistage experiment. We distinguish this from what we term a two-stage experiment by the assumption that the related prior information is vague or unrelated enough that it is only useful in roughly characterizing the response curve and does not consist of raw data that can be grouped with the results of the second stage in a combined analysis.

If a percentile of H is of particular interest, a natural design criterion, called c-optimality, is to minimize the asymptotic variance of the percentile estimate $AV(\hat{L}_p) = c^{\mathrm{T}} I^{-1}(\mu, \beta) c$, where $\hat{L}_p = \hat{\mu} + \gamma_p/\hat{\beta}$, $c = (1, \{\mu - \gamma_p\}/\beta)^{\mathrm{T}}$, and $\gamma_p = H^{-1}(p)$ (Wu, 1988). Alternatively, one can use the length of a Fieller interval as a design criterion (Finney, 1971, Chapter 8; Sitter and Wu, 1993; Sitter and Fainaru, 1997). If estimation of μ and β are of equal interest, various optimality criteria based on $I(\mu, \beta)$ have been suggested. The most common example is D-optimality, which entails choosing the design to minimize the determinant of $I^{-1}(\mu, \beta)$ and amounts to maximizing $D = S_0 S_2 - S_1^2$ (Sitter and Wu, 1993; Sitter and Fainaru, 1997). These criteria all yield one-, two-, or three-point designs.

There are a number of major concerns with strictly adopting this approach to designing experiments: (a) often, good initial estimates of μ and β are not available, and these 1–3-point designs are not robust to poor initial values; (b) 1–3-point designs may not allow adequate model checking; and (c) the choice of optimal design depends on the assumed model, which may be incorrect. To address points (a) and (b), attempts have been made to incorporate the initial lack of knowledge about the parameters into the unistage design framework. Two of these are (i) Bayesian techniques (Chaloner and Larntz, 1989) and (ii) a minimax approach (Sitter, 1992). In the first, prior distributions are assumed on μ and β and computer intensive techniques are used to generate Bayesian designs. In the second, initial values μ_0 and β_0 are assumed to be the best guesses and the design is chosen to be robust over some region containing this point. These robust criteria tend to spread out the support to protect against different possible parameter values. Unfortunately, they usually do not perform much better even if the best initial guesses of the parameters are perfectly correct since the design is forced to protect against other possibilities.

A remedial measure is to conduct a follow-up study. If the response y (or nonresponse $1 - y$) can be observed in a short time, then a fully sequential design can be implemented, which determines the next dose level x_{t+1} based on the information in y_i and x_i, $i = 1, \ldots, t$ (Wu, 1985; Young and Easterling, 1994). Note that, in studies like sensitivity testing, education testing, or psychophysical research, a short response may be obtainable. But in many other quantal response studies, this is simply unrealistic. For example, in most clinical trials, the patient's response will take days to weeks to be observed. A compromise between efficiency and time is to conduct a two-stage (or multistage) design, which can take advantage of the information from the initial study to design a follow-up study and still not unduly prolong the study's duration. The purpose of this paper is to propose one such strategy.

One motivation for the present research is to improve the conduct of phase II trials. Insufficient knowledge about a new therapeutic substance may lead to the choice of a poor dosage regimen for a phase II trial. Since the doses in the follow-up phase III trial are influenced by these results, it may be discovered only after an expensive and time-consuming phase III trial has been started that the doses are improperly chosen and a phase II trial has to be repeated. This would result in wasted resources and delay approval times (McDonald, 1993). In this scenario, a better approach, as we advocate in this paper, is to design a two-stage study that uses the best knowledge available to design the first-stage study, analyzes its data, and then decides whether there is sufficient information in the data to proceed to a phase III trial. If not, a second-stage study that was already planned will be implemented so that more useful information on effective dosage can be obtained. The second stage is an option the investigator may forego if there is sufficient information in the first-stage study. So the proposed two-stage approach includes the traditional unistage approach as a special case. We should point out that the two-stage methodology is not limited to the conduct of phase II trials. Any study that shares the same features as described above may benefit from adopting this approach. One prominent example is animal studies with expensive subjects like monkeys. Another is economic valuation studies such as the one described subsequently.

Two-stage or multistage designs are not new. In linear models, it is a key aspect of response surface methodology and is often used to break confounding in fractional factorial designs (Box, Hunter, and Hunter, 1978). In nonlinear models, the situation is more complex since the information matrix depends on the unknown parameters. [See Minkin (1987) and Abdelbasit and Plackett (1983) for further discussion.]

The paper is organized as follows. The next section describes a 4-month study on tidal sport fishing that was redesigned after 2 months because the original design was poor. Section 3 proposes a two-stage procedure. In Section 4, this procedure is illustrated by considering gains that might have been made if it had been used in the sport fishing study. Advantages and disadvantages of the proposed methodology are discussed in the concluding section.

2. A Fishery Study

To attempt to value the tidal sport fishery in the Canadian province of British Columbia (BC) for use in making public policy decisions, the Department of Fisheries and Oceans (DFO) contracted the DPA Group Inc. to perform a large-scale study ("Economic Valuation of the BC Tidal Sport Fishery," prepared by the DPA Group Inc. for DFO in March 1985). Fishermen were interviewed as they returned to launch sites in four areas of the south coast of BC (Victoria, Port Alberni, Campbell River, and Sechelt) from July to October 1984. One of the primary questions that each fisherman was asked is:

> 16. Now imagine that the cost of fishing in BC tidal waters increased. If the cost of your fishing trip had been ___z___ dollars higher today, would you still have gone fishing?
> No ___ Yes ___.

This question was asked for various values of z and the number of yes and no responses was recorded. Thus, z is the stimulus level and the response is binary. Logistic regression was used for analysis with the main focus being estimation of the ED50. This estimate was then multiplied by an estimate of the number of angler days in a year to estimate the total value of the sport fishery.

DFO specified the same design for each region and pretested the questionaires. The original design consisted of 30 different dose levels ranging from \$1 to \$50 with an approximately equal number of subjects at each dose level. However, "[a]s the survey progressed, it became apparent that at the upper range of the dollar amounts, a substantial number of people ... would still have gone fishing at the increased fishing cost amount (Question 16). Consequently, effective September 1, 1984, the range of offer amounts was expanded to ... \$1 to \$100. ..." This was done by replacing 10 of the existing 30 dose levels by 10 new higher dose levels. Table 1 gives the orginal design, the revised design, and responses for the Victoria and Port Alberni areas.

This is not an example of a preplanned two-stage designed experiment. However, this survey exemplifies a situation where the proposed two-stage (or multistage) design strategy would have been ideal. Operational considerations driven by the large number of sites and interviewers spread over a large geographical region and considerations of randomization precluded the possibility of a fully sequential approach. But it is clear that performing the study in stages was quite feasible. Also, having decided to redesign on the basis of data analyzed up to that time, the investigators could have used the strategy that we will propose subsequently to choose the second-stage design.

Suppose we are in the position that the Stage I design has been run as in Table 1 with resulting data therein, and a second-stage design was to be chosen. Let us analyze the information that would be available for each region. We use logistic regression, as was done in the study. The parameter estimates and their respective estimated 95% confidence intervals are given in Table 2. We should note that, in both cases, comparing the deviance to a chi-squared distribution yielded a p-value between 0.05 and 0.1, suggesting that the model fit the data only marginally well. The estimated response curves $\hat{p}(x)$ from Stage I for both Victoria and Port Alberni are given in Figure 1. For Victoria, the $\hat{p}(x_{\mathrm{I}i})$ are $(0.042, 0.049, \ldots, 0.91, 0.93, 0.98)$, where $x_{\mathrm{I}i}$ refers to the Stage I design points, with 20 of the 30 dose levels representing 257 of 391 observations in the moderate range (0.1, 0.9) of the estimated response curve. Of these 20 dose levels, 7 were above the estimated ED50, $\hat{\mu}_{\mathrm{I}}$, and 13 below. For Port Alberni, the $\hat{p}(x_{\mathrm{I}i})$ are $(0.015, 0.016, 0.018, \ldots, 0.34, 0.38, 0.61)$, with only 10 of the 30 dose levels representing 308 of 994 observations in the moderate range (0.1, 0.9) of the response curve. Of these 10 dose levels, only the last, corresponding to 11 observations, was above the estimated ED50. One might presume that this poor coverage of the response curve and the resulting difficulties in checking the model were the reasons for redesigning this study. The question is how a redesign should be done.

Table 1
BC tidal sport fishery data

	Victoria				Port Alberni			
	Stage I		Stage II		Stage I		Stage II	
x_i	n_i	r_i	n_i	r_i	n_i	r_i	n_i	r_i
1	13	0	17	0	35	2	34	0
2	14	0	17	1	35	1	29	1
3	16	0			36	0	3	0
4	12	0	15	0	27	0	29	0
5	15	1	1	0	32	1	4	0
6	13	0	14	1	34	0	32	3
7	18	0			37	0		
8	12	0	12	0	33	0	33	2
9	11	0			31	1	7	1
10	12	1	13	5	41	0	28	1
11	14	2	1	0	33	0	5	0
12	16	5	16	7	27	4	31	3
13	13	4			32	2	4	0
14	9	2	18	9	33	0	28	2
15	13	4			32	0	2	0
16	17	8	12	8	39	2	31	1
17	11	5	1	0	35	3	3	0
18	13	7	12	10	38	4	31	4
20	15	7	15	8	39	6	30	6
22	15	9	14	8	37	4	35	8
24	9	5	17	12	37	3	31	5
26	13	12	16	10	35	3	35	3
28	15	11	11	9	27	4	39	4
30	10	6	17	16	34	11	28	3
32	12	9	17	14	31	7	32	5
34	14	10	15	12	34	12	31	5
36	13	11			34	7	3	0
38	13	9	16	14	30	12	33	5
40	16	15	1	1	35	10	3	1
42			16	14			32	12
46			15	12			28	9
50	4	4	17	17	11	6	30	11
55			16	15			29	6
60			16	15			31	11
65			10	9			29	7
70			13	13			28	13
75			14	13			28	17
80			12	12			29	14
90			15	15			29	16
100			13	13			31	19

Table 2
Stage I parameter estimates and 95% confidence intervals

Region	$\hat{\mu}_{\mathrm{I}}$	95% CI for μ	$\hat{\beta}_{\mathrm{I}}$	95% CI for β
Victoria	22.11	(20.21, 24.01)	0.15	(0.12, 0.18)
Port Alberni	45.23	(40.91, 49.55)	0.10	(0.08, 0.11)

3. A Two-Stage Design Procedure

We will now present a two-stage design procedure. The basic philosophy is that one should use two (and possibly three) stages in many experiments of this type. Noting that a unistage experiment is a special case of a two-stage experiment with zero observations in the second stage, any sound methodology for two-stage experimentation should allow the experimenter to evaluate the necessity of a second stage after the completion of the first-stage analysis. Also, if the first-stage data indicates that a different model is more appropriate, the second stage can be designed using this new model. This addresses point (c) given in Section 1.

Stage I. (1) Choose a robust design

$$\mathbf{d}_{\mathrm{I}} = \begin{cases} x_{\mathrm{I}1}, \ldots, x_{\mathrm{I}k_{\mathrm{I}}} & \text{design points} \\ n_{\mathrm{I}1}, \ldots, n_{\mathrm{I}k_{\mathrm{I}}} & \text{sample sizes} \end{cases}$$

using only part of available resources. Use a design with three to five points depending on the anticipated model and amount of prior information from subject knowledge and results from previous related experiments. For example, under a logit model, one could use a kk-design based on the minimax criterion chosen from Tables 1–3 of Sitter (1992, pp. 1149–1151).

(2) Perform the first stage of experimentation and fit a model to obtain estimates $\hat{\mu}_{\mathrm{I}}$ and $\hat{\beta}_{\mathrm{I}}$, estimated confidence intervals or regions, and estimated probabilities of response at the first-stage design points $\hat{p}(x_{\mathrm{I}1}), \ldots, \hat{p}(x_{\mathrm{I}k_{\mathrm{I}}})$.

Stage II. Now one must consider the cost versus gain of various second-stage designs, including the option of not performing a second-stage experiment at all. We will take the view that the experimenter's goal is to obtain estimates $\hat{\mu}$ and $\hat{\beta}$ (or some function of them) that are within c units of the truth with high confidence.

(1) Most important at this point is to ascertain the level of reliance that should be placed on first-stage estimates $\hat{\mu}_{\mathrm{I}}$ and $\hat{\beta}_{\mathrm{I}}$. One must realize that simply considering estimated confidence intervals or regions might be misleading since they themselves depend on $\hat{\mu}_{\mathrm{I}}$ and $\hat{\beta}_{\mathrm{I}}$.

We suggest first considering $\hat{p}(x_{\mathrm{I}1}), \ldots, \hat{p}(x_{\mathrm{I}k_{\mathrm{I}}})$ and the corresponding $n_{\mathrm{I}1}, \ldots, n_{\mathrm{I}k_{\mathrm{I}}}$. The better the first-stage design appears to have covered the moderate portion ($0.1 \leq p \leq 0.9$) of the response curve, the more reliance one should place on $\hat{\mu}_{\mathrm{I}}$ and $\hat{\beta}_{\mathrm{I}}$. There are two aspects to this: the number of moderate $\hat{p}(x_{\mathrm{I}i})$ and the total number of observations these points represent. It is also important to consider the balance of these points over the moderate portion of the curve, with ideally some points in both the upper and lower range. Few moderate $\hat{p}(x_{\mathrm{I}i})$, a small amount of data at these points, or a lack of balance indicate a high probability that $\hat{\mu}_{\mathrm{I}}$ and $\hat{\beta}_{\mathrm{I}}$ are poor estimates despite any indication to the contrary from estimated confidence intervals. This is demonstrated through simulation in Sitter and Wu (1998). In such cases, the experimenter should take a very conservative approach in steps 2 and 3 below by choosing a more robust second-stage design with more support points and should forgo step 4 below in favor of using all resources that were held back for use in Stage II.

If the first-stage design appears to have covered the response curve well, then not only is more reliance on $\hat{\mu}_{\mathrm{I}}$ and $\hat{\beta}_{\mathrm{I}}$ warranted but confidence intervals and/or regions based on them yield reliable measures of their precision and can be used to aid in decisions regarding the sample size n_{II} to be used in the second-stage design. This will entail repeating steps 2 and 3 below for a few values of n_{II} between zero and the maximum number allowable given resources.

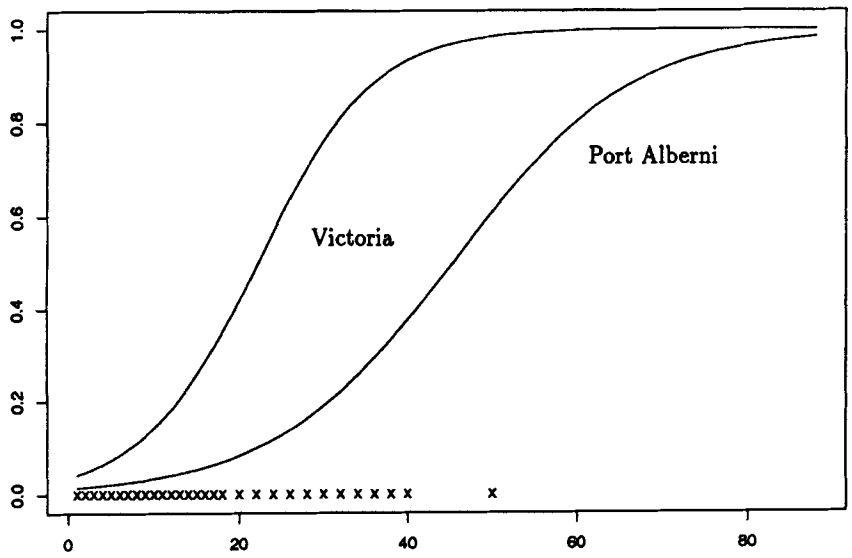

Figure 1. Estimated response curves from Stage I, with ×'s denoting the Stage I design points.

(2) For fixed n_{II}, use an optimality criterion to obtain an optimal second-stage design

$$\mathbf{d}_{II} = \begin{cases} x_{II1}, \ldots, x_{IIk_{II}} & \text{design points} \\ n_{II1}, \ldots, n_{IIk_{II}} & \text{sample sizes.} \end{cases}$$

Include the first-stage design points as fixed. For example, for the D-criterion, obtain $\mathbf{d}_{II}$ by

$$\max_{\mathbf{d}_{II}} \det I(\hat{\mu}_I, \hat{\beta}_I \mid \mathbf{d}_I, \mathbf{d}_{II}), \tag{2}$$

where $(\hat{\mu}_I, \hat{\beta}_I)$ are estimates based on $\mathbf{d}_I$ and the corresponding r_i values and det I is the determinant of the information matrix I in (1).

This optimization can be easily implemented in some common situations. The D-criterion is appropriate if some precision requirement on the parametric curve $H[\beta(\cdot - \mu)]$ is of interest. According to the theoretical results in Sitter and Forbes (1997) for the logit and probit models, we need only consider second-stage designs with two points (i.e., $k_{II} = 2$) of the form (λ, z), $(1-\lambda, -z)$ for some $0 \leq \lambda \leq 1$ and $z > 0$. Note that if $\lambda = 0$ or 1, this collapses to a one-point design. This reduces the problem to choosing

$$\lambda = h(z) = \begin{cases} 0 & \text{if } (1-\xi)/2 < 0, \\ (1-\xi)/2 & \text{if } 0 \leq (1-\xi)/2 \leq 1, \\ 1 & \text{if } (1-\xi) > 1, \end{cases} \tag{3}$$

where $\xi = \epsilon S_1'/\{(1-\epsilon)zw(z)\}$ and $\epsilon = n_I/n$, and maximizing

$$D(z) = [\epsilon S_0' + (1-\epsilon)w(z)][\epsilon S_2' + (1-\epsilon)z^2 w(z)] - [\epsilon S_1' + (1-\epsilon)(2h(z)-1)zw(z)]^2 \tag{4}$$

over $0 < z < \infty$. If z^* is the resulting maximizing value, the optimal second-stage design becomes $\{(\lambda^*, z^*), (1-\lambda^*, -z^*)\}$, where $\lambda^* = h(z^*)$.

It is interesting to consider $h(z)$. First, assuming that the true values of (μ, β) are $(\hat{\mu}_I, \hat{\beta}_I)$, one can see that, if the first-stage design is symmetric about $\hat{\mu}_I$, i.e., $S_1' = 0$, the D-optimal second-stage design will also be symmetric about $\hat{\mu}_I$. If the first-stage design is skewed left (right), i.e., $S_1' < 0$ (> 0), then $\lambda = h(z^*) > 1/2$ $(< 1/2)$, which implies a skewed right (left) second-stage design to compensate for the lesser information on the right (left). This parallels intuition.

Sitter and Forbes (1997) show that for the logit and probit models and any of the most commonly used criteria, i.e., A, c, D, E, F, G, the optimal second-stage design has two points. They also obtain a characterization of the resulting design similar to the above for the A, c, and E criteria.

(3) The number k_{II} of dose levels in optimal second-stage designs will usually be small (Sitter and Forbes, 1997). One may wish to increase k_{II}

(a) if there is a perception that one cannot rely on $\hat{\mu}_I$ and $\hat{\beta}_I$ (see step 1).
(b) if there is concern about the fit of the model. Then the experimenter may wish to ensure that the resulting combined design has enough points in the moderate range of the response curve to allow adequate model checking.
(c) after considering how the second-stage design changes when $\hat{\mu}_I$ and $\hat{\beta}_I$ are purturbed over some region. Then one may wish to choose the second-stage design as a compromise among the various optimal second-stage designs for each perturbation to protect against poor first-stage estimates.

(4) If it is felt that $\hat{\mu}_I$ and $\hat{\beta}_I$ are reasonably reliable (see step 1), one may also choose $n_{II} = \Sigma_i n_{IIi}$ considering the potential increase in accuracy and its resulting gains versus the additional cost of a second stage of experimentation. For example, if a particular function $\theta = g(\mu, \beta)$ is of interest, $\hat{\theta}_I = g(\hat{\mu}_I, \hat{\beta}_I)$ is its first-stage estimate and $\text{LEN}_{1-\alpha}(\theta, \mathbf{d})$ denotes the length of a $1-\alpha$ confidence interval for θ based on design $\mathbf{d}$, n_{II} can be determined by comparing $\text{LEN}_{1-\alpha}(\hat{\theta}_I, \mathbf{d}_I)$ (using only the 1st stage) with $\text{LEN}_{1-\alpha}(\hat{\theta}_I, \mathbf{d}_I, \mathbf{d}_{II})$ (using both stages) for various values of n_{II}, where $\mathbf{d}_{II}$ is obtained using steps 2 and 3.

A point that has not been discussed is the possibility of a stage effect due to performing the experiment at different times. One would hope that, by designing for two stages at the outset of an experiment, the possibility of inducing a stage effect can be minimized. If after the first stage there is concern that a stage effect may exist, one should consider this in choosing the second-stage design. In principle, a stage effect could be included in the model as a nuissance parameter and the second-stage design could be chosen based on some optimality criterion applied to the asymptotic variance–covariance matrix of μ and β only. Unfortunately, this would typically require prior knowledge of the size of the stage effect, which is not likely to be available. More informally, one should try to avoid a situation where $\mathbf{d}_I$ and $\mathbf{d}_{II}$ have no overlap in their coverage of the design space. If this is so, it becomes difficult to distinguish between a situation where each stage can be represented by the same functional form of model H with a stage effect present and the situation where the data has been generated by a different functional form for the model H.

4. An Illustration Using the Fishery Study

In order to do step 2, we must decide on n_{II}. For this experiment we may not have full control over this value as it is a function of the number of fishermen returning to the sites per unit time versus workload on an interviewer. However, since the redesign occurred near the middle of the planned duration of the survey, assuming no information on month-to-month fishing rates, we would assume that, if the number of interviewers in each region remained the same, n_{II} would be approximately equal to n_I. The actual n_I and n_{II} values turned out to be $n_I = 391$, $n_{II} = 445$ in Victoria region and $n_I = 994$, $n_{II} = 958$ in Port Alberni region. We will assume that the primary goal is to obtain an estimate of the ED50 that we are quite confident is within \$ 1.50 of the truth. That is an estimated confidence interval we can rely on that has length less than 3.0. We choose this value arbitrarily to aid in our illustration of the technique. Since more general information on the response curve is also desired, the D-criterion will be used.

Victoria. (1) Examining the $\hat{p}(x_{Ii})$, we see that the design did reasonably well in covering the moderate portion of the response curve with fair balance above and below the estimated ED50. This suggests that we can place a reasonably high level of reliance on $\hat{\mu}_I$ and $\hat{\beta}_I$ and the estimated confidence intervals in Table 2, and we may wish to proceed with step 4.

(2) For this illustration, we will first use $n_{II} = 445$, the actual observed value, to aid comparisons to the design that was actually used. In practice, we would likely have used n_{II} approximately equal to n_I, i.e., $\epsilon = 0.5$ in Section 3, but since the above choice yields $\epsilon = 0.51$, the results would be similar. Using the D-criterion and equations (3) and (4), we get the D-optimal Stage II design $\mathbf{d}_{II}$: $(n_1^*, n_2^*) = (167, 278)$ and $(x_1^*, x_2^*) = (11.5, 32.7)$.

(3) We will take the position that we are sure enough of $\hat{\mu}_I$ and $\hat{\beta}_I$ and that we feel the resulting combined design will have enough design points for adequate model checking if we use the recommended D-optimal design. If one is concerned with the fact that the logit model fits only moderately well and that the upper part of the response curve has fewer design points, one may decide to split the $n_2^* = 278$ observations among two or three points around $x_2^* \doteq 33$ chosen to ensure high efficiency to the optimal design.

(4) Note that the estimated 95% confidence interval for μ in Table 2 has a length of approximately \$4. Thus, if we chose not to perform a Stage II, we would not have attained our stated goal. The initial design allocated two interviewers to each region. One obvious question is, "Can we attain our stated goal with only one interviewer in Stage II?" If this is so, it may be possible to increase the number of interviewers in another region. It is this kind of option that may require planning of a two-stage study at the outset because one can build this aspect into the hiring of interviewers. If one has committed to hiring the interviewers in their specific regions for the entire duration, this option may not exist. With only one interviewer, we might anticipate n_{II} to be halved to 223, which implies $\epsilon = 0.64$. If we redo step 2 with this choice, we obtain a D-optimal design $\mathbf{d}'_{II}$: $(n_1^*, n_2^*) = (56, 167)$ and $(x_1^*, x_2^*) = (11.4, 32.8)$. The expected length of a 95% confidence interval using $\mathbf{d}_{II}$ is \$2.45 and using $\mathbf{d}'_{II}$ is \$2.87. Thus, we can likely attain our goal using only one interviewer in the Victoria region for the remainder of the study. This will be especially important in light of the results to follow for Port Alberni. We should note that the combined design used in the actual study had 65% efficiency in terms of the D-criterion relative to $\mathbf{d}_{II}$.

Port Alberni. (1) Examining the $\hat{p}(x_{Ii})$, we see that the design did very poorly in covering the moderate portion of the response curve and was highly unbalanced above and below the estimated ED50. This suggests that we cannot rely on $\hat{\mu}_I$ and $\hat{\beta}_I$ and the estimated confidence intervals in Table 2. Thus, we should be very conservative and use all the resources available. Given the results for Victoria, the experimenter might even consider having only one interviewer in that region while increasing to three interviewers in Port Alberni. Also, we should not use step 4.

(2) For this illustration, we will again use the actual observed value $n_{II} = 958$, which yields $\epsilon = 0.47$. The D-optimal Stage II design is $\mathbf{d}_{II}$: $(n_1^*, n_2^*) = (168, 790)$ and $(x_1^*, x_2^*) = (30.5, 59.9)$.

(3) A simple way to consider the effect on the second-stage design if the first-stage parameter estimates are incorrect is to repeat step 2 for different hypothetical $\hat{\mu}_I$ and $\hat{\beta}_I$. If one tries ranging $\hat{\mu}_I$ from 35 to 55 and $\hat{\beta}_I$ from 0.07 to 0.12, the optimal design points range from 11 to 47 for the lower point and from 48 to 74 for the upper point with the weight on the lower point between 0.12 and 0.34. The ranges for $\hat{\mu}_I$ and $\hat{\beta}_I$ were chosen to extend beyond the estimated 95% confidence bounds. One could then try various designs with, say, about 20% of the available observations spread over a few points in the lower range and the remaining spread over a few points in the upper range. For example, $\mathbf{d}'_{II}$: $\{n_{IIi}\} = (64, 64, 64, 255, 255, 256)$ and $x_{IIi} = (20, 30, 40, 50, 60, 70)$. The resulting combined design has 91% efficiency in terms of the D-criterion (D-efficiency) relative to the optimal design. One could instead use a more formal approach along the lines of Sitter (1992). The combined design in Table 1, which was used in the study, has a D-efficiency of only about 50% relative to the optimal design.

We have presented the above illustration using only information available after Stage I. We can now use the actual results from the second stage to emphasize some of the advantages and inherent difficulties with choosing a second-stage design in nonlinear situations. To do this, we fit a logistic regression model to the combined first- and second-stage data in Table 1. This is the model used in the original analysis. Since this is merely for illustration, we did not attempt any further model fitting. The estimated parameter values were Victoria, $\hat{\mu} = 21.11$ and $\hat{\beta} = 0.12$; and Port Alberni, $\hat{\mu} = 73.18$ and $\hat{\beta} = 0.042$. Let us assume that these are in fact the true parameter values and that a logistic model is correct. We can now reevaluate the expected performance of the above designs to see how they would likely have done. For Victoria, the $\hat{\mu}_I$ and $\hat{\beta}_I$ values were close enough to the truth that the expected performance would be similar to that presented above. For Port Alberni, this is not the case. For example, consider the length of 95% confidence intervals for μ. If we had ignored the warning signs and trusted the estimates from Stage I, the expected lengths of 95% confidence intervals for μ based on the first-stage data from $\mathbf{d}_{II}$ in step 2 and $\mathbf{d}'_{II}$ in step 3 would have been \$2.70 and \$2.73, respectively. Their true expected lengths using $\mu = 73.18$ and $\beta = 0.042$ would have been \$9.22 and \$9.59. Thus, we could have been mislead into believing a second stage was unnecessary when in fact this was far from the case. In fact, in this case, if the stated goal of a 95% confidence interval of length 3.0 is truly to be achieved, it will likely be necessary to perform a third stage. To choose a third stage, one can merely treat the combined first- and second-stage data as Stage I and reuse the procedure of Section 3.

5. Conclusion

It is common to use a unistage approach to design quantal response studies. While these designs are simple to administer, they may not be effective if prior knowledge is poor. In this paper, we have proposed a two-stage design strategy that has the following advantages: the second stage can "fix" mistakes made at the first stage by choosing stimulus levels to compensate for the lack of information afforded by the initial design; the two-stage approach can allocate more (less) resources to the first-stage design if there is less (more) confidence in the initial design choice; if the first stage is sufficiently informative, the second stage does not need to be invoked, thus the unistage design approach is included as a special case; by borrowing strength from data in both stages, the precision of estimates is enhanced; with the same resources, it can study more stimulus levels and achieve better quantification of the response curve. The performance of the two-stage procedure

relative to unistage strategies was investigated through simulation in Sitter and Wu (1998). The two-stage procedure performed better in terms of the relative bias and stability of the parameter estimates and the coverage and length of resulting confidence intervals.

A disadvantage is the possible time required between the two stages for amending protocol and securing investigators' commitments. Since the second stage is only invoked if it is determined from the data that there is insufficient information in the first-stage study, the time required for a second stage should not be viewed as a disadvantage. For example, as argued in Section 1, if such a study is not conducted in a Stage II trial, its downstream loss can be much more severe.

Acknowledgements

RRS was supported by the Natural Sciences and Engineering Research Council of Canada, and CFJW was supported by the National Science Foundation DMS-9704649. The authors thank Gordon Gislason and A. J. Petkau for providing the data in Table 1.

Résumé

Les connaissances nécessaires pour un choix adéquat des doses d'un essai par tout ou rien peuvent faire défaut. L'information obtenue avec de telles études peut être insuffisante, voire douteuse. On propose ici une méthode en deux étapes qui peut d'abord déterminer si il faut compléter une première expérience insuffisamment informative. La méthode indique ensuite comment préciser les niveaux additionnels de stimulus à sélectionner. On utilise pour cela la théorie des dispositifs optimaux appliquée aux résultats du modèle ajusté sur les données de la première étude. Les avantages de la méthode sont illustrés par une application à des données de pêcherie.

References

Abdelbasit, K. M. and Plackett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association* **78,** 90–98.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters, An Introduction to Design, Data Analysis, and Model Building.* New York: John Wiley.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference* **21,** 191–208.

Finney, D. J. (1971). *Probit Analysis,* 3rd edition. Cambridge: The Cambridge University Press.

McDonald, M. (1993). Dose-ranging studies: The key to registration. *Applied Clinical Trials* **2,** 50–58.

Minkin, S. (1987). Optimal design for binary data. *Journal of the American Statistical Association* **82,** 1098–1103.

Sitter, R. R. (1992). Robust designs for binary data. *Biometrics* **48,** 1145–1156.

Sitter, R. R. and Fainaru, I. (1997). Optimal designs for the logit and probit models for binary data. *Canadian Journal of Statistics* **25,** 175–189.

Sitter, R. R. and Forbes, B. (1997). Optimal two-stage designs for binary response experiments. *Statistica Sinica* **7,** 941–956.

Sitter, R. R. and Wu, C. F. J. (1993). Optimal designs for binary response experiments: Fieller, D, and A criteria. *Scandinavian Journal of Statistics* **20,** 329–342.

Sitter, R. R. and Wu, C. F. J. (1998). *Two-stage design of quantal response studies.* Research Report 98-1, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia, Canada.

Wu, C. F. J. (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association* **80,** 974–984.

Wu, C. F. J. (1988). Optimal design for percentile estimation of a quantal response curve. In *Optimal Design and Analysis of Experiments,* Y. Dodge, V. Fedorov, and H. P. Wynn (eds), 213–223. Amsterdam: Elsevier Science Publishers B.V. (North Holland).

Young, L. J. and Easterling, R. G. (1994). Estimation of extreme quantiles based on sensitivity tests: A comparative study. *Technometrics* **36,** 48–60.

Received June 1995. *Revised May* 1998.
Accepted June 1998.