Should we take measurements at an intermediate design point?*

Andrew Gelman†

January 5, 2000

Abstract

It is well known that, for estimating a linear treatment effect with constant variance, the optimal design divides the units equally between the two extremes of the design space. If the dose-response relation may be nonlinear, however, intermediate measurements may be useful in order to estimate the effects of partial treatments. We consider the decision of whether to gather data at an intermediate design point: do the gains from learning about nonlinearity outweigh the loss in efficiency in estimating the linear effect? Under reasonable assumptions about nonlinearity, we find that, unless sample size is very large, the design with no interior measurements is best, because with moderate total sample sizes, any nonlinearity in the dose-response will be difficult to detect. We discuss in the context of a simplified version of the problem that motivated this work—a study of pest-control treatments intended to reduce asthma symptoms in children.

Keywords: asthma, Bayesian inference, dose response experimental design, pest control, statistical significance

1 Introduction

1.1 An experimental design problem

It is well known that, when estimating a linear treatment effect with constant measurement variance, the optimal design is to find the two most extreme points on the design space (e.g., no treatment and some maximum feasible treatment level) and take half the measurements at each of extreme. If the dose-response relation may be nonlinear, however, intermediate measurements may be useful in order to estimate the effects of partial treatments. We consider the decision of whether to gather data at an intermediate design point: do the gains from learning about nonlinearity outweigh the loss in efficiency in estimating the linear effect? Setting up this problem goes beyond the usual paradigm of optimal experimental design because we must consider multiple inferential goals.

The decision of how to spread the measurements clearly depends on the sample size—or, equivalently, the ratio between the treatment effect and the measurement standard deviation. If the sample size is large enough, then there is power available to estimate a curved dose-response relation. Conversely, with small sample size, the estimation uncertainty will be so large that, even with

*To appear in *Biostatistics*. We thank David H. Krantz, Paul Meier, Yongzhao Shao, and the editors for helpful comments and the U.S. National Science Foundation for support through grant SBR-9708424 and Young Investigator Award DMS-9796129.

†Department of Statistics, Columbia University, New York, USA

intermediate measurements, any nonlinearity in the dose-response curve will most likely be undetectable, in the sense of being not statistically significant, in which case it makes sense to devote all data collection effort to the endpoints so as to estimate the linear trend in the treatment effect most efficiently.

We study this problem by setting up a simple model for a potentially nonlinear treatment effect and then consider the bias, variance, and mean squared error of various design/estimator combinations (as in Box and Draper, 1959, Jones and Mitchell, 1978, and Welch, 1983). We find that, with reasonable departures from linearity and moderate sample sizes, the simple design with no interior measurements has lowest mean squared error.

In a mathematical sense, the methods in this paper are straightforward. Our contribution is to use prior considerations to set up a range of reasonable sample sizes and departures from linearity over which to compare the competing designs. In addition to the relevance of our particular findings, we hope that this general approach will be useful in evaluating designs for other statistical problems.

1.2 Motivating application

This research was motivated by a study of an integrated pest management plan for reducing cockroach infestation and allergic sensitivity for inner-city children with asthma, supervised by Dr. Patrick Kinney of the Division of Environmental Health Sciences at Columbia University.

The treatment intervention, applied to cockroach-infested apartments, involves laying out poison and then returning a few weeks later to clean the apartment, seal off possible entry points, and instruct the residents on roach control measures. By far the most expensive part of the treatment is the cleaning and sealing of the apartment, which requires the services of several laborers for a day, at a total cost of about $700. The partial treatment would be to do a less effective half-day job.

The plan is to study 36 pairs of children, with a new pair enrolled in the study every 10 days, thus taking a year for the entire study, which conveniently averages over seasonal effects. In each pair, the treated child's apartment gets cleaned immediately and the control child's is cleaned eight months later. Units are followed up from enrollment until one year after intervention. The total costs of all the treatments is small compared to the budget of the entire study, so in our analysis we consider the sample size as fixed even if some units receive only the half treatment.

The experiment has further complications, but for the purposes of this paper, we consider only the basic analysis of the randomized experiment comparing outcomes between units. In fact, we consider an idealized version of this analysis, ignoring covariates such as pairing information, season, and measurements on pre-treatment variables such as child's asthma symptoms and apartment infestation level. This information would be included in a regression model, at which point our results here are relevant if we take them to refer to the treatment effects conditional on the predictor variables.

Finally, the endpoints of the applied analysis will include immediate outcomes such as cockroach infestation and indirect outcomes such as child's allergic sensitivity and asthma symptoms. For this paper, we assume that only a single continuous outcome is being measured.

2 Model

2.1 Notation

Consider a dose response of the form $g(x)$, where x can range from 0 (no treatment) to 1 (maximum possible treatment); see Figure 1. Assume we can afford to take n measurements with independent errors: $y_i \sim \mathrm{N}(g(x_i), \sigma^2)$, with the treatment levels x_i chosen by the experimenter. We consider designs with n_0, n_1, n_2 measurements at $x = 0, 0.5, 1$, respectively. The means have independent distributions $\bar{y}_x \sim \mathrm{N}(g(x), \sigma^2/n_x)$.

Our estimands of interest are the full treatment effect,

$$\theta_1 = g(1) - g(0),$$

which we assume is nonzero, and the effect of a half treatment,

$$\theta_{0.5} = g(0.5) - g(0).$$

It will be useful to define the relative nonlinearity of the treatment effect,

$$\delta = \frac{0.5\theta_1 - \theta_{0.5}}{\theta_1}. \tag{1}$$

We shall also be working with a quadratic model for the dose response, which we parameterize as

$$g(x) = \beta_0 + \beta_1(x - 0.5) + \beta_2(x - 0.5)^2. \tag{2}$$

We assume that we are interested only in estimating efficacy as a function of treatment level x; we thus ignore complications, such as multiple outcomes and toxicity, that would introduce other design criteria.

2.2 Designs and estimators

We consider a family of designs, indexed by a weight w between 0 and 1, with measurements at the two endpoints and the center:

$$n_0 = n_1 = (1 - w)n/2, \quad n_{0.5} = wn. \tag{3}$$

The special case $w = 0$ corresponds to the design with all data at the endpoints; $w = 1$ corresponds to the (nonsensical) design with all data at the center. We will focus on the designs $w = 0$ (equal sample sizes at 0 and 1) and $w = 1/3$ (equal sample sizes at 0, 0.5, and 1). The symmetry of the problem and the equal-variance assumption allows us to constrain $n_0 = n_1$.

As a function of w, we determine the mean squared errors of the following four estimates of the full and partial treatment effects, θ_1 and $\theta_{0.5}$.

1. **Linear regression.** If we fit a linear regression to the data, the treatment effects are determined directly from the regression slope, which, from the symmetry of the design, has a least-squares estimate of $\bar{y}_1 - \bar{y}_0$; thus:

$$\begin{aligned} \hat{\theta}_1^{\text{lin}} &= \bar{y}_1 - \bar{y}_0 \\ \hat{\theta}_{0.5}^{\text{lin}} &= 0.5(\bar{y}_1 - \bar{y}_0). \end{aligned} \tag{4}$$

2. **Quadratic regression.** Fitting a quadratic regression to the three data points is is equivalent to estimating $g(x)$ by $\bar{y}_x$ for $x = 0, 0.5, 1$, and thus yields

$$\begin{aligned} \hat{\theta}_1^{\text{quad}} &= \bar{y}_1 - \bar{y}_0 \\ \hat{\theta}_{0.5}^{\text{quad}} &= \bar{y}_{0.5} - \bar{y}_0. \end{aligned} \tag{5}$$

3. **Quadratic-if-significant.** A method that mimics standard practice is to fit a quadratic regression and then, if the quadratic term is statistically significant (that is, more than two standard errors away from 0), use the estimates (5), otherwise using the estimates (4) based on the linear regression fit. For our symmetric design, the least-squares estimate of the quadratic term is simply

$$\hat{\beta}_2 = 2\bar{y}_0 + 2\bar{y}_1 - 4\bar{y}_{0.5}, \tag{6}$$

with estimation variance

$$\text{var}(\hat{\beta}_2) = \frac{16}{w(1-w)} \frac{\sigma^2}{n} \tag{7}$$

The variance in (7) comes from evaluating the variance of (6) with the sample sizes in (3).

4. **Bayesian.** Fitting a Bayesian quadratic regression with a prior distribution on the curvature; as discussed in Section 3, a normal prior distribution on the nonlinearity parameter δ (see (1)) with mean 0 and standard deviation 1/4 might be reasonable. Since we are focusing on mean squared error, it makes sense to define the Bayes estimates of θ_1 and $\theta_{0.5}$ as their posterior means.

Computation for this Bayesian model is not trivial; expression (1) is a nonlinear function of the θ_x's and thus of the regression coefficients, and so a normal prior distribution on δ is not conjugate to the normal regression likelihood. Simulation-based computation of this model would be possible, but for the purposes of this paper it is enough to construct an approximate conjugate prior distribution by plugging the a point estimate $\hat{\theta}_1 = \bar{y}_1 - \bar{y}_0$ (note that this is both $\hat{\theta}_1^{\text{lin}}$ and $\hat{\theta}_1^{\text{quad}}$ above) into the denominator of (1) to yield a prior distribution for $0.5\theta_1 - \theta_{0.5}$, which is in fact equal to 1/4 times the quadratic term β_2 in (2). Assuming a $\mathrm{N}(0, (1/4)^2)$ prior distribution for δ then yields

$$\beta_2 \approx \mathrm{N}(0, (\bar{y}_1 - \bar{y}_0)^2).$$

The approximate posterior estimate of β_2 is a precision-weighted average of the estimate (6) from the data and the prior mean, which is 0; thus:

$$\hat{\beta}_2^{\text{Bayes}} = (2\bar{y}_0 + 2\bar{y}_1 - 4\bar{y}_{0.5})\lambda,$$

where λ is the Bayes shrinkage factor,

$$\lambda = \frac{(\bar{y}_1 - \bar{y}_0)^2}{(\bar{y}_1 - \bar{y}_0)^2 + \frac{16}{w(1-w)}\frac{\sigma^2}{n}}. \tag{8}$$

The approximate posterior mean estimates are then

$$\begin{aligned}
\hat{\theta}_1^{\text{Bayes}} &= \bar{y}_1 - \bar{y}_0 \\
\hat{\theta}_{0.5}^{\text{Bayes}} &= 0.5\hat{\beta}_1 - 0.25\hat{\beta}_2 = 0.5(\bar{y}_1 - \bar{y}_0) - 0.25(2\bar{y}_0 + 2\bar{y}_1 - 4\bar{y}_{0.5})\lambda.
\end{aligned}$$

The Bayes estimate depends on λ and thus σ (see (8)), which can be estimated from the data by the pooled variance of measurements within the three groups. In our analysis, we shall assume σ is known.

In the special case $w = 0$, the linear estimate is the only reasonable possibility: the quadratic estimate of $\theta_{0.5}$ is undefined, and the quadratic-if-significant and Bayes estimates reduce to the linear estimates.

2.3 Computing mean squared errors

We evaluate the designs and estimators by their mean squared errors (mse), which we will express as multiples of σ^2/n.

As noted above, the estimate for θ_1 is simply $\bar{y}_1 - \bar{y}_0$ for all methods; this has a bias of 0 and a variance (and, thus, mse) of $\frac{4}{1-w}\sigma^2/n$. Clearly, any motivation for setting $w > 0$ will come from the estimation of $\theta_{0.5}$.

The mse of $\theta_{0.5}$ has simple analytic forms for the linear and quadratic estimators. For the linear estimator, the bias is $0.5\theta_1 - \theta_{0.5} = \delta\theta_1$ and the variance is $\frac{1}{1-w}\sigma^2/n$, and so

$$\text{mse}(\hat{\theta}_{0.5}^{\text{lin}}) = \delta^2\theta_1^2 + \frac{1}{1-w}\frac{\sigma^2}{n} = \left(\delta^2 T^2 + \frac{1}{1-w}\right)\frac{\sigma^2}{n}.$$

For the quadratic estimator, the bias is 0 and the variance is $\frac{1}{w}\sigma^2/n + \frac{2}{1-w}\sigma^2/n$, and so the mse is, after simplifying,

$$\text{mse}(\hat{\theta}_{0.5}^{\text{quad}}) = \frac{1+w}{w(1-w)}\frac{\sigma^2}{n}$$

The mse's of the quadratic-if-significant and Bayes estimators can be most easily evaluated by simulation. We simply draw 1000 data vectors $(y_0, y_{0.5}, y_1)$ under the model, with means and variances determined by $g(0)$, $g(0.5)$, $g(1)$, σ^2/n, and w. Without loss of generality in evaluating mse as a multiple of σ^2/n, we can set $g(0) = 0$, $g(1) > 0$, and $\sigma^2/n = 1$. We can then write the dose response at the design points as a function of δ and T: $g(0) = 0$, $g(0.5) = (0.5 - \delta)T$, and $g(1) = T$, with measurement variances $\text{var}(\bar{y}_0) = \text{var}(\bar{y}_1) = \frac{2}{1-w}\sigma^2/n$ and $\text{var}(\bar{y}_{0.5}) = \frac{1}{w}\sigma^2/n$. Once we have simulated the data, we compute the estimates from each simulation draw, then compute the mse for each estimate as the mean squared difference between the estimate and the postulated true value, $\theta_{0.5} = (0.5 - \delta)T$. Note that, in computing the mse of the Bayes estimate, we are *not* averaging over the prior distribution or assuming that it is true.

3 Prior information

The properties of the estimates depend on the nonlinearity parameter $\delta = (0.5\theta_1 - \theta_{0.5})/\theta_1$ and the ratio of the treatment effect to the estimation uncertainty,

$$T = \frac{|\theta_1|}{\sigma}\sqrt{n}.$$

What prior information is available on δ and T?

We first consider δ. If the treatment effect is monotone, then δ must be between $-1/2$ and $1/2$ (see Figure 1). If we further constrain the treatment effect to be quadratic and monotone in the range $(0, 1)$, then δ must lie between $-1/4$ and $1/4$. The quadratic assumption is stronger than it may seem at first, however, since the assignment of an intermediate treatment to the value of $x = 0.5$ is somewhat arbitrary in many practical examples with qualitative treatments. It also seems reasonable, in the absence of other prior information, to set a unimodal prior distribution for δ centered at 0. If, for mathematical convenience, we wish to use a normal prior distribution, it seems reasonable to set the mean to 0 and the standard deviation at $1/4$ (which puts 95% of the prior mass in $[-1/2, 1/2]$).

We use our prior inference on δ in two ways. First, it is used in constructing the estimate $\hat{\theta}_{0.5}^{\text{Bayes}}$ above. Second,, the statistical properties of our estimators under the various designs depend crucially on δ; based on our prior considerations, we need only consider the range $|\delta| \leq 1/2$.

We do not put a formal prior distribution on T, but we shall attempt to get an idea of its order of magnitude by supposing that the experiment as initially designed (with $n_0 = n_1 = n/2$) has a sample size that is just about large enough that θ_1 can be statistically distinguished from zero. Suppose that means that the absolute value of θ_1 is in the range of 2 to 4 times as large as the standard deviation of its estimate. The variance of the simple estimate of θ_1 is $\frac{2}{n}\sigma^2 + \frac{2}{n}\sigma^2 = 4\sigma^2/n$, and so we are assuming

$$2(2\sigma/\sqrt{n}) < |\theta_1| < 4(2\sigma/\sqrt{n});$$

that is, $|T|$ is between 4 and 8. In other settings, an experiment may be designed to estimate parameters such as θ_1 very precisely, in which case $|T|$ could be much higher than 8.

4 Results

We compute the mean squared errors (mse) of $\hat{\theta}_1$ and $\hat{\theta}_{0.5}$, as a function of δ and T, for four design/estimator combinations: (1) $w = 0$ (measurements only at the endpoints) with the linear estimator, (2) $w = 1/3$ (equal sample sizes at all three design points) with the quadratic estimator, (3) $w = 1/3$ with the quadratic-if-significant estimator, and (4) $w = 1/3$ with the Bayes estimator. For $w = 0$, only the linear estimator is possible, since the others use $\bar{y}_{0.5}$, which does not exist if $w = 0$. Conversely, for $w \neq 0$, we need not consider the linear estimator since, if one were planning to use it, there would be no reason to gather data at the center point.

Figure 2 displays the mse's as a function of $|\delta| \in [0, 0.5]$, for $T = 4$ and 8, which correspond to a full treatment effect that is two or four standard deviations away from zero. In reading these graphs, we focus on the range $|\delta| \leq 1/4$, which corresponds to a quadratic and monotone treatment effect.

For $T = 4$—that is, a treatment effect that is on the border of statistical significance—the $w = 0$ design dominates as long as $|\alpha| < 0.25$. If $|\alpha|$ lies between 0.25 and 0.5, the $w = 1/3$ design performs better, but only if the Bayes estimate is used. The quadratic and quadratic-if-significant estimates perform much worse.

For $T = 8$, so that the linear treatment effect is clearly statistically significant, the $w = 0$ design dominates if $|\alpha| < 0.15$, and the $w = 1/3$ with Bayes estimator dominates if $|\alpha|$ lies between 0.15 and 0.5.

As T increases beyond 8, the $w = 1/3$ design eventually becomes best even for small values of α.

We repeated these calculations for other values of w and found that, when $w = 1/3$ is preferred to $w = 0$, it also does approximately as well as or better than other values of w between 0 and 1/3.

Thus we are satisfied with treating this problem as a comparison between the two choices of $w = 0$ and $w = 1/3$.

Given these results, we can make some recommendations:

- If the sample size is small, so that the treatment effect is expected to be on the border of statistical significance, then we recommend the simple design with data only at the extreme points. In addition to providing the best estimate of $\theta_{0.5}$ under the reasonable assumption that $|\delta| < 0.25$, this design is uniformly optimal for estimating the main treatment effect θ_1.

- If the sample size is large enough that the treatment effect can be estimated with high precision, *and* the dose-response might be far from linear, then we recommend collecting data at $x = 0$, 0.5, and 1, and using the Bayes estimate.

- In any case, if the sample size is moderate and measurements are taken at an intermediate design point, then the Bayes estimate with $\mathrm{N}(0, (1/4)^2)$ prior distribution on δ is preferable to the quadratic or quadratic-if-significant estimates. If it is known in advance that the Bayes estimate will *not* be used, then we do not recommend collecting data at the intermediate point.

5 Conclusions

5.1 General recommendations

We have compared two design for estimating the effects of a continuous treatment: the simple two-point design with data only at the extremes of design space, and a three-point design with data at the extremes and center of design space. Assuming fixed total sample size and constant measurement variance, the decision to collect data at an intermediate design point sacrifices efficiency in the estimation of the full treatment effect, θ_1, with the goal of more efficiently estimating $\theta_{0.5}$, the effect of a partial treatment.

Under reasonable assumptions on the magnitude of the nonlinearity of the dose-response, we find that, even for the goal of estimating $\theta_{0.5}$, the simple design outperforms the design with intermediate data unless the nonlinearity is large *and* the sample size is large enough that the main treatment effect is several standard deviations away from 0. Thus, we recommend that the simple design be used unless there is a large sample size and a suspicion of nonlinearity, in which case the three-point design is superior, as long as the Bayes estimate is used (as discussed in Section 4).

A natural area of further research is to consider multifactor designs and designs with more than three treatment levels.

5.2 Application to our example

In the experiment that motivated this project, described in Section 1.2, our sample size was fairly small, and variability between individuals was large enough that we doubted we would get highly significant treatment effect estimates. Thus, we chose the simple design with no data at the intermediate point. If the results of our experiment are promising, we may recommend large-scale public intervention to clean the apartments of inner-city children with asthma. In this case, we would expect different municipalities and different individuals to apply different levels of treatments (based on budgetary constraints if nothing else), and a large sample would be available for follow-up at a variety of different treatment levels.

References

Box, G. E. P., and Draper, N. R. (1959). A basis for the selection of a response surface design. *Journal of the American Statistical Association* **54**, 622–654.

Jones, E. R., and Mitchell, T. J. (1978). Design criteria for detecting model inadequacy. *Biometrika* **65**, 541–551.

Welch, W. J. (1983). A mean squared error criterion for the design of experiments. *Biometrika* **70**, 205–213.

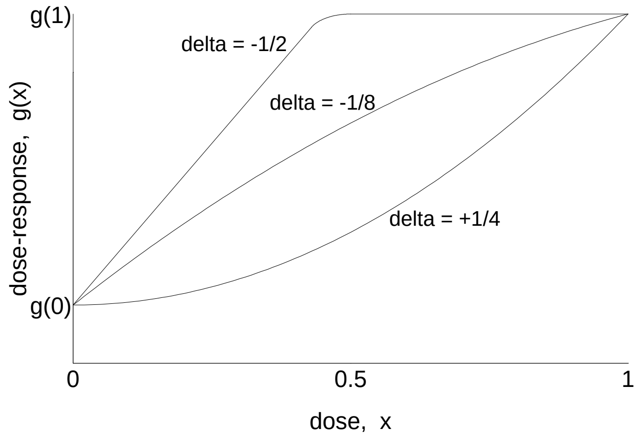

Figure 1: Dose response functions for different values of the nonlinearity parameter δ. If the curve is monotonic in $[0,1]$, then δ must be between $-1/2$ and $1/2$. If we further assume that the curve is quadratic, then δ must be between $-1/4$ and $1/4$.

Figure 2: Mean squared error (as a multiple of σ^2/n) for four design/estimator combinations of $\hat{\theta}_{0.5}$ as a function of $|\delta|$, the relative magnitude of nonlinearity of the dose-response. The plots show $T = 4$ and $T = 8$, which correspond to a treatment effect that is 2 or 4 standard deviations away from 0. The design $w = 0$ (all the data collected at the two extreme points) dominates unless both $|\delta|$ and T are large. When the design $w = 1/3$ (data evenly divided between the three design points) is chosen, the Bayes estimate has lowest mean squared error for the range of δ and T considered here.