Optimal Experimental Design for Another's Analysis

Ruth Etzioni, Joseph B. Kadane

Journal of the American Statistical Association, Volume 88, Issue 424 (Dec., 1993), 1404-1411.

Your use of the JSTOR database indicates your acceptance of JSTOR's Terms and Conditions of Use. A copy of JSTOR's Terms and Conditions of Use is available at http://www.jstor.org/about/terms.html, by contacting JSTOR at jstor-info@umich.edu, or by calling JSTOR at (888)388-3574, (734)998-9101 or (FAX) (734)998-9113. No part of a JSTOR transmission may be copied, downloaded, stored, further transmitted, transferred, distributed, altered, or otherwise used, in any form or by any means, except: (1) one stored electronic and one paper copy of any article solely for your personal, non-commercial use, or (2) with prior written permission of JSTOR and the publisher of the article or other text.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Journal of the American Statistical Association is published by American Statistical Association. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/asa.html.

Journal of the American Statistical Association
©1993 American Statistical Association

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2000 JSTOR

http://www.jstor.org/
Wed Mar 15 22:02:05 2000

Optimal Experimental Design for Another's Analysis

Ruth Etzioni and Joseph B. Kadane*

We consider the optimal design of experiments in which estimation and design are performed by different parties. The parties are assumed to share similar goals, as reflected by a common loss function, but they may have different prior beliefs. After presenting a few motivating examples, we examine the problem of optimal sample size selection under a normal likelihood with constant cost per observation. We also consider the problem of optimal allocation for given overall sample sizes. We present results under both squared-error loss and a logarithmic utility, paying attention to the differences between one- and two-prior optimal designs. An asymmetric discrepancy measure features repeatedly in our development, and we question the extent of its role in optimal two-prior design.

KEY WORDS: Adversarial design; Bayesian analysis; Conjugate prior; Optimal allocation; Posterior distribution; Predictive distribution.

1. INTRODUCTION

Dan and Edward will do an experiment together. Dan will choose the design, and Edward will choose an estimator. To simplify issues, assume that Dan and Edward share a loss function and normal likelihood with known precision and have different, but conjugate, prior distributions. Also, suppose that Dan knows Edward's prior beliefs. Under these circumstances, what is the best design for Dan to choose?

In practice, one often encounters experiments in which the data produced are evaluated by a decision maker other than the experimenter, with possibly differing prior beliefs. Kadane and Seidenfeld (1989) contrasted experiments to learn, which involve only one decision maker, with experiments to prove, which involve several decision makers. They cited the example of an inventor and proponent of a medical treatment who wished to design a study to convince a skeptical colleague that his treatment is more effective than the standard. Lindley and Singpurwalla (1991) considered the problem of a manufacturer who wished to perform an experiment to convince a consumer to accept his product. In a discussion of Bayesian approaches to clinical trials, Spiegelhalter and Freedman (1988) emphasized that reviewers and consumers, rather than experimenters, ultimately determine whether new treatments are adopted in clinical practice. They argued that inferences should convince those evaluating medical trials, despite the prior opinions of those performing the trials. A further example is that of statistical inference in legal contexts, where the prosecution gathers information to convince an impartial or objective court of the liability of the defendant (see, for example, Kadane 1990).

The earliest reference we have found that acknowledges the possibility of using different priors for estimation and design is Tsutakawa (1972), who considered a bioassay problem. But Tsutakawa did not incorporate at the design stage the knowledge that once the data have been collected, estimation will be conducted according to a different prior distribution. Burt (1990) and DasGupta and Studden (1991) considered design in a Bayesian context against a set of prior distributions.

If the prior beliefs of the party evaluating the data are known by the experimenter, then the experimeter can use this information to develop an optimal design; that is, one that will minimize his expected loss. The resulting design can be very different from the optimal design under the assumption that the party evaluating the data share the prior beliefs of the experimenter. Such situations were described as adversarial by Jackson, Novick, and DeKeyrel (1980) and Lindley and Singpurwalla (1991). Jackson et al. were concerned with the computation of one decision maker's expectation of his adversary's posterior distribution, but they did not formally address the design question. Lindley and Singpurwalla did address the design problem of optimal sample size selection; they considered a two-action scenario that is truly adversarial in that the experimenter and the decision maker have different utility functions as well as different prior distributions.

Denote Dan's prior density on θ by $\pi_d(\theta)$ and Edward's prior density by $\pi_e(\theta)$. Corresponding posterior distributions are $\pi_d(\theta|\mathbf{x})$ and $\pi_e(\theta|\mathbf{x})$, where $\mathbf{x}$ represents the sample of n observations and θ may be multidimensional. The goal of the study is to compute a, an estimate of θ. Because Edward is choosing the estimator, a is the Bayes estimate of θ with respect to $\pi_e(\theta|\mathbf{x})$; if the loss incurred in estimating θ by a is $L(\theta, a)$, a minimizes $E^{\pi_e(\theta|\mathbf{x})}L(\theta, a)$. But because Dan is designing the experiment, the optimal design minimizes over all possible designs $E^{p_d(\mathbf{x})}E^{p_d(\theta|\mathbf{x})}L(\theta, a)$, where $p_d(\mathbf{x})$ is, according to Dan's prior beliefs, the marginal distribution of the sample. If the two prior distributions coincide ($\pi_e = \pi_d$), then the estimate a and both expectations are computed with respect to the common prior distribution. In fact, one can think of the two-prior problem as a one-prior problem, where the prior is π_d, and Dan is attempting to design the sample optimally to estimate suboptimally.

This article examines two different design problems. First, we consider sample size selection for a normal likelihood. We assume conjugate prior distributions and constant cost,

* Ruth Etzioni is Associate in Biostatistics, Fred Hutchinson Cancer Research Center, Seattle, WA 98104. Joseph B. Kadane is Leonard J. Savage Professor of Statistics and Social Sciences, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213. The first author was supported by a National Institutes of Health Biostatistics training grant no. 5T32CA09168-17 while a senior fellow at the Department of Biostatistics, University of Washington while working on this article. The second author was supported by National Science Foundation Grants SES-8900025 and DMS-9005858 and Office of Naval Research Grant N00014-89-J-1851. The authors thank Kathryn Chaloner, Madhumita Lodh, and Isabella Verdinelli for helpful conversations and early readings.

© 1993 American Statistical Association

c, per observation and explore two different forms for $L(\theta, a)$. Section 2.1 considers squared-error loss, and Section 2.2 considers a logarithmic utility. We determine optimal sample sizes and derive conditions under which the optimal sample sizes exceed optimal sample sizes under only one prior distribution. The conditions are essentially inequalities on the quantity

$$r = \frac{\tau_e(\tau_e - \tau_d)}{\tau_d} + \tau_e^2(\mu_d - \mu_e)^2, \tag{1}$$

where μ_i and τ_i are prior means and precisions ($i = e, d$). Note that $r = 0$ when the two priors are the same and that r can be negative. Also, given τ_e, r is increasing in $|\mu_d - \mu_e|$ and decreasing in τ_d.

In Section 3 we consider the problem of optimal allocation of observations to k groups, given an overall sample size n. We give a general expression for the predictive expected loss to be minimized with respect to the proportion of observations in each group. We then extend results obtained in Section 2 to handle the multivariate case. A multivariate version of r is key in this development.

The designs we present are approximate in the sense that they assume that sample sizes may be real numbers rather than integers. In all the examples that we consider, the approximate designs can be rounded to the closest integer without significant increase in expected loss.

In Section 4 we summarize our findings and explore the extension of our results to different loss structures and further design problems.

2. DETERMINING OPTIMAL SAMPLE SIZES

First, consider sample size selection for a normal likelihood with known variance σ^2, and unknown mean θ. Without loss of generality, σ is taken to be equal to 1. Assume conjugate prior distributions—that is, $\pi_i(\theta) = N(\mu_i, 1/\tau_i)$—for $i = d, e$. The posterior distributions are

$$\pi_i(\theta|\bar{x}) = N\left(\frac{n}{n+\tau_i}\bar{x} + \frac{\tau_i}{n+\tau_i}\mu_i, \frac{1}{n+\tau_i}\right) \text{ for } i = d, e,$$

where $\bar{x}$ is the sample mean. Denote the posterior means and variances by $\mu_i(\bar{x})$ and σ_i^2.

2.1 Quadratic Loss

2.1.1 Optimal Sample Size. Under squared-error loss, the Bayes decision is the posterior mean. Thus Edward's estimate of θ is $\mu_e(\bar{x})$; however, Dan's posterior distribution is $N(\mu_d(\bar{x}), \sigma_d^2(\bar{x}))$. The posterior expected loss relevant for the sample size calculation is

$$E^{\pi_d(\theta|x)}(\theta - \mu_e(\bar{x}))^2$$
$$= E^{\pi_d(\theta|x)}(\theta - \mu_d(\bar{x}))^2 + (\mu_d(\bar{x}) - \mu_e(\bar{x}))^2. \tag{2}$$

Thus Dan's posterior expected loss of Edward's optimal action is simply Dan's posterior expected loss of his own optimal action plus a positive penalty. Dan's predictive risk of sampling n observations is then given by the expectation of (2) with respect to a $N(\mu_d, (1/\tau_d) + (1/n))$ distribution and can be written as $R(n) = cn + 1/(n + \tau_e) + r/[(n + \tau_e)^2]$, where r is defined in 1. Here cn is the sampling cost incurred by Dan; any setup costs are assumed to be independent of n. Although $R(n)$ is defined so far only on the nonnegative integers n, it is useful analytically to consider it to be defined on the whole nonnegative real line.

When $n = 0$, $R(n)$ reduces to Dan's prior expected loss of Edward's prior mean; that is, $R(0) = (1/\tau_d) + (\mu_d - \mu_e)^2$. The optimal sample size, n^*, minimizes $R(n)$ over $n \geq 0$. Let $y = \sqrt{c}(n + \tau_e)$. Then, equivalently, one may minimize

$$g(y) = y + \frac{1}{y} + \frac{\sqrt{c}r}{y^2} \tag{3}$$

over $y \geq \sqrt{c}\tau_e$, where r is defined in (1). If y minimizes $g(y)$ over this region, then $n^* = (y - \sqrt{c}\tau_e)/\sqrt{c}$ minimizes $R(n)$. This reexpression shows that although the original problem is expressed in terms of μ_e, μ_d, τ_e, and τ_d, only $\tilde{r} = \sqrt{c}r$ and $\tilde{\tau}_e = \sqrt{c}\tau_e$ matter for finding $\tilde{n}^* = \sqrt{c}n^*$.

Taking the first derivative of $R(n)$, the first order condition for n^* reduces to solving a cubic equation on a subset of the positive real line. From the properties of cubic equations, one can infer the value of n^*. Let $y(\tilde{r})$ denote the largest real root of the cubic $y^3 - y - 2\tilde{r} = 0$.

Theorem 1. If $\tilde{r} > (-1/3\sqrt{3})$ and $(-\tilde{r}/y^2(\tilde{r})) < \tilde{\tau}_e < y(\tilde{r})$, then $\tilde{n}^* = y(\tilde{r}) - \tilde{\tau}_e$. Otherwise, $\tilde{n}^* = 0$.

The proof is given in the Appendix. Figure 1 plots contours of $\tilde{n}^*$ as a function of $\tilde{r}$ and $\tilde{\tau}_e$. The region between the dotted lines is the region of positive $\tilde{n}^*$. Although the domain

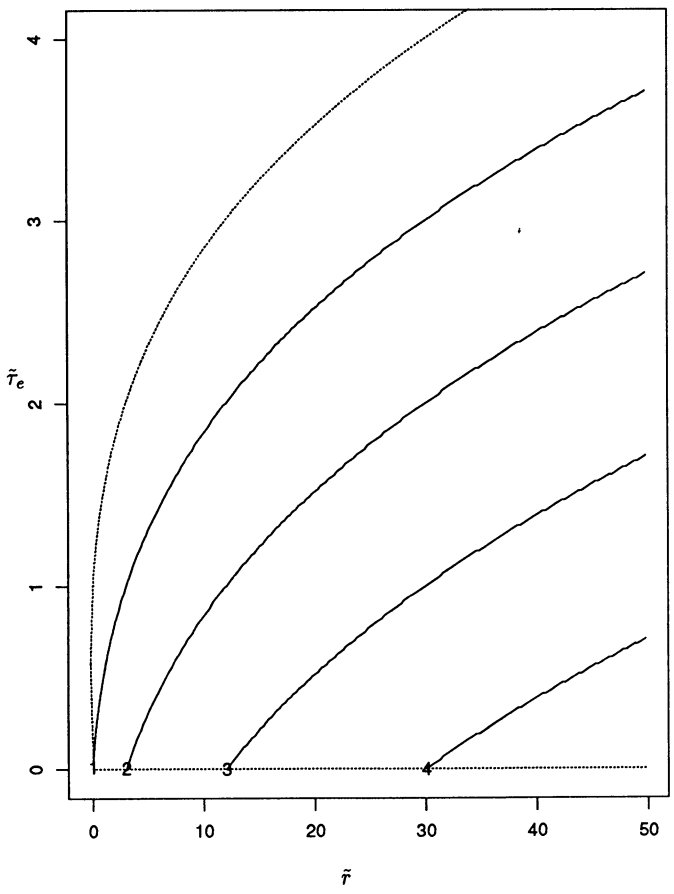

Figure 1. Contours of $\tilde{n}^*$ as a Function of $\tilde{r}$ and $\tilde{\tau}_e$. The region between the dotted lines is the region of non-0 $\tilde{n}^*$.

of $\tilde{r}$ in Figure 1 extends to $-1/(3\sqrt{3})$, the scale of $\tilde{r}$ is so large that it appears that $\tilde{r}$ extends only to 0. Several interesting aspects of the solution are revealed by this figure.

First, one can see the region in which the optimal value of $\tilde{n}^*$ is 0, toward the upper left of the figure. This corresponds to a region of large $\tilde{\tau}_e$ (i.e., high prior precision on Edward's part) and small $\tilde{r}$ (i.e., reasonable agreement between the two parties and high prior precision on Dan's part). The optimal sample size is also 0 when $\tilde{r}$ is sufficiently negative; that is, when τ_d exceeds τ_e, τ_e is bounded away from 0, and the prior means are close. Under these circumstances, it may not be worth Dan's while to do the experiment. For example, in the extreme case of ageement between the two parties ($\tilde{r} = 0$), the prior precision, interpretable as a prior sample size, may be so high as to make sampling simply unnecessary. In this case, $y(\tilde{r}) = y(0) = 1$. From the theorem, if $\tilde{\tau}_e < 1$, then $n^* = 1 - \tilde{\tau}_e$. If $\tilde{\tau}_e \geq 1$, then $n^* = 0$. Hence for $\tilde{\tau}_e > 1$, it becomes optimal not to sample.

If $\tilde{\tau}_e$ is fixed, then $\tilde{n}^*$ increases with $\tilde{r}$. In other words, $\tilde{n}^*$ increases as the prior means diverge and/or Dan becomes less certain about his prior mean. This can be understood intuitively by viewing the two-prior problem as a one-prior problem in which Dan is attempting to sample optimally to estimate suboptimally. Edward's posterior mean is a suboptimal estimate for Dan, but Dan can reduce his estimation penalty by bringing Edward's posterior mean close to his own. But doing so will require more observations, leading to higher sampling costs. For Dan, the crucial question is this: How many observations does he think he will need to bring the posterior means into reasonable agreement?

For fixed $\tilde{\tau}_e$, the answer is a function of both the difference in prior means and Dan's prior precision. Certainly, prior means that are further apart will require more data to bring them together a posteriori. In addition, as τ_d gets smaller, Dan's *predictive* distribution becomes more diffuse about his prior mean; this increases the number of observations that he thinks he will need to bring Edward's posterior mean into reasonable agreement with his own. Hence the relationship between $\tilde{r}$ and $\tilde{n}^*$ in Figure 1. Note that Figure 1 cannot be used to inform about the behavior of $\tilde{n}^*$ as a function of $\tilde{r}$ and $\tilde{\tau}_e$ independently, as $\tilde{r}$ is a function of $\tilde{\tau}_e$. In fact, $\tilde{r} = \tilde{\tau}_e\{(\tau_e/\tau_d) - 1 + (\mu_d - \mu_e)^2\}$, so that as $\tilde{\tau}_e \to 0$, $\tilde{r} \to 0$, unless $\tau_d \to 0$ (or $(\mu_d - \mu_e)^2 \to \infty$) at a rate greater than or equal to that of $\tilde{\tau}_e$.

The case $\tilde{\tau}_e \to 0$ is in fact of special interest, as many commonly used non-Bayesian inference rules coincide with Bayes estimates for a prior distribution with precision tending to 0. When $\tilde{\tau}_e \to 0$, $\tilde{r} \to 0$ so long as $|\mu_e - \mu_d|$ and τ_d are held fixed. By the preceding analysis, $\tilde{n}^* \to 1$, so $n^* \to 1/\sqrt{c}$, which happens to be the optimal one-prior sample size under a flat prior. This suggests that if Edward's prior is diffuse, then Dan can well approximate the optimal sample size by designing as if he agreed with Edward.

2.1.2 Comparison With One-Prior Strategies. When Dan and Edward agree, the risk to Dan is $R(n) = cn + 1/(n+\tau)$. If $\tau < 1/\sqrt{c}$, then $n^* = (1/\sqrt{c}) - \tau$. Otherwise, $n^* = 0$. In the two-prior problem, it is possible for one or both prior precisions to exceed the threshold $1/\sqrt{c}$ and to still obtain a positive optimal sample size. In some cases the optimal two-prior sample size may be somewhat larger than the largest possible optimal sample size under one prior. For example, if $c = .0625$, then the optimal one-prior sample size is always less than or equal to 4, which corresponds to a prior precision of 0. But in the two-prior case, if $|\mu_e - \mu_d| = 2$, then n^* is as high as 43 for $\tau_e = 6$ and $\tau_d = .01$. Thus when τ_e is far from 0, two-prior sample sizes can be very different from the sample sizes that would be optimal for either prior distribution.

More formal statements about the difference between one- and two-prior optimal sample sizes can be made. Let $n^*_{e,d}$ denote the optimal sample size for the problem in which π_e is used for estimation and π_d is used for design ($n^*_{e,d} = n^*$). Then for fixed sampling costs, $n^*_{e,d} \geq n^*_{e,e}$ if and only if $r > 0$. This is a direct consequence of the fact that $y(\tilde{r})$ is nondecreasing in $\tilde{r}$ (see the Appendix), and hence $n^*_{e,d}$ is nondecreasing in r for fixed τ_e.

A special case of this result occurs when Dan and Edward have common prior means but different prior precisions. Suppose that Edward's prior is more precise than Dan's ($\tau_e > \tau_d$). Then r is positive, and the optimal sample size for the experiment in which Edward estimates and Dan designs is greater than or equal to that for the experiment in which Edward both estimates and designs. The reverse is true when Dan's prior is more precise than Edward's. Note, however, that the condition $\tau_e > \tau_d$ is neither necessary nor sufficient for $n^*_{e,d}$ to be greater than $n^*_{d,d}$.

2.2 The Loss Function $-\log(a)$

Bernardo (1979) suggested the logarithmic utility function when the object of a study is to report the distribution of the parameter of interest (see also Lindley 1956). The development here is similar to that for squared-error loss. Edward's optimal decision is the a that minimizes $-\int \pi_e(\theta \mid \bar{x}) \log(a)\, d\theta$ over the space of distributions on θ, and is given by $a = \pi_e(\theta \mid \bar{x})$. The posterior expected loss relevant for the two-prior sample size calculation is $-\int \pi_d(\theta \mid \bar{x}) \log(\pi_e(\theta \mid \bar{x}))\, d\theta$, and Dan's predictive expectation of this expression, plus sampling costs, can be written as

$$R(n) = cn + \frac{1}{2}\log\left(\frac{2\pi}{n+\tau_e}\right) + \frac{1}{2} + \frac{r}{2(n+\tau_e)},$$

to be minimized over n. When $n = 0$,

$$R(0) = \frac{1}{2}\log\left(\frac{2\pi}{\tau_e}\right) + \frac{\tau_e}{2}\left[\frac{1}{\tau_d} + (\mu_d - \mu_e)^2\right].$$

Equivalently, one may minimize over $y \geq \tilde{\tau}_e$: $g(y) = y - \log y + (\tilde{r}/y)$, where $\tilde{r} = 2cr$ and $\tilde{\tau}_e = 2c\tau_e$. Let $\tilde{n}^* = 2cn^*$, where n^* minimizes g. Then the problem of finding the first-order condition for $\tilde{n}^*$ reduces to that of solving a quadratic over a subset of the positive real line; here too one can plot the contours of $\tilde{n}^*$ as a function of $\tilde{\tau}_e$ and $\tilde{r}$ (Fig. 2). The region between the dotted lines is the region of positive $\tilde{n}^*$. The corresponding mathematical expressions are given in the Appendix.

The zero boundary in Figure 2 lies outside that in Figure 1, indicating the increased desirability of sampling under the

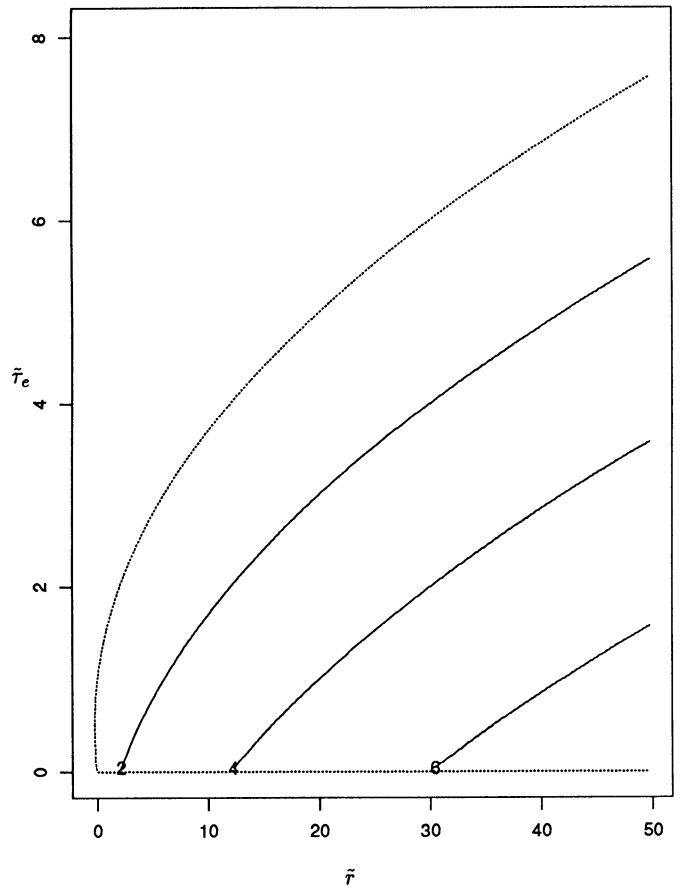

Figure 2. Contours of $\tilde{n}^*$ as a function of $\tilde{\tau}$ and $\tilde{\tau}_e$: Logarithmic Utility. The region between the dotted lines is the region of non-0 $\tilde{n}^*$.

logarithmic loss. (But note the change in cost-normalization between squared-error loss and logarithmic loss.) Optimal sample sizes are higher under this loss structure than under quadratic loss, but they follow the same pattern for fixed τ_e, reaching their highest values when Dan's prior is relatively diffuse about a mean that conflicts with Edward's.

As $\tau_e \to 0$, the expected loss of not performing the experiment becomes infinite, while the optimal sample size and associated risk tend to a finite constant. Thus as τ_e tends to 0, it is always optimal for Dan to sample, and the optimal sample size is simply the one-prior sample size corresponding to $\tau_e = 0$, namely $1/2c$. Thus in this case too, Dan would do well to design in accordance with Edward's prior if Edward's prior is diffuse.

Once again one can compare the optimal sample size when the two priors differ, to the optimal sample size under a common prior distribution, and the same result as in the quadratic loss case obtains. Fixing Edward's prior precision, $n^* > n^*_{e,e}$ if and only if $r > 0$.

3. OPTIMAL ALLOCATION

This section examines the problem of optimal allocation of n observations to k groups, where n and k are given and the n observations $\mathbf{y}$ are independent and normally distributed with common variance, which we take to be equal to 1 without loss of generality: $\mathbf{y}|\theta \sim N(X\theta, I)$.

Here θ is a k vector of unknown group means, $\mathbf{X}$ is an $n \times k$ design matrix, and $\mathbf{I}$ is the identity matrix. In the one-way layout, $X^TX = \text{diag}\{np_i\}$; np_i is the number of observations allocated to group i. Conjugate normal priors on θ are $\pi_i(\theta) = N(\mu_i, T_i^{-1})$, $i = d, e$. The posterior distributions are normal with means $\mu_i(\mathbf{y}) = (T_i + X^TX)^{-1}(T_i\mu_i + X^T\mathbf{y})$, $i = d, e$, and precision matrices $\mathbf{A}_i = (T_i + X^TX)$, $i = d, e$. The goal is to solve for Dan's optimal allocations $\{p_i\}$ given that the estimate of θ will be selected according to Edward's prior distribution. For instance, suppose (as in Smith and Verdinelli 1980) that the first group, with mean θ_1, corresponds to a control treatment, and Dan believes that the other treatment means are exchangeable, but not with the control. If Edward is skeptical and believes that all the groups are exchangeable, then $\mu_d = (\mu, \mu', \ldots, \mu')^T$, $T_d = \text{diag}\{\tau, \tau', \ldots, \tau'\}$, and $\mu_e = (\mu, \mu, \ldots, \mu)^T$, $T_e = \text{diag}\{\tau, \tau, \ldots, \tau\}$.

3.1 Quadratic Loss

Under the quadratic loss function $L(\theta, \mathbf{a}) = (\theta - \mathbf{a})^T \times (\theta - \mathbf{a})$, Dan's posterior expected loss is

$$E^{\pi_d(\theta|\mathbf{y})}(\theta - \mu_e(\mathbf{y}))^T(\theta - \mu_e(\mathbf{y})) = \text{tr}(A_d^{-1}) + (\mu_d(\mathbf{y}) - \mu_e(\mathbf{y}))^T(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y})). \quad (4)$$

Dan's predictive distribution on $\mathbf{y}$ is $N(X\mu_d, I + X\mathbf{T}_d^{-1}X^T)$; the Appendix shows that Dan's predictive expectation of (4) is given by

$$\mathbf{R}(n, \mathbf{p}) = \text{tr}(\mathbf{A}_e^{-1}) + \text{tr}(\mathbf{A}_e^{-1}\mathbf{R}\mathbf{A}_e^{-1}), \quad (5)$$

where $\mathbf{R}$ is a matrix version of r in (1): $\mathbf{R} = \mathbf{T}_e\mathbf{T}_d^{-1}(\mathbf{T}_e - \mathbf{T}_d) + \mathbf{T}_e(\mu_d - \mu_e)(\mu_d - \mu_e)^T\mathbf{T}_e$. If $\mathbf{T}_e = \text{diag}\{\tau_i\}$, then $\mathbf{A}_e = \text{diag}\{z_i\}$, where $z_i = \tau_i + np_i$, and the optimization problem is to minimize $g(\mathbf{z}) = \sum (1/z_i) + \sum (r_{ii}/z_i^2)$ subject to the constraints $\sum z_i = \sum \tau_i + n$ and $z_i \geq \tau_i$. This optimization problem is solved by setting equal to 0 the partial derivatives with respect to z_i and λ of the Lagrangian $g(\mathbf{z}) + \lambda(\sum z_i - [\sum \tau_i + n])$ and then solving for $z_i \geq \tau_i$ and λ. The values of z_i and λ thus obtained will be denoted λ^* and z_i^*.

In practice one may fix λ equal to an initial value, λ_0. Solve the equations obtained by equating the derivatives of the Lagrangian with respect to z_i to 0. The solutions obtained will depend on λ_0 and are denoted $z_i^*(\lambda_0)$. Then obtain an updated value for λ by substituting the $z_i^*(\lambda_0)$ into the equation obtained by equating the derivative of the Lagrangian with respect to λ to 0. Iterate through these two steps until λ converges. The computations are aided by the fact that for given λ, the z_i solutions are solutions to a cubic equation that is almost identical to that obtained in the optimal sample size problem. By similar arguments, one may derive conditions under which observations will be allocated to group i. These conditions are given in the Appendix, together with an argument that proves the existence and uniqueness of λ^*.

Figure 3 shows contours of the optimal allocation to a treatment group in the case of two groups, control and treatment. The prior parameter settings reflect agreement with

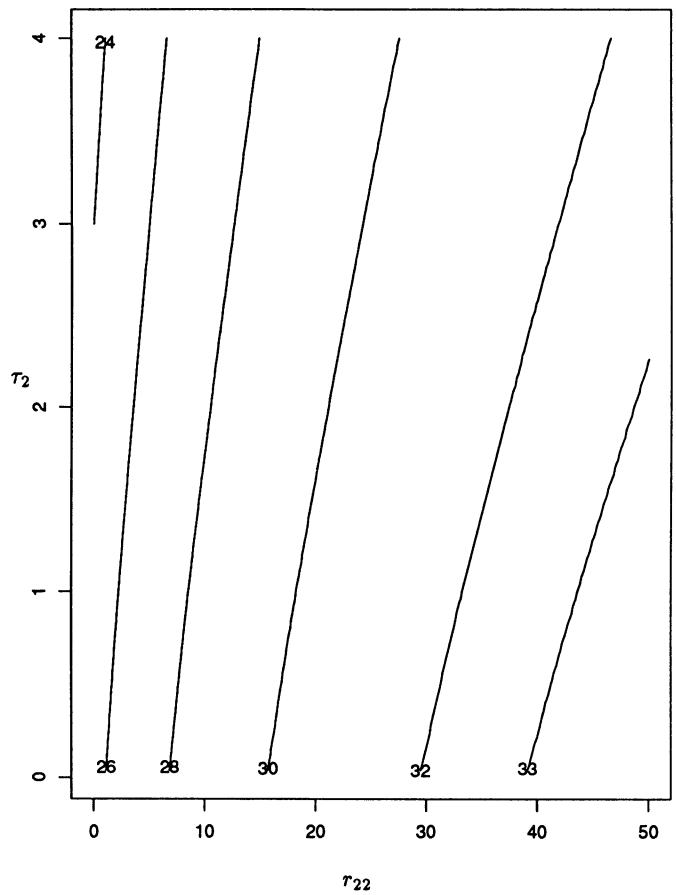

Figure 3. Optimal Allocation to Treatment as a Function of r_{22} and τ_2. Prior distributions on the control mean are normal with equal means and variance 1. The overall sample size is 50.

respect to the control group, which is labeled group 1; $r_{11} = 0$ and $\tau_1 = 1$. The plot shows the number of subjects allocated to the treatment group as a function of r_{22} and τ_2 when the overall sample size, n, is 50.

Note that although Figure 3 does not show a curve for 25, half the observations are allocated to the treatment group when $r_{22} = 0$ and $\tau_2 = 1$, and this is the optimal allocation whenever the treatment prior means are equal (although possibly different from the control prior mean) and the treatment prior precisions are equal to the common control prior precision. This is simply a case of the result that under one prior, allocation is a function of the prior precisions and not of the prior means (see, for example, Smith and Verdinelli 1980).

The patterns observed in Figure 3 are very similar to those in Figure 1. For fixed τ_2, the number of observations allocated to treatment increases with r_{22}; that is, as the treatment prior means diverge and/or Dan becomes less certain about his treatment prior mean. Conversely, the smallest allocations occur when Dan and Edward have high prior precision about a common treatment prior mean; in the upper left corner of the figure, less than half of the observations are allocated to treatment. These features are expected to extend to the case of more than two groups. For instance, in the case of k groups with common τ_i, $i = 1, \ldots, k$, one would expect the r_{ii} to induce an ordering on the group allocations, so that the group with the highest value of r_{ii} would receive the most subjects and so forth. This has been confirmed computationally in several examples.

Increasing the common control prior precision leads to a general increase in allocation to the treatment group (contours not shown here). Prior information for the control group is more precise, and thus fewer control observations are required.

As a final point, we note that the optimal allocation depends on λ^*, which in turn depends on n. In general, the effect of increasing n is to move the proportions allocated towards 50%, yielding a more balanced design. For example, for $\tau_2 = 1$ and $r_{22} = 20$, 67% of observations are allocated to treatment when $n = 20$, but only 60% of observations are allocated to treatment when $n = 50$.

3.2 Logarithmic Utility

Under the loss function $-\log(a)$, the posterior expected loss relevant for Dan's design problem is

$$E^{\pi_d(\theta|\mathbf{y})}\{-\log(\pi_e(\theta|\mathbf{y}))\}$$

$$= -\log|\mathbf{A}_e| + \frac{1}{2}(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y}))^T\mathbf{A}_e(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y})) + \frac{1}{2}\operatorname{tr}\mathbf{A}_e\mathbf{A}_d^{-1}. \quad (6)$$

By the calculations in the Appendix, Dan's predictive expected loss can be simplified to

$$-\frac{k}{2} - \log|\mathbf{A}_e| + \frac{1}{2}\operatorname{tr}(\mathbf{R}\mathbf{A}_e^{-1}), \quad (7)$$

and the optimization problem is to minimize $g(\mathbf{z}) = -\sum \log(z_i) + \frac{1}{2}\sum (r_{ii}/z_i)$, subject to the constraints $\sum z_i = \sum \tau_i + n$ and $z_i \geq \tau_i$. The Lagrangian is $g(\mathbf{z}) + \lambda(\sum z_i - [\sum \tau_i + n])$, and here also, for given λ, the values $z_i^*(\lambda)$ are solutions to quadratic equations that are almost identical to those obtained in the sample size problem. A similar argument to that given in the quadratic loss case establishes the uniqueness of λ^*, as shown in the Appendix.

Plots of the optimal allocation to treatment under a log utility are very similar to those observed under the quadratic loss and are not given here. Software to compute the allocations is available from the authors.

4. DISCUSSION

Most experiments are designed under the assumption that the party performing the experiment and the party evaluating the experimental data are one and the same (Pilz 1991). This research is motivated by providing some important examples in which this is not the case. Even if both parties share a loss function and have common goals, their prior beliefs may well differ. This article presents a theory that is flexible enough to allow for this contingency and that includes as a special case two parties (Dan and Edward) with identical prior distributions.

The results are striking in the sense that the optimal two-prior designs are in many cases very different to the optimal

one-prior designs. Optimal sample sizes (and allocations) are highest when Edward's prior is relatively peaked about a mean that contradicts Dan's prior mean, and they are lowest when Dan's prior is relatively peaked about a mean that agrees with Edward's. In the former case, the optimal sample sizes may greatly exceed the optimal one-prior sample sizes. The accompanying reduction in predictive risk leads us to conclude that in such extreme cases, if the person designing an experiment knows the prior distribution of the party evaluating the data, he would be wise to take it into account. For example, consider the case of squared-error loss with $c = .0625$. If $\tau_d = .01$, $\tau_e = 6$, and $|\mu_e - \mu_d| = 2$, then $n^* = 43$, and Dan's corresponding predictive risk is 6.1. Given that Edward will be estimating, the predictive risk to Dan is 36 if he ignores Edward's prior opinion, and uses a sample size of 4 (one-prior optimum, corresponding to a prior precision of .01)—a six-fold increase! In practice, however, one would rarely expect to find such vague prior opinion on the part of the person designing (and paying for) the experiment. It is more common to have an estimation prior that is more diffuse than the design prior as in the legal and medical examples presented in Section 1. The results indicate that in these cases, optimal two-prior designs will not be too different from the optimal designs corresponding to the estimation prior; one would expect this to be true for other likelihoods as well.

This analysis leads to the question of whether the resulting designs are compromises in some sense, particularly for intermediate values of the precision parameters. Lindley and Singpurwalla (1991) provided an example in which the optimal sample size in their adversarial problem falls between the two participants' optima, balancing the cost to the party performing the experiment of providing the data and the enthusiasm of the party evaluating the data for much of it. We have shown that in point estimation, the optimal two-prior sample size can exceed both one-prior sample sizes, but it may be interesting to examine conditions under which the optimal sample size falls between the individual optimal sample sizes.

Under a normal likelihood with known variance, is the extension from fixed sample sizes to sequential sampling as straightforward with two decision makers as with one decision maker? It is well known that with one decision maker having a normal likelihood with known variance, the optimal sequential sample size is equal to the optimal fixed sample size. But this is a result of the fact that in this case the posterior expected loss is independent of the sample data, at least for the loss functions considered in this article. With Dan and Edward, however, one must evaluate Dan's posterior expected loss of Edward's optimal action, which is not independent of the sample data. Thus the extension to sequential sampling seems to be nontrivial in the case of two decision makers.

Perhaps the most intriguing aspect of our development is the appearance of r as a critical quantity in the characterization and evaluation of the optimal two-prior sample sizes, both in the univariate and in the multivariate cases. Can this quantity be interpreted as a measure of discrepancy between the two priors? Its asymmetry reflects the different roles of Dan and Edward in design and estimation, and it is certainly not a distance measure, but perhaps it might be calibrated to yield a meaningful measure of comparability of the prior distributions. An extension of this work would be to examine other likelihoods and loss functions to check whether r or some analog thereof plays a more extensive role than that already observed.

The work presented here is rich in potential extensions. For example, expressions (5) and (7) are quite general expressions in the sense that they hold for design matrices other than those implied by the one-way ANOVA and for nondiagonal prior covariance matrices. This suggests an extension to other design problems, such as linear regression. In the case of allocation problems, once the optimal allocation to k groups has been determined, it may be useful to consider the cost per observation, possibly different for different groups, and to solve for the optimal overall sample size n. The extension to other loss functions and likelihoods may not be straightforward, but we expect to see some similar results, and the work here should provide some useful guidelines in these more-complex models.

We believe that experimental design is best modeled as a two-prior problem. We hope to see more two-prior designs in the future.

APPENDIX: PROOFS

Squared-Error Loss

Proof of Theorem 1. The cubic equation arises from differentiating $g(y)$ with respect to y. By the properties of cubic equations (see, for example, Littlewood 1950), note that this equation has one real root if $\tilde{r} < -(1/3\sqrt{3})$ or $\tilde{r} > (1/3\sqrt{3})$, it has two real roots if $\tilde{r} = -(1/3\sqrt{3})$ or $\tilde{r} = (1/3\sqrt{3})$, and it has three real roots if $-(1/3\sqrt{3}) < \tilde{r} < (1/3\sqrt{3})$. By studying the behavior of the function $y_3 - y$, one can ascertain whether the real roots of $y^3 - y = 2\tilde{r} = 0$ are positive or negative. The function $y^3 - y$ decreases to $-\infty$ as y decreases to $-\infty$, and it increases to ∞ as y increases to ∞. It has a local maximum of $(2/3\sqrt{3})$ at $y = -(1/\sqrt{3})$ and a local minimum of $-(2/3\sqrt{3})$ at $y = 1/\sqrt{3}$. Thus $\tilde{r} \le -(1/3\sqrt{3})$ implies that $y^3 - y - 2\tilde{r}$ is nonnegative on $y > 0$. In this case $g(y)$ is increasing in y for $y > 0$, and the optimal sample size is 0.

Similarly, if $\tilde{r} \ge 0$, then $y^3 - y - 2\tilde{r}$ has only one positive root, which provides a local minimum, because $g''(y)$ is nonnegative on $y > 0$ for $\tilde{r} \ge 0$. Thus for $\tilde{r} \ge 0$ there exists a unique, real, positive solution, $y(\tilde{r})$. The optimal sample size is $(y(\tilde{r}) - \tilde{\tau}_e)/\sqrt{c}$ if $y(\tilde{r}) > \tilde{\tau}_e$, and 0 otherwise. Note that $\tilde{r} \ge 0$ and $y(\tilde{r}) > \tilde{\tau}_e$ together imply that the second condition in the theorem is automatically satisfied.

Now consider the case $-(1/3\sqrt{3}) < \tilde{r} < 0$, which has two positive roots. The smaller yields a local maximum; the larger yields a local minimum. The larger positive root is $y(\tilde{r})$. Is it sufficient to compare the values of $\tilde{\tau}_e$ and $y(\tilde{r})$ when deciding on the optimal sample size? Because $g(y)$ tends to $-\infty$ as y decreases to 0, it is not automatically true that $g(y(\tilde{r})) < g(\tilde{\tau}_e)$ if $y(\tilde{r}) > \tilde{\tau}_e$. In fact, for sufficiently small $\tilde{\tau}_e$, $g(\tilde{\tau}_e)$ may be less than $g(y(\tilde{r}))$. In this case the optimal sample size will be 0, not $(y(\tilde{r}) - \tilde{\tau}_e)/\sqrt{c}$.

For fixed $\tilde{r}$, and thus $y(\tilde{r})$, it is possible to obtain a closed-form expression for the values of $\tilde{\tau}_e$ such that $g(\tilde{\tau}_e)$ is less than $g(y(\tilde{r}))$. Clearly, if $\tilde{\tau}_e$ is greater than or equal to $y(\tilde{r})$, then $g(y)$ is increasing on $[\tilde{\tau}_e, \infty)$, so that the optimal sample size is 0 in this case. Consider $0 < \tilde{\tau}_e < y(\tilde{r})$. There exists a t^* such that $\tilde{\tau}_e \le t^* \Rightarrow g(\tilde{\tau}_e) \le g(y(\tilde{r})) \Rightarrow n^* = 0$ and $\tilde{\tau}_e > t^* \Rightarrow g(\tilde{\tau}_e) > g(y(\tilde{r})) \Rightarrow n^*$

$= (y(\tilde{r}) - \tilde{\tau}_e)/\sqrt{c}$. Clearly, t^* is a solution to the equation $g(\tilde{\tau}_e) = g(y(\tilde{r}))$. This is a cubic equation in $\tilde{\tau}_e$, but it factors as $(\tilde{\tau}_e - y(\tilde{r}))^2(\tilde{\tau}_e + (\tilde{r}/y^2(\tilde{r}))) = 0$, indicating that the t^* sought is $t^* = -[\tilde{r}/y^2(\tilde{r})]$.

Conditions Necessary for Positive n^*: −Log Loss

Under the loss function $-\log(a)$, the following result characterizes the optimal sample size n^*. Let $y(\tilde{r}) = (1 + \sqrt{1 + 4\tilde{r}})/2$. Let $t^*(\tilde{r})$ be the solution of

$$t^* - \log(t^*) + \frac{\tilde{r}}{t^*} = g\left(\frac{1 + \sqrt{1 + 4\tilde{r}}}{2}\right) \quad \text{if } -\frac{1}{4} < \tilde{r} < 0,$$

and $t^*(\tilde{r}) = 0$ otherwise.

If $\tilde{r} > -\frac{1}{4}$ and $t^*(\tilde{r}) < \tilde{\tau}_e < y(\tilde{r})$,

$$\text{then} \quad \tilde{n}^* = y(\tilde{r}) - \tilde{\tau}_e; \quad \text{otherwise,} \quad \tilde{n}^* = 0.$$

The reasoning here is similar to that under quadratic loss but somewhat more straightforward. This is due to the fact that $y(\tilde{r})$ is a solution to the quadratic equation $y^2 - y - \tilde{r} = 0$, which has roots $(1 - \sqrt{1 + 4\tilde{r}})/2$ and $(1 + \sqrt{1 + 4\tilde{r}})/2$. When $\tilde{r} \le -(1/4)$, the quadratic has no positive roots, and $g(y)$ is increasing for positive y. When $\tilde{r} \ge 0$, there is only one positive root, and this provides a local minimum. When $-(1/4) < \tilde{r} < 0$, the quadratic has two positive roots. The smaller root is a local maximum; the larger is a local minimum. Because $g(y)$ decreases to $-\infty$ as y tends to 0, there exists $t^*(\tilde{r})$ for given $\tilde{r}$ such that $t^*(\tilde{r}) \in (0, (1 - \sqrt{1 + 4\tilde{r}})/2)$, and t^* solves the equation

$$t^* - \log(t^*) + \frac{\tilde{r}}{t^*} = g\left(\frac{1 + \sqrt{1 + 4\tilde{r}}}{2}\right).$$

Then

$$\tilde{\tau}_e \le t^* \Rightarrow g(\tilde{\tau}_e) \le g(y(\tilde{r})) \Rightarrow \tilde{n}^* = 0$$

and

$$\tilde{\tau}_e > t^* \Rightarrow g(\tilde{\tau}_e) > g(y(\tilde{r})) \Rightarrow \tilde{n}^* = y(\tilde{r}) - \tilde{\tau}_e.$$

The Multivariate Case: Quadratic Loss

Under the quadratic loss function $L(\theta, \mathbf{y}) = (\theta - a)^T(\theta - a)$, the posterior expected loss relevant for the sample size calculation is

$$E^{\pi_d(\theta|\mathbf{y})}(\theta - \mu_e(\mathbf{y}))^T(\theta - \mu_e(\mathbf{y})) = \operatorname{tr}(\mathbf{A}_d^{-1}) + (\mu_d(\mathbf{y}) - \mu_e(\mathbf{y}))^T(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y})).$$

But $(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y}))^T(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y}))$ can be expressed as

$$(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y}))^T(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y})) = (\mathbf{A}_d^{-1} - \mathbf{A}_e^{-1})X^T(X\mu_d - \mathbf{y}) + \mathbf{A}_e^{-1}\mathbf{T}_e(\mu_e - \mu_d). \quad \text{(A.1)}$$

Using the symmetry of $\mathbf{A}_d$, $\mathbf{A}_e$, $\mathbf{T}_d$, and $\mathbf{T}_e$, and the fact that $\mathbf{A}_e^{-1} - \mathbf{A}_d^{-1} = \mathbf{A}_d^{-1}(\mathbf{A}_d - \mathbf{A}_e)\mathbf{A}_e^{-1} = \mathbf{A}_d^{-1}(\mathbf{T}_d - \mathbf{T}_e)\mathbf{A}_e^{-1}$, Dan's predictive expectation of (4) is

$$\mathbf{R}(\mathbf{p}) = \operatorname{tr}(\mathbf{A}_e^{-1}[(\mathbf{A}_d^{-1} - \mathbf{A}_e^{-1})X^TX(\mathbf{T}_d^{-1}\mathbf{T}_e - I) + \mathbf{A}_e\mathbf{A}_d^{-1}]) + \operatorname{tr}(\mathbf{A}_e^{-1}\mathbf{T}_e(\mu_e - \mu_d)(\mu_e - \mu_d)^T\mathbf{T}_e\mathbf{A}_e^{-1}). \quad \text{(A.2)}$$

This can be simplified using the following lemma.

Lemma. Let $\mathbf{Z}$ be an arbitrary conformable matrix. Then

$$\operatorname{tr}\mathbf{Z}[(\mathbf{A}_d^{-1} - \mathbf{A}_e^{-1})(X^TX)(\mathbf{T}_d^{-1}\mathbf{T}_e - I) + \mathbf{A}_e\mathbf{A}_d^{-1} + \mathbf{T}_e(\mu_e - \mu_d)(\mu_e - \mu_d)^T\mathbf{T}_e\mathbf{A}_e] = \operatorname{tr}(\mathbf{Z}\mathbf{A}_e^{-1}\mathbf{R}) + \operatorname{tr}\mathbf{Z}.$$

Proof. Let $S = \operatorname{tr}(\mathbf{Z}\mathbf{T}_e(\mu_e - \mu_d)(\mu_e - \mu_d)^T\mathbf{T}_e\mathbf{A}_e)$. Then

$$\begin{aligned}
&\operatorname{tr}(\mathbf{Z}[(\mathbf{A}_d^{-1} - \mathbf{A}_e^{-1})(X^TX)(\mathbf{T}_d^{-1}\mathbf{T}_e - I) + \mathbf{A}_e\mathbf{A}_d^{-1}]) + S \\
&= \operatorname{tr}(\mathbf{Z}[\mathbf{A}_d^{-1}(X^TX\mathbf{T}_d^{-1}\mathbf{T}_e - X^TX + \mathbf{A}_e)]) \\
&\quad - \operatorname{tr}(\mathbf{Z}\mathbf{A}_e^{-1}X^TX(\mathbf{T}_d^{-1}\mathbf{T}_e - I)) + S \\
&= \operatorname{tr}(\mathbf{Z}[\mathbf{A}_d^{-1}(X^TX\mathbf{T}_d^{-1}\mathbf{T}_e + \mathbf{T}_e)]) \\
&\quad - \operatorname{tr}(\mathbf{Z}\mathbf{A}_e^{-1}(\mathbf{A}_e - \mathbf{T}_e)(\mathbf{T}_d^{-1}\mathbf{T}_e - I)) + S \\
&= \operatorname{tr}(\mathbf{Z}\mathbf{A}_d^{-1}(X^TX + \mathbf{T}_d)\mathbf{T}_d^{-1}\mathbf{T}_e) - \operatorname{tr}(\mathbf{Z}\mathbf{T}_d^{-1}\mathbf{T}_e) \\
&\quad + \operatorname{tr}\mathbf{Z} + \operatorname{tr}(\mathbf{Z}\mathbf{A}_e^{-1}\mathbf{R}) \\
&= \operatorname{tr}(\mathbf{Z}\mathbf{A}_d^{-1}\mathbf{A}_d\mathbf{T}_d^{-1}\mathbf{T}_e) - \operatorname{tr}(\mathbf{Z}\mathbf{T}_d^{-1}\mathbf{T}_e) + \operatorname{tr}\mathbf{Z} + \operatorname{tr}(\mathbf{Z}\mathbf{A}_e^{-1}\mathbf{R}) \\
&= \operatorname{tr}\mathbf{Z} + \operatorname{tr}(\mathbf{Z}\mathbf{A}_e^{-1}\mathbf{R}).
\end{aligned}$$

This lemma is used with $\mathbf{Z} = \mathbf{A}_e^{-1}$ in quadratic loss and with $\mathbf{Z} = I$ for $-\log d$ loss.

Conditions Necessary and Sufficient for Positive Allocation to Group i: Quadratic Loss

Given $\lambda > 0$, let $z(r_{ii}, \lambda)$ denote the largest real root of the cubic, $\lambda z^3 - z - 2r_{ii} = 0$. If $r_{ii} > -(1/3\sqrt{3\lambda})$ and $-(r_{ii}/\lambda z^2(\lambda)) < \tau_i < z(r_{ii}, \lambda)$, then $z_i^*(\lambda) = z(r_{ii}, \lambda)$; otherwise, $z_i^*(\lambda) = \tau_i$. This result is analogous to the univariate case.

Uniqueness of λ

If λ is negative, then the equation $\lambda z^3 - z - 2r_{ii} = 0$ does not have any positive roots for $r_{ii} > 0$ and has a unique positive root, $z(r_{ii}, \lambda)$, for negative r_{ii}. But this root corresponds to a local maximum (in variable i) of the Lagrangian. Thus for $\lambda < 0$, $z_i^*(\lambda) = \infty$, and it suffices to consider only $\lambda \ge 0$. When $\lambda = 0$, the problem reduces to a minimization of $g(\mathbf{z})$ over the region $\{z_i \ge \tau_i\}$. But note that $r_{ii} \ge -\tau_i$ so that $g(\mathbf{z}) \ge 0$ on this region $[(1/z_i) + (r_{ii}/z_i^2) \ge (1/z_i) - (\tau_i/z_i^2) = (1/z_i)(1 - (\tau_i/z_i)) \ge 0]$. Thus when $\lambda = 0$, the optimal allocation to group i is infinite, and $\sum z_i^*(\lambda) - T = \infty$. At the other extreme, when $\lambda = \infty$, we have $\sum z_i^*(\lambda) - T < 0$. This implies the existence of λ satisfying

$$\sum z_i^*(\lambda) = \sum \tau_i + n, \quad \text{(A.3)}$$

which is the equation that results from equating to 0 the derivative of the Lagrangian with respect to λ. This solution is in fact unique, as it is possible to show that for given $\mathbf{R}$, $\sum z_i^*(\lambda) - T$ is decreasing in λ for $\lambda > 0$. We summarize the argument briefly as follows: For given r_{ii}, $z(r_{ii}, \lambda)$ is decreasing in λ. Moreover, the region over which $z_i^*(\lambda) = z(r_{ii}, \lambda)$ gets smaller as λ increases, and the region over which $z_i^*(\lambda) = \tau_i$ gets larger as λ increases. Consequently, $\sum z_i^*(\lambda)$ decreases with λ, and hence $\sum z_i^*(\lambda) - T$ decreases with λ. So λ^* is unique and positive and can be found quite easily, for instance, by a bisection algorithm.

The Multivariate Case: −Log Loss

Here the relevant posterior expected loss is a function of $(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y}))^T A_e(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y}))$. Using (A.1), Dan's predictive expectation of $\mu_d(\mathbf{y}) - \mu_e(\mathbf{y})$ is

$$E^{p_d(\mathbf{y})}(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y})) = (I - \mathbf{A}_e^{-1}X^TX)(\mu_e - \mu_d),$$

and the predictive variance–covariance matrix of $\mu_d(\mathbf{y}) - \mu_e(\mathbf{y})$ is

$$V^{p_d(\mathbf{y})}(\mu_d(\mathbf{y}) - \mu_e(\mathbf{y})) = (\mathbf{A}_d^{-1} - \mathbf{A}_e^{-1})X^TX\mathbf{T}_d^{-1}\mathbf{A}_d(\mathbf{A}_d^{-1} - \mathbf{A}_e^{-1}).$$

Using a formula for the expected value of a quadratic form (Graybill 1976, p. 139), one can obtain the predictive expectation of (6). Some algebraic manipulation yields the following expression to be

minimized over p_i, $i = 1, \ldots k$:

$$\mathbf{R}(\mathbf{p}) = -\log|\mathbf{A}_e| + \frac{1}{2}\operatorname{tr}\mathbf{A}_e\mathbf{A}_d^{-1} + \frac{1}{2}\operatorname{tr}(\mathbf{A}_d^{-1} - \mathbf{A}_e^{-1})X^TX(\mathbf{T}_d^{-1}\mathbf{T}_e - I) - \frac{1}{2}\operatorname{tr}(\mu_e - \mu_d)^T\mathbf{T}_e\mathbf{A}_e^{-1}\mathbf{T}_e(\mu_e - \mu_d).$$

Compare this to (A.2). The preceding lemma yields (7).

Conditions Necessary for Positive Allocation to Group i: −Log Loss

Given $\lambda > 0$, let $w_i = r_{ii}/2$, and let $z(w_i, \lambda) = (1 + \sqrt{1 + 4\lambda w_i})/2\lambda$. Let $t^*(w_i, \lambda)$ be the solution of

$$(\lambda t^* - \log(t^*) + \frac{w_i}{t^*} = g\left(\frac{1 + \sqrt{1 + 4\lambda w_i}}{2\lambda}\right)$$

$$\text{if } -\frac{1}{4\lambda} < w_i < 0, \quad \text{and} \quad t^*(w_i, \lambda) = 0$$

otherwise. If $w_i > -(1/4\lambda)$ and $t^*(w_i, \lambda) < \tau_i < z(w_i, \lambda)$, then $z_i^*(\lambda) = z(w_i, \lambda)$; otherwise, $z_i^*(\lambda) = \tau_i$.

Here too, it suffices to consider only positive λ by an argument similar to that given in the squared-error loss case. If λ is negative, then $\lambda z^2 - z - w_i$ does not have any positive roots for $w_i > 0$, and it is concave on $\{z_i > \tau_i\}$ for $w_i < 0$. In fact, if λ is negative, then $z_i^*(\lambda) = \infty$. In the unconstrained problem ($\lambda = 0$), we find also that $z_i^*(\lambda) = \infty$, because $w_i \geq -\tau_i/2$. When $\lambda = +\infty$, $z_i^*(\lambda) - T < 0$, which proves the existence of a solution to (A.3) on the positive real line. The uniqueness argument mimics that for quadratic loss.

[Received February 1992. Revised October 1992.]

REFERENCES

Bernardo, J. M. (1979), "Expected Information as Expected Utility," *The Annals of Statistics,* 7, 686–690.

Burt, J. (1990), "Towards Agreement: Bayesian Experimental Design," Technical Report 90-41, Purdue University, Dept. of Statistics.

Chaloner, K., and Larntz, K. (1989), "Optimal Bayesian Design Applied to Logistic Regression Experiments," *Journal of Statistical Planning and Inference,* 21, 191–208.

DasGupta, A., and Studden, W. J. (1991), "Robust Experimental Designs in Normal Linear Models," *The Annals of Statistics,* 19, 1244–1256.

Graybill, F. A. (1983), *Matrices with Applications in Statistics* (2nd ed.), Belmont, CA: Wadsworth.

Jackson, P. H., Novick, M. R., and DeKeyrel, D. F. (1980), "Adversary Preposterior Analysis for Simple Parametric Models," in *Bayesian Analysis in Econometrics and Statistics,* ed. Arnold Zellner, Amsterdam: North-Holland, pp. 113–132.

Kadane, J. B. (1990), "Statistical Analysis of Adverse Impact on Employer Decisions," *Journal of the American Statistical Association,* 85, 925–933.

Kadane, J. B., and Seidenfeld, T. (1989), "Randomization in a Bayesian Perspective," *Journal of Statistical Planning and Inference,* 25, 329–345.

Lindley, D. V. (1956), "On a Measure of Information Provided by an Experiment," *Annals of Mathematical Statistics,* 27, 986–1005.

Lindley, D. V., and Singpurwalla, N. (1991), "On the Evidence Needed to Reach Agreed Action Between Adversaries With Application to Acceptance Sampling," *Journal of the American Statistical Association,* 86, 933–937.

Littlewood, D. E. (1950), *A University Algebra,* London: Heinemann.

Pilz, J. (1991), *Bayesian Estimation and Experimental Design in Linear Models,* Chichester, U.K.: John Wiley.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal,* 27, 379–423, 623–656.

Smith, A. F. M., and Verdinelli, I. (1980), "A Note on Bayes Designs for Inference Using a Hierarchical Linear Model," *Biometrika,* 67, 613–619.

Spiegelhalter, D. J., and Freedman, L. S. (1988), "Bayesian Approaches to Clinical Trials," in *Bayesian Statistics 3,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Oxford University Press.

Tsutakawa, R. K. (1972), "Design of Experiment for Bioassay," *Journal of the American Statistical Association,* 67, 584–590.