GENERALIZED DEFINITION OF THE STATISTICAL POPULATION DESIGN

Maciej Patan and Basia Bogacka

Institute of Control and Computation Engineering, University of Zielona Góra, Poland

e-mail: M.Patan@issi.uz.zgora.pl

School of Mathematical Sciences, Queen Mary, University of London
e-mail: B.Bogacka@qmul.ac.uk

PODE'2006

MODEL

We consider a nonlinear regression model

$$y = \eta(t, \theta) + \epsilon,$$

where

$t \in T = [0, t_{max}]$, $t_{max} < \infty$, denotes an explanatory variable,

$\theta \in \Theta$ a p-dimensional vector of unknown parameters,

Θ a set of admissible values of θ,

$\eta(t, \theta)$ the expected response at t,

ϵ a random error of observations.

There is a population of N individuals for each of which n_i measurements are gathered.

The model for each observation can be written as

$$y_i^k = \eta(t_i^k; \theta_i^k) + \varepsilon_i^k, \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, N$$

where

y_i^k is an observation at time $t_i^k \in T$,

ε_i^k are i.i.d. random errors with a known density f, $\mathrm{E}(\varepsilon_i^k) = 0, \ \mathrm{var}(\varepsilon_i^k) = \sigma^2$.

We assume that the parameter vectors θ_i^k are random, and that

$$\mathrm{E}(\theta_i^k) = \theta \qquad \mathrm{Var}(\theta_i^k) = \mathrm{diag}\{\sigma_1^2, ..., \sigma_p^2\}$$

Efficient estimation of the population parameter vector

$$\psi = (\theta_1, \ldots, \theta_p, \sigma_1^2, ..., \sigma_p^2, \sigma^2)^{\mathrm{T}}$$

is our primary interest.

EXPERIMENTAL DESIGN

We assume that the population of N patients consists of G groups;

all individuals in the same group follow the same schedule of measurements (design).

We construct the population experimental design in two stages:

Individual level

$$\xi_j = \left\{ \begin{array}{ccc} t_1^j & \dots & t_{s_j}^j \\ w_1^j & \dots & w_{s_j}^j \end{array} \right\}; \quad w_i^j \in (0,1], \quad \sum_{i=1}^{s_j} w_i^j = 1.$$

The whole experimental system per individual is described by the pair (ξ_j, n_j),

where n_j is the number of observations per individual.

Population level

We define the population design as

$$\zeta = \left\{ \begin{array}{ccc} (\xi_1, n_1) & \ldots & (\xi_G, n_G) \\ \alpha_1 & \ldots & \alpha_G \end{array} \right\}; \quad \sum_{j=1}^{G} \alpha_j = 1,$$

where α_j is the proportion of individuals in the whole population who follow plan (ξ_j, n_j).

FISHER INFORMATION MATRIX

The assumption of independent observations allows us to sum up the FIMs for all single observations.

$$M(\zeta, N) = N \sum_{j=1}^{G} \alpha_j M(\xi_j, n_j) = N \sum_{j=1}^{G} \alpha_j n_j \sum_{i=1}^{s_j} w_i^j M(t_i^j),$$

where

$$M(t_i^j) = \mathrm{E}\left\{ -\frac{\partial^2 \ell(\psi|y_i^j)}{\partial\psi \partial\psi^{\mathrm{T}}} \right\}$$

is the elementary FIM for the observation made at time instant t_i^j and

$$\ell(\psi|y_i^j) = \log \int_{\Theta} g(y_i^j|\theta) h(\theta;\psi)\, \mathrm{d}\theta,$$

$g(y|\theta)$ is the conditional probability density of y given θ,

$h(\theta;\psi)$ is the probability density of θ.

PROBLEM FORMULATION

The design problem here is to optimize a functional Ψ operating on $\mathcal{M}$, a set of FIMs:

$$\Psi : \mathcal{M} \longrightarrow \mathcal{R} \qquad \text{or} \qquad \Psi[M(\zeta, N)] \longrightarrow \min .$$

We look for a design $\zeta^\star$ which gives the optimum FIM for some initially chosen values of the population parameters.

Here, we make the following assumptions:

1. T is compact,
2. $M(t)$ is continuous on T,
3. Ψ is convex,
4. If $M_1 \underset{L}{\leq} M_2$, then $\Psi(M_1) \geq \Psi(M_2)$.

We constrain the total number of observations to be not greater than N_0:

$$N \sum_{j=1}^{G} \alpha_j n_j \leq N_0.$$

In general:

$$N \sum_{j=1}^{G} \alpha_j c(\xi_j, n_j) \leq C_0,$$

where $c(\xi_j, n_j)$ is a nonnegative cost function, C_0 is the total cost.

If N has to be estimated, it is convenient to allow it to take any positive real value.

Then, the optimal solution $(\zeta^\star, N^\star)$ satisfies the cost constraint on the boundary, i.e., the inequality becomes an equality at $(\zeta^\star, N^\star)$.

Further, we introduce the so-called average per total cost (normalized) FIM:

$$M(\upsilon) = \frac{N}{C_0} \sum_{j=1}^{G} \alpha_j c(\xi_j, n_j) \sum_{i=1}^{s_j} w_i^j M(t_i^j) = \sum_{j=1}^{G} \beta_j M(\xi_j),$$

where

$$\beta_j = \frac{N}{C_0} \alpha_j c(\xi_j, n_j) \; ; \quad M(\xi_j) = \sum_{i=1}^{s_j} w_i^j M(t_i^j)$$

and

$$\upsilon = \left\{ \begin{array}{ccc} \xi_1 & \dots & \xi_G \\ \beta_1 & \dots & \beta_G \end{array} \right\} ; \quad \beta_j \in (0, 1], \quad \sum_{j=1}^{G} \beta_j = 1.$$

Instead of minimizing $\Psi(M(\zeta, N))$ subject to the cost constraints we can equivalently minimize $\Psi(\upsilon)$ subject to $\sum_{j=1}^{G} \beta_j = 1$.

However, due to the independence of all the observations the average FIM may be rewritten in the form

$$M(\upsilon) = \sum_{j=1}^{G} \sum_{i=1}^{s_j} \frac{N}{C_0} \alpha_j c(\xi_j, n_j) w_i^j M(t_i^j) = \sum_{j=1}^{G} \sum_{i=1}^{s_j} \gamma_i^j M(t_i^j)$$

where

$$\gamma_i^j = \frac{N}{C_0} \alpha_j c(\xi_j, n_j) w_i^j = \beta_j w_i^j \; ; \quad \sum_{j=1}^{G} \sum_{i=1}^{s_j} \gamma_i^j = 1.$$

$M(\upsilon)$ can also be written as

$$M(\upsilon) = \sum_{k=1}^{s} \gamma_k M(t_k) = M(\omega),$$

where $\gamma_1, \ldots, \gamma_s$ are the sums of γ_i^j's for the repeated time instants.

The design

$$\omega = \left\{ \begin{array}{ccc} t_1 & \dots & t_s \\ \gamma_1 & \dots & \gamma_s \end{array} \right\}; \quad \sum_{k=1}^{s} \gamma_k = 1.$$

is called a ***global design***.

Note

- Such a reformulation simplifies the problem of finding the two level hierarchical optimal population design to that of finding the equivalent one level design.
- It significantly reduced the problem of dimensionality.
- The information about groups is included in γ_i^j and so in γ_k. This information is later recovered after an optimum design ω has been found.

NUMERICAL ALGORITHM

- Substantial difficulties in determining the population designs arises from the fact that they are not unique.
- The criterion Ψ is most often strictly convex on $\mathcal{M}(\Xi)$, and this guarantees that the optimal FIM is unique.
- But this does not mean that $(\zeta, N) \mapsto \Psi[M(\zeta, N)]$ is strictly convex in (ζ, N).
- Multiple global solutions $(\zeta^\star, N^\star)$ may yield the same minimum value of $\Psi(M(\zeta, N))$.

The determination of a final solution can be achieved in three steps:

Step 1. Solve the optimization problem:

$$\omega^\star = \arg\min_{\omega \in \Xi} \Psi(M(\omega)).$$

Step 2. Transform $\omega^\star$ into an equivalent design $\upsilon^\star \in \Upsilon$, which satisfies

$$\upsilon^\star = \arg\min_{\upsilon \in \Upsilon} \Psi(M(\upsilon)).$$

Step 3. Transform $\upsilon^\star$ into an equivalent design pair $(\zeta^\star, N^\star)$.

In **Step 2** we allow zero weights and we solve the following system of equations:

$$\begin{cases} \beta_j w_i^j - \gamma_i^j = 0, \ \ i = 1, \ldots, s, \ \ \ j = 1, \ldots, G & (sG \text{ nonlinear equations}) \\ \sum_{i=1}^{s} w_i^j = 1, \quad j = 1, \ldots, G & (G \text{ linear equations}) \\ \sum_{j=1}^{G} \gamma_i^j = \gamma_i^\star, \ \ j = 1, \ldots, s & (s \text{ linear equations}) \end{cases}$$

There are $s(G-1)$ more variables than equations.

Treating $s(G-1)$ variables γ_i^j as nonnegative parameters the solution becomes simple:

$$\begin{cases} \beta_j^\star = \sum_{i=1}^{s} \gamma_i^j, \ \ j = 1, \ldots, G, \\ w_i^{j\star} = \gamma_i^j / \beta_j^\star, \quad i = 1, \ldots, s, \ \ \ j = 1, \ldots, G, \end{cases}$$

These values are further used in **Step 3**.

The optimal values of the population parameters $\alpha_j^\star, n_j^\star,\ j = 1, \ldots, G$ and $N^\star$ can be retrieved solving the system of equations:

$$\begin{cases} \dfrac{N}{C_0}\alpha_j c(\xi_j^\star, n_j) = \beta_j^\star,\ j = 1, \ldots, G & (G \text{ nonlinear equations}) \\ \sum_{i=1}^{G} \alpha_j = 1, & (1 \text{ linear equation}) \end{cases}$$

Here we have $G + 1$ equations (and $2G + 1$ variables) which can be solved numerically up to G parameters.

For example, if the numbers of observations per individual in each group, n_j, are known then the optimal solution is

$$N^\star = C_0 \sum_{j=1}^{G} \frac{\beta_j^\star}{c(\xi_j^\star, n_j)}, \qquad \alpha_j^\star = \frac{C_0}{N^\star}\frac{\beta_j^\star}{c(\xi_j^\star, n_j)},\ \ j = 1, \ldots, G.$$

Two special forms of population designs which minimize $\Psi[M(\zeta, N)]$ are the following:

(i) one-group design

$$\zeta^\star = \begin{Bmatrix} \omega^\star \\ 1 \end{Bmatrix}, \qquad N^\star = \frac{C_0}{c(\omega^\star, n_1)}, \qquad n_1 > 0$$

that is

$$\zeta^\star = \left\{ \left(\begin{Bmatrix} t_1 & \dots & t_s \\ \gamma_1 & \dots & \gamma_s \end{Bmatrix}, n_1 \right) \right\}; \quad N^\star$$
$$1$$

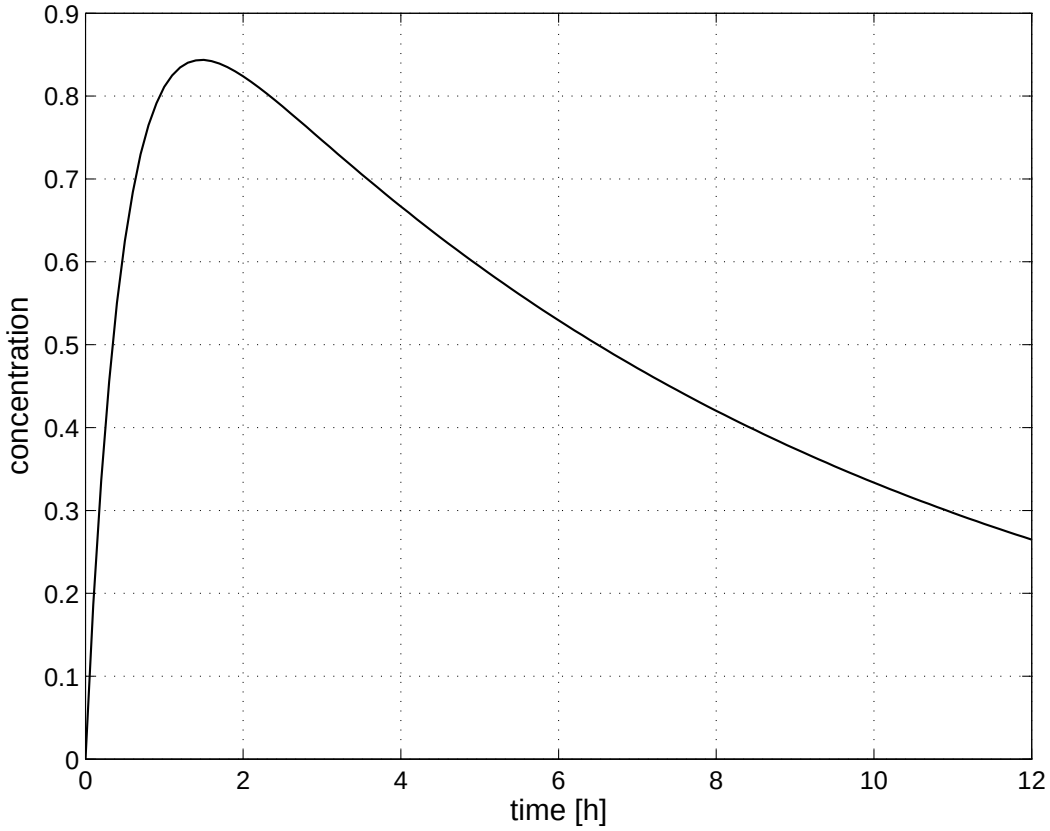
(ii) one-point s-group design

$$\zeta^\star = \begin{Bmatrix} \omega_1^\star & \dots & \omega_s^\star \\ q_1^\star & \dots & q_s^\star \end{Bmatrix}, \quad N^\star = C_0 \sum_{j=1}^{s} \frac{q_j^\star}{c(\omega_j^\star, n_j)}, \quad \omega_j^\star = \begin{Bmatrix} t_j^\star \\ 1 \end{Bmatrix}, \quad n_j > 0, \;\; j = 1, \dots, s.$$

That is

$$\zeta^\star = \begin{Bmatrix} \left(\begin{Bmatrix} t_1^\star \\ 1 \end{Bmatrix}, n_1 \right) & \dots & \left(\begin{Bmatrix} t_s^\star \\ 1 \end{Bmatrix}, n_s \right) \\ q_1^\star & \dots & q_s^\star \end{Bmatrix}; \quad N^\star$$

There are other possibilities, depending on what information we put into the reformulating systems of equations.



EXAMPLE: ONE-COMPARTMENT PK MODEL

Model

$$y = \frac{Dk_a}{V(k_e - k_e)}\left(e^{-k_e t} - e^{-k_a t}\right) + \varepsilon,$$

where k_a and k_e are the first-order absorption and elimination rates, V is the apparent volume of distribution and D is a known dose.

The regression parameters $\theta = (V, k_a, k_e)^{\mathrm{T}}$ are independent and normally distributed.

The initial values of the population parameters are:

$$\begin{aligned}\psi^0 &= \left(\mathrm{E}(V)^0, \mathrm{E}(k_a)^0, \mathrm{E}(k_e)^0, \mathrm{var}(V)^0, \mathrm{var}(k_a)^0, \mathrm{var}(k_e)^0, \mathrm{var}(\varepsilon)^0\right)^{\mathrm{T}} \\ &= (100, 2.08, 0.1155, 0.3, 0.3, 0.03, 0.15)^{\mathrm{T}}\end{aligned}$$

We are looking for a **D-optimum population design**

in time domain $T = [0.25, 12]$, for cost function $c(\xi, n) = n$ and $C_0 = 900$.

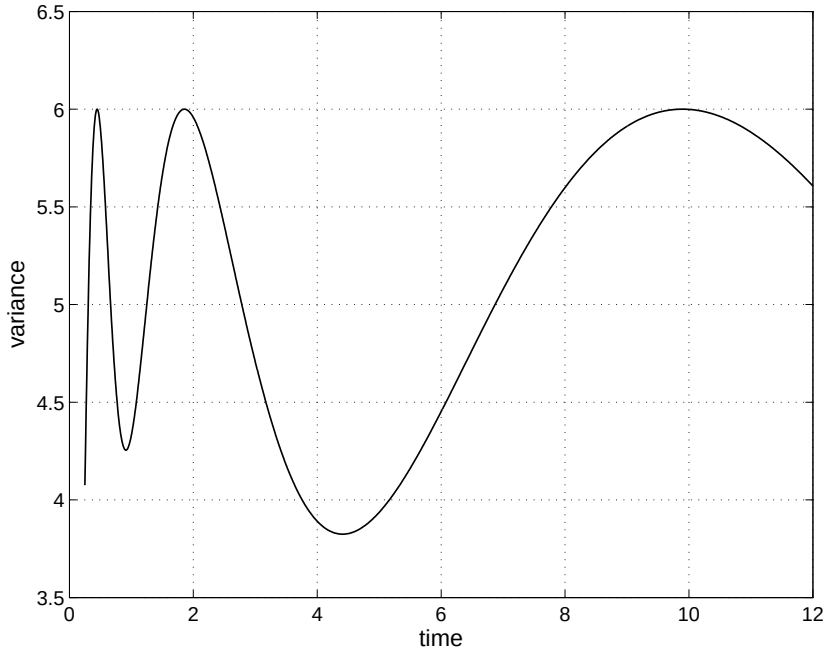


Figure 1: Variance of the response prediction.

The global design from the **Step 1** of the algorithm is:

$$\omega^{\star} = \left\{ \begin{array}{ccc} 0.45 & 1.86 & 9.90 \\ 0.3334 & 0.3334 & 0.3333 \end{array} \right\}$$

Relationships among the unknowns

$$\begin{array}{cccc|c} \gamma_1^1 & \gamma_1^2 & \cdots & \gamma_1^G & \gamma_1^\star \\ \gamma_2^1 & \gamma_2^2 & \cdots & \gamma_2^G & \gamma_2^\star \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_s^1 & \gamma_s^2 & \cdots & \gamma_s^G & \gamma_s^\star \\ \hline \beta_1 & \beta_2 & \cdots & \beta_G & 1 \end{array} \qquad \begin{array}{cccc} w_1^1 & w_1^2 & \cdots & w_1^G \\ w_2^1 & w_2^2 & \cdots & w_2^G \\ \vdots & \vdots & \vdots & \vdots \\ w_s^1 & w_s^2 & \cdots & w_s^G \\ \hline 1 & 1 & \cdots & 1 \end{array} \qquad w_i^j = \frac{\gamma_i^j}{\beta_j}$$

(1) One group design, $G = 1$

(a) In **Step 2**:

$$\Gamma = [\gamma_i^1] = [\gamma_i^\star] = \begin{bmatrix} 0.3333 \\ 0.3334 \\ 0.3333 \end{bmatrix} \Longrightarrow \beta = \Big[1\Big], \quad W = \Gamma$$

We have

$$N = C_0 \sum_{j=1}^{G} \frac{\beta_j}{n_j}, \quad \alpha_j = \frac{C_0}{N}\frac{\beta_j}{n_j}, \quad j = 1, \ldots, G.$$

(b) Assume $n_1 = 9$, $C_0 = 900$; then in **Step 3** we obtain: $\alpha = 1, \ N = 100$

with the final population design:

$$\zeta^\star = \left\{ \left(\begin{Bmatrix} 0.45 & 1.86 & 9.90 \\ 0.3333 & 0.3334 & 0.3333 \end{Bmatrix}, 9 \right) \right\}; \quad N^\star = 100.$$

$$1$$

Here, for each patient we have to conduct exactly three measurements at each time instant.

(2) one-point 3-group population design, $s = G = 3$

(a) From **Step 2**:

$$\Gamma = \begin{bmatrix} 0.3333 & 0.0 & 0.0 \\ 0.0 & 0.3334 & 0.0 \\ 0.0 & 0.0 & 0.3333 \end{bmatrix} \Longrightarrow \beta = \begin{bmatrix} 0.3333 & 0.3334 & 0.3333 \end{bmatrix}, W = I$$

(b) Assume $n_1 = n_2 = n_3 = 10$, $C_0 = 900$; then from **Step 3** we have:

$$\alpha = \beta = \begin{bmatrix} 0.3333 & 0.3334 & 0.3333 \end{bmatrix}, \quad N = 90$$

and the final population design is:

$$\zeta^\star = \begin{Bmatrix} \left(\left\{ \begin{smallmatrix} 0.45 \\ 1 \end{smallmatrix} \right\}, 10 \right) & \left(\left\{ \begin{smallmatrix} 1.86 \\ 1 \end{smallmatrix} \right\}, 10 \right) & \left(\left\{ \begin{smallmatrix} 9.90 \\ 1 \end{smallmatrix} \right\}, 10 \right) \\ 0.3333 & 0.3334 & 0.3333 \end{Bmatrix}; \quad N^\star = 90$$

Here, each group has 30 patients, each patient has 10 samples taken at the same time instant.

(3) unstructured design, $G = 3$

(a) In **Step 2** the weights of the global design are randomly split into the required number of groups and number of support points in individual group designs, e.g.

$$\Gamma = \begin{bmatrix} 0.2298 & 0.1036 & 0 \\ 0 & 0.2089 & 0.1245 \\ 0.1710 & 0.0855 & 0.0768 \end{bmatrix},$$

then

$$\beta = \begin{bmatrix} 0.4008 & 0.3979 & 0.2013 \end{bmatrix}, \quad W = \begin{bmatrix} 0.5733 & 0.2603 & 0 \\ 0 & 0.5248 & 0.6184 \\ 0.4267 & 0.2149 & 0.3816 \end{bmatrix}$$

(b) Assume $n_1 = n_2 = n_3 = 10$, $C_0 = 900$; then from **Step 3** we have:

$$\alpha = \beta = \begin{bmatrix} 0.4008 & 0.3979 & 0.2013 \end{bmatrix}, \quad N = 90.$$

and the population design:

$$\zeta^\star = \left\{ \begin{array}{ccc} \left(\left\{ \begin{array}{cc} 0.45 & 9.90 \\ 0.5733 & 0.4267 \end{array} \right\}, 10 \right) & \left(\left\{ \begin{array}{ccc} 0.45 & 1.86 & 9.90 \\ 0.2603 & 0.5248 & 0.2149 \end{array} \right\}, 10 \right) & \left(\left\{ \begin{array}{cc} 1.86 & 9.90 \\ 0.6184 & 0.3816 \end{array} \right\}, 10 \right) \\ 0.4008 & 0.3979 & 0.2013 \end{array} \right\}, \quad N^\star = 90$$

More realistic design:

$$\zeta = \left\{ \begin{array}{ccc} \left(\left\{ \begin{array}{cc} 0.45 & 9.90 \\ 0.6 & 0.4 \end{array} \right\}, 10 \right) & \left(\left\{ \begin{array}{ccc} 0.45 & 1.86 & 9.90 \\ 0.3 & 0.5 & 0.2 \end{array} \right\}, 10 \right) & \left(\left\{ \begin{array}{cc} 1.86 & 9.90 \\ 0.6 & 0.4 \end{array} \right\}, 10 \right) \\ 0.4 & 0.4 & 0.2 \end{array} \right\}, \quad N^\star = 90$$

Efficiency:

$$E_\zeta = \left(\frac{\det M(\zeta, N^\star)}{\det M(\zeta^\star, N^\star)} \right)^{1/6} = 0.9984$$

Note:

- Large total number of measurements ensures small decrease of the efficiency caused by rounding.
- However, the flexibility in constructing an optimal design allows us to reduce the influence of rounding on the efficiency of the final design.

(4) unstructured design, $G = 3$

(a) In **Step 2** we can set the elements of matrix W to be appropriate ratios, e.g.

$$W = \begin{bmatrix} 0 & 0 & \frac{3}{10} \\ 0 & \frac{2}{5} & \frac{7}{10} \\ 1 & \frac{3}{5} & 0 \end{bmatrix}.$$

Then we obtain

$$\Gamma = \begin{bmatrix} 0 & 0 & 0.1000 \\ 0 & 0.1334 & 0.2333 \\ 0.3333 & 0.2000 & 0 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0.3333 & 0.3334 & 0.3333 \end{bmatrix}$$

(b) Assume $n_1 = n_2 = n_3 = 10$, $C_0 = 900$; then from **Step 3** we get:

$$\alpha = \beta = \begin{bmatrix} 0.3333 & 0.3334 & 0.3333 \end{bmatrix}, \quad N = 90.$$

Final population design:

$$\zeta^\star = \left\{ \begin{array}{ccc} \left(\left\{ \begin{array}{c} 9.90 \\ 1 \end{array} \right\}, 10 \right) & \left(\left\{ \begin{array}{cc} 1.86 & 9.90 \\ \frac{2}{5} & \frac{3}{5} \end{array} \right\}, 10 \right) & \left(\left\{ \begin{array}{cc} 0.45 & 1.86 \\ \frac{3}{10} & \frac{7}{10} \end{array} \right\}, 10 \right) \\ 0.3333 & 0.3334 & 0.3333 \end{array} \right\}, \quad N^\star = 90$$

Rounding the global weights to $\frac{1}{3}$ we obtain

$$\zeta = \left\{ \begin{array}{ccc} \left(\left\{ \begin{array}{c} 9.90 \\ 1 \end{array} \right\}, 10 \right) & \left(\left\{ \begin{array}{cc} 1.86 & 9.90 \\ \frac{2}{5} & \frac{3}{5} \end{array} \right\}, 10 \right) & \left(\left\{ \begin{array}{cc} 0.45 & 1.86 \\ \frac{3}{10} & \frac{7}{10} \end{array} \right\}, 10 \right) \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \right\}, \quad N^\star = 90$$

and

$$E_\zeta = \left(\frac{\det M(\zeta, N^\star)}{\det M(\zeta^\star, N^\star)} \right)^{1/6} \approx 1.0000$$

CONCLUDING REMARKS

- The definition of the optimal population design we have presented leads to non-unique solutions.
- It gives room for tailoring optimum designs to practical requirements.
- It allows the use of additional information an experimenter may have.
- It gives an experimenter some freedom to impose additional constraints on the design variables.
- It incorporates the maximum cost of the experiment into the design.
- The Equivalence Theorem works for the global design as well as for the “intermediate state” design.

OPEN PROBLEMS

- Make groups meaningful.
- Introduce correlations among observations (within groups, within patients)
- Produce user-friendly software.
- ?