

BAYESIAN LOOK AHEAD ONE-STAGE SAMPLING  
ALLOCATIONS FOR SELECTION OF THE BEST POPULATION

by

Shanti S. Gupta  
Purdue University

and

Klaus J. Miescke  
University of Illinois  
at Chicago

Technical Report #94-30

Department of Statistics  
Purdue University

November 1994  
Revised January 1995

**BAYESIAN LOOK AHEAD ONE-STAGE SAMPLING ALLOCATIONS  
FOR SELECTION OF THE BEST POPULATION \***

SHANTI S. GUPTA  
Department of Statistics  
Purdue University  
Math Sciences Bldg  
West Lafayette, IN 47907-1399  
U.S.A.

KLAUS J. MIESCKE  
Department of Mathematics, Statistics,  
and Computer Science (M/C 249)  
University of Illinois at Chicago  
851 South Morgan  
Chicago, IL 60607-7045  
U.S.A.

Revised January 1995

**AMS(1980) Subject Classification.** Primary 62F07; Secondary 62F15.

**Key words and phrases.** Bayesian look ahead; selection and ranking; optimum sampling allocations; normal populations.

\* This research was supported in part by NSF Grant DMS-8923071 at Purdue University.

## SUMMARY

From  $k$  independent normal populations with unknown means and a common known variance, Bayesian selection procedures are considered for finding that population which has the largest mean. Suppose that a first stage has been completed already, where  $k$  samples have been observed, which may be of different sizes. Let there be  $m$  additional observations allowed to be taken at a future second stage. The problem of interest treated here is how to allocate these  $m$  observations in an optimum way among the  $k$  populations, given all of the information, prior and first stage observations, gathered so far. This allocation problem will be formulated and discussed in a more general framework, and specific results will be presented for the normal case with independent conjugate priors under linear loss.

**Abbreviated title.** Bayesian Sampling Allocations.

## 1. Introduction

Let  $\mathcal{P}_1, \dots, \mathcal{P}_k$  be  $k \geq 2$  given populations which are associated with unknown real parameters  $\theta_1, \dots, \theta_k \in \mathbb{R}$  from a common underlying exponential family  $\mathcal{F}$ . Suppose we want to find that population which has the largest  $\theta$ -value, using a Bayes selection rule which is based on a known prior density  $\pi(\boldsymbol{\theta})$  for  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ , and a given loss function  $L(\boldsymbol{\theta}, i)$  for selecting  $i \in \{1, 2, \dots, k\}$  at  $\boldsymbol{\theta} \in \mathbb{R}^k$ . The purpose of this paper is to provide an introduction to and an overview of multi-stage selection procedures in the Bayesian approach. Recent results on Bayesian look ahead one-stage selection procedures for means of normal populations are described in details, and directions of future research are indicated.

Multi-stage selection procedures have been studied extensively in the past. In their pioneering monograph, Bechhofer, Kiefer, and Sobel (1968) derived, in the frequentist approach, similar to Wald's SPRT, optimum sequential rules based on vector-at-a-time sampling, the natural terminal selection, and an optimum stopping rule. However, elimination of populations, i.e. excluding populations from further sampling and selection, at the intermediate stages were not allowed there. Later work, which includes elimination, was done in the Bayesian approach by Gupta and Miescke (1984), this time, however, only for truncated sequential procedures, which are procedures with a predetermined maximum number of stages. Overviews of this area are provided by Gupta and Panchapakesan (1979) and by Miescke (1984). The spirit behind all these efforts is to select the best population most efficiently in terms of both, risk and sampling costs. The vehicle that allowed to derive strong results was 'symmetry', delivered through vector-at-a-time sampling of not yet eliminated populations and permutation symmetric priors.

Elimination of populations, as defined above, can lead to a conflicting situation when the data collected from a population before it was eliminated later make it look better again, after further sampling has been done from other populations which then turn out to be not so favorable for those. Especially in the Bayesian approach, since Bayes procedures make use of all the data collected, this is a realistic possibility. Elimination of populations from future selections is thus seen to possibly have undesirable effects. Therefore, not

incorporating elimination but considering instead allocation of unequal sample sizes at various stages seems to be more appropriate. The latter acts like temporary eliminations, but leaves all  $k$  populations in the pool for the final selection decision. This approach allows also the use of nonsymmetric priors, which occur in a natural way as updated priors at various stages.

Getting unequal sample sizes at some stage of an ongoing experiment is a matter of life. Be it a patient who does not return to a scheduled exam, a laboratory animal which is lost, a testing equipment that breaks down, a shortened deadline for the experimental phase - symmetric models often are hard to enforce. To deal with such situations, let us assume that  $k$  independent samples of sizes  $n_1, \dots, n_k$ , respectively, have been observed already at the first stage, say, which may represent the combined outcomes of several previous stages, and that  $m$  additional observations are allowed to be taken at a future second stage. The problem of interest treated here is how to allocate these  $m = m_1 + \dots + m_k$  observations in an optimum way among the  $k$  populations, given all the information, prior and first stage observations, gathered so far. This was done previously by Gupta and Miescke (1993) for simultaneous selection and estimation in the binomial case, including costs of sampling. For simplicity of presentation, the latter, which would involve stopping rules, will not be considered here. It should be noted that the special case of  $n_1 = \dots = n_k = 0$  represents the analogous problem of how to allocate  $m = m_1 + \dots + m_k$  observations at a first stage.

Looking ahead one stage using the expected posterior Bayes risk, given the prior and all the observations collected so far, and then minimizing it, does not only provide an optimum allocation of observations in the future. It also allows to assess how much better the final decision can be expected to be after further sampling has been done, following this optimum allocation. If costs of sampling at the second stage would be incorporated in the loss, an optimum stopping rule would have to be implemented, too. Modifications of the allocations considered later on to this case are straightforward and thus omitted in the discussions for brevity.

In many empirical studies in marketing research (e.g. direct marketing), medical research (e.g. clinical trials, cf. Whitehead 1991), and social research (e.g. survey sampling, cf. Govindarajulu and Katehakis, 1991), there are interim analyses performed at certain

stages to decide if sampling should be continued, and if so, how to allocate new observations. The latter cannot be done before it is known which terminal selection rule will be used at the end. Thus, the first optimization step is to find optimum (Bayes) single-stage selection rules under unequal sample sizes. This has been done for the binomial case in Abughalous and Miescke (1989), and for the normal case in Gupta and Miescke (1988). Once the optimum terminal selection rules are fixed, however, one can look ahead with various sampling allocations, compare the associated expected posterior risks, and then go for the smallest possible value. There is, however, one more complication that has to be taken care of. Why allocating all  $m$  observations at once, rather than allocating only a few (or just one), learning more through them (it), and then having a stronger basis for the remaining allocations?

In a fully sequential approach, the optimum allocation could be constructed, at least in principle, by finding first the Bayesian terminal selection rule for every possible allocation  $m = m_1 + \dots + m_k$ , and then use backward optimization (or possibly backward induction) to optimize successively every single allocation before. Although the former are not too hard to find, the latter appears to be infeasible to be carried out in practice. Therefore, one reasonable procedure proposed in this paper later on is to allocate in an optimum way one observation at a time, pretending that it is the last one to be drawn before final selection, and then iterate this process until  $m$  observations have been taken. Allocating more than one observation at a time, on the other hand, appears to be less appealing since with each new observation more is learnt about the unknown parameters, which improves the basis for further decisions. In certain applications (e.g. in some clinical trials) it may be desired, or even mandatory, to choose each next observation (treatment) according to the "state-of-the-art". This would require to select the best population first, and then allocate the next observation (treatment) to it, with iterations of this process. This policy is different from the present one under consideration, where emphasis is made on optimum terminal selections. Look ahead procedures, which have been utilized previously by Govindarajulu and Katehakis (1991) in survey sampling, and which are described and discussed in various other settings in Berger (1985), Chapter 7.4, will be considered in Section 2.

For the normal means case, under the assumption of  $k$  independent normal priors, and either the linear loss or the 0-1-loss, a solution to the problem has been obtained previously by Gupta and Miescke (1994) for the case of  $k = 2$ , which turns out to be already rather involved. It allocates the  $m = m_1 + m_2$  observations in such a way that the posterior gets as close as possible to being decreasing in transposition (DT). Moreover, somewhat surprising, it does not depend at all on the observations gathered so far at the first stage. This fact implies that one can allocate, in an optimum way, one new observation at a time, until all  $m$  have been drawn, thereby arriving at the same allocation as in the former approach. For  $k \geq 3$  populations, this is no longer true! It turns out that the observations from Stage 1 are relevant for further allocations, and this in an interesting manner. This will be shown and discussed in Section 3. Similar considerations for the 0-1-loss will not be pursued in this paper, since they are much harder to make, and since this loss has the undesired discontinuity property around parameter configurations with ties.

Selecting the population with the largest (overall) sample mean is usually called the natural selection rule, since it is the uniformly best permutation invariant selection procedure, in the frequentist sense, for a general class of loss functions, provided that the sample sizes are equal. However, for unequal sample sizes, the natural selection rule loses much of its quality, although it still remains intuitively appealing. Because of the latter, optimum sample size allocations for the natural selection rule have been considered in the frequentist approach by Bechhofer (1969), Dudewicz and Dalal (1975), and Bechhofer, Hayter, and Tamhane (1991). On the other hand, Bayes selection rules with normal priors have complicated forms which typically cannot be represented in closed form, except for those situations where the posterior is DT, as it has been shown in Gupta and Miescke (1988). Bayes rules for similar but more involved models have been studied by Berger and Deely (1988) and Fong and Berger (1993). Earlier ideas of sampling allocations for Bayes rules under normality and linear loss have been presented and discussed thoroughly by Dunnett (1960).

For the present setting, it has been recommended previously by Gupta and Miescke (1988) to plan the experiment's sampling allocation in such a way that the posterior is DT. For the normal means case, this is corroborated by the optimum allocation for  $k = 2$ .

However, for more than two normal populations, it turns out that there is some tradeoff necessary between this suggestion and the values observed at the first stage, except for those two populations which have yielded the largest two posterior means. This will be explained and discussed in Section 3.

## 2. Bayesian Look Ahead Sampling Allocations

After a standard reduction of the data by sufficiency, the model assumptions can be summarized as follows: At  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$ , let  $X_i$  and  $Y_i$  be sufficient statistics of the samples from population  $\mathcal{P}_i$  at Stage 1 and Stage 2, resp.,  $i = 1, \dots, k$ , which altogether are assumed to be independent. A priori, the parameters are considered as realizations of a random variable  $\boldsymbol{\Theta} = (\Theta_1, \dots, \Theta_k)$  which follows a given prior distribution. Let the loss for selecting  $\mathcal{P}_i$ , at both stages, be  $L(\boldsymbol{\theta}, i)$ , at  $\boldsymbol{\theta} \in \mathbb{R}^k$ ,  $i = 1, \dots, k$ . Costs of sampling, which would require the incorporation of a stopping rule, are not included in the loss in favor of a simpler presentation of basic ideas. Modification of the allocations discussed below to this more general setting are straightforward.

Having observed  $\mathbf{X} = \mathbf{x} \in \mathbb{R}^k$  at Stage 1, based on samples of sizes  $n_1, \dots, n_k$ , every (it may not be unique) Bayes selection rule  $d_1^*(\mathbf{x})$  satisfies

$$E\{L(\boldsymbol{\Theta}, d_1^*(\mathbf{x})) | \mathbf{X} = \mathbf{x}\} = \min_{i=1, \dots, k} E\{L(\boldsymbol{\Theta}, i) | \mathbf{X} = \mathbf{x}\}. \quad (1)$$

Likewise, after  $\mathbf{Y} = \mathbf{y} \in \mathbb{R}^k$  has been observed at Stage 2, every Bayes rule  $d_2^*(\mathbf{x}, \mathbf{y})$  satisfies

$$E\{L(\boldsymbol{\Theta}, d_2^*(\mathbf{x}, \mathbf{y})) | \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} = \min_{i=1, \dots, k} E\{L(\boldsymbol{\Theta}, i) | \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\}. \quad (2)$$

There will be no need to consider randomized Bayes selection rules in the following, since minimaxity and invariance concepts will not be used here. Whenever at least one Bayes rule exists, one can choose a nonrandomized version.

Many results for Bayes selection rules in symmetric models can be found in the literature. An overview is provided by Gupta and Panchapakesan (1979, 1991). Only recently, however, attention has been given also to nonsymmetric models. The binomial case has been treated in Abughalous and Miescke (1989), the normal case in Gupta and Miescke



(1988), and more involved models have been considered in Berger and Deely(1988) and in Fong and Berger (1993). Rather than studying the properties of such Bayes selection rules  $d_1^*$  and  $d_2^*$  in more details, let us assume here that they have been derived already, ready to be used by the experimenter, and that all that is left to do is to allocate sample sizes in an optimum manner.

Let us now consider fixed total sample size allocation problems. Before entering each of the two stages, similar allocation problems of this type arise, which are closely related. Before entering Stage 1, by looking ahead one stage, we would like to minimize the expected posterior risk subject to  $n_1 + \dots + n_k = n$ , where  $n$  is the total number of observations allowed to be taken at Stage 1. This leads to the following criterion for  $n_1, \dots, n_k$ .

$$\min_{\substack{n_1, \dots, n_k \\ n_1 + \dots + n_k = n}} E\left(\min_{i=1, \dots, k} E\{L(\Theta, i)|X\}\right). \quad (3)$$

Likewise, at the end of Stage 1, the criterion for  $m_1, \dots, m_k$ , with a total number of  $m$  observations allowed at Stage 2, is the following.

$$\min_{\substack{m_1, \dots, m_k \\ m_1 + \dots + m_k = m}} E\left\{\min_{i=1, \dots, k} E\{L(\Theta, i)|X = \mathbf{x}, Y\}|X = \mathbf{x}\right\}, \quad (4)$$

where the outer expectation is with respect to the conditional distribution of  $Y$ , given  $X = \mathbf{x}$ .

Formally, or by using a standard sequential updating argument as it is described in Berger (1985) on page 445, one can get solutions of (3) from those of (4) by setting  $n_1 = \dots = n_k = 0$  and then relabel  $m_i$  by  $n_i$ ,  $i = 1, \dots, k$ , and  $m$  by  $n$ . Thus, we need to consider only criterion (4) in the sequel, since it is more general.

Taking into account that optimum future sampling allocations do depend on the past observations, it seems more appropriate to break down the allocation of  $m$  observations into more than one step. Several sampling allocation schemes are reasonable. Let  $\mathcal{R}_t$ , for  $t \leq m$ , denote the fixed total number  $t$  of observations allocation determined by (4), with  $m$  replaced by  $t$  there. Furthermore, let  $\mathcal{R}_{t,1}$  allocate any one, but only one, observation to one of the populations suggested by  $\mathcal{R}_t$ . Finally, let  $\mathcal{B}_1$  stand for the optimum allocation of one observation, knowing all future actions, whatever they may be in a particular

situation. The allocation rule which allocates the last  $m - 1$  observations using (4), with  $m$  replaced by  $m - 1$  there, and the first observation through backward optimization, can thus be represented by  $(\mathcal{B}_1, \mathcal{R}_{m-1})$ . Likewise,  $(\mathcal{B}_1, \mathcal{R}_{m-1,1}, \mathcal{R}_{m-2})$  allocates the last  $m - 2$  observations using  $\mathcal{R}_{m-2}$ , the second observation to one of the populations suggested by  $\mathcal{R}_{m-1}$  after the first observation has been drawn, and the first observation by backward optimization. It should be pointed out that  $\mathcal{R}_t$  and  $\mathcal{R}_{t,1}$  are stand-alone procedures, i.e. procedures which can be used directly without knowing future actions, whereas  $\mathcal{B}_1$  is only meaningful in connection with completely specified future actions.

The best sampling allocation of  $m$  observations is determined by an optimum allocation of the  $m$ -th (i.e. the last) allocation, and then by using backward optimization (or if possible backward induction), i.e. the procedure  $(\mathcal{B}_1, \mathcal{B}_1, \dots, \mathcal{B}_1, \mathcal{R}_1)$ , with  $m - 1$  repetitions of  $\mathcal{B}_1$ . For the purpose of comparisons, let us consider the following partial ordering of allocation schemes: let  $\mathcal{S} \prec (=) \mathcal{T}$  mean that procedure  $\mathcal{T}$  has a smaller (the same) overall Bayes risk than (as) procedure  $\mathcal{S}$ .

Using the basic fact that, except for trivial cases where equality may occur, every relevant (conditional) expectation of a minimum is smaller than the minimum of the respective (conditional) expectations, some interesting chains of preferences can be established. The proofs are straightforward and therefore omitted.

**Lemma 1.** *The following chain of preferences holds.*

$$\begin{aligned} (\mathcal{R}_m) \prec (\mathcal{R}_{m,1}, \mathcal{R}_{m-1}) \prec (\mathcal{B}_1, \mathcal{R}_{m-1}) \prec (\mathcal{B}_1, \mathcal{R}_{m-1,1}, \mathcal{R}_{m-2}) \prec \\ \prec (\mathcal{B}_1, \mathcal{B}_1, \mathcal{R}_{m-2}) \prec \dots \prec (\mathcal{B}_1, \mathcal{B}_1, \dots, \mathcal{B}_1, \mathcal{R}_1). \end{aligned} \quad (5)$$

Since backward optimization, i.e. using  $\mathcal{B}_1$ , appears to be infeasible in most situations, one may be inclined to capitalize on one of the following sampling allocation rules, which may have good performance characteristics:

**Lemma 2.** *The following chain of preferences holds.*

$$\begin{aligned} (\mathcal{R}_m) \prec (\mathcal{R}_{m,1}, \mathcal{R}_{m-1}) \prec (\mathcal{R}_{m,1}, \mathcal{R}_{m-1,1}, \mathcal{R}_{m-2}) \prec \dots \prec \\ \prec (\mathcal{R}_{m,1}, \mathcal{R}_{m-1,1}, \dots, \mathcal{R}_{2,1}, \mathcal{R}_1). \end{aligned} \quad (6)$$

There is one problem, however, that arises when one attempts to use procedure  $\mathcal{R}_{t,1}$ . It is not clear from which one of the populations, suggested by  $\mathcal{R}_t$ , the next observation should be drawn. The natural alternative is to use instead one of the following simpler allocation rules, which are easier to determine:

$$(\mathcal{R}_1, \mathcal{R}_{m-1}), (\mathcal{R}_1, \mathcal{R}_1, \mathcal{R}_{m-2}), \dots, (\mathcal{R}_1, \mathcal{R}_1, \dots, \mathcal{R}_1), \quad (7)$$

where the last allocation rule consists of  $m$  repetitions of  $\mathcal{R}_1$ . Also here some conflict arises. Whenever an  $\mathcal{R}_{t,1}$  is replaced by an  $\mathcal{R}_1$ , it remains unclear whether the latter picks one population from those available for  $\mathcal{R}_{t,1}$  from  $\mathcal{R}_t$ , i.e. whether  $\mathcal{R}_1$  is a version of  $\mathcal{R}_{t,1}$ . This conflict, however, appears to be minor, considering the simplicity in usage one has gained. Further research in this direction, however, seems to be appropriate to corroborate this standpoint.

The allocation rule  $(\mathcal{R}_1, \mathcal{R}_1, \dots, \mathcal{R}_1)$  allocates at each step one more observation in an optimum manner, pretending that it would be the last one before final selection. Thus for every single new observation (treatment), it "selects" that population (treatment) which appears to be the best to improve the pretended final selection decision after it has been drawn (applied). As it has been mentioned already in Section 1, this allocation policy is different from the "state-of-the-art" allocation, which would first select that population which appears to be the best at the present moment, and then allocate the next observation (treatment) to it.

### 3. The Normal Case Under Linear Loss

Let  $\mathcal{P}_1, \dots, \mathcal{P}_k$  be  $k \geq 3$  given normal populations with means  $\theta_1, \dots, \theta_k \in \mathbb{R}$  and a common known variance  $\sigma^2 > 0$ . For convenience, and to facilitate comparisons with Gupta and Miescke (1994), the same notation used there will be also adopted here. After a standard reduction of the data by sufficiency, the model assumptions can be summarized as follows: At  $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$ ,  $X_i \sim N(\theta_i, p_i^{-1})$  with  $p_i^{-1} = \sigma^2/n_i$ , and  $Y_i \sim N(\theta_i, q_i^{-1})$  with  $q_i^{-1} = \sigma^2/m_i$ , are the independent sample means of the samples from  $\mathcal{P}_i$  at Stage 1 and Stage 2, respectively,  $i = 1, \dots, k$ , which altogether are assumed to be independent. A priori, the parameters  $\Theta = (\Theta_1, \dots, \Theta_k)$  are considered to be random and follow a given

prior distribution where  $\Theta_i \sim N(\mu_i, \nu_i^{-1})$ ,  $i = 1, \dots, k$ , are independent. We assume from now on that the loss is of the following form:  $L(\boldsymbol{\theta}, i) = \theta_{[k]} - \theta_i$ ,  $i = 1, \dots, k$ ,  $\boldsymbol{\theta} \in \mathbb{R}^k$ , where  $\theta_{[k]} = \max\{\theta_1, \dots, \theta_k\}$ . Although it is not a linear function of  $\boldsymbol{\theta} \in \mathbb{R}^k$ , it is traditionally called linear loss, cf. Gupta and Panchapakesan (1979). Costs of sampling are not incorporated into the loss to simplify the presentation of basic ideas. The stopping rule is considered here to be deterministic: the sampling process stops after  $m$  observations have been drawn.

To evaluate the inner conditional expectation of (4), note that the conditional distribution of  $\boldsymbol{\Theta}$ , given  $\mathbf{X} = \mathbf{x}$  and  $\mathbf{Y} = \mathbf{y}$ , has to be used, which is:

$$\Theta_i \sim N\left(\frac{\alpha_i \mu_i(\mathbf{x}) + q_i y_i}{\alpha_i + q_i}, \frac{1}{\alpha_i + q_i}\right), \quad i = 1, \dots, k, \text{ independent}, \quad (8)$$

where  $\alpha_i = p_i + \nu_i$  represents the relative combined (prior plus sampling) information gained from population  $\mathcal{P}_i$ , and where  $\mu_i(\mathbf{x}) = (\nu_i \mu_i + p_i x_i)/(\nu_i + p_i)$ , represents the relative strength or quality of population  $\mathcal{P}_i$  at the present moment,  $i = 1, \dots, k$ . The latter would be the criterion used for "state-of-the-art" allocations.

To evaluate the outer conditional expectation of (4), the conditional distribution of  $\mathbf{Y}$ , given  $\mathbf{X} = \mathbf{x}$ , has to be used, which is:

$$Y_i \sim N\left(\mu_i(\mathbf{x}), \frac{\alpha_i + q_i}{\alpha_i q_i}\right), \quad i = 1, \dots, k, \text{ independent}. \quad (9)$$

Under the linear loss, criterion (4) reduces to minimize as a function of  $m_1, \dots, m_k$ , subject to  $m_1 + \dots + m_k = m$ , the look ahead expected posterior risk

$$E\{\Theta_{[k]}|\mathbf{X} = \mathbf{x}\} - E\{\max_{i=1, \dots, k} E\{\Theta_i|\mathbf{X} = \mathbf{x}, \mathbf{Y}\}|\mathbf{X} = \mathbf{x}\}. \quad (10)$$

This criterion reduces further, by using first (8) and then (9), to maximize, subject to the side condition  $m_1 + \dots + m_k = m$ , the following quantity:

$$E\{\max_{i=1, \dots, k} \frac{\alpha_i \mu_i(\mathbf{x}) + q_i Y_i}{\alpha_i + q_i} | \mathbf{X} = \mathbf{x}\} = E\left(\max_{i=1, \dots, k} \left[ \mu_i(\mathbf{x}) + \left(\frac{q_i}{\alpha_i(\alpha_i + q_i)}\right)^{\frac{1}{2}} N_i \right]\right), \quad (11)$$

where  $N_1, \dots, N_k$  are generic independent standard normal random variables. The quantity in (11) will be called 'look ahead expected posterior gain' in the sequel.

Let us recall that a distribution with density (conditional density)  $h(\mathbf{u}|\mathbf{v})$ ,  $\mathbf{u} \in \mathbb{R}^k$ , where  $\mathbf{v} \in \mathbb{R}^k$  is a parameter (value of a random vector), is said to have the decreasing in transposition (DT) property if the following two conditions hold:  $h(\pi(\mathbf{u})|\pi(\mathbf{v})) \equiv h(\mathbf{u}|\mathbf{v})$  for every permutation  $\pi$ , and  $h(\mathbf{u}|\mathbf{v}) \leq h(\mathbf{u}|\pi_{(i,j)}\mathbf{v})$  if  $(u_i - u_j)(v_i - v_j) \leq 0$  and  $\pi_{(i,j)}$  is the permutation which exchanges the  $i$ -th and  $j$ -th coordinate. Functions decreasing in transposition have been studied thoroughly by Hollander, Proschan, and Sethuraman (1977).

Now, the posterior distribution of  $\Theta_1, \dots, \Theta_k$  at Stage 2, as given by (8), can be seen to be DT if and only if

$$\alpha_1 + q_1 = \alpha_2 + q_2 = \dots = \alpha_k + q_k . \quad (12)$$

It was recommended by Gupta and Miescke (1988) to plan the experiment in such a that  $\alpha_1 + q_1, \dots, \alpha_k + q_k$  are equal. The Bayes rule is then in a simple and closed form, and the posterior information about  $\theta_1, \dots, \theta_k$  is equally balanced. The solutions of criterion (4) for  $k = 2$  under linear and under 0-1-loss, as shown in Gupta and Miescke (1994), turn out to be the same: Choose  $m_1$  and  $m_2$  in such a way that  $|\alpha_1 + q_1 - \alpha_2 - q_2|$  is minimum, subject to  $m_1 + m_2 = m$ . This solution does not depend on any previous observations, and it coincides with both, the rule  $(\mathcal{R}_1, \mathcal{R}_1, \dots, \mathcal{R}_1)$  and the optimum rule  $(\mathcal{B}_1, \mathcal{B}_1, \dots, \mathcal{B}_1, \mathcal{R}_1)$ . As it will be seen below, this is no longer true for  $k \geq 3$  populations.

To study the properties  $E(\max_{i=1, \dots, k} [a_i + b_i N_i])$  as a function of  $a_i \in \mathbb{R}$  and  $b_i > 0$ ,  $i = 1, \dots, k$ , it proves useful to introduce the following auxiliary function  $T$ , given by

$$T(w) = w \Phi(w) + \varphi(w) = \int_{-\infty}^w \Phi(u) du , \quad w \in \mathbb{R}, \quad (13)$$

where  $\Phi$  and  $\varphi$  denote the standard normal c.d.f. and density, respectively. This function has been introduced and studied previously in Miescke (1979), but there only for the case where  $b_1, \dots, b_k$  are equal.  $T$  is positive, strictly increasing, and convex. Moreover, it has the following basic properties, which can be verified easily:  $T(w) > w$  , for  $w > 0$  ;  $T(w) = T(-w) + w$  , for  $w \in \mathbb{R}$  ;  $\gamma T(w/\gamma)$  is strictly increasing in  $\gamma \in \mathbb{R}$  for  $w \in \mathbb{R}$ , with a discontinuity jump from 0 to  $w$  ( from  $w$  to 0 ) at  $\gamma = 0$  if  $w > 0$  (  $w < 0$  ) ;  $T(w) = E(\max[N_1, w])$  , for  $w \in \mathbb{R}$  . To simplify notation, let  $M_k = \max_{i=1, \dots, k} [a_i + b_i N_i]$  with  $a_i \in \mathbb{R}$  and  $b_i > 0$  ,  $i = 1, \dots, k$  , in the following.

**Lemma 3.** *The following equations hold.*

- (i)  $E(T(M_1)) = (b_1^2 + 1)^{1/2} T(a_1/(b_1^2 + 1)^{1/2}).$
- (ii)  $E(M_2) = a_1 + (b_1^2 + b_2^2)^{1/2} T((a_2 - a_1)/(b_1^2 + b_2^2)^{1/2}).$
- (iii)  $E(M_k) = a_k + b_k E(T((M_{k-1} - a_k)/b_k)) > a_k + b_k T((E(M_{k-1}) - a_k)/b_k).$
- (iv)  $E(M_k)$  is increasing in  $a_i \in \mathbf{R}$  and  $b_i > 0$ ,  $i = 1, \dots, k$ .

**Proof:** Let  $M_1, \dots, M_k$  be as given above.

(i) Let  $a_1 = a$  and  $b_1 = b$  for brevity. Then one can see that

$$\begin{aligned} E(T(M_1)) &= a \int_{\mathbf{R}} \Phi(a + bz) \varphi(z) dz + b \int_{\mathbf{R}} \Phi(a + bz) z \varphi(z) dz \\ &\quad + \int_{\mathbf{R}} \varphi(a + bz) \varphi(z) dz \\ &= a \Phi(a/(b^2 + 1)^{1/2}) + (b^2 + 1) \int_{\mathbf{R}} \varphi(a + bz) \varphi(z) dz. \end{aligned}$$

Using integration by parts, the second integral could be brought into the form of the third. Combining now the two  $\varphi$ -functions into one exponential function, standard calculations show that (i) holds.

(iii) From the representation  $M_k = a_k + b_k E(\max[N_k, (M_{k-1} - a_k)/b_k])$  and the identity  $T(w) \equiv E(\max[N_1, w])$ , one can see that the conditional expectation of  $M_k$ , given  $N_1, \dots, N_{k-1}$ , is equal to  $a_k + b_k T((M_{k-1} - a_k)/b_k)$ . The inequality follows from Jensen's inequality applied to the convex function  $T$ .

(ii) This is seen by applying (i) to the equation in (iii) with  $k = 2$ .

(iv) The monotonicity with respect to  $a_1, \dots, a_k$  is obviously correct. As to  $b_1, \dots, b_k$ , consider without loss of generality  $b_k$ . We know that  $\gamma T(w/\gamma)$  is increasing in  $\gamma$  for every  $w \in \mathbf{R}$ . An application of this fact to the equation in (iii) completes the proof.  $\square$

**Remark 1.** The equation and the inequality in (iii) of the lemma can be used iteratively to get lower bounds on  $E(M_k)$  in terms of  $E(M_r)$  for  $r = 2, \dots, k - 1$ .  $E(M_2)$  has a closed form representation given by (ii) in the lemma.

The remainder of this section is devoted to a detailed study of the properties of the allocation rule  $\mathcal{R}_1$ , and finally to some discussion of optimum allocations for an experiment which is planned for a total of  $n$  observations.  $\mathcal{R}_1$  is found by maximizing the look ahead expected posterior gain given in (11), subject to one of the values  $q_i = m_i/\sigma^2, i = 1, \dots, k$ , being equal to  $q = 1/\sigma^2$  and all others being equal to zero. All of the results derived below are given in terms of  $q$ , but it should be pointed out that they also hold true for any  $q = h/\sigma^2$  with  $1 \leq h \leq m$ , i.e. for the analogous allocation rule which optimizes allocation of  $h$  observations to exactly one population, where  $h \in \{1, \dots, m\}$ . For this case, the formula for the look ahead expected posterior gain given in (11) simplifies to a closed form as given in the following lemma, which will serve as the main tool to set up a useful algorithm for finding the desired optimum sampling allocation.

**Lemma 4.** *For  $q_i = q$ , for  $q_j = 0, j \neq i$ , and  $i$  fixed, we have*

$$E\left\{ \max_{s=1, \dots, k} \frac{\alpha_s \mu_s(\mathbf{x}) + q_s Y_s}{\alpha_s + q_s} \mid \mathbf{X} = \mathbf{x} \right\} = \mu_i(\mathbf{x}) + \sigma_i T(\Delta_i(\mathbf{x})/\sigma_i) \quad (14)$$

$$= \mu_i(\mathbf{x}) + E(\max[N_i, \max_{j \neq i} [\mu_j(\mathbf{x})]]),$$

$$\text{where } \sigma_i = \left( \frac{q}{\alpha_i(\alpha_i + q)} \right)^{\frac{1}{2}}, \text{ and } \Delta_i(\mathbf{x}) = \max_{j \neq i} [\mu_j(\mathbf{x}) - \mu_i(\mathbf{x})]$$

**Proof:** Under the assumptions above, the left hand side of (11), after some standard calculations, can be seen to be equal to  $\mu_i(\mathbf{x}) + \sigma_i E(\max[N_i, \Delta_i(\mathbf{x})/\sigma_i])$ . The rest follows from the identity  $T(w) \equiv E(\max[N_1, w])$ .  $\square$

In the sequel, let  $\mu_{[1]}(\mathbf{x}) < \mu_{[2]}(\mathbf{x}) < \dots < \mu_{[k]}(\mathbf{x})$  denote the ordered posterior means from Stage 1. Let  $\mathcal{P}_{(i)}$  be the population which is associated with  $\mu_{[i]}(\mathbf{x})$ , and let  $\alpha_{(i)}$ ,  $\sigma_{(i)}$ , and  $\Delta_{(i)}(\mathbf{x})$  be defined analogously. Furthermore, let  $\mathcal{R}^{(i)}(\mathbf{x})$  denote the rule which assigns, at  $\mathbf{X} = \mathbf{x}$ , the next allocation to population  $\mathcal{P}_{(i)}$ ,  $i = 1, \dots, k$ .

First, let us consider the two populations  $\mathcal{P}_{(k-1)}$  and  $\mathcal{P}_{(k)}$ , since they turn out to play a very special role in this situation: these are the only two populations between which a preference in terms of the order relation " $\prec$ " can be established which does not depend on  $\mu_1(\mathbf{x}), \dots, \mu_k(\mathbf{x})$ . In fact the following theorem shows that the next allocation is not assigned to that one of the two population for which more prior plus sampling information has been gathered so far.

**Theorem 1.** *At every  $\mathbf{X} = \mathbf{x}$ , the following holds.*

$$\alpha_{(k-1)} > (=, <) \alpha_{(k)} \quad \text{if and only if} \quad \mathcal{R}^{(k-1)}(\mathbf{x}) < (=, >) \mathcal{R}^{(k)}(\mathbf{x}).$$

**Proof:** Consider the look ahead expected posterior gain, given  $\mathbf{X} = \mathbf{x}$ , of rule  $\mathcal{R}^{(k)}(\mathbf{x})$ , which is based on one more observation from population  $\mathcal{P}_{(k)}$ . Let it be denoted by  $g_k(\mathbf{x})$ . Starting with its representation given by (14), we can see that it satisfies

$$g_k(\mathbf{x}) = \mu_{[k]}(\mathbf{x}) + \sigma_{(k)}T(\Delta_{(k)}(\mathbf{x})/\sigma_{(k)}) = \mu_{[k-1]}(\mathbf{x}) + \sigma_{(k)}T(\Delta_{(k-1)}(\mathbf{x})/\sigma_{(k)}), \quad (15)$$

where the second equation follows from  $\Delta_{(k)}(\mathbf{x}) = \mu_{[k-1]}(\mathbf{x}) - \mu_{[k]}(\mathbf{x}) = -\Delta_{(k-1)}(\mathbf{x})$ , and from the identity  $T(w) \equiv T(-w) + w$ . On the other hand, the look ahead expected posterior gain  $g_{k-1}(\mathbf{x})$  of rule  $\mathcal{R}^{(k-1)}(\mathbf{x})$  is seen to be

$$g_{k-1}(\mathbf{x}) = \mu_{[k-1]}(\mathbf{x}) + \sigma_{(k-1)}T(\Delta_{(k-1)}(\mathbf{x})/\sigma_{(k-1)}). \quad (16)$$

The rest follows from the fact that  $\gamma T(\Delta/\gamma)$  is strictly increasing in  $\gamma$  for every  $\Delta \in \mathbb{R}$ .  $\square$

Let  $\mathcal{R}^{(*)}(\mathbf{x})$  be the better of the two rules  $\mathcal{R}^{(k-1)}(\mathbf{x})$  and  $\mathcal{R}^{(k)}(\mathbf{x})$  at  $\mathbf{X} = \mathbf{x}$ , as determined by Theorem 1, and let  $\alpha_{(*)} = \min\{\alpha_{(k-1)}, \alpha_{(k)}\}$  and  $\sigma_{(*)} = \max\{\sigma_{(k-1)}, \sigma_{(k)}\}$  be the parameters associate with that population  $\mathcal{P}_{(*)}$ , say, which is chosen by  $\mathcal{R}^{(*)}(\mathbf{x})$ . This rule has the look ahead expected posterior gain

$$g_*(\mathbf{x}) = \mu_{[k-1]}(\mathbf{x}) + \sigma_{(*)}T((\mu_{[k]}(\mathbf{x}) - \mu_{[k-1]}(\mathbf{x}))/\sigma_{(*)}), \quad (17)$$

which has to be compared now with the maximum look ahead posterior expected gain rules  $\mathcal{R}^{(1)}(\mathbf{x}), \dots, \mathcal{R}^{(k-2)}(\mathbf{x})$ , i.e. with the maximum of

$$g_t(\mathbf{x}) = \mu_{[t]}(\mathbf{x}) + \sigma_{(t)}T((\mu_{[k]}(\mathbf{x}) - \mu_{[t]}(\mathbf{x}))/\sigma_{(t)}), \quad t = 1, \dots, k-2. \quad (18)$$

Comparing any two rules  $\mathcal{R}^{(i)}(\mathbf{x})$  and  $\mathcal{R}^{(j)}(\mathbf{x})$  with  $1 \leq i < j \leq k-2$  will necessarily involve  $\mu_{[i]}(\mathbf{x})$ ,  $\mu_{[j]}(\mathbf{x})$ , and  $\mu_{[k]}(\mathbf{x})$ . This situation is summarized in the next theorem. For brevity, the argument  $\mathbf{x}$  will be suppressed in Theorem 2, in its proof, and in the subsequent Corollary.



**Theorem 2.** For  $1 \leq i < j \leq k-2$  the following holds at every  $\mathbf{X} = \mathbf{x}$ .

- (i) If  $\alpha_{(i)} \geq \alpha_{(j)}$ , i.e. if  $\sigma_{(i)} \leq \sigma_{(j)}$ , then  $\mathcal{R}^{(i)} \prec \mathcal{R}^{(j)}$ .
- (ii) If  $\alpha_{(i)} < \alpha_{(j)}$ , and  $\sigma_{(i)} < (=, >) \sigma_{ij}$  then  $\mathcal{R}^{(i)} \prec (=, >) \mathcal{R}^{(j)}$ ,

where  $\sigma_{ij}$  is determined by

$$\mu_{[j]} + \sigma_{(j)}T((\mu_{[k]} - \mu_{[j]})/\sigma_{(j)}) = \mu_{[i]} + \sigma_{ij}T((\mu_{[k]} - \mu_{[i]})/\sigma_{ij}). \quad (19)$$

Moreover, the threshold  $\sigma_{ij}$  for  $\sigma_{(i)}$  satisfies

$$\sigma_{(j)} < \sigma_{ij} < \frac{\mu_{[k]} - \mu_{[i]}}{\mu_{[k]} - \mu_{[j]}} \sigma_{(j)}. \quad (20)$$

**Proof:** Let  $\mathbf{X} = \mathbf{x}$  be observed, and let  $i$  and  $j$  satisfy the assumptions of the theorem. Note, that in particular we have  $\mu_{[i]} < \mu_{[j]} < \mu_{[k-1]} < \mu_{[k]}$ .

(i) Suppose that  $\alpha_{(i)} \geq \alpha_{(j)}$ . Then the following can be shown to hold true:

$$\begin{aligned} g_i(\mathbf{x}) &= \mu_{[i]} + \sigma_{(i)}T(([\mu_{[k]} - \mu_{[j]}] + [\mu_{[j]} - \mu_{[i]}])/\sigma_{(i)}) \\ &< \mu_{[i]} + (\mu_{[j]} - \mu_{[i]}) + \sigma_{(i)}T((\mu_{[k]} - \mu_{[j]})/\sigma_{(i)}) \\ &= \mu_{[j]} + \sigma_{(i)}T((\mu_{[k]} - \mu_{[j]})/\sigma_{(i)}) \\ &\leq \mu_{[j]} + \sigma_{(j)}T((\mu_{[k]} - \mu_{[j]})/\sigma_{(j)}) = g_j(\mathbf{x}). \end{aligned}$$

The first inequality follows from the fact that

$$T(u+v) - T(u) = \int_u^{u+v} \Phi(z) dz < v, \text{ for } v > 0 \text{ and } u \in \mathbf{R}.$$

The second inequality follows from  $\sigma_{(i)} \leq \sigma_{(j)}$  and from the fact that  $\gamma T(\Delta/\gamma)$  is increasing in  $\gamma$ . This completes the proof of (i).

(ii) Suppose now that  $\alpha_{(i)} < \alpha_{(j)}$ , and consider the following difference as a function of  $\sigma_{(i)}$ :

$$g_j(\mathbf{x}) - g_i(\mathbf{x}) = [\mu_{[j]} + \sigma_{(j)}T((\mu_{[k]} - \mu_{[j]})/\sigma_{(j)})] - [\mu_{[i]} + \sigma_{(i)}T((\mu_{[k]} - \mu_{[i]})/\sigma_{(i)})], \quad (21)$$

which is positive (negative) if  $\sigma_{(i)}$  tends to  $\sigma_{(j)}$  (infinity), strictly decreasing in  $\sigma_{(i)}$ , and equal to zero at  $\sigma_{ij}$ , say, which is determined by (19).

Finally, we have to show that the threshold  $\sigma_{ij}$  for  $\sigma_{(i)}$  is located in the range given at the end of the theorem. As mentioned above, if  $\sigma_{(i)}$  tends to  $\sigma_{(j)}$  then (21) is positive, and thus  $\sigma_{ij} > \sigma_{(j)}$ . On the other hand, if  $\sigma_{(i)} = \sigma_{(j)}(\mu_{[k]} - \mu_{[i]})/(\mu_{[k]} - \mu_{[j]})$  then (21) reduces to

$$g_j(\mathbf{x}) - g_i(\mathbf{x}) = \mu_{[j]} - \mu_{[i]} - \frac{\mu_{[j]} - \mu_{[i]}}{\mu_{[k]} - \mu_{[j]}} \sigma_{(j)} T((\mu_{[k]} - \mu_{[j]})/\sigma_{(j)}) ,$$

which is negative, since  $T(w) > w$  for  $w \in \mathbb{R}$ . This completes the proof of the theorem.  $\square$

Let  $\mathcal{R}^{(\bullet)}(\mathbf{x})$  be the best of the rules  $\mathcal{R}^{(t)}(\mathbf{x})$  for  $t = 1, \dots, k-2$  at  $\mathbf{X} = \mathbf{x}$ , as determined by Theorem 2, which has to be compared now with  $\mathcal{R}^{(*)}(\mathbf{x})$ . Let  $\alpha_{(\bullet)}$ ,  $\sigma_{(\bullet)}$ , and  $\mu_{[\bullet]}(\mathbf{x})$  be the parameters associated with that population chosen by  $\mathcal{R}^{(\bullet)}(\mathbf{x})$ . This rule has the look ahead expected posterior gain

$$g_{\bullet}(\mathbf{x}) = \mu_{[\bullet]}(\mathbf{x}) + \sigma_{(\bullet)} T((\mu_{[k]}(\mathbf{x}) - \mu_{[\bullet]}(\mathbf{x}))/\sigma_{(\bullet)}) . \quad (22)$$

This has to be compared now with (17). The result is summarized in the following.

**Corollary.** *The comparison of the look ahead expected posterior gains of procedures  $\mathcal{R}^{(\bullet)}$  and  $\mathcal{R}^{(*)}$  is the same as (i) and (ii) in Theorem 2, with (i), (j),  $\sigma_{ij}$ ,  $\mu_{[i]}$ , and  $\mu_{[j]}$  replaced by  $(\bullet)$ ,  $(*)$ ,  $\sigma_{\bullet*}$ ,  $\mu_{[\bullet]}$ , and  $\mu_{[k-1]}$ , respectively.*

**Remark 2.** To apply allocation rule  $\mathcal{R}_1$  at  $\mathbf{X} = \mathbf{x}$ , a quick and simple algorithm is to just find the maximum of (15), (16), and (18), i.e. the maximum of the values in (14) for  $i = 1, \dots, k$ , and then to take the next observation from that population which has yielded the maximum. The results presented in the two theorems and in the Corollary are not needed for this purpose.

**Remark 3.** The importance of the two theorems and the Corollary is the insight which they provide into how and why the  $k$  populations are ordered with respect to " $\prec$ ". Interpreting  $\alpha_i = \nu_i + n_i/\sigma^2$  as the total of prior plus first stage sampling information gathered from population  $\mathcal{P}_i$ ,  $i = 1, \dots, k$ , we can see that the allocation rule  $\mathcal{R}_1$  favors populations  $\mathcal{P}_i$  which have, at Stage 1, larger yields  $\mu_i(\mathbf{x})$  and smaller sampling information  $\alpha_i$ ,  $i = 1, \dots, k$ . One exception, as revealed by Theorem 1, is that for the two populations

$\mathcal{P}_{(k-1)}$  and  $\mathcal{P}_{(k)}$  with the largest yields  $\mu_{[k-1]}(\mathbf{x})$  and  $\mu_{[k]}(\mathbf{x})$ , that one with the smaller of the two values  $\alpha_{(k-1)}$  and  $\alpha_{(k)}$  is preferred. Besides this exception, however, both goals cannot always be met simultaneously: there is a tradeoff between the two goals which is provided by Theorem 2 and by the Corollary.

**Example.** To illustrate the use of allocation rule  $\mathcal{R}_1$ , consider the data given in Table 7.3.3. on page 282 of Bickel and Doksum (1977). Blood cholesterol levels of men in three different socioeconomic groups  $\mathcal{P}_1$  (high end),  $\mathcal{P}_2$  (middle group), and  $\mathcal{P}_3$  (low end) are listed there with respective sample sizes  $n_1 = 5, n_2 = 10, n_3 = 6$ , and sample means  $x_1 = 315.6, x_2 = 320.1, x_3 = 302.3333$ . It may be worth noting that the p-value of the one-way layout analysis of variance F-test is 0.8823, and that none of the standard multiple comparison procedures (Duncan, LSD, Tukey-HSD, Student-Newman-Keuls, Tukey-B, Modified LSD, Sheffe') detect a difference between any two of the groups at significance level  $\alpha = 0.05$ .

Assume that apriori,  $\Theta_1, \Theta_2, \Theta_3$  are i.i.d.  $N(\mu, \nu^{-1})$ , and that the data given in the Table on page 309 of Bickel and Doksum (1977), which have been drawn from the same Los Angeles Heart Study, can be utilized to derive estimates for  $\sigma^2, \mu$ , and  $\nu$ . Hereby, no claims are made that this is in fact appropriate for that particular study, it is only done to demonstrate how to use allocation rule  $\mathcal{R}_1$ . The estimates turn out to be  $\sigma^2 = 3386.64, \mu = 265.97$  and  $\nu^{-1} = 1508.39$ . With these parameter values, one can calculate  $\alpha_1 = .0021, \alpha_2 = .0036, \alpha_3 = .0024$ , and  $\mu_1(\mathbf{x}) = 300.2203, \mu_2(\mathbf{x}) = 310.1751, \mu_3(\mathbf{x}) = 292.4314$ . Finally, formula (14) yields the look ahead expected posterior gains 310.5020, 310.1988, and 310.1831 for populations  $\mathcal{P}_1, \mathcal{P}_2$ , and  $\mathcal{P}_3$ , respectively. Therefore, the next observation  $Y_*$ , say, has to be taken from population  $\mathcal{P}_1$ .

This analysis can be carried one step further by considering  $Y_*$  as part of the sample drawn from  $\mathcal{P}_1$  at Stage 1, i.e. by updating  $p_1$  and  $x_1$  accordingly. Thereby, the following can be seen to hold: If  $Y_* < (>, =) 235.9998$  then the next observation  $Y_{**}$ , say, has to be taken from population  $\mathcal{P}_3$  ( $\mathcal{P}_1, \mathcal{P}_1$  or  $\mathcal{P}_3$ ). That no allocation for these first two future observations are made in favor of population  $\mathcal{P}_2$ , although it has the largest yield  $\mu_2(\mathbf{x}) = \mu_{[3]}(\mathbf{x})$ , is due to its relatively large prior plus sampl

**Remark 4.** The case of an improper prior distribution can be obtained by letting the prior variances of  $\Theta_1, \dots, \Theta_k$  tend to infinity, i.e. by letting  $\nu_1, \dots, \nu_k$  tend to zero. Hereby,  $\mu_i(\mathbf{x})$  tends to  $x_i$  and  $\alpha_i$  tends to  $n_i/\sigma^2$ ,  $i = 1, \dots, k$ , and the results of the two theorems and the Corollary can then be interpreted in terms of the sample means  $x_1, \dots, x_k$  and the sample sizes  $n_1, \dots, n_k$  of the  $k$  populations at Stage 1.

To conclude this study, it is natural to consider also the case of an experiment which is planned for a total of  $n$  observations, where the allocation procedure  $(\mathcal{R}_1, \mathcal{R}_1, \dots, \mathcal{R}_1)$ , listed last in (7), has been chosen to be used. The allocation of the very first observation is of the type of a Stage 1 allocation. As mentioned below of (4), the results derived in this section are more general and apply also to this case. All that has to be done is to set  $n_1 = n_2 = \dots = n_k = 0$  and then relabel  $m_i$  by  $n_i$ , for  $i = 1, \dots, k$ , and  $m$  by  $n$ , which is now equal to one in this case. The relevant parameters reduce then to  $\alpha_i = \nu_i$  and  $\mu_i(\mathbf{x}) = \mu_i$ ,  $i = 1, \dots, k$ , and the results for allocation rule  $\mathcal{R}_1$  derived above hold analogously.

The very first observation is allocated by  $\mathcal{R}_1$  to that population which is associated with the largest of the values  $E(\max[\mu_i + \sigma_i N_1, \max_{j \neq i} \{\mu_j\}])$ ,  $i = 1, \dots, k$ . Especially, in the case of i.i.d. priors, i.e. where  $\Theta_i \sim N(\mu, \nu^{-1})$ ,  $i = 1, \dots, k$ , are independent,  $\mathcal{R}_1$  allows to start with any of the  $k$  populations. However, the second and all consecutive allocations made by  $\mathcal{R}_1$  are more involved, since updating of the prior has to be performed before the criterion based on (11), using (14), can be utilized. In other words, allocation of the first observation is a Stage 1 type optimization, whereas allocation of the  $s$ -th observation for  $s = 2, \dots, k$  is a Stage 2 type optimization, where the first  $s - 1$  observations are considered as data collected at a Stage 1.

Suppose that an allocation of these  $n$  observations can be made in such a way that  $\nu_1 + n_1/\sigma^2 = \dots = \nu_k + n_k/\sigma^2$  holds, i.e. that the posterior, after these  $n$  observations have been drawn, is DT. Then allocating  $n$  times iteratively one observation to that population with the smallest  $\alpha$ -value, without updating the prior, would lead to this configuration. Iterating rule  $\mathcal{R}_1$   $n$  times, i.e. using the allocation procedure  $(\mathcal{R}_1, \mathcal{R}_1, \dots, \mathcal{R}_1)$  considered above, would of course lead to other configurations, since each newly allocated observation

would provide additional information which is utilized. However, one might speculate whether the fixed sample size  $n$  allocation rule  $\mathcal{R}_n$  would use here the DT-configuration. This question can be settled with the method of Lagrangian multipliers. For  $k = 2$  populations it turns out to be true: a fact which has been shown already in Gupta and Miescke (1994) with a different method. For  $k = 3$ , on the other hand, it can be seen with that method of Lagrangian multipliers that the DT-configuration is not a solution for  $\mathcal{R}_n$  to the problem at hand, except for the case of i.i.d. normal priors. Dealing with  $k \geq 3$  populations turns out to be structurally different and much more challenging than treating only  $k = 2$  populations, a typical contrast well known for ranking and selection problems.

The case of  $k$  normal populations with known but different variances can be treated analogously. However, the results will be more involved, since then  $\alpha_{(i)} \geq \alpha_{(j)}$  is no longer equivalent to  $\sigma_{(i)} \leq \sigma_{(j)}$  for  $i \neq j$ , which has been crucial in the proofs of the two theorems. The case of a common unknown variance, which requires a proper extension of the prior to account for a random variance, and similarly the case of  $k$  unknown variances, will be subject to future research.

**Acknowledgment:** The authors are grateful to two Referees whose comments and suggestions have been very helpful.

### References

- ABUGHALOUS, M.M. and MIESCKE, K. J. (1989): On selecting the largest success probability under unequal sample sizes, *Journal of Statistical Planning and Inference*, **21**, 53–68.
- BECHHOFER, R.E. (1969): Optimal allocation of observations when comparing several treatments with a control, in: *Multivariate Analysis*, P.R.KRISHNAIAH (ed.), Academic Press, New York, 465-473.
- BECHHOFER, R.E., KIEFER, J. and SOBEL, M. (1968): *Sequential Identification and Ranking Procedures*, The University of Chicago Press, Chicago.

- BECHHOFER, R.E., HAYTER, A.J. and TAMHANE, A.C. (1991): Designing experiments for selecting the largest normal mean when the variances are known and unequal: Optimal sample size allocation, *Journal of Statistical Planning and Inference*, **28**, 271–289.
- BERGER, J.O. (1985): *Statistical Decision Theory and Bayesian Analysis*, (2nd ed.), Springer, New York.
- BERGER, J.O. and DEELY, J. (1988): A Bayesian approach to ranking and selection of related means with alternatives to AOV methodology, *Journal of the American Statistical Association*, **83**, 364–373.
- BICKEL, P.J. and DOKSUM, K.A. (1977): *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden–Day, San Francisco.
- DUDEWICZ, E.J. and DALAL, S.R. (1975): Allocation of observations in ranking and selection with unequal variances, *Sankhyā*, **B-37**, 28–78.
- DUNNETT, C.W. (1960): On selecting the largest of  $k$  normal population means, *Journal of the Royal Statistical Society*, **B-22**, 1–40.
- FONG, D.K.H. and BERGER, J.O. (1993): Ranking, estimation and hypothesis testing in unbalanced two-way additive models - A Bayesian approach, *Statistics and Decisions*, **11**, 1–24.
- GOVINDARAJULU, Z. and KATEHAKIS, M.N. (1991): Dynamic allocation in survey sampling, *American Journal of Mathematical and Management Sciences*, **11**, 199–221.
- GUPTA, S.S. and MIESCKE, K.J. (1994): Bayesian look ahead one stage sampling allocations for selecting the largest normal mean, *Statistical Papers*, **35**, 169–177.
- GUPTA, S.S. and MIESCKE, K.J. (1993): On combining selection and estimation in the search of the largest binomial parameter, *Journal of Statistical Planning and Inference*, **36**, 129–140.
- GUPTA, S.S. and MIESCKE, K.J. (1988): On the problem of finding the largest normal mean under heteroscedasticity, in: *Statistical Decision Theory and Related Topics*, S.S. GUPTA and J.O. BERGER (eds.), **Vol. 2**, Springer, New York, 37–49.
- GUPTA, S.S. and MIESCKE, K.J. (1984): Sequential selection procedures - a decision theoretic approach, *Annals of Statistics*, **12**, 336–350.

GUPTA, S.S. and PANCHAPAKESAN, S. (1991): Sequential ranking and selection procedures, in: *Handbook of Sequential Analysis*, B.K. GHOSH and P.K.SEN (eds.), Marcel Dekker, New York, 363–379.

GUPTA, S.S. and PANCHAPAKESAN, S. (1979): *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*, John Wiley, New York.

HOLLANDER, M., PROSCHAN, F., and SETHURAMAN, J. (1977): Functions decreasing in transposition and their applications in ranking problems. *Annals of Statistics*, 5, 722–733.

MIESCCKE, K.J. (1984): Recent results on multi-stage selection procedures, in: *Proceedings of the Seventh Conference on Probability Theory*, M. IOSIFESCU (ed.), Editura Academiei Republicii Socialiste Romania, Bucharest, 259–268.

MIESCCKE, K.J. (1979): Bayesian subset selection for additive and linear loss functions, *Communications in Statistics*, A-8, 1205–1226.

WHITEHEAD, J. (1991): Sequential methods in clinical trial, in: *Handbook of Sequential Analysis*, B.K. GHOSH and P.K. SEN (eds.), M. Dekker, New York, 593–611.