Submitted to *Journal American Statistical Association*

Sequentially Deciding Between Two Experiments for Estimating a Common Success Probability

Janis Hardwick[1]
University of Michigan

Connie Page
Michigan State University

Quentin F. Stout[1]
University of Michigan

Abstract: To estimate a success probability p, two experiments are available: individual Bernoulli(p) trials or the product of r individual Bernoulli(p) trials. This problem has its roots in reliability where either single components can be tested or a system of r identical components can be tested. A total of N experiments can be performed, and the problem is to sequentially select some combination (allocation) of these two experiments, along with an estimator of p, to achieve low mean square error of the final estimate. This scenario is similar to that of the better-known group testing problem, but here the goal is to estimate failure rates rather than identify defective units. The problem also arises in epidemiological applications such as estimating disease prevalence.

Information maximization considerations, and analysis of the asymptotic mean square error of several estimators, leads to the following adaptive procedure: use the maximum likelihood estimator to estimate p, and if this estimator is below (above) the cut-point a_r, then observe an individual (product) trial at the next stage. An exact analysis of this adaptive procedure for fixed sample sizes shows that it behaves roughly as the asymptotics predict, and that several other estimators and procedures behave far worse than their asymptotics indicate. Further, the adaptive procedure exhibits negative regret over a portion of the parameter range.

Keywords: batch testing, risk assessment, infection rate, grouped data, omniscient allocation

 Last modified: 18 Sep 1995.

[1]Research supported in part by National Science Foundation under grants DMS-9157715 and DMS-9504980.

1 Introduction

Consider a sequence of Bernoulli trials with success probability p. To estimate p, two experiments can be performed: either an individual trial outcome can be observed (the *p-experiment*) or the product of r individual trial outcomes can be observed (the *p^r-experiment*), where r is an integer ≥ 2. A total of N experiments (tests) can be performed, and the problem is to select some combination (allocation) of these two experiments along with an estimator of p to achieve low mean square error of the terminal estimator. The experiments can be selected sequentially, so that at each stage, information available at that stage can be used to determine which experiment to carry out at the next stage.

Before continuing, some comments on the origin and application of this problem are in order. The p^r-experiment is a slightly disguised version of the well-studied grouped data experiment with groups of size r. In the grouped data setting, the goal is to estimate the failure probability, $q = 1-p$, and using groups of sizes other than 1 can reduce the cost of testing (Sobel and Elashoff, 1975), and can lower the variance of the resulting estimator (Chen and Swallow, 1990). There are also privacy reasons for using group sizes greater than 1 (Gastwirth and Hammick, 1989). Finally, the group testing problem can arise in a reliability setting in which either individual components can be tested or a system of r identical components can be tested in series (Easterling and Prairie, 1971).

The problem considered here assumes that there is a "natural" r to use for testing, and the choices are then restricted to only two experiments, the p-experiment and the p^r-experiment. An example of a scenario in which such an assumption is realistic is a reliability setting in which a system requires r units in series to function. The choice of an optimal r was recently addressed in the context of two-stage sampling procedures by Oliver-Hughes and Swallow (1994).

In Section 2, notation is given and the problem is precisely defined. In Section 3, allocation that maximizes information is derived, and in Section 4, pooling data across experiments is discussed and several estimators are analyzed. In Section 5, allocation that maximizes asymptotic mean square error is derived for each of the estimators. In Section 6, an ad-hoc adaptive allocation procedure is proposed, and it is shown that not all of the estimators are consistent when combined with an arbitrary allocation procedure. In Section 7, an exact approach is taken to evaluating and optimizing allocation policies for fixed values of N. Here it is shown that the ad-hoc adaptive rule of Section 6 has negative regret with respect to the optimal best fixed sample size rule that can be generated when the parameter is known. It is also shown that, for moderate sample sizes, several of the estimators perform far worse than their asymptotic behavior would predict.

2 Notation and Problem Statement

The problem will be set up in its fully sequential form although much of the development in the next sections will not use all of this notation. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed Bernoulli(p) random variables, independent of $Y_1, Y_2, \ldots$, independent and identically distributed Bernoulli(p^r) random variables. Success probability p is restricted to (0,1) throughout the paper.

A total of N tests (experiments) will be done, where at each stage the decision to observe an X or a Y can be made based on past observations. More precisely, an allocation rule is a sequence $\mathbf{d} = (d_1, \ldots, d_N)$ such that for $k = 1, \ldots, N$, d_k takes values 0 or 1, and is measurable $\{Z_1, \ldots, Z_k\}$, where $Z_i = d_i X_i + (1 - d_i) Y_i$. Thus, d_i indicates where the i^{th} observation or test is from, with "1" indicating an X-observation or the p-experiment, and "0" indicating a Y-observation or the p^r-experiment. A *terminal estimator of* p must be measurable $\{Z_1, \ldots, Z_N\}$. Finally, let $m_k = \sum_{i=1}^{k} d_i$ be the total number of observations taken from the p-experiments at stage k, and let $n_k = k - m_k$ be the total number of observations taken from the p^r-experiment at stage k, where $k = 1, 2, \ldots, N$. Sometimes the k subscript will be dropped.

Since the final goal is estimation of p, an allocation scheme and estimator will be evaluated as a pair by the mean square error of the terminal estimator. That is, the mean square error of using allocation $\mathbf{d}$ and estimator $\hat{p}$ is given by $\mathrm{MSE}_p(\mathbf{d}, \hat{p}) = E_p(\hat{p} - p)^2$. In the Bayesian framework, the problem of selecting both the allocation rule and the estimator reduces to selecting only the allocation rule and using the Bayes estimator. However, in the non-Bayesian framework, the choice of estimator is not so obvious, and as will be noted in Section 6, the allocation rule and the estimator can interact.

3 Maximum Information Allocation

In this section, the problem of selecting an estimator of p is ignored, and only allocation is considered. The criterion used for allocation will be maximizing the Fisher information, and the best nonrandom allocation will be found. As is typical with such optimal nonrandom allocations, the rule will depend on the unknown parameter p. However, this parameter-dependent rule will suggest the form of an adaptive rule, and the relationship between the MSE's of estimators and Fisher information will make the adaptive rule efficient.

In typical sequential allocation problems, the different experiments give information about different parameters. However, this problem is unusual in that both experiments give information about the same parameter, although one experiment will give more information depending on the actual value of the parameter.

In particular, the Fisher information about p contained in a single observation of the p-experiment is

$$I_X(p) = \frac{1}{pq}, \qquad \text{where} \quad q = 1 - p,$$

and the Fisher information about p contained in a single observation of the p^r-experiment is

$$I_Y(p) = \frac{r^2 p^{r-2}}{1 - p^r}.$$

It is easy to show that $I_X(p) > I_Y(p)$ if and only if $p < a_r$, where a_r is the unique root in (0,1) of the equation in p,

$$p^r(1 - r^2) + r^2 p^{r-1} - 1 = 0.$$

Note that a_r is a function only of r, which is known, and hence a_r can be determined explicitly.

Proposition 3.1 *If N tests are available, then the allocation that maximizes the information about p is*

$$m_N = \begin{cases} N \ \textit{(observe all X's)} & \textit{if } p < a_r \\ 0 \ \textit{(observe all Y's)} & \textit{if } p > a_r \\ \textit{arbitrary} & \textit{if } p = a_r \end{cases}$$

(The informations are equal at $p = a_r$).

r	2	5	10	20	50	100
a_r	0.333	0.536	0.679	0.792	0.892	0.937

Table 1: Cut-point Values

Proof. If m X's and n Y's are observed, then because of the independence among X's and among Y's and between X's and Y's, the information about p is $mI_X(p) + nI_Y(p)$. This is obviously maximized at either $m = N$ or $m = 0$ if $p < a_r$ or $> a_r$, respectively. □

Thus, the maximum information allocation is to observe only X's (the p-experiment) if $p < a_r$, or only Y's (the p^r-experiment) if $p > a_r$. This will be denoted as *a_r-cut allocation*, and recall that a_r depends only on r. However, the region where one experiment is better than the other depends on the unknown parameter p. Thus, the obvious adaptive rule is suggested where p is estimated at each stage, and the next observation is allocated depending the relationship between the estimated p and the cut a_r.

Some of the values of a_r (first reported in Loyer, 1983) are noted in Table 1. In general, as r increases to infinity, the cut-point a_r tends to 1, and the region over which the p^r-experiment is better shrinks.

This section is concluded by evaluating the need for sequential allocation. This is done by comparing the information in each experiment. Consider the ratio

$$\frac{\min\{I_X(p), I_Y(p)\}}{\max\{I_X(p), I_Y(p)\}}$$

over the range of p. Were this ratio bounded below by, say, .98, then using the nonoptimal experiment would never result in more than a 2% loss of information, greatly limiting the worth of adaptive allocation. Since adaptive allocation is somewhat more complicated than fixed allocation, there needs to be sufficient benefit to justify its utilization.

Proposition 3.2

$$\frac{I_X(p)}{\max\{I_X(p), I_Y(p)\}} = \begin{cases} 1 & \text{for } p \le a_r \\ (p^{1-r} + p^{2-r} + \ldots + 1)/r^2 \quad \textit{(which is } \ge 1/r\textit{)} & \text{for } p > a_r \end{cases}$$

$$\frac{I_Y(p)}{\max\{I_X(p), I_Y(p)\}} = \begin{cases} r^2 q p^{r-1}/(1-p^r) \quad \textit{(which tends to 0 as p tends to 0)} & \text{for } p \le a_r \\ 1 & \text{for } p > a_r \end{cases}$$

Proof. This is simply algebra. □

Proposition 3.2 indicates that if the p^r-experiment is used, then for p sufficiently small, the information obtained can be arbitrarily close to 0% of that possible when the p-experiment is used. On the other hand, if the p-experiment is used, then the information obtained never falls below (100/r)% of the maximal information obtainable, approaching this bound when p tends to 1. Thus, adaptive allocation can be worthwhile for increasing information. Also, using the p-experiment would be the more conservative approach since one never loses more than (100/r)% of the maximal information obtainable. Note that the bound decreases as r increases, so that at $r = 2$, no more than 50% of optimal can be lost, but at $r = 10$, one might get only 10% of the optimal.

4 Estimators of p

Since both experiments give information about p, and an adaptive allocation procedure would typically allocate to both experiments, there is the question of how to combine or pool data across experiments. Several estimators are presented here and their properties are derived.

Throughout this section, m is the number of observations from the p-experiment, n is the number of observations from the p^r-experiment, and $m + n = N$ is the fixed total number of experiments.

If m observations from only the p-experiment are used to estimate p, then the best estimator (uniform minimum variance unbiased and the maximum likelihood) is the sample mean $\bar{X}$, and its mean square error is equal to its variance, pq/m. Let $\hat{p}_x$ denote this estimator.

If n observations from only the p^r-experiment are used to estimate p, then there is no unbiased estimator. However, the maximum likelihood estimator is $\bar{Y}^{1/r}$, the rth root of the sample mean, which is equivalent to the usual estimator of p in grouped data experiments using groups of size r. Let $\hat{p}_y$ denote this estimator. The MSE of $\hat{p}_y$ can be computed for different values of p and n using binomial probabilities, and as n tends to infinity, $n \cdot \mathrm{MSE}_p(\hat{p}_y)$ tends to $1/I_Y(p) = (1-p^r)/r^2p^{r-2}$. This has been noted and studied by Sobel and Elashoff (1975). Some comparisons of the exact mean square and this asymptotic form are made in Loyer (1983).

Next, the maximum likelihood estimator is derived for the general situation with observations of both experiments.

Theorem 4.1 *Given m observations of the p-experiment and n observations of the p^r-experiment, the maximum likelihood estimator of p, denoted $\hat{p}_{ml}$, is the unique root in $[0,1]$ of the equation*

$$p^r(rn+m) + (m-x)(p^{r-1} + \ldots + p) - (x+ry) = 0,$$

where $x = \sum_{i=1}^m X_i$ and $y = \sum_{j=1}^n Y_j$.

Proof. The joint likelihood function is equal to $p^x(1-p)^{m-x}p^{ry}(1-p^r)^{n-y}$, and taking the logarithm and differentiating with respect to p gives the equation above. Let $H(p)$ be the polynomial in p in the equation, and note that if $m - x = 0$, then the solution of the equation $H(p) = 0$ is $p = \{(m+ry)/(m+rn)\}^{1/r}$. Thus, suppose that $m - x > 0$. Then $H(1) = r(n-y) + r(m-x) > 0$. But $H(0) = -(x+ry)$, which is < 0 if $x + y > 0$ and $= 0$ if $x = y = 0$. Furthermore, $H'(p) = rp^{r-1}(rn+m) + (m-x)\{(r-1)p^{r-2} + \ldots + 2p + 1\}$, which is strictly greater than 0 for $p > 0$. This implies the existence and uniqueness of the root in [0,1]. □

For the case of $r = 2$, the maximum likelihood estimator can be given in closed form:

$$\hat{p}_{ml} = \frac{\sqrt{a^2(1-\hat{p}_x)^2 + 4a\hat{p}_x + 4(1-a)\hat{p}_y{}^2} \; - \; a(1-\hat{p}_x)}{2},$$

where $a = m/(2n+m)$. For $r > 2$, no such closed form is available, and it is for this reason that other types of estimators are sought. If there isn't much performance difference between the maximum likelihood estimator and some other estimator, then one might prefer an estimator that had a simple form to one that is merely the solution of an equation.

Other natural estimators of p are weighted averages of $\hat{p}_x$ and $\hat{p}_y$, where the weights could depend on m, n, and N. Two particular weight choices are considered below:

- The *constant weights estimator:* $\hat{p}_\alpha = (\alpha)\hat{p}_x + (1-\alpha)\hat{p}_y$.

- The *weight proportional to sample size estimator:* $\hat{p}_s = (m/N)\hat{p}_x + (n/N)\hat{p}_y$.

Finally we consider *ratio estimators.* For the case of $r = 2$, the obvious estimator is $\hat{p}_R = \bar{Y}/\bar{X}$. The best generalization of this estimator for larger values of r is not obvious, although a natural choice is the following:

$$\hat{p}_R = \left(\frac{\bar{Y}}{\bar{X}}\right)^{\frac{1}{r-1}}$$

There are some special cases that should be considered when implementing a ratio estimator. When the ratio $\bar{Y}/\bar{X}$ is larger than 1 (including the case $\bar{X} = 0 < \bar{Y}$), then the estimator is set to 1, and when both the numerator and denominator are zero, then the estimator is set to zero. The primary reason for considering ratio estimators is because they have been used previously, although their role was more for model testing than estimation (Chen and Swallow, 1990). Note that Chen and Swallow use a ratio estimator $\hat{p}_R^*$ which is slightly modified from [4]. This estimator, which handles the division by zero problems a little differently, is defined as

$$\hat{p}_R^* = \left[\frac{(y+.5)/(n+.5)}{(x+.5)/(m+.5)}\right]^{\frac{1}{r-1}} .$$

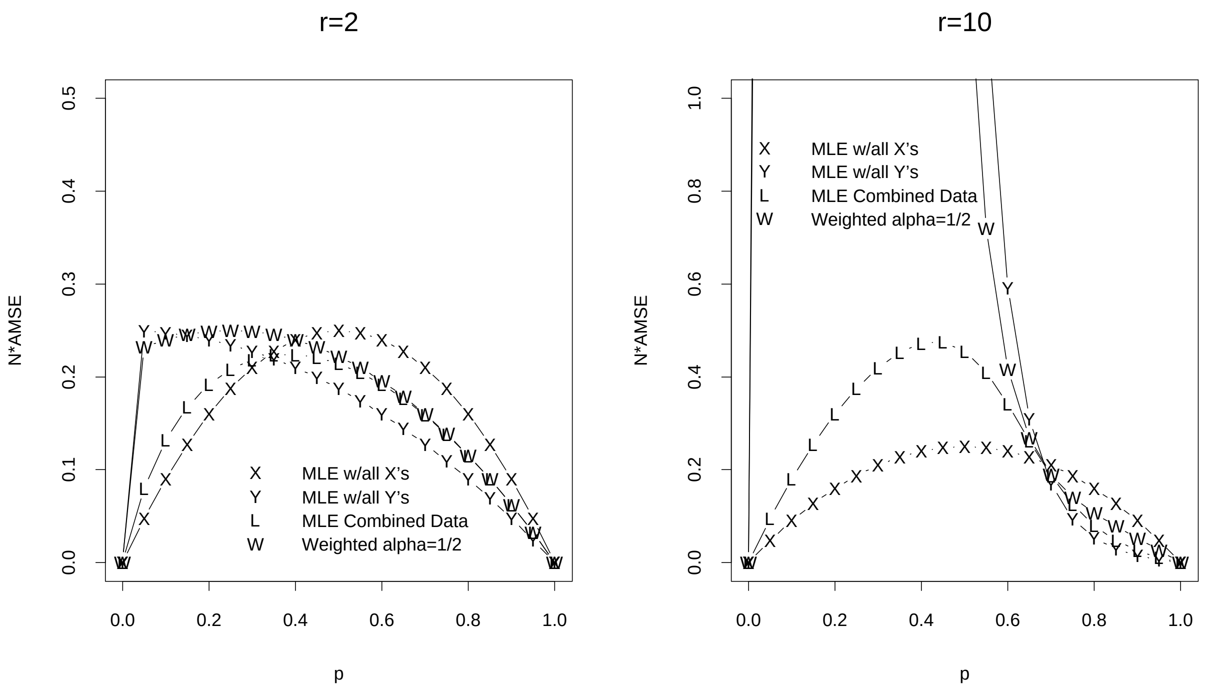
As will be shown in Section 7.3 ratio estimators perform quite poorly regardless of the adjustments described.

To determine the MSE's of the estimators, note that for the weighted average estimators, since $\hat{p}_x$ is unbiased for p, $\text{MSE}_p(\hat{p}_\alpha) = \alpha^2 pq/m + (1-\alpha)^2 \text{MSE}_p(\hat{p}_y)$. For the other estimators, their MSE's can be computed exactly for any values of m, n, and p, but there is no apparent closed form for them. However, their asymptotic expressions are very tractable. Define *asymptotic mean square error of an estimator* $\hat{p}$ as $\text{AMSE}_p(\hat{p}) = \text{MSE}_p(\hat{p}) + o(1/n) + o(1/m)$ for p in (0,1).

Theorem 4.2 *The asymptotic MSE's of the estimators are as follows:*

$$\begin{aligned}
\text{AMSE}_p(\hat{p}_x) &= \frac{pq}{m} = \text{MSE}_p(\hat{p}_x) \\
\text{AMSE}_p(\hat{p}_y) &= \frac{1-p^r}{nr^2p^{r-2}} \\
\text{AMSE}_p(\hat{p}_{ml}) &= \frac{pq(1-p^r)}{m(1-p^r)+nr^2p^{r-1}q} \\
\text{AMSE}_p(\hat{p}_\alpha) &= \frac{\alpha^2 pq}{m} + \frac{(1-\alpha)^2(1-p^r)}{nr^2p^{r-2}} \\
\text{AMSE}_p(\hat{p}_s) &= \frac{mpq + n(1-p^r)/r^2p^{r-2}}{N^2} \\
\text{AMSE}_p(\hat{p}_R) &= \frac{1}{(r-1)^2}\left[\frac{pq}{m} + \frac{1-p^r}{np^{r-2}}\right].
\end{aligned}$$

Proof. The MSEs for the $\hat{p}_x$ and $\hat{p}_y$ estimators had been noted earlier, and the AMSEs for these estimators imply the results for the weighted average estimators. For the maximum likelihood estimator, its AMSE is simply the reciprocal of the information, $mI_X(p)+nI_Y(p)$, and this problem



Allocation for L and W is 50/50 on p and p^r.

Figure 1: Normalized $N*$AMSEs of Estimators, $r = 2, 10$. Note change in vertical scales.

---

satisfies all the usual regularity conditions to insure that the asymptotic variance of the maximum likelihood estimator is equal to this reciprocal. Finally, for the ratio estimator, this is a first order expansion of the MSE, and details can be found in Noble (1992). □

To illustrate the relative performance of the estimators when standard fixed allocations have been used, normalized AMSE's are presented for several of the estimators in Figure 1. The values plotted in Figure 1 are $N*$AMSE for $\hat{p}_x$ $\hat{p}_y$ $\hat{p}_{ml}$ and $\hat{p}_\alpha$ for the cases where $r = 2$ and 10. Note that for the combined maximum likelihood and weighted estimators, the proportions allocated to each population are equal. These are merely examples of fixed allocations and others may be of interest. Note that the ratio estimator is omitted. It appears in Figure 5, because its AMSE is so large that plotting it diminishes the resolution among the AMSE's of the other estimators.

5 Allocation Minimizing Asymptotic Mean Square Error

While fixed allocations performance of different estimators may be important in some instances, the problem of interest here is the pairing of an estimator and an allocation rule to lower the MSE of the "terminal" estimator of p. In this section, the estimators proposed in Section 4 are considered. Since their exact MSE's are analytically intractable while their first order asymptotic forms are simple, allocation rules that minimize the asymptotic MSE among nonrandom allocations are considered for each estimator. As is typical with optimal nonrandom rules, the allocation will depend on the unknown p, but will suggest an adaptive rule.

For each estimator in Section 4, the following theorem gives a nonrandom allocation, as a function of p, that minimizes the AMSE.

Theorem 5.1 *1. The AMSE's of estimators $\hat{p}_{ml}$ and $\hat{p}_s$ are minimized by the a_r-cut allocation described in Section 3.*

2. $\mathrm{AMSE}_p(\hat{p}_\alpha)$ *is minimized by allocating m/n in proportion $\{r\alpha/(1-\alpha)\}\sqrt{p^{r-1}q/(1-p^r)}$.*

3. $\mathrm{AMSE}_p(\hat{p}_R)$ *is minimized by allocating m/n in proportion $\sqrt{p^{r-1}q/(1-p^r)}$.*

Proof. The result follows from minimizing the AMSE formulae in Theorem 4.2, as functions of m and n, under the constraint that $m + n = N$. Note that for both the maximum likelihood estimator and the sample size weighted average estimator, the minimizing allocation satisfies the same conditions as the allocation that maximized the information (Section 3). □

For each estimator considered, the allocations reported in Theorem 5.1 will give the lowest AMSE for that estimator. These are referred to as *asymptotic omniscient fixed allocations* corresponding to the estimators. To compare the estimators based on the omniscient fixed allocations, the minimum AMSE's are needed. Let "*" on the AMSE denote the minimum asymptotic mean square error when the allocations of Theorem 5.1 are used.

Corollary 5.2 *The normalized* AMSE_p^* *values are as follows:*

$$\begin{aligned} N * \mathrm{AMSE}_p^*(\hat{p}_{ml}) = \mathrm{AMSE}_p^*(\hat{p}_s) &= \min\{pq, (1-p^r)/r^2p^{r-2}\} \\ N * \mathrm{AMSE}_p^*(\hat{p}_\alpha) &= \left[\alpha\sqrt{pq} + (1-\alpha)\sqrt{(1-p^r)/(r^2p^{r-2})}\right]^2 \\ N * \mathrm{AMSE}_p^*(\hat{p}_R) &= \frac{\left[\sqrt{pq} + \sqrt{(1-p^r)/p^{r-2}}\right]^2}{(r-1)^2}. \end{aligned}$$

Proof. Simply substitute the allocations from Theorem 5.1 into the AMSE formulas of Theorem 4.2. □

Corollary 5.3 *For p in $(0,1)$, for any weight α in $(0,1)$,*

$$\mathrm{AMSE}_p^*(\hat{p}_{ml}) = \mathrm{AMSE}_p^*(\hat{p}_s) \leq \min\{\mathrm{AMSE}_p^*(\hat{p}_\alpha), \mathrm{AMSE}_p^*(\hat{p}_R)\},$$

with equality occurring only at $p = a_r$.

Proof. This follows easily from Corollary 5.2. □

Corollary 5.3 shows that the AMSE* for the maximum likelihood and weighted average estimators is the lowest among those considered. Thus, according to the asymptotic MSE criterion, the a_r-cut allocation, along with either the maximum likelihood estimator or the sample size weighted average estimator, should be used.

Plots of AMSE* for $\hat{p}_{ml}$, $\hat{p}_\alpha$ and $\hat{p}_R$ are shown in Figure 2 for $r = 2$ and 10. For $\hat{p}_{ml}$, the a_r-cut allocation uses all X observations or all Y observations depending on whether p is above or below a_r. The relative proportions sampled for the estimators $\hat{p}_\alpha$ and $\hat{p}_R$ are $N/2$ from each population. While it is difficult to discern from Figure 2, for $r = 10$ in the region above $p = 0.8$ the ratio estimator does better than the weighted average estimator with weight $\alpha = 1/2$.

Note that caution should be exercised when using the AMSE formulas. These are first order approximations, ignoring terms $o(1/m) + o(1/n)$, and thus are valid when both m and n tend to

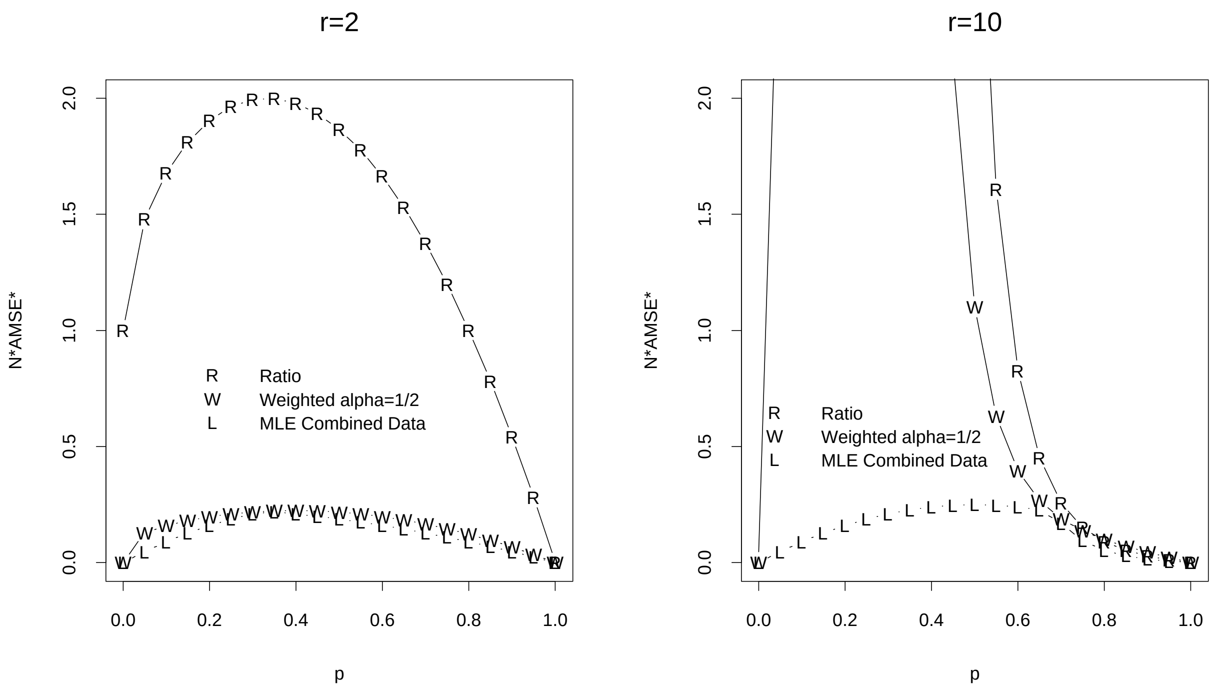


Figure 2: Normalized $N{*}\mathrm{AMSE}_p^*$ Values.

infinity. However, the point of allocation here is to eventually end up on the better experiment. As will be seen in Section 6, the AMSE for the maximum likelihood estimator and the sample size weighted average estimator remains valid if one sample size does not tend to infinity. However, the constant weighted average and the ratio estimators are not even consistent if one of the sample sizes does not tend to infinity.

6 Adaptive Allocation

As mentioned, the allocation rules considered thus far are motivated by the nonrandom "optimal" allocations of Sections 3 and 5. In Section 3, the maximum information rule was the a_r-cut rule. In Section 5, this same allocation rule came out of minimizing the asymptotic MSE. Thus, the a_r-cut allocation rule is a natural choice for an adaptive rule, where an estimator of p is inserted in the a_r-cut form, and that estimator is updated at each stage for sequential allocation. Other results in Section 5 indicated that the estimator of p and the allocation should be compatible in the sense that the a_r-cut allocation should minimize the asymptotic MSE of the estimator in use.

To define an adaptive a_r-cut allocation, consider an estimator of p. By *estimator*, a sequence of estimators $\{\hat{p}_k\}_{k=1}^{\infty}$ is implied, where $\hat{p}_k$ depends on observations up to and including the k^{th}. For $k = 1, \ldots, N$, $m_k + n_k = k$, where m_k and n_k are the number of observations on the p-experiment and on the p^r-experiment, respectively, at stage k. The estimators suggested in Section 4 fill this requirement since they are defined for each pair (m, n). Once an estimator is selected, the ad-hoc a_r-cut allocation is defined in the obvious way:

The a_r-cut allocation with estimated p: At stage 1, take an observation from the p-experiment. At stage k, $1 < k \leq N$, observe from the p-experiment if and only if $\hat{p}_{k-1} \leq a_r$.

The tie at a_r has been decided in favor of the p-experiment because the worst of the consequences of a wrong decision is less (see Proposition 3.2).

The a_r-cut allocation aims eventually to allocate to the better experiment depending on the value of p. This requires that the estimator of p used with the allocation be consistent even if the number of observations on one of the experiments does not tend to infinity. This consistency is called "allocation" consistency:

Estimator $\{\hat{p}_k\}$ is (strongly) allocation consistent if $\hat{p}_k$ tends to p a.s. as k tends to infinity, for all p in $(0, 1)$.

Theorem 6.1 *Estimators $\hat{p}_{ml}$ and $\hat{p}_s$ are allocation consistent; estimators $\hat{p}_\alpha$ and $\hat{p}_R$ are not.*

Proof. The lack of consistency of $\hat{p}_\alpha$ and $\hat{p}_R$ occurs when one of the sample sizes does not tend to infinity. That is, if, say, $1 \leq n < n_o < \infty$, then $\bar{Y}^{1/r}$ does not converge in probability or a.s., and the weight $1 - \alpha$ does not tend to zero. On the other hand, the $\hat{p}_s$ estimator has weights m/N and n/N and if either sample size is bounded as N tends to infinity, the weight tends to zero, implying consistency.

The consistency of $\hat{p}_{ml}$ requires more work. Let $x = \sum_{i=1}^m X_i$ and $y = \sum_{j=1}^n Y_j$. Then the logarithm of the joint likelihood function is equal to

$$h_{m,n}(p) = f_m(p) + g_n(p), \text{ where}$$

$$f_m(p) = x\log(p) + (m - x)\log(1 - p) \quad \text{and} \quad g_n(p) = y\log(p^r) + (n - y)\log(1 - p^r).$$

The derivative with respect to p is

$$h'_{m,n}(p) = f'_m(p) + g'_n(p), \text{ where}$$

$$f'_m(p) = \tfrac{x}{p} - \tfrac{m-x}{1-p} \quad \text{and} \quad g'_n(p) = r\tfrac{y}{p} - \tfrac{(n-y)p^{r-1}}{1-p^r}.$$

The maximum likelihood estimator is the unique root in [0,1] of $h'_{m,n}(p) = 0$, and because f'_m and g'_n are both decreasing in p, it follows that the root of $h'_{m,n}$ is between the roots of f'_m and g'_n. Thus, $\min\{\hat{p}_x, \hat{p}_y\} \leq \hat{p}_{ml} \leq \max\{\hat{p}_x, \hat{p}_y\}$. Therefore, if both m and n tend to infinity, then both $\hat{p}_x$ and $\hat{p}_y$ tend to p a.s., and thus $\hat{p}_{ml}$ tends to p a.s.

Now, suppose that $n \leq n_o < \infty$. Then $\frac{-rn_o}{1-p} \leq g'n(p) \leq \frac{rn_o}{1-p}$, which in turn bounds $h'_{m,n}$ by

$$A_-(p) = \frac{x}{p} - \frac{m - x + rn_o}{q} \;\leq\; h'_{m,n}(p) \;\leq\; \frac{x + rn_o}{p} - \frac{m - x}{q} = A_+(p).$$

But $A_-(p) = 0$ at $p_- = \frac{x}{m+rn_o}$ and $A_+(p) = 0$ at $p_+ = \frac{x+rn_o}{m+rn_o}$. Then, $p_- \leq \hat{p}_{ml} \leq p_+$. But p_- and p_+ both tend to p a.s. as m tends to infinity, and thus $\hat{p}_{ml}$ tends to p a.s. as m tends to infinity even though $n \leq n_o$. A similar argument holds for $m \leq m_o < \infty$. □

The estimator used with the a_r-cut should be allocation consistent. However, while use of a consistent estimator is prudent, it does not guarantee low MSE of the terminal estimator. That requires efficiency of an estimator, as indicated below.

Ideally, an adaptive allocation should select the better experiment quickly (high precision of the estimator), and once the better experiment is being used, the estimator should approximate the individual maximum likelihood estimator for that experiment. That is, the goal is to use an estimator $\hat{p}$ with the property that

$$\mathrm{MSE}_p(\hat{p}) \approx \min\{\mathrm{MSE}_p(\hat{p}_x), \mathrm{MSE}_p(\hat{p}_y)\}.$$

Thus, $\hat{p}$ should approximate $\hat{p}_x$ over the range where the p-experiment is better, and should approximate $\hat{p}_y$ over the range where the p_r-experiment is better. Of course, adaptive allocation requires that some of the N observations be used to identify the better experiment. Thus, the desired MSE would not be attained. However, the larger the N, the smaller the proportion expected on the inferior experiment, and the MSE would tend to be nearer to the desired MSE.

Asymptotically, (as N tends to infinity), $\mathrm{MSE}_p(\hat{p}_y)$ can be replaced by $(1-p^r)/Nr^2p^{r-2}$, and the desired limiting MSE can be given as

$$\mathrm{MSE}_p(\hat{p}) \approx \frac{\min\{pq, (1-p^r)/r^2p^{r-2}\}}{N}. \tag{1}$$

This is the lower envelope with which exact MSE will be compared in the next section (Figure 6).

7 Fixed Sample Size Behavior

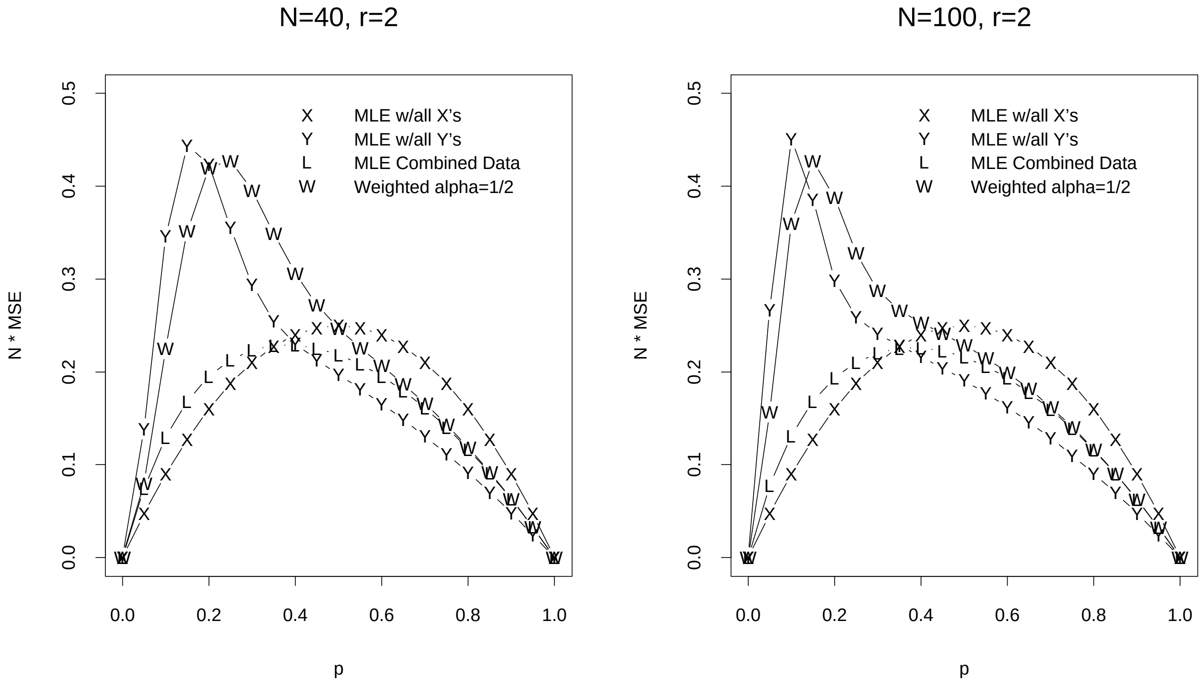
Until now, the development and evaluation of allocation rules and estimators has been based solely on asymptotic arguments. In this section, the estimators and allocation rules are examined for their behavior based on fixed sample sizes. For moderate sample sizes, some of the estimators do not behave as expected, and there are several adjustments that need to be made when implementing the estimators and allocation rules. Despite these problems, it will be shown that the a_r-cut allocation rule using the MLE estimator does very well, and in fact out-performs omniscient fixed allocation rules for some values of p.

7.1 Adjusting Cut-points

A note on the a_r value is appropriate here, since it was derived from asymptotic considerations. If the exact MSE for the estimator $\hat{p}_y$ is compared with the exact MSE for $\hat{p}_x$ using the same number of tests, the cut-point below which the p-experiment does better depends on N as well as r. Determining the cut-point is a straightforward computation, but it does not have a simple closed form. This cut-point differs from a_r, and N may need to be quite large (say ≥ 100) before it is approximately equal to its asymptotic value of a_r. For example, Loyer (1983) showed that for $r = 2$ the value of a_2 is $1/3$, but for $N = 20$, the p-experiment is better when $p < .445$. For $r = 5$, the value of a_5 is .536, but for $N = 20$, the p-experiment is better when $p < .729$. Thus, if the total sample size is moderate, then the a_r-cut should be modified to account for the lack of asymptotic fit. This adjustment is more pronounced the larger r is.

7.2 Exact Mean Squared Errors of Estimators Using Fixed Allocations

As noted earlier, the MSE's of most of the estimators under consideration don't have convenient analytic forms. However, given p and N, the MSE's can be calculated for each estimator. In this section, performance of the estimators based on fixed allocations are reviewed. Representative



Allocations for L and W are 50/50 on p and p^r.

Figure 3: Exact N∗MSE Values Using Fixed Allocations.

---

behavior for all but the ratio estimator is illustrated in Figure 3 for $N = 40$ and $N = 100$, for $r = 2$. (The ratio estimator, illustrated in Section 7.3, performs far worse than the estimators in Figure 3.) The labels and relative sample sizes used for the different estimators are the same as in Figure 1. The values in Figure 3 have been scaled by a factor of N, so they can be compared across sample sizes as well as between estimators. It seems clear from both figures that the best estimators for small and large values of p, respectively, are $\hat{p}_x$ and $\hat{p}_y$. As noted, as the total sample size increases, the point at which $\hat{p}_y$ begins to improve on $\hat{p}_x$ moves towards the value $a_2 = 1/3$.

Note that, even with predetermined allocations of $N/2$ observations from each experiment, the maximum likelihood estimator, $\hat{p}_{ml}$, does very well across the entire range of parameter values regardless of the total sample size. Also, while it may seem odd that the MSE of $\hat{p}_\alpha$ or $\hat{p}_s$ could be larger than the MSE's of either of the two components that comprise the estimators, the plots indicate that this is true. However, there is no contradiction here because the MSE's for $\hat{p}_x$ and $\hat{p}_y$ are each based on N observations while the $\hat{p}_x$ and $\hat{p}_y$ components of the averaged estimators are each based on fractions of N, and thus are not averages of the X and Y values shown in the plots.

Comparing Figure 3 to Figure 1, one sees a decrease in the MSE of $\hat{p}_y$ and the estimators that depend on it as N goes to infinity. Unfortunately, the MSE's of these estimators don't converge very rapidly to the AMSE's. We examined MSE's for $N = 200$ and they were not much closer to the AMSE's than are the MSE's for $N = 100$. On the positive side, the MSE of the maximum likelihood estimator rapidly converges to its AMSE. The MSE's and AMSE's for $\hat{p}_x$ are, of course, the same.

It is also worth noting that the performance of $\hat{p}_y$ for small p is many times worse than the performance of $\hat{p}_x$ for large p, as is predicted by AMSE considerations. However, there is some

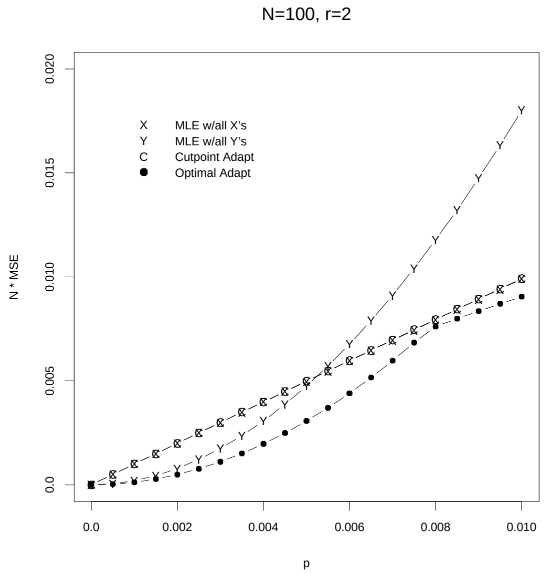


Figure 4: $N*$MSE Near Zero.

---

interesting behavior in neighborhoods that approach zero as $N \to \infty$ (a region which cannot be predicted by AMSEs). If p is sufficiently small relative to N ($<\approx 1/N$) it turns out that $\hat{p}_y$ is the preferred estimator. While this may seem surprising, it occurs because the MSE of $\hat{p}_x$ is approximately ϵ/N when $p = \epsilon$, while the MSE of $\hat{p}_y$ is approximately $\epsilon^2 + \epsilon^r N^{1-2/r}$ (this can be seen by considering the first few terms of the binomial expansion of the exact MSE). Thus, for ϵ sufficiently small relative to N, the MSE of $\hat{p}_y$ is less than the MSE of $\hat{p}_x$. This behavior can be seen in Figure 4 for $r = 2$. (The omniscient sequential allocation displayed there is discussed in Section 7.4.) However, this behavior is of little practical interest. Only very small MSEs are involved, and it is not reasonable to use only N tests to estimate values smaller than $1/N$.

7.3 Ratio Estimators

As mentioned, ratio estimators have not been included in most of the earlier plots because their performance is off by as much as an order of magnitude. The behavior of ratio estimators is nevertheless quite interesting. For $r = 2$, normalized MSE's for $\hat{p}_R$ and $\hat{p}_R^*$ are exhibited in Figure 5. The figure shows the values of $N*$MSE for $N = 10$, 20, 50, and 100, as well as the values of $N*$AMSE. In each instance, 50% of the N observations come from each of the two populations. Clearly neither $\hat{p}_R$ nor $\hat{p}_R^*$ has acceptable performance for $r = 2$. Also, notice that the vertical axes for $\hat{p}_R$ and $\hat{p}_R^*$ are different and that this is an indication that $\hat{p}_R^*$ performs significantly worse than $\hat{p}_R$.

One of the interesting features of these MSE plots for the ratio estimators is the way the values shoot up near $p = 0$. While the functions are continuous, and converge pointwise to the better-behaved AMSE, the continuity and convergence is not uniform in p. Further, even for values of p closer to one, the ratio estimators perform badly as compared with the estimators shown in Figure 3. While only values for $r = 2$ are shown, similarly poor behavior occurs for larger values of r. Thus ratio estimators do not appear to be a prudent choice.

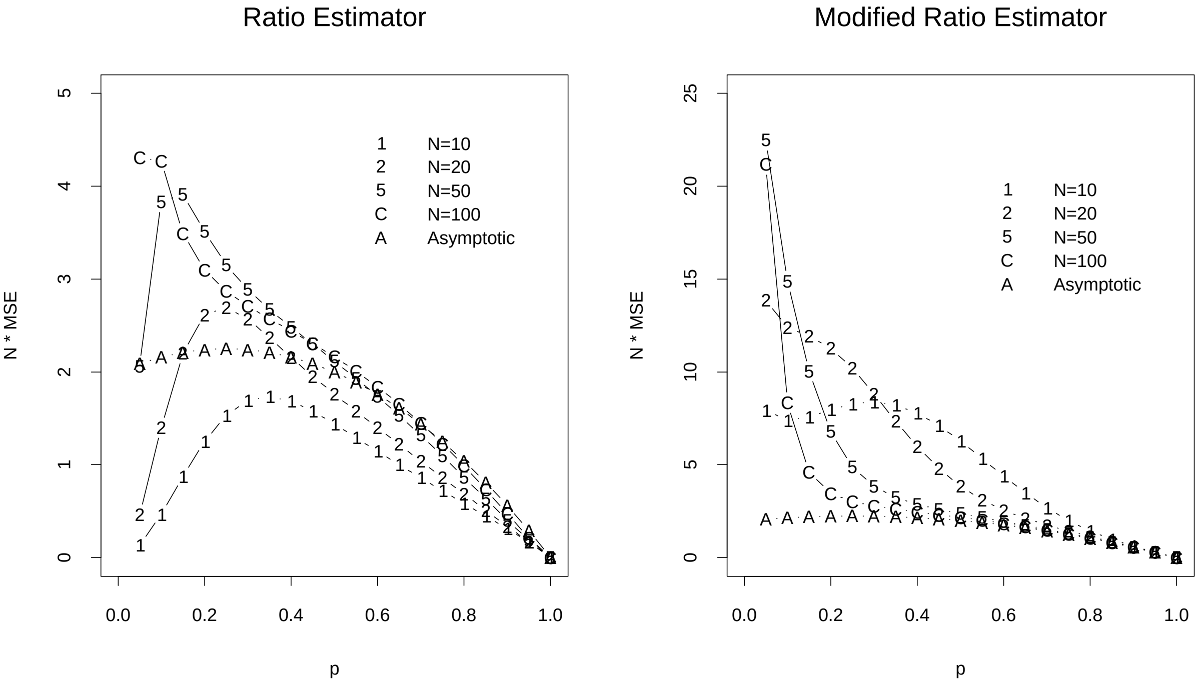


Figure 5: Ratio Estimators, 50/50 Allocation, $r = 2$. (Note change of vertical scale.)

7.4 MSEs of Adaptive Procedures

The performance of the sequential a_r-cut allocation rule is illustrated in Figure 6 for $N = 100$ and $a_2 = 1/3$. In this figure, **C** denotes the exact MSE of the adaptive a_r-cut rule and **E** represents the normalized lower envelope in Equation (1) of Section 6, namely $\min\{\mathrm{MSE}_p(\hat{p}_x), \mathrm{MSE}_p(\hat{p}_y)\}$.

Note that the cut-point rule has essentially the same MSE as the lower envelope of the AMSE's, except for the region of approximately $0.3 < p < 0.4$. In this interval, the MSE of $\hat{p}_{ml}$ actually improves upon the best values gotten using the AMSE's. This behavior leads to questions regarding optimal performance by any allocation rule based on the maximum likelihood estimator. A natural issue to examine is how well one could do if one were prescient, i.e., if one knew the value of p in question. Just as with the asymptotic MSE's, one can derive omniscient fixed allocations that minimize the MSE's for various estimators. In this section, the focus is on allocations used with maximum likelihood estimators because these estimators seem to perform the best overall. For $N = 100$, MSE's arising from use of the omniscient fixed allocation with $\hat{p}_{ml}$ were computed and are presented in Figure 7. These values are denoted with **B**'s. The plot on the right provides MSE's over the entire range of p and the one on the left is a blow-up of the parameter region in which the MSE of the omniscient fixed rule differs from, and is in fact, better than, the asymptotic lower envelope, denoted by **E**'s. The region in which the omniscient fixed rule improves on the asymptotic lower bound is quite small (and shrinks to zero width as $N \to \infty$), but it shows that allocating wholly from the p-experiment or wholly from the p^r-experiment is not optimal for all values of p and N. For example, when $N = 100$ and $p = .35$, the omniscient fixed allocation indicates that one should sample 53% from the p-experiment and the remainder from the p^2-experiment.

Also shown in Figure 7 is the adaptive cut-point rule. This rule performs even better than the omniscient fixed allocation rule in the area of the cut-point, 1/3. This is a region of *negative*

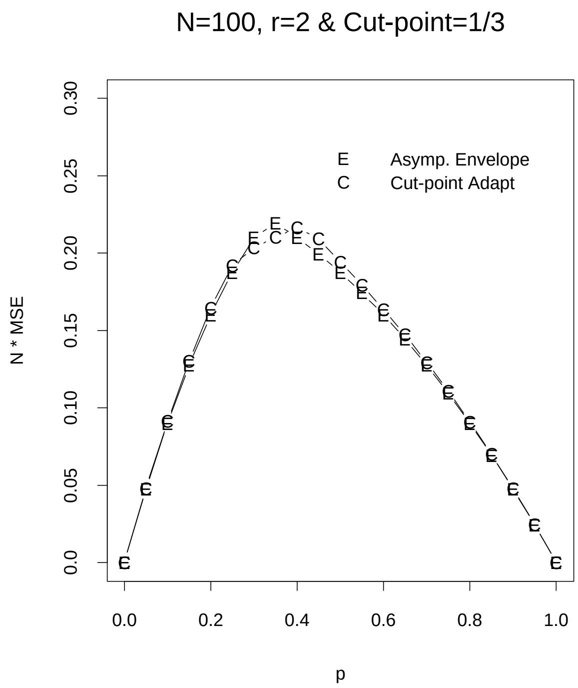


Figure 6: Adaptive Allocation vs. Asymptotic Lower Bound.

---

regret, as discussed, for example, in Woodroofe (1977) and Martinsek (1983), where the ability to adapt, without prior knowledge of the parameter, is so beneficial that it overtakes allocation that has advance knowledge of the parameter but which must be fixed in advance of any experiments.

To determine the true lower bound of the MSE for all rules using the MLE as the terminal estimator, an *omniscient sequential rule* (indicated by • in the figures), was computed. This is the rule that knows p and N in advance (but does not know the outcomes of tests in advance) and uses fully sequential adaptation to minimize the MSE of the terminal estimator (the MLE). While such rules are difficult to analyze, they can be determined computationally, for fixed p and N, through dynamic programming. The MSE of the omniscient sequential rule appears in Figure 7, along with the MSEs of several other rules. This rule had also appeared earlier in Figure 4.

Finally, computations similar to Figure 7 are shown in Figure 8 for $r = 10$ and $N = 100$. The cut-point used was 0.76, rather than $a_{10} = 0.679$, because this is the point at which the p^{10}-experiment is superior to the p-experiment when $N = 100$. The basic behavior is similar to that seen for $r = 2$, but many of the differences between asymptotic behavior and that for moderate N are more pronounced. For example, for small p, the extent to which the p^r-experiment is worse than the p-experiment is far more extreme. For the same value of N, the region in which the omniscient fixed allocation rule is a mixture of p- and p^r-experiments is far larger than it was for $r = 2$, and the ability of the omniscient sequential rule to improve upon the omniscient fixed rule is greatly enhanced. The MSE of the omniscient sequential rule is a more irregular function of p, and, though not evident in Figure 8, it is no longer unimodal.

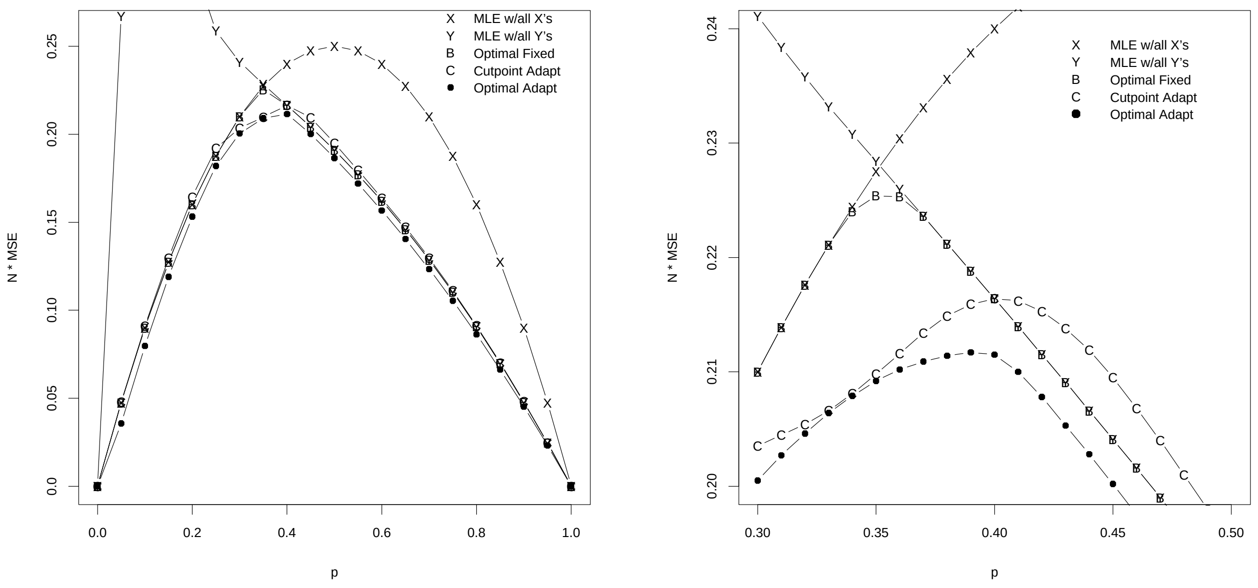


Figure 7: Adaptive Allocation using the MLE, $N = 100$, $r = 2$.

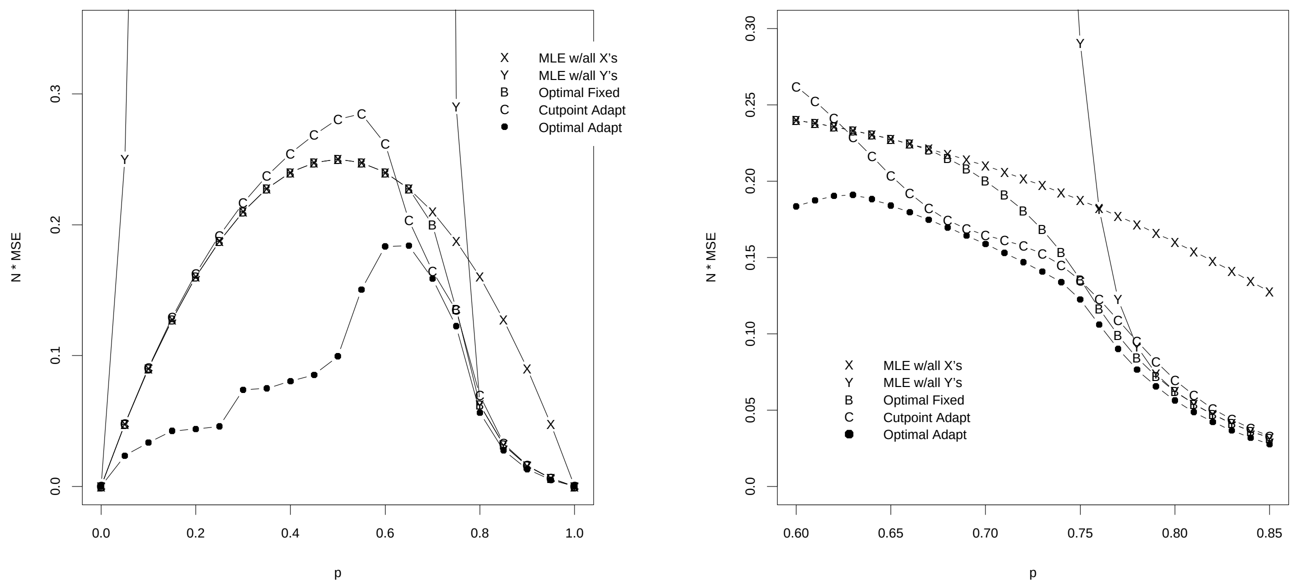


Figure 8: Adaptive Allocation using the MLE, $N = 100$, $r = 10$.

8 Conclusions

The adaptive procedure recommended in Section 6 uses a single cut-point, a_r, with the maximum likelihood estimator of p at each stage, and allocates to the p-experiment if the estimator of p is below the cut-point. The single cut form of the allocation rule was derived from maximum information considerations, and also arose when the asymptotic mean square error of selected estimators of p was considered. In the latter case, estimators and allocations were shown to interact, and the single cut-point allocation, along with either the maximum likelihood estimator or the sample size weighted average estimator, was shown to have asymptotic mean square error equal to the asymptotic lower bound.

While the proposed rule was designed to have excellent asymptotic properties, one is naturally concerned about its performance for practical sample sizes as well. Recall, for example, that some of the estimators considered were shown to have fixed sample behavior far worse than the MLE, and far from their asymptotic limits. The fixed sample size behavior of the adaptive cut-point rule, however, was quite close to its asymptotic behavior, though several unexpected points were discovered. First, it was determined that, if the two experiments are compared using the exact mean square error for the maximum likelihood estimator for each experiment separately, then the region where the p-experiment is favored is not always of the form $[0, a)$ for some a, as it is in the asymptotic case. Instead, there is a tiny interval $[0, a^*)$, depending on N, where the p^r experiment out-performs the p experiment. This indicates that the optimal allocation for the maximum likelihood estimator is not a single cut procedure. While the region where the MSE for p^r is less than that for p shrinks to zero as N tends to infinity, it is somewhat surprising that it exists at all.

Furthermore, when the exact mean square error for the maximum likelihood estimator is computed as a function of (m, n) for fixed $N = m + n$ in an interval of p around the point where the two experiments give the same MSE, the choice of (m, n) giving the lowest MSE is not all on one experiment or all on the other, but rather a mixture of observations on both experiments. However, near the cut-point this omniscient fixed allocation, which knows the values of p and N, does not perform as well as the adaptive cut allocation which does not know the parameters. (Note that the cut-point allocation rule is myopic, and is thus unaffected by knowledge of N.)

To summarize, the single cut adaptive allocation rule, using the MLE as the estimator, is simple to use and well motivated asymptotically. While some non-monotonic behavior of the omniscient fixed allocation is present in any fixed sample size, this behavior does not warrant a change in the form of the single cut rule. Perhaps the strongest property of the adaptive cut-point rule is the fact that it remains close to the omniscient fixed allocation rule throughout the entire parameter space, exceeding it near the cut-point, and nearly equals the performance of the unattainable omniscient sequential rule. (The omniscient sequential allocation knows the parameters p and N, and can exploit adaptive allocation to minimize the MSE of its terminal estimator, the maximum likelihood estimator.) Since adaptive procedures are derived by optimizing objective functions dependent on unknown parameters, then estimating the parameters, their performance is typically evaluated by how close they get to the optimum value of the function when the parameter is known. Adaptive procedures usually do well compared to fixed sample size, known parameter procedures, though the phenomena of negative regret observed here, whereby the adaptive procedure exceeds the omniscient fixed procedure, is reported only rarely. Further, the class of all sequential procedures is much richer, and the good behavior of the adaptive cut rule compared to the omniscient fixed procedure in this wider class is especially impressive, and indeed recommends its usage.

A major goal of this research was to derive an efficient adaptive allocation procedure for allocating observations between two experiments when estimating a common success probability. While related problems have been discussed in the group testing setting, there are still few analyses of situations where more than one test can give information about a single parameter. However, as more diagnostic tests are developed for medical or industrial applications, one can expect that such scenarios will become more commonplace. Directions for future research include incorporating cost models, tests which yield more than binary outcomes, and more flexible allocation routines. For example, recently Oliver-Hughes and Swallow (1994) analyzed an analogous two-stage setting which would permit simultaneous tests during each stage, as opposed to the requirement that the outcome of each test be available before the next is allocated. Other related work that we are pursuing incorporates a Bayesian model, rather than the frequentist model used herein. In the Bayesian setting the choice of estimator is simplified, but the computational analyses are more complicated because the posterior distributions are mixtures.

References

[1] Chen, C. L. and Swallow, W. H. (1990), "Using Group Testing to Estimate a Proportion, and to Test the Binomial Model", *Biometrics* **46**, 1035–1046.

[2] Easterling, R. G. and Prairie, R. R. (1971), "Combining Component and System Information", *Technometrics* **13**, 271–280.

[3] Gastwirth, J. and Hammick, P. (1989), "Estimation of the Prevalence of a Rare Disease, Preserving the Anonymity of the Subjects by Group Testing: Application to Estimating the Prevalence of AIDS Antibodies in Blood Donors", *Journal of Statistical Planning and Inference* **22**, 15–27.

[4] Loyer, M. W. (1983), "Bad Probability, Good Statistics, and Group Testing for Binomial Estimation", *The American Statistician* **37**, 57–59.

[5] Martinsek, A. (1983), "Second Order Approximation to the Risk of a Sequential Procedure", *Annals of Statistics* **11**, 827–836.

[6] Noble, W. (1990), *First Order Allocation*, Ph. D. Thesis, Michigan State University.

[7] Oliver-Hughes, J. and Swallow, W. H. (1994), "A Two-Stage Adaptive Group-Testing Procedure for Estimating Small Proportions", *Journal of the American Statistical Association* **89**, 982–993.

[8] Sobel, M. and Elashoff, R. M. (1975), "Group Testing with a New Goal, Estimation", *Biometrika* **62**, 181–193.

[9] Woodroofe, M. (1977), "Second Order Approximations for Sequential Point and Interval Estimation", *Annals of Statistics* **5**, 984–995.