

## **Ranked Set Sampling for a Population Proportion: Allocation of Sample Units to Each Judgment Order Statistic**

Jessica K. Kohlschmidt  
Department of Statistics  
The Ohio State University, OH  
USA

Elizabeth A. Stasny  
Department of Statistics  
The Ohio State University, OH  
USA

Douglas A. Wolfe  
Department of Statistics  
The Ohio State University, OH  
USA  
dw@stat.osu.edu

### **Abstract**

Ranked set sampling is an alternative to simple random sampling that has been shown to outperform simple random sampling in many situations by reducing the variance of an estimator, thereby providing the same accuracy with a smaller sample size than is needed in simple random sampling. Ranked set sampling involves preliminary ranking of potential sample units on the variable of interest using judgment or an auxiliary variable to aid in sample selection. Ranked set sampling prescribes the number of units from each rank order to be measured.

Balanced ranked set sampling assigns equal numbers of sample units to each rank order. Unbalanced ranked set sampling allows unequal allocation to the various ranks, but this allocation may be sensitive to the quality of information available to do the allocation. In this paper we use a simulation study to conduct a sensitivity analysis of optimal allocation of sample units to each of the order statistics in unbalanced ranked set sampling. Our motivating example comes from the National Survey of Families and Households.

**Keywords:** Optimal Allocation, Sensitivity Analysis, Simulation Study, Unbalanced Ranked Set Sampling.

### **1. Introduction**

Ranked set sampling (RSS), originally proposed by McIntyre (1952), is an alternative method of sample selection that has been shown to improve on simple random sampling (SRS). RSS uses judgment ranking of the characteristic of interest to improve estimation of a population parameter. For a general introduction to RSS, see Wolfe (2004). Theoretical results have shown that in many settings RSS estimators are unbiased with precisions at least as small as those of the corresponding SRS estimators (see, for

example, Patil, 1995). The improvement in RSS over SRS is especially evident in situations where the sample units can be easily ranked but actual measurement of units is costly in time and/or effort. In this section, we provide background information on balanced and unbalanced RSS.

### 1.1 Balanced RSS

The most basic version of RSS is balanced RSS. In this setting, each judgment order statistic is allotted the same number of sample units. Under balanced RSS, we first select  $m^2$  items from the population at random. These items are then randomly divided into  $m$  sets of  $m$  units each. Within each set, we rank the  $m$  units according to the characteristic of interest by judgment or through the use of an auxiliary variable or variables. From the first set, we select the item with the smallest ranking,  $X_{[1]}$ , for measurement. From the second set, we select the item with the second smallest ranking,  $X_{[2]}$ . We continue in this manner until we have ranked the items in the  $m^{\text{th}}$  set and selected the item with the largest ranking,  $X_{[m]}$ . This complete procedure, called a cycle, is repeated independently  $k$  times to obtain a RSS of size  $n = mk$ . As is evident, a total of  $m^2k$  items are selected randomly but only  $mk$  units are measured.

Let  $X_{[r]i}$  denote the quantified  $r^{\text{th}}$  judgment order statistic from the  $i^{\text{th}}$  cycle. The RSS estimator of the population mean  $\mu$  is the average of these RSS observations; that is,

$$\hat{\mu} = \frac{1}{mk} \sum_{i=1}^k \sum_{r=1}^m X_{[r]i}.$$

We note that if the  $X_{[r]i}$  are binary variables then  $\hat{\mu}$  estimates the population proportion. The RSS estimator,  $\hat{\mu}$ , is an unbiased estimator of  $\mu$  and is at least as precise as the SRS estimator based on the same number of measured observations (see, for example, Dell and Clutter, 1972; Bohn, 1996; Patil, 2002). There are a number of factors that affect how much more precise the RSS estimator is than the SRS estimator. For example, the more accurate the ranking is within each set, the more precise the RSS estimator will be. In cases where the ranking is based on a concomitant or auxiliary variable, Chen *et al.* (2005) show that the amount of improvement in the precision of the RSS estimator of a population proportion is directly related to the correlation between the concomitant variable and the variable of interest.

### 1.2 Unbalanced RSS

Another option is that of unbalanced RSS under which possibly different numbers of each ranked order statistic are selected for measurement. Neyman allocation may be used to allocate sample units for each order statistic proportionally according to its standard error. This is the optimal form of unbalanced allocation in that it leads to minimum variance among the class of all such RSS estimators. Chen *et al.* (2006) discuss the general properties of unbalanced RSS and describe the Neyman allocation method for assigning sampling units to each judgment order statistic when the goal is to estimate a population proportion. We let  $n_r$  denote the number of observations allocated to the  $r^{\text{th}}$  order statistic. Then, for each  $r = 1, \dots, m$ , we sample  $n_r$  sets of size  $m$  units each from the population and obtain rankings of the variable of interest within each set as before.

Instead of measuring equal numbers of the various judgment ordered units, however, we take  $n_r$  measurements of the  $r^{\text{th}}$  judgment order statistic, for  $r = 1, \dots, m$ . The total number of measured units is then  $n = \sum_{r=1}^m n_r$ . Under unbalanced RSS, it is not always the case that the RSS estimator will have greater precision than the balanced RSS or the SRS estimator.

In this paper, we examine the sensitivity of unbalanced RSS to departures from optimal allocation when the goal is to estimate a population proportion. Neyman allocation is only optimal if the true population proportion is known in advance of performing the allocation. In practice, of course, we only have an estimate of the population proportion based on a previous study or an educated guess. Thus, it is important to know how such “approximate” Neyman allocation performs. In our study, we vary the sample sizes from the optimal sample sizes found using Neyman allocation and examine the effect this has on the standard error of the RSS estimator. We also study the effect of imperfect rankings on this standard error.

For additional discussion of the use of RSS with binary data, see Lacayo *et al.* (2002), Terpstra (2004), Terpstra and Liudahl (2004), Chen *et al.* (2005, 2007), Terpstra and Nelson (2005), Terpstra and Miller (2006), and Chen (2007). Theoretical results for the sample mean apply immediately to this situation as a sample proportion is simply the sample mean for binary data. Terpstra and Liudahl (2004) and Chen *et al.* (2005, 2007) studied the use of logistic regression based on auxiliary variables that are either readily available or easy to obtain from potential sampling units to estimate probabilities of success that are, in turn, used to improve the accuracy of the within-sets ranking process. This approach can lead to considerable gains in precision of the RSS estimator over the SRS estimator. In the simulations for this paper, we use only a single auxiliary variable to accomplish our within-sets ranking, which is, of course, equivalent to using the estimated probabilities from a logistic regression model based on only this one auxiliary variable.

Another issue in RSS is that of perfect versus imperfect rankings. If rankings are perfect, the judgment order statistics equal the true order statistics. In the case of a binary variable, when rankings are perfect we can express the probability of success for each rank order statistic as a function of the underlying population proportion only. When the ranking procedure lies somewhere between random ordering and perfect rankings (that is, we have imperfect rankings), there is concern as to how well Neyman allocation will perform. We study the potential loss of precision in the unbalanced RSS estimator if our stipulated unbalanced allocation (derived under the assumption of perfect rankings) deviates from the true optimal allocation.

In section 2, we discuss how we expect the sample allocation to affect the precision of the estimator. In section 3, we describe the data that we use for our simulation. The simulation results are discussed in section 4. Section 5 presents conclusions and future work.

## 2. Departures from Optimal Allocation

In this paper, we address the sensitivity of the estimator of a population proportion to departures from the optimal allocation in unbalanced RSS. If we knew the true population proportion, then we would be able to determine the exact Neyman optimal allocation of the total sample to the various judgment order statistics. Chen *et al.* (2006) provide the following results for determining probabilities of success,  $p_{(r)}$ , within rank  $r$ ,  $r = 1, 2, \dots, m$  when the true population proportion is known:

$$p \geq 0.5 \Rightarrow p_{(r)} = \sum_{t=m-r+1}^m \binom{m}{t} p^t (1-p)^{m-t} = 1 - F(m-r) \quad (1)$$

$$p < 0.5 \Rightarrow p_{(r)} = \sum_{t=r}^m \binom{m}{t} p^t (1-p)^{m-t} = 1 - F(r-1) \quad (2)$$

The function  $F(\cdot)$  in equations (1) and (2) is the c.d.f. for the Binomial( $m, p$ ) distribution. Neyman allocation then specifies the rank order sample sizes as follows:

$$n_r = n \frac{\sqrt{p_{(r)}(1-p_{(r)})}}{\sum_{r=1}^m \sqrt{p_{(r)}(1-p_{(r)})}}, \quad r = 1, \dots, m. \quad (3)$$

We do not know the true population proportion, however, since that is what we wish to estimate with our sample. Thus, it is necessary to use a rough preliminary estimate of the population proportion to determine an approximate “optimal” allocation of the sample units.

We anticipate that there is some flexibility in how close our approximate allocation has to be to the optimal allocation to still achieve a degree of precision close to that of the RSS estimator with optimal allocation. If the precision of the RSS estimator is relatively insensitive to departures from optimal allocation, then errors in the preliminary rough estimate of the population proportion used to obtain approximate Neyman allocation should not result in large increases in the variance of the RSS estimator.

We examine the sensitivity of the standard errors of the RSS estimators to departures from the optimal allocations through a simulation study. This is accomplished by first determining the Neyman allocation based on a known population proportion,  $p$ , using a set size of three. Then we use this optimal allocation to simulate the sampling distribution of the RSS estimator of  $p$ . This provides us with an estimate of the best possible improvement (over SRS) in precision from RSS. Then we conduct similar simulations with allocations differing from the optimal Neyman allocation to assess the resulting effect on the precision of the unbalanced RSS estimators. We estimate the relative precision of the unbalanced RSS estimator to the balanced RSS estimator by the standard error of the balanced RSS estimator divided by the standard error of the unbalanced RSS estimator. A similar relative precision is also used to compare the unbalanced RSS estimator to the SRS estimator. Plots of the estimated relative precisions of these various allocations and their associated RSS estimators will be used to evaluate the robustness of Neyman allocation to misspecification.

### **3. The Data**

To make our simulated RSS realistic, we choose a public-use data set, the National Survey of Families and Households (NSFH), and treat it as our population of interest. The NSFH was funded by the Center for Population Research of the United States' National Institute of Child Health and Human Development. The fieldwork was completed by the Institute for Survey Research at Temple University. The NSFH data were collected in three waves using a national probability sample of 13,008 individuals aged 19 and over (see Sweet, *et al.*, 1988, for a detailed description of the NSFH). The sample includes a main, cross-sectional sample of 9,637 households plus an over-sampling of African Americans, Puerto Ricans, Mexican Americans, single-parent families, families with step-children, cohabiting couples and recently married persons. One adult per household was randomly selected as the primary respondent. Several portions of the main interview were self-administered to facilitate the collection of sensitive information and to ease the flow of the interview. The average interview lasted one hour and forty minutes. In addition, a shorter self-administered questionnaire was given to the spouse or cohabiting partner of the primary respondent.

Respondents were first interviewed between March 1987 and March 1988. They were re-contacted between 1992 and 1994 for a follow-up interview, and a third interview was conducted in 2001-2003. A considerable amount of life-history information was collected, including: the respondent's family living arrangements in childhood, departures and returns to the parental home, and histories of marriage, cohabitation, education, fertility, and employment. The design permits the detailed description of past and current living arrangements and other characteristics and experiences, as well as the analysis of the consequences of earlier patterns on current states, marital and parenting relationships, kin contact, and economic and psychological well-being on employment, age, and receipt of public assistance.

We consider data collected in the first wave of the NSFH as our population so that we know the population proportions exactly. This permits us to determine the optimal Neyman allocation in a variety of situations. In practice, estimates of the population proportion from the first wave of the survey could be used to determine the sample allocation for follow-up surveys.

### **4. The Simulation Study**

#### **4.1 Perfect Rankings**

We discuss the details of our simulation for the case of perfect rankings with set sizes of  $m = 3, 4$ , and 5. A total sample size of two hundred observations was used. We denote the sample size allocated to the  $i^{\text{th}}$  judgment order statistic as  $n_i$ ,  $i = 1, 2, 3$ . For this setting, we take the NSFH females to be our population of interest and consider the women's age as the variable of interest. We construct three binary variables based on age range to provide data sets with three different population proportions. The category of age over 23 yields a population proportion of  $p = 0.897$ , the category of age over 30 yields a population proportion of  $p = 0.743$ , and the third category of age over 35 has associated  $p = 0.575$ . We use the actual age of the respondent as the ranking variable for  $p =$  probability of a given age range. Thus we guarantee that the rankings are correct.

**Table 1: Optimal Allocation for Various Values of the Population Proportion,  $p$** 

$P$	$\%n_1$	$\%n_2$	$\%n_3$
0.575	34.5	42.5	23.0
0.743	49.5	37.5	13.0
0.897	69.0	26.0	5.0

Note:  $\% n_r$  = percentage of total sample size allocated to the  $r^{\text{th}}$  order statistic

Table 1 shows the optimal allocation of the RSS sample found using equation (3) for the three values of  $p$ . For each different population proportion, there is one judgment order statistic that is least likely to have a success. We measure a larger number of observations from that order statistic so that we have a sufficient number of successes to accurately estimate the probability of success in that judgment rank. As the population proportion  $p$  varies, the allocations to the three judgment order statistics change. We see from Table 1 that for  $p$  close to 0.5 the percentages of the sample allocated to the various order statistics are similar. On the other hand, when  $p$  is close to 1, a very small portion of the sample is allocated to the largest judgment order statistic; the majority of the sample is allocated to the smallest judgment order statistic.

In our simulation study for  $m = 3$ , we first select 3 units at random from the NSFH “population”. We rank these 3 units by some characteristic or auxiliary variable depending on the situation. In our case, we want two hundred measured observations in our unbalanced RSS. The optimal proportion of the sample allocated to each order statistic is given in Table 1. In our simulations, we vary the number allotted to each order statistic. First, we vary by one unit out to five units from the optimal allocation and then we vary by increments of five units out another twenty-five units from optimal allocation. Within each rank, we calculate the sample proportion of the variable of interest. For  $\hat{p}_{RSS}$ , our RSS estimate of the population proportion,  $p$ , we simply take the average of these within-ranks sample proportions. This simulation is repeated 10,000 times to enable us to estimate the mean and standard error of  $\hat{p}_{RSS}$  for each possible allocation that we are considering. The exact standard error for the SRS estimator,  $\hat{p}_{SRS}$ , is given by

$$SE(\hat{p}_{SRS}) = \sqrt{\frac{p(1-p)}{n}},$$

where  $p$  is the true population proportion. Note that in our case, the finite population correction factor is 0.98 and so it can be omitted from the calculations without changing the results significantly.

We take a similar approach in our simulation for set sizes  $m = 4$  and 5. In the figures that appear in the appendix, we plot  $SE(\hat{p}_{BalRSS})/SE(\hat{p}_{UnbalRSS})$  and  $SE(\hat{p}_{SRS})/SE(\hat{p}_{UnbalRSS})$ . The horizontal line at one that appears in graphs corresponds to the sampling methods having the same standard errors, meaning their performance is the same. Above this line, unbalanced RSS yields higher precision and below this line balanced RSS or SRS has greater precision.

For perfect rankings, Figures 1 – 9 show plots of the relative precision of the unbalanced RSS estimator to both the balanced RSS estimator and the SRS estimator for set sizes 3,

4, and 5, with  $p$ 's corresponding to the various proportions of women in the three age groups. The general shape of the graphs of the relative precision is a parabola. The lack of smoothness in the maximum portions of the parabolas is due to simulation error. (Note the small scale on the axis for the relative precision.) The top curve ( $\circ$ ) in each graph shows the relative precision of unbalanced RSS to SRS and the bottom curve ( $\Delta$ ) shows the relative precision of unbalanced RSS to balanced RSS.

Figures 1 - 3 display graphs of relative precision for estimating the proportion of females older than 35, corresponding to  $p = 0.575$ , for set sizes  $m = 3, 4$ , and  $5$ , respectively. It is evident from Figure 1 that when we hold  $n_1$  fixed the attained precision of the RSS estimator remains close to the relative precision under optimal allocation even if we allow the allocations to the other two ranks to vary by as much as plus or minus ten observations. Similarly, when  $n_2$  or  $n_3$  is held fixed, the allocation to the other two ranks may change by as much as plus or minus ten observations from the optimal allocation with only minimal loss of precision for the estimator. These results suggest that for set size 3 and  $p$  close to 0.5 the number of sample units allocated to each judgment order statistic does not have to be exactly at the optimal allocation for the precision of the RSS estimator to remain close to optimal. From Figures 2 and 3, we see a similar pattern for set sizes 4 and 5, respectively. Note that the top curve on the plots ( $\circ$ ) is the relative precision of unbalanced RSS to SRS. Chen *et al* (2006) discuss how unbalanced RSS significantly outperforms SRS and that is evident from our graphs as well.

Next, we consider females older than 30, corresponding to a population proportion of  $p = 0.743$ . In this case, the first judgment order statistic is assigned almost half the observations under optimal allocation. Figures 4 – 6 show the effect that changing the sample allocations has on the relative precision of the RSS estimator for this value of  $p$  and for set sizes 3, 4, and 5, respectively. For set size  $m = 3$ , holding  $n_1$  fixed, we see from Figure 4 that we can vary the sample allocations to the other two order statistics by fifteen in either direction and still have nearly optimal precision. Holding  $n_2$  fixed, we can vary the other two sample allocations from the optimal allocation by plus or minus twenty observations without much loss in precision. When we hold  $n_3$  fixed, there is even more flexibility. We can vary by plus or minus thirty observations without unbalanced RSS performing worse than balanced RSS. We see from Figures 5 and 6 that the same pattern also holds for set sizes 4 and 5.

Lastly, we consider females older than 23, corresponding to a population proportion of  $p = 0.897$ . In this case, optimal allocation assigns 69% of the sample to the first order statistic. Figures 7 – 9 show the effect of changing the sample allocations on the relative precision of the RSS estimator for this value of  $p$  and set sizes 3, 4, and 5, respectively. In this case, due to the heavy skewness of the binomial distribution with  $p = 0.897$ , unbalanced RSS clearly outperforms both balanced RSS and SRS. For set size  $m = 3$ , holding  $n_1$  fixed, we can vary the sample size from the optimal allocation by about thirty in either direction if possible and still outperform balanced RSS and SRS. When  $n_2$  or  $n_3$  is fixed, we see that we can vary from the optimal allocation by thirty observations if possible without substantial reduction in precision. The pattern holds for set sizes 4 and 5 as well.

## 4.2 Imperfect Rankings

Next, we address the case of imperfect rankings. In this setting, we use auxiliary variables for the ranking of units. Here we look at the proportions of males and females working: 0.743 and 0.544, respectively. For both males and females we consider separately the three possible auxiliary variables, “public assistance”, “age”, and “hours worked last week” to obtain our RSS rankings. (The public assistance variable is an indicator of whether or not the respondent’s family received public assistance when the respondent was a child.) We could, of course, incorporate all three of these auxiliary variables simultaneously by using a single logistic regression model based on the three variables to estimate probabilities of success and then use these estimated probabilities to obtain our within-sets rankings, as proposed in Terpstra and Liudahl (2004) and Chen *et al.* (2005, 2007). This would certainly lead to improvement in precision over the three separate RSS estimators based on one of these auxiliary variables at a time. One of the items of interest to us in this paper, however, is the effect of the correlation between a single ranking variable and the variable of interest on the robustness of the optimal allocation of sample units in unbalanced RSS. Thus, we choose to consider each of these three variables separately in our simulation studies.

We expect that the gain in precision for the unbalanced RSS estimator over the balanced RSS and SRS estimators will be an increasing function of the absolute magnitude of the correlation between the variable of interest and the auxiliary variable used to obtain the RSS rankings. The relevant Kendall correlations between working (a binary variable) and the auxiliary variables mentioned above for both females and males are shown in Table 2.

**Table 2: Correlations of Auxiliary Variables with Working**

Auxiliary Variable	Females	Males
Public Assistance	0.039	0.035
Age	0.238	0.319
Number of Hours Worked	0.758	0.597

We again consider sample allocations different from the optimal allocation to each judgment order statistic to see what effect this has on the standard error of the RSS estimator of  $p$ .

We discuss our findings for the women’s data (the results are similar for the male data set), for which the proportion of women reporting that they are working is  $p = 0.544$ . The effect of adjusting sample allocations on the relative precision of the unbalanced RSS estimator is displayed in Figures 10 – 18 for set sizes 3, 4, and 5. Once again, the top curve ( $\circ$ ) is the relative precision of unbalanced RSS to SRS and the bottom curve ( $\Delta$ ) is the relative precision of unbalanced RSS to balanced RSS.

We first consider the case where “public assistance” (correlation of 0.039 with working) is used for the rankings. Since this ranking variable is virtually uncorrelated with the variable of interest, we see from Figures 10 – 12 that the ranking does not improve the precision of the unbalanced RSS estimator. In this case, we always do worse with



unbalanced RSS than with balanced RSS or SRS even with the “optimal allocation” under perfect rankings. This phenomenon is common across all three set sizes and illustrates the danger of using unbalanced RSS when the rankings are not very reliable. Only balanced RSS guarantees that we will not do worse than SRS. This leads us to the conclusion that it is extremely important to have a rough idea of the correlation that the ranking variable has with the variable of interest before using unbalanced RSS.

Next, consider the case where “age” (correlation 0.238 with working) is used for ranking. We again look at what happens when we hold the “optimal” sample allocations associated with each of the judgment order statistics fixed. The results are displayed in Figures 13 – 15. Since the ranking variable is not highly correlated with the variable of interest, it is not surprising that the precision of the unbalanced RSS estimator is less than the precision of the balanced RSS or the SRS estimator unless the sample allocations are nearly optimal. When we hold  $n_3$  fixed, unbalanced RSS does not outperform balanced RSS. As we vary the sample allocations for the judgment order statistics, both the balanced RSS and the SRS estimator quickly outperform the unbalanced RSS estimator with non-optimal sample allocations. This general pattern also holds as we increase the set size. With weak correlation, it is difficult for unbalanced RSS to outperform balanced RSS and SRS for larger set sizes due at least partly to the fact that there are more units that can be ranked incorrectly as we increase the set size.

Finally, consider the ranking variable “hours worked last week”, which has correlation 0.758 with the variable of interest. When we vary the sample allocations to each judgment order statistic for this setting, we see from Figures 16 – 18 that there is substantial improvement in precision with the unbalanced RSS estimator in comparison to that from balanced RSS and SRS. It is interesting to note that balanced RSS and SRS out performs unbalanced RSS only when there is a small number of observations in one of the order statistics. It is necessary to depart from optimal allocation by more than twenty in a given order statistic for the precision of the unbalanced RSS estimator to be worse than that of the balanced RSS and the SRS estimators when we hold  $n_1$  or  $n_2$  fixed. When we hold  $n_3$  fixed, we can vary from optimal allocation by thirty and still have unbalanced RSS perform better than either balanced RSS and SRS. A similar pattern holds for set sizes 4 and 5.

In a few settings when we deviate from optimal allocation a problem occurs when the sample size for one of the judgment order statistics gets too small. In these cases, we quickly do worse with overly unbalanced RSS than with balanced RSS or SRS, since it is necessary even in unbalanced RSS to sample a minimal number of units from each of the judgment order statistics to effectively estimate a population proportion.

## **5. Conclusions and Further Work**

In this paper, we have studied the sensitivity of unbalanced RSS estimators to deviations from the optimal allocation of the sample to the judgment order statistics. We concluded that under perfect rankings the optimal allocation is not crucial to insure that the unbalanced RSS estimator has greater precision than both the balanced RSS and the SRS estimators. In the case where we have imperfect rankings, however, there is not as much flexibility in departing from the optimal allocation of the sample. When the correlation

between the ranking variable and the variable of interest is low, deviating too far from the optimal allocation results in the unbalanced RSS estimator being worse than both the balanced RSS and the SRS estimators. As this correlation increases to 0.5 and above, we once again have considerable flexibility in how the sample is allocated. In such settings even if we differ from the optimal allocation by 10% in one of the judgment order statistics, the unbalanced RSS estimator still has greater precision than both the balanced RSS and the SRS estimators.

In this paper we considered set sizes three, four, and five. We did this because no clear pattern emerged for set size two. We note, of course, that as the set size increases the probability of imperfect rankings also increases in any practical setting. The other point to make is that if the overall sample size stays the same and the set size increases, then the number of units in each rank decreases, which provides less flexibility in our rank allocations.

We were concerned with how much flexibility we have in varying the sample allocation under unbalanced RSS and still improving on the balanced RSS and SRS estimators. We showed that in most cases, unbalanced RSS outperforms balanced RSS and SRS even if we are not using optimal allocation of the sample units. Clearly, if it is possible to use optimal allocation then that is the best for minimizing standard error. It has already been shown that optimal allocation of unbalanced RSS will do considerably better than balanced RSS and SRS. Here we wanted to show that if there was uncertainty about the proportion  $p$  we could still improve on the balanced RSS and SRS estimators using unbalanced RSS. This gives us the freedom to use unbalanced RSS even in situations where there is uncertainty about the likely value of  $p$ .

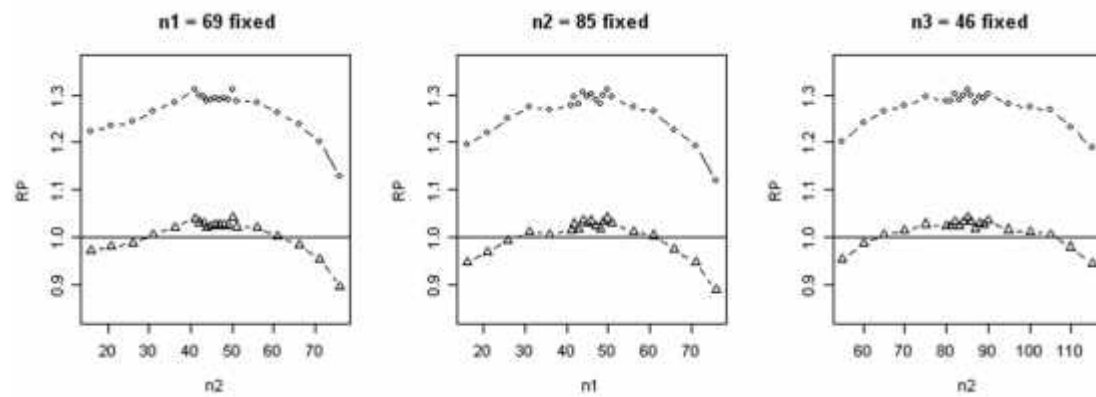
## References:

1. Bohn, L. L., 1996. A Review of Nonparametric Ranked Set Sampling Methodology, *Communications in Statistics, Theory and Methods*, **25**, 2675-2685.
2. Chen, H., 2007. Alternative Ranked Set Sampling Estimators for the Variance of a Sample Proportion, *Journal of Applied Statistical Science*, **15(1)**, 35-38.
3. Chen, H., Stasny, E. A., and Wolfe, D. A., 2005. Ranked Set Sampling for Efficient Estimation of a Population Proportion, *Statistics in Medicine*, **24**, 3319-3329.
4. Chen, H., Stasny, E. A., and Wolfe, D. A., 2006. Unbalanced Ranked Set Sampling for Estimating a Population Proportion, *Biometrics*, **62**, 150-158.
5. Chen, H., Stasny, E. A., and Wolfe, D. A., 2007. Improved Procedures for Estimation of Disease Prevalence Using Ranked Set Sampling, *Biometrical Journal*, **49(4)**, 530-538.
6. Dell, T. R. and Clutter, J. L., 1972. Ranked Set Sampling Theory with Order Statistics Background, *Biometrics*, **28**, 545-555.
7. Lacayo, H., Neerchal, N. K., and Sinha, B. K., 2002. Ranked Set Sampling from a Dichotomous Population, *Journal of Applied Statistical Science*, **11(1)**, 83-90.

8. McIntyre, G. A., 1952. A Method for Unbiased Selective Sampling, Using Ranked Sets, *Australian Journal of Agricultural Research*, **3**, 385-390.
9. Patil, G. P., 1995. Editorial: Ranked set Sampling, *Environmental and Ecological Statistics*, **2**, 271-285.
10. Patil, G. P., 2002. Ranked Set Sampling, *Encyclopedia of Environmetrics*, **3**, 1684-1690.
11. Sweet, J., Bumpass, L., and Call, V.R.A., 1988. The Design and Content of the National Survey of Families and Households, NSFH Working Paper No. 1, Center for Demography and Ecology, University of Wisconsin, Madison.
12. Terpstra, J.T., 2004. On Estimating a Population Proportion via Ranked Set Sampling, *Biometrical Journal*, **2**, 264-272.
13. Terpstra, J.T. and Liudahl, L.A., 2004. Concomitant-based Rank Set Sampling Proportion Estimates, *Statistics in Medicine*, **23**, 2061-2070.
14. Terpstra, J.T. and Miller, Z.A., 2006. Exact Inference for a Population Proportion Based on a Ranked Set Sample, *Communications in Statistics: Simulation and Computation*, **35(1)**, 19-27.
15. Terpstra, J.T. and Nelson, E. J., 2005. Optimal Rank Set Sampling Estimates for a Population Proportion, *Journal of Statistical Planning and Inference*, **127**, 309-321.
16. Wolfe, D. A., 2004. Ranked Set Sampling: An Approach to More Efficient Data Collection, *Statistical Science*, **4**, 636-643.

## Appendix of Figures

**Figure 1:** Relative Precision for Females with age over 35,  $m = 3$ ,  $p = 0.575$ , with Departures from Optimal Allocation

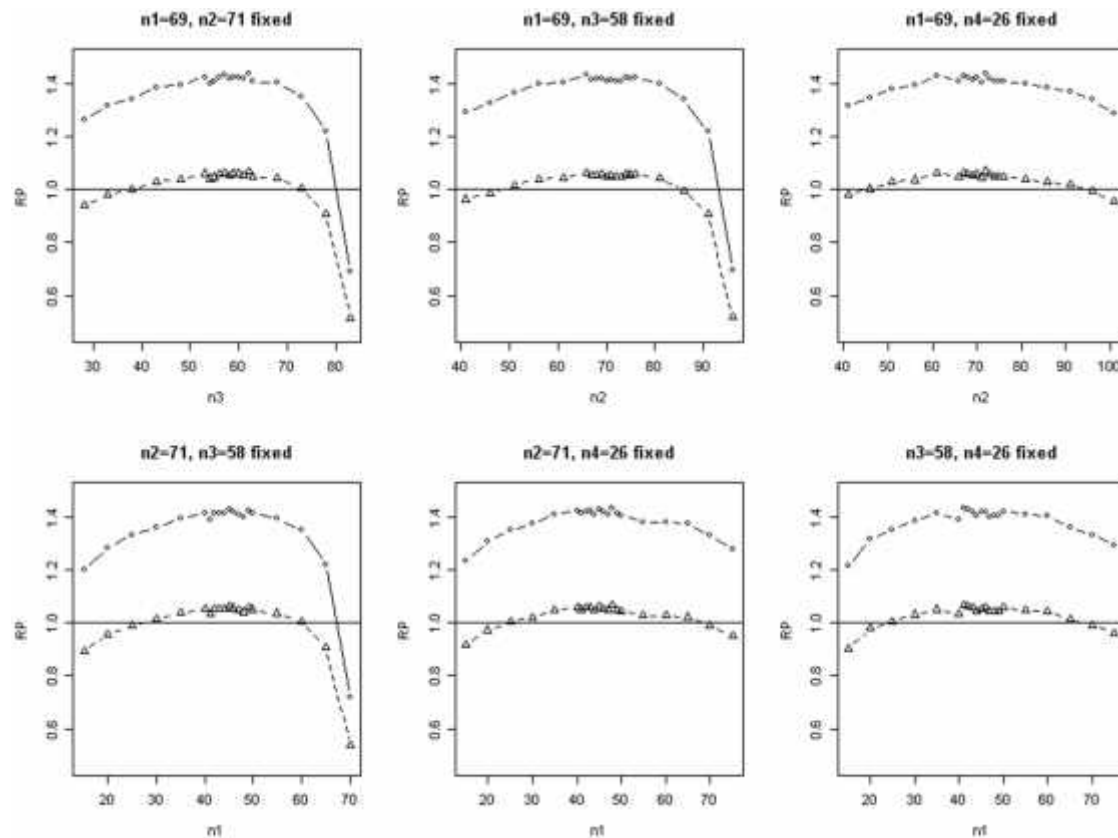


Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 69$ ,  $n_2 = 85$ , and  $n_3 = 46$ .

( $\circ$ ) denotes the relative precision of unbalanced RSS to SRS.

( $\Delta$ ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 2:** Relative Precision for Females with age over 35,  $m = 4$ ,  $p = 0.575$ , with Departures from Optimal Allocation

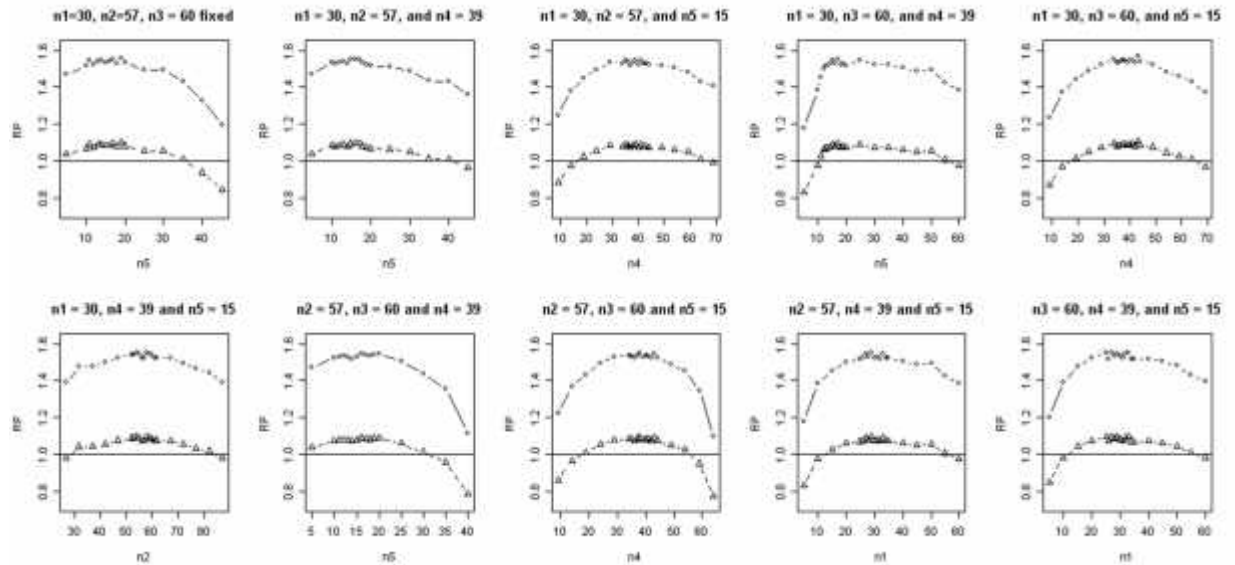


Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 69$ ,  $n_2 = 71$ ,  $n_3 = 58$ , and  $n_4 = 26$ .

( $\circ$ ) denotes the relative precision of unbalanced RSS to SRS.

( $\Delta$ ) denotes the relative precision of unbalanced RSS to balanced RSS.

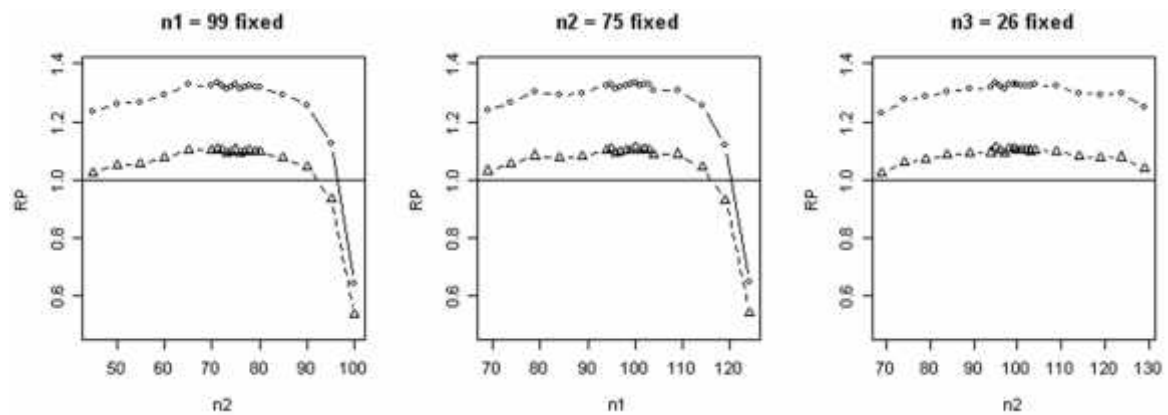
**Figure 3:** Relative Precision for Females with age over 35,  $m = 5$ ,  $p = 0.575$ , with Departures from Optimal Allocation



Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 30, n_2 = 57, n_3 = 60, n_4 = 39$ , and  $n_5 = 15$ .  
 (○) denotes the relative precision of unbalanced RSS to SRS.

(△) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 4:** Relative Precision for Females with age over 30,  $m = 3$ ,  $p = 0.743$ , with Departures from Optimal Allocation

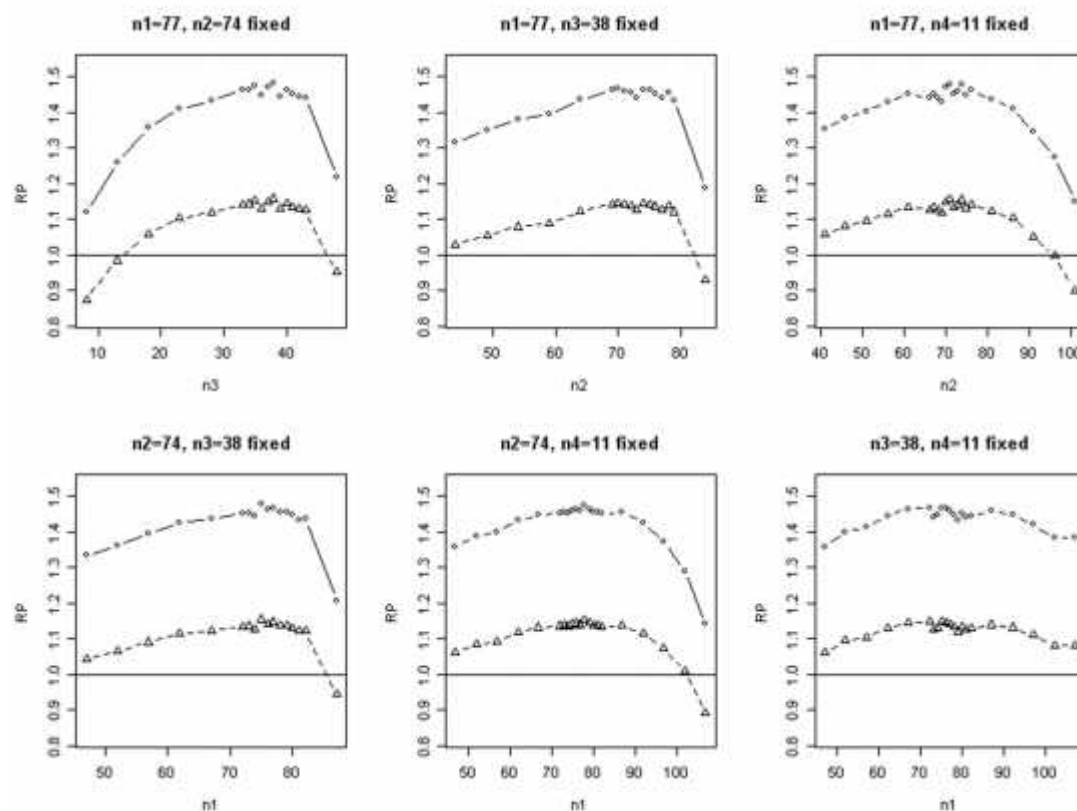


Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 99, n_2 = 75$ , and  $n_3 = 26$ .

(○) denotes the relative precision of unbalanced RSS to SRS.

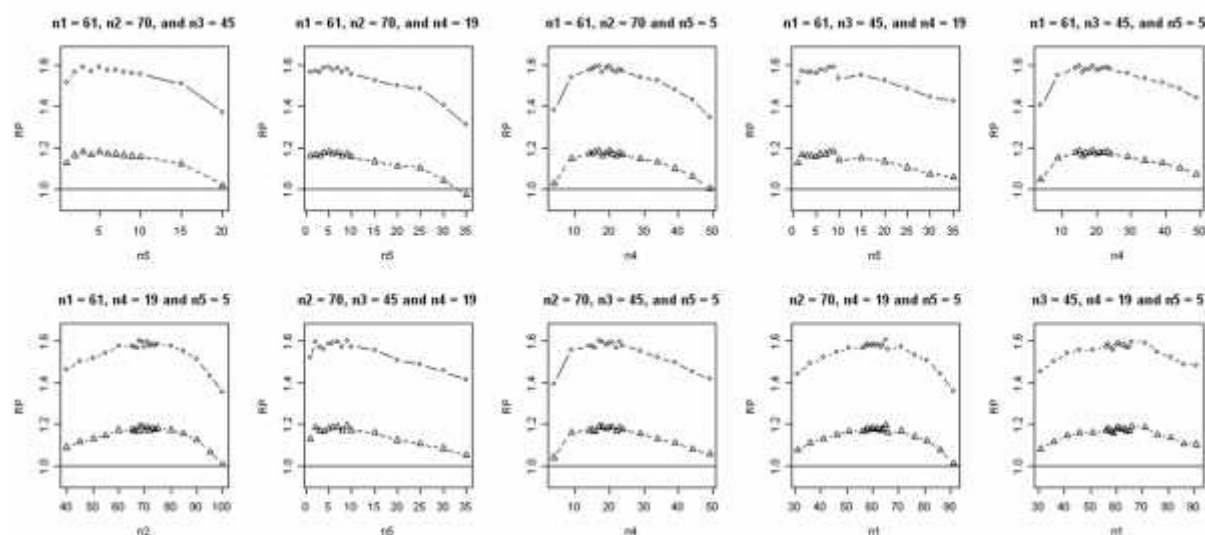
(△) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 5:** Relative Precision for Females with age over 30,  $m = 4$ ,  $p = 0.743$ , with Departures from Optimal Allocation



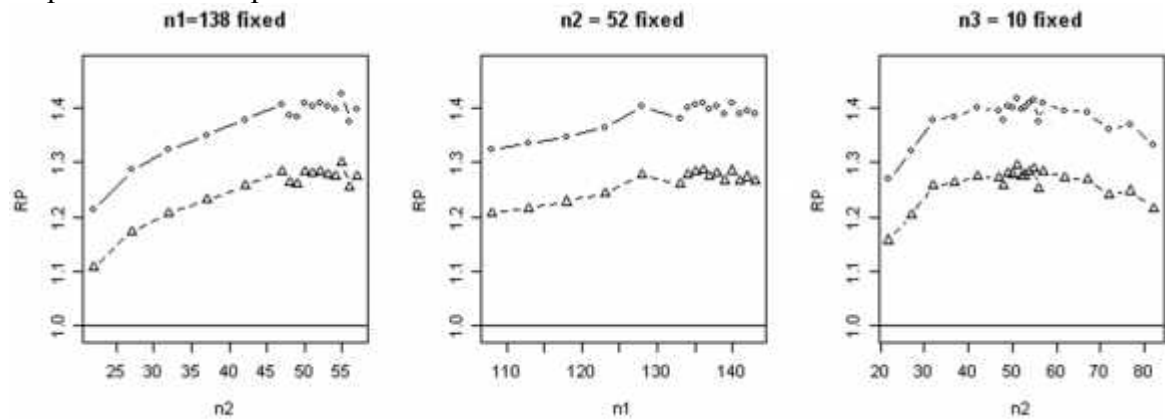
Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 77$ ,  $n_2 = 74$ ,  $n_3 = 38$ , and  $n_4 = 11$ .  
 (○) denotes the relative precision of unbalanced RSS to SRS.  
 (Δ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 6:** Relative Precision for Females with age over 30,  $m = 5$ ,  $p = 0.743$ , with Departures from Optimal Allocation



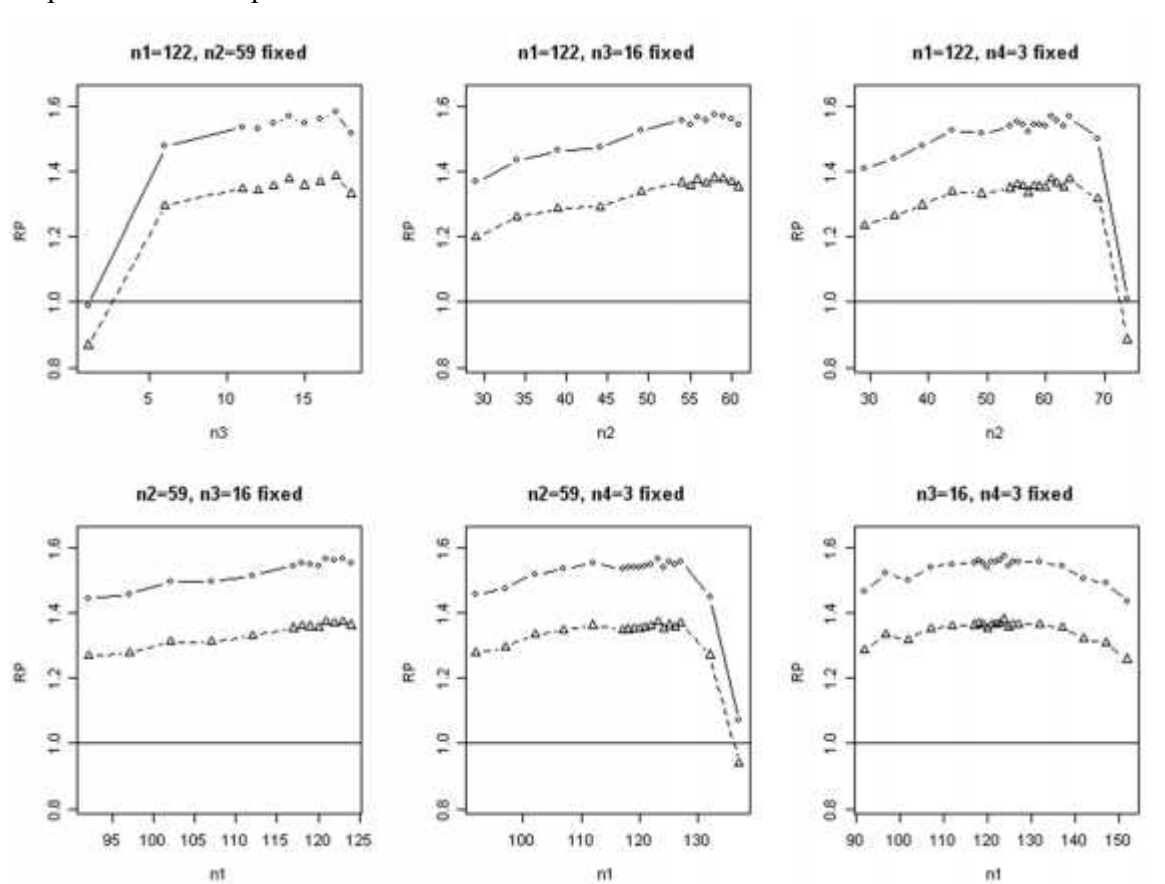
Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 61$ ,  $n_2 = 70$ ,  $n_3 = 45$ ,  $n_4 = 19$ , and  $n_5 = 5$ .  
 (○) denotes the relative precision of unbalanced RSS to SRS.  
 (Δ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 7:** Relative Precision for Females with age over 23,  $m = 3$ ,  $p = 0.898$ , with Departures from Optimal Allocation



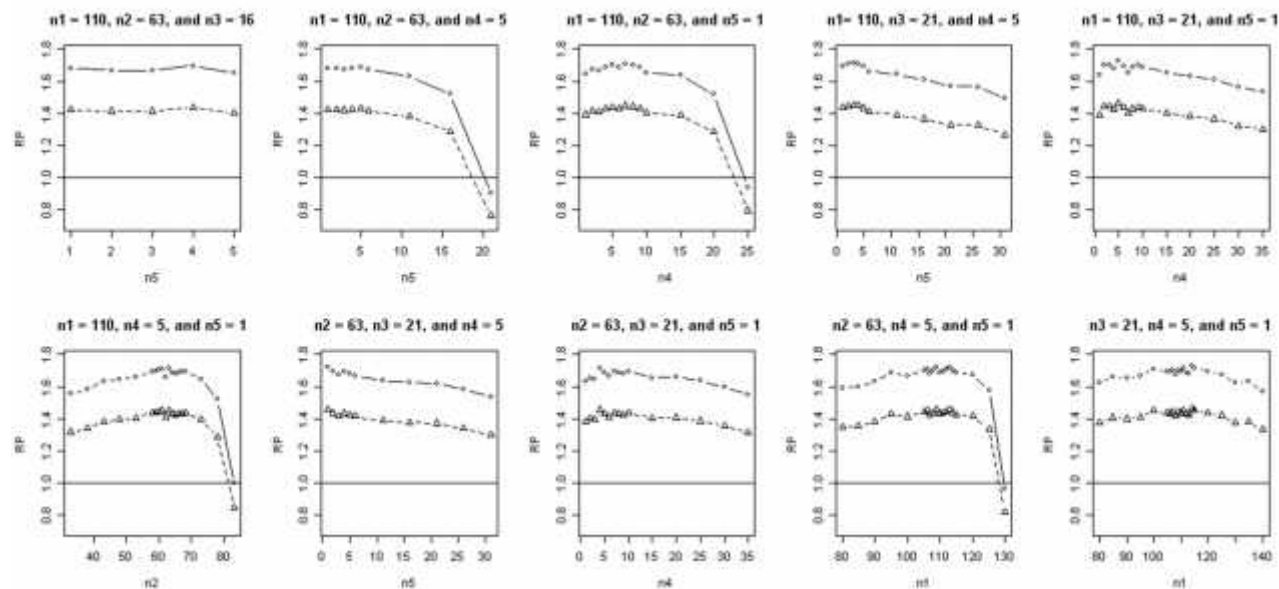
Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 138$ ,  $n_2 = 52$ ,  $n_3 = 10$ .  
 (○) denotes the relative precision of unbalanced RSS to SRS.  
 (△) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 8:** Relative Precision for Females with age over 23,  $m = 4$ ,  $p = 0.898$ , with Departures from Optimal Allocation



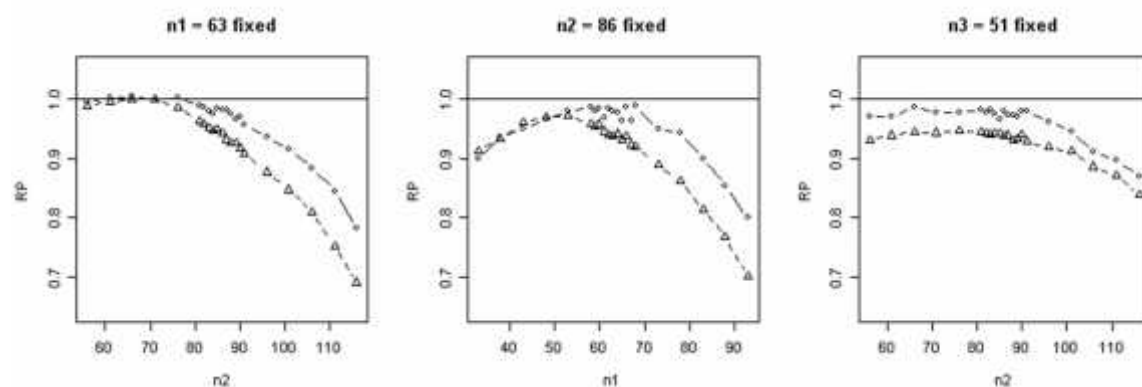
Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 122$ ,  $n_2 = 59$ ,  $n_3 = 16$ , and  $n_4 = 3$ .  
 (○) denotes the relative precision of unbalanced RSS to SRS.  
 (△) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 9:** Relative Precision for Females with age over 23,  $m = 5$ ,  $p = 0.898$ , with Departures from Optimal Allocation



Note: The ranking variable here is the continuous variable age. The optimal allocation is  $n_1 = 110$ ,  $n_2 = 63$ ,  $n_3 = 21$ ,  $n_4 = 5$ , and  $n_5 = 1$ .  
 (○) denotes the relative precision of unbalanced RSS to SRS.  
 (Δ) denotes the relative precision of unbalanced RSS to balanced RSS.

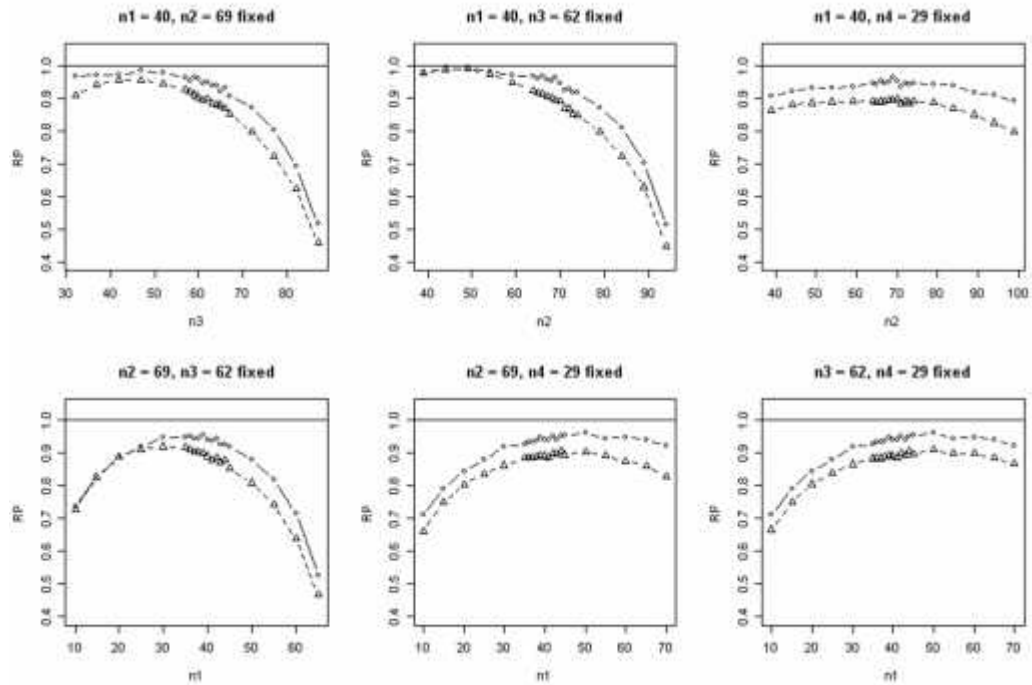
**Figure 10:** Relative Precision for Females Working,  $m = 3$ ,  $p = 0.544$ , with Departures from Optimal Allocation



Note: The ranking variable here is “public assistance”, which has correlation of 0.039 with the female working variable.  
 The optimal allocation assuming perfect rankings is  $n_1 = 63$ ,  $n_2 = 86$ , and  $n_3 = 51$ .  
 (○) denotes the relative precision of unbalanced RSS to SRS.  
 (Δ) denotes the relative precision of unbalanced RSS to balanced RSS.



**Figure 11:** Relative Precision for Females Working,  $m = 4$ ,  $p = 0.544$ , with Departures from Optimal Allocation



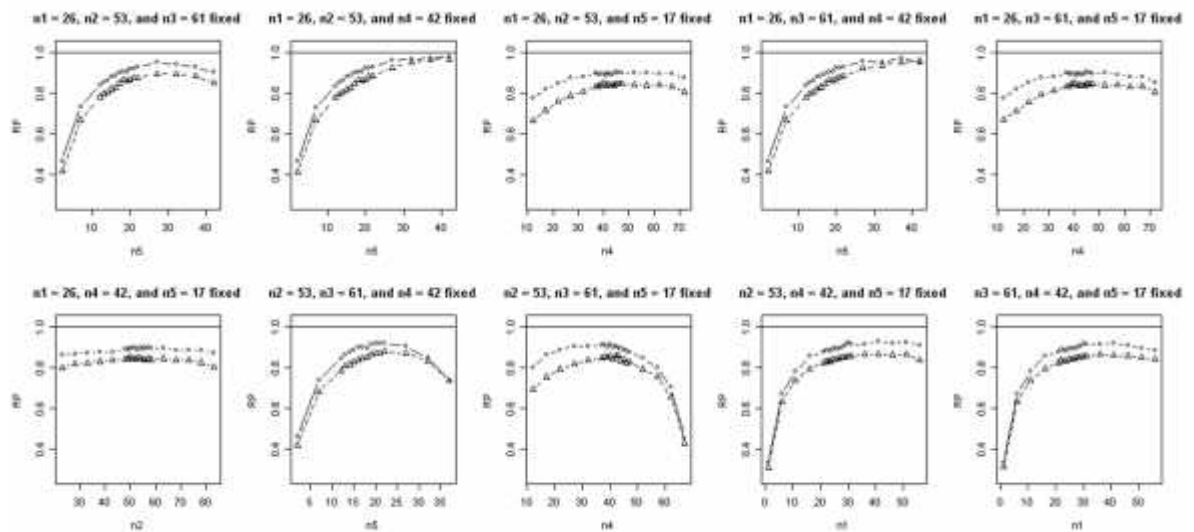
Note: The ranking variable here is “public assistance”, which has correlation of 0.039 with the female working variable.

The optimal allocation assuming perfect rankings is  $n_1 = 40$ ,  $n_2 = 69$ ,  $n_3 = 62$ , and  $n_4 = 29$ .

( $\circ$ ) denotes the relative precision of unbalanced RSS to SRS.

( $\Delta$ ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 12:** Relative Precision for Females Working,  $m = 5$ ,  $p = 0.544$ , with Departures from Optimal Allocation



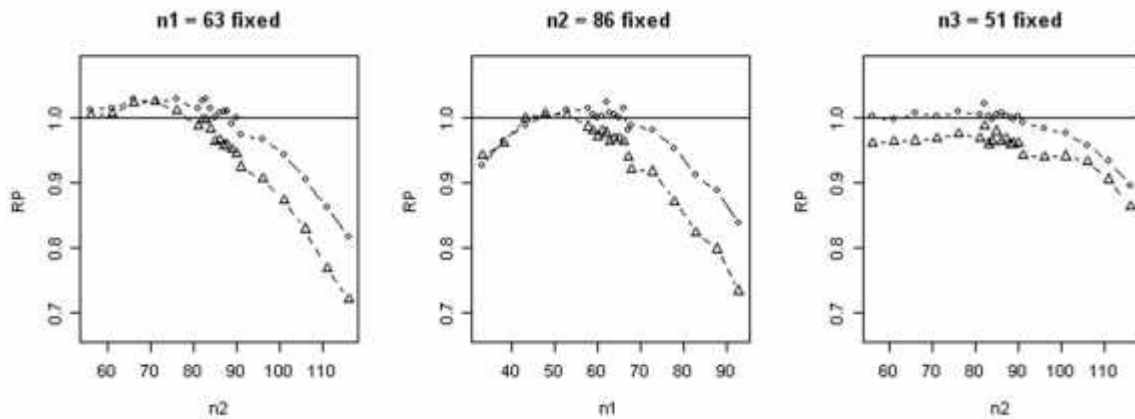
Note: The ranking variable here is “public assistance”, which has correlation of 0.039 with the female working variable.

The optimal allocation assuming perfect rankings is  $n_1 = 26$ ,  $n_2 = 53$ ,  $n_3 = 61$ ,  $n_4 = 42$ , and  $n_5 = 17$ .

( $\circ$ ) denotes the relative precision of unbalanced RSS to SRS.

( $\Delta$ ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 13:** Relative Precision for Females Working,  $m = 3$ ,  $p = 0.544$ , with Departures from Optimal Allocation



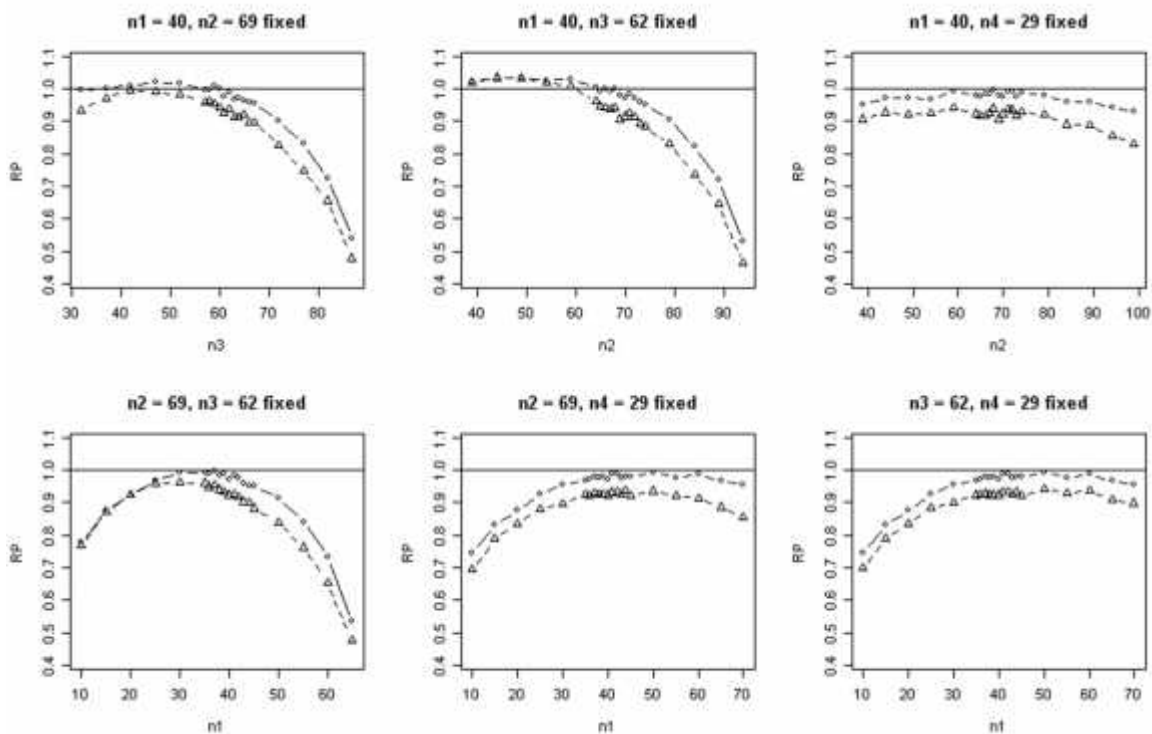
Note: The ranking variable here is age, which has correlation of 0.238 with the female working variable.

The optimal allocation assuming perfect rankings is  $n_1 = 63$ ,  $n_2 = 86$ , and  $n_3 = 51$ .

( $\circ$ ) denotes the relative precision of unbalanced RSS to SRS.

( $\Delta$ ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 14:** Relative Precision for Females Working,  $m = 4$ ,  $p = 0.544$ , with Departures from Optimal Allocation



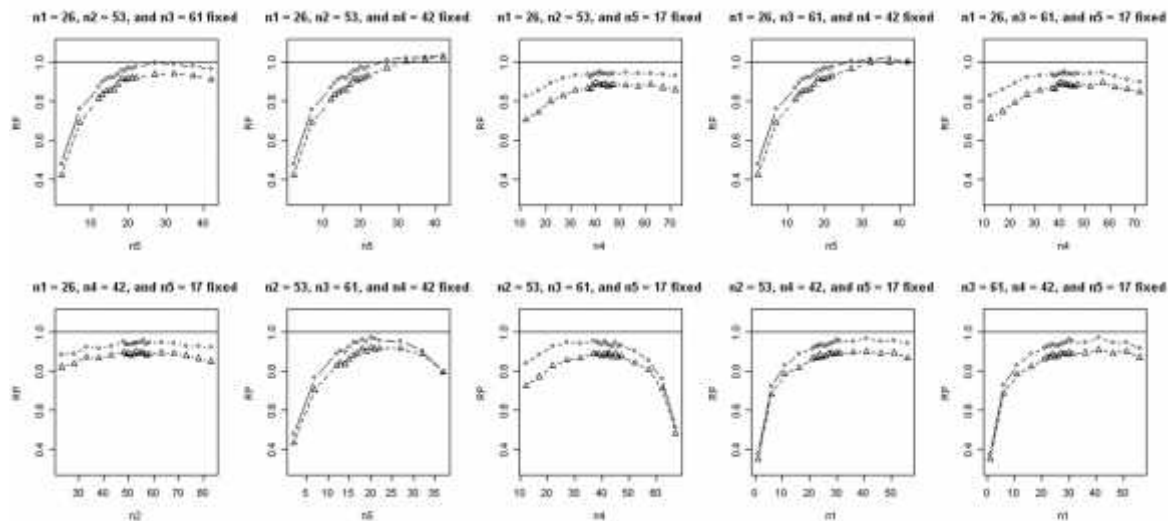
Note: The ranking variable here is age, which has correlation of 0.238 with the female working variable.

The optimal allocation assuming perfect rankings is  $n_1 = 40$ ,  $n_2 = 69$ ,  $n_3 = 62$ , and  $n_4 = 29$ .

( $\circ$ ) denotes the relative precision of unbalanced RSS to SRS.

( $\Delta$ ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 15:** Relative Precision for Females Working,  $m = 5$ ,  $p = 0.544$ , with Departures from Optimal Allocation



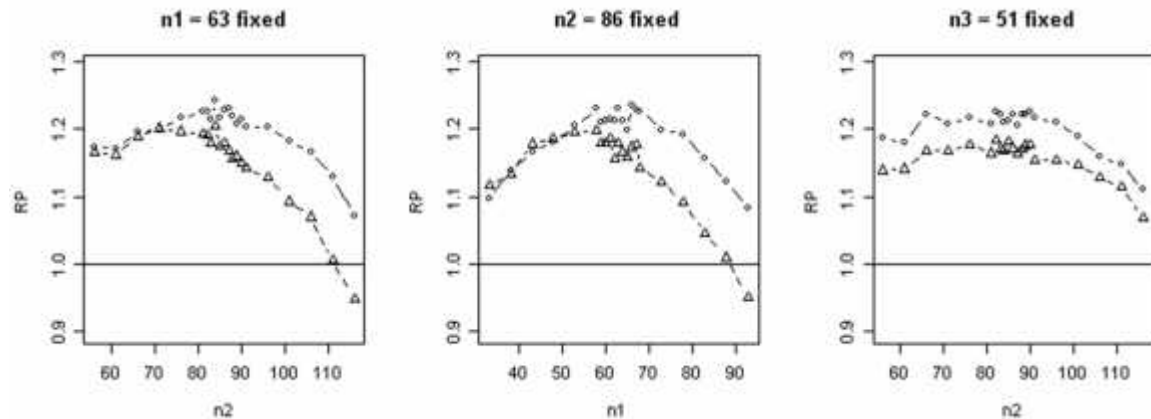
Note: The ranking variable here is age, which has correlation of 0.238 with the female working variable.

The optimal allocation assuming perfect rankings is  $n_1 = 26$ ,  $n_2 = 53$ ,  $n_3 = 61$ ,  $n_4 = 42$ , and  $n_5 = 17$ .

(o) denotes the relative precision of unbalanced RSS to SRS.

(Δ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 16:** Relative Precision for Females Working,  $m = 3$ ,  $p = 0.544$ , with Departures from Optimal Allocation



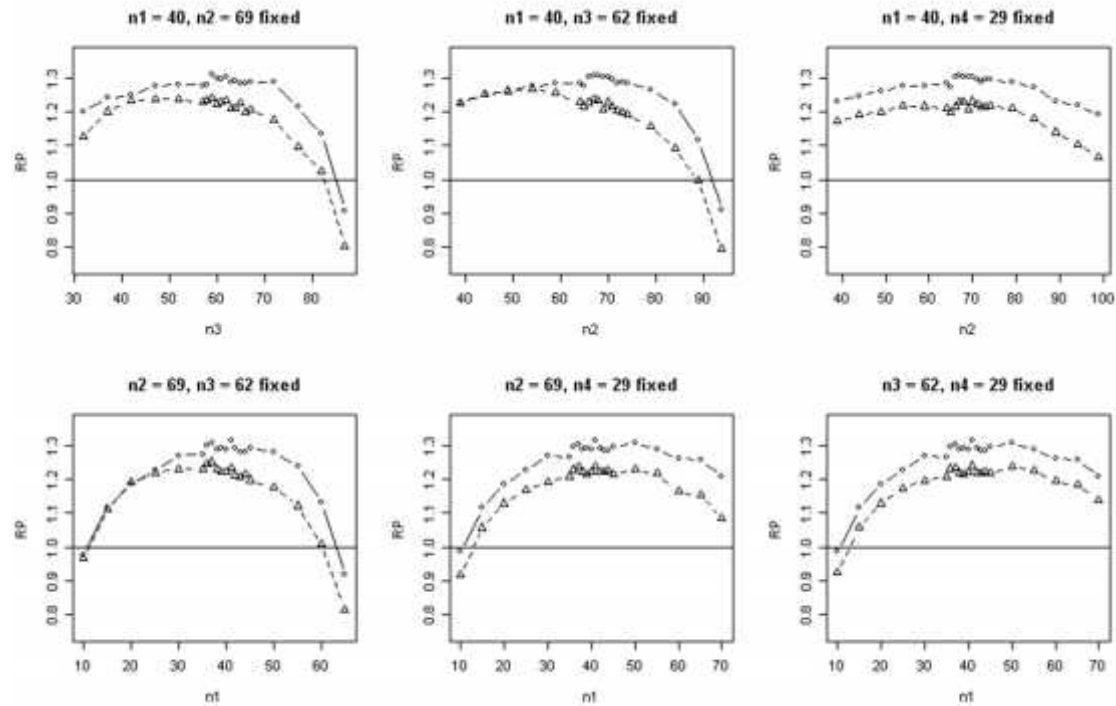
Note: The ranking variable here is "Hours Worked Last Week", which has correlation of 0.758 with the female working variable.

The optimal allocation assuming perfect rankings is  $n_1 = 63$ ,  $n_2 = 86$ , and  $n_3 = 51$ .

(o) denotes the relative precision of unbalanced RSS to SRS.

(Δ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 17:** Relative Precision for Females Working,  $m = 4$ ,  $p = 0.544$ , with Departures from Optimal Allocation



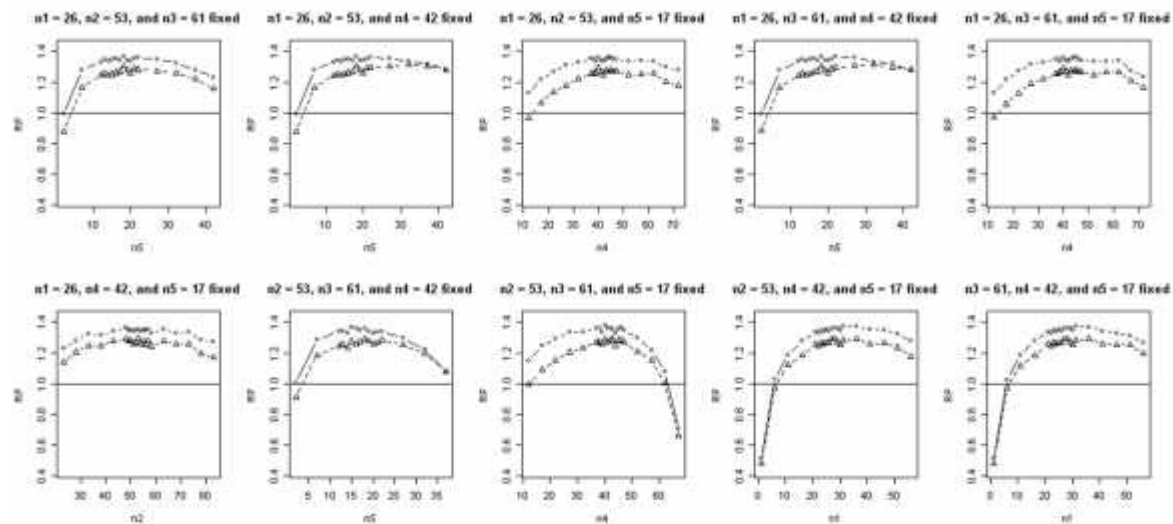
Note: The ranking variable here is “Hours Worked Last Week”, which has correlation of 0.758 with the female working variable.

The optimal allocation assuming perfect rankings is  $n_1 = 40$ ,  $n_2 = 69$ ,  $n_3 = 62$ , and  $n_4 = 29$ .

( $\circ$ ) denotes the relative precision of unbalanced RSS to SRS.

( $\Delta$ ) denotes the relative precision of unbalanced RSS to balanced RSS.

**Figure 18:** Relative Precision for Females Working,  $m = 5$ ,  $p = 0.544$ , with Departures from Optimal Allocation



Note: The ranking variable here is “Hours Worked Last Week”, which has correlation of 0.758 with the female working variable.

The optimal allocation assuming perfect rankings is  $n_1 = 26$ ,  $n_2 = 53$ ,  $n_3 = 61$ ,  $n_4 = 42$ , and  $n_5 = 17$ .

( $\circ$ ) denotes the relative precision of unbalanced RSS to SRS.

( $\Delta$ ) denotes the relative precision of unbalanced RSS to balanced RSS.