Ranked Set Sampling: Allocation of Sample Units to Each Judgment Order Statistic

Jessica K. Kohlschmidt, Elizabeth A. Stasny, and Douglas A. Wolfe
The Ohio State University

Abstract

Ranked set sampling is an alternative to simple random sampling that has been receiving considerable attention in the statistics literature. Researchers have shown that ranked set sampling outperforms simple random sampling in many situations by reducing the variance of a parameter 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 that are to be measured.

In this paper we conduct a sensitivity analysis of optimal allocation of sample units in unbalanced ranked set sampling to each of the order statistics. We use a simulation study to examine the sensitivity of the optimal allocation. 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 data collection that has been shown to improve on simple random sampling (SRS). RSS uses judgment ranking of a 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 the corresponding SRS estimators (see, for example, Patil, 1996). 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 paper, we consider the use of RSS for estimating a population proportion.

1.1 Balanced RSS

The most basic version of RSS in balanced RSS. In this form, 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 by judgment or through an auxiliary variable according to the characteristic of interest. We select the item with the smallest ranking, $X_{[1]}$, for measurement from the first set. 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^{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 ranked set sample 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^{th} judgment order statistic in the i^{th} cycle. The RSS estimator of the population mean μ is then the average of these RSS observations; that is,

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

This RSS estimator is an unbiased estimator of μ 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. 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.* (2005b) show that the amount of increase in the precision of the RSS 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 to its standard deviation. 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*. (2005b) discuss the general properties of unbalanced RSS and describe the Neyman allocation method for assigning sampling units to each judgment order statistic. We will use the following notation: let n_1 denote the number of observations allocated to the first order statistic, n_2 the number of observations allocated to the second order statistic, and so forth, until n_m denotes the m^{th} judgment order statistic. Then, for each r = 1,…, 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_i measurements of the i^{th} judgment order statistic, for i = 1,…, m. The total sample size 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 SRS estimator.

In this paper, we examine the sensitivity of unbalanced ranked set sampling to departures from optimal allocation when the goal is to estimate a population proportion. Neyman allocation is only optimal if the population proportion is known in advance of doing the allocation. In practice, of course, we only have an estimate of the population proportion based on a previous study or an informed guess. Thus, it is important to know how Neyman allocation performs if the allocation is not truly optimal. We will vary the sample sizes from the optimal sample sizes found using Neyman allocation and examine the effect this has on the standard deviation of the RSS estimator. We will also study the effect of imperfect rankings on this standard deviation.

For a discussion of the use of RSS with binary data, see Lacayo (2002) and Chen (2005a). Theoretical results for the sample mean apply immediately to this situation as a sample proportion is just the sample mean for binary data. To accomplish the within-set ranking we can use a logistic regression model to estimate probabilities of success and then use those estimates for ranking the sample items. Data that are readily available or easy to obtain from potential sampling units can be used in the logistic regression model for ranking. Chen *et al.* (2005a) have shown that using logistic regression in ranking improves the accuracy of the ranking process and therefore leads to considerable gains in precision of the RSS estimator over the SRS estimator. We will be using a single concomitant variable in the logistic regression for prediction, which has been researched independently by Chen *et al.* (2003) and Terpstra and Liudahl (2004).

Another issue in RSS is that of perfect versus imperfect rankings. In the case of perfect rankings the judgment order statistics equal the true order statistics. When rankings are perfect, we can express the probabilities of success for the various ordered items as functions only of the underlying population proportion. 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 Neyman allocation.

In Section 2, we discuss how we expect sample size allocation to effect the precision of the estimator. In Section 3, we describe the data set that is used in the simulation study. The simulation results are presented in Section 4.

2. Effect of Sample Size Allocation on the Precision of the Estimator

In this paper, we address the sensitivity of optimal allocation in unbalanced RSS. If we knew the population proportion, then we would be able to determine the exact optimal allocation of the total sample to the various judgment order statistics. In this situation, the optimal allocation produces the greatest precision in the RSS estimator. 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 to the optimal allocation our allocation needs to be to maintain a degree of precision similar 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 will examine the sensitivity of variances of RSS estimators to departures from the optimal allocations through a simulation study. This will be accomplished by first determining the Neyman allocation based on a known population proportion, p, using a set size of three. Then we will use this optimal allocation to simulate the sampling distribution of the RSS estimator of p. This will provide us with an estimate of the best possible improvement (over SRS) in precision from RSS. Then we will conduct similar simulations with allocations differing from the optimal Neyman allocation to assess the resulting effect on the precision of the RSS estimators. Plots of the estimated relative (to SRS) precisions of these various RSS estimators will be used to evaluate how robust Neyman allocation is to misspecification of p. For the purpose of this comparison, the estimated relative precision of the RSS estimator relative to the SRS estimator is just the standard error of the SRS estimator divided by the standard error of the RSS estimator.

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 field work was completed by the Institute for Survey Research at Temple University. The NSFH data were collected in two 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 as well as 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 recontacted between 1992 and 1994 for a follow-up interview. Responses to the second wave of the survey provide information on marriage transitions or union dissolutions since the first wave of the survey. The design permits the detailed description of past and current living arrangements and other characteristics and experiences, as well as the analysis of the possible consequences of earlier living arrangements on current states, marital and parenting relationships, kin contact, and economic and psychological well-being.

We will consider the observations collected by the NSFH as our population so that we know the population proportions exactly. This will permit us to determine the optimal Neyman allocation in a variety of situations.

4. Simulation Specifications

4.1 Perfect Rankings

We first consider the case of perfect rankings and set size m = 3. The total sample size of interest is 200 observations. We denote the sample size allocated to the i^{th} judgment order statistic as n_i, i = 1, 2, 3. For this setting, we take the NSFH females to be our population and consider the variable age. We construct three binary variables to provide data sets with three different population proportions. The category of age over 20 yields a population proportion of p = 0.994, 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.

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

p	$\%n_1$	$\%n_2$	$\%n_3$
0.575	34.5	42.5	23
0.743	49.5	37.5	13
0.964	81.5	16.5	2

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

Table 1 shows the optimal allocation of the sample for the three values of p. Notice that, for each different population proportion, there is one judgment order statistic that provides the most information if the measured observation is a success. This is the one to which we allocate the largest portion of the sample. As the population proportion p varies, this most informative judgment order statistic changes. If we hold the Neyman allocation fixed for the judgment order statistic allotted the most observations we expect that the RSS standard deviation will not be affected seriously by varying the sample sizes allotted to the other two judgment order statistics. On the other hand, we expect the RSS standard deviation will move more quickly away from its optimal level if we vary the number allocated to the group that is designated to have the most under Neyman allocation.

In the figures that appear in the Appendix, we plot $SD(\hat{p}_{SRS}) \big/ SD(\hat{p}_{RSS})$. The horizontal line that appears in some of the graphs corresponding to $SD(\hat{p}_{RSS}) = SD(\hat{p}_{SRS})$. Below this line, simple random sampling yields higher precision and above this line ranked set sampling is preferred as it has greater precision. For perfect rankings, Figure 1 shows plots of the relative precision of the RSS estimator to the SRS estimator. We provide such plots for the p's representing the proportions in the three age groups so that we can see if varying p has any effect on the robustness of the allocation. The general shape for the graphs of relative precision is a parabola. There is some variation in the maximum portion of the parabola, due to simulation error. For p = 0.575, it is evident that when holding 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 order statistics to vary by as much as plus or minus ten observations. Similarly for p = 0.575, it is clear that when holding n_2 fixed there is also some flexibility in allowing the allocation between the other two order statistics to change by as much as plus or minus five observations from the optimal allocation without losing precision for the estimator. Finally, when holding n_3 fixed for p = 0.575, we can allow the allocations to the other two order statistics to vary by as much as plus or minus ten observations from the optimal allocation without seriously affecting the precision. These results suggest that the sample allocated to each judgment order statistic does not have to be exactly at the optimal allocation for the precision of the estimator to remain close to optimal when p is near 0.5.

Next, we examine the situation when p = 0.743, in which case the first judgment order statistic has the largest allocation of the sample. Figure 2 shows the effect of changing the sample allocations on the relative precision of the RSS estimator for this setting. Holding n_1 fixed, we can vary the sample size allocation to the other two order statistics by five in either direction and still have nearly optimal relative precision. When examining the situation where n_2 is fixed, we see that we can vary from the optimal sample allocation by about seven observations in either direction without substantial reduction in precision. Lastly, fixing n_3, we can vary from the optimal allocation by plus or minus seven observations without much loss in precision.

Lastly, we consider the case where p = 0.964. The effect of changing the sample allocations on the relative precision of the RSS estimator in this setting is examined as well. Holding n_1 fixed, we can vary the sample size from the optimal allocation by about three in either direction and still have good relative precision. When examining the situation where n_2 is fixed, we see that we can vary from the optimal allocation by about five observations without substantial reduction in precision. Lastly, fixing n_3, we can vary from the optimal allocation by plus or minus five observations without much loss of precision.

An analysis was done with the males from the NSFH as the population and considered the variable age as well. This analysis yielded similar results to the females. As p gets closer to 0 or 1, it becomes more important that our sample allocations to the judgment order statistics are close to the optimal allocations in order not to lose much in relative precision. On the other hand, the allocations are not as critical if p is close to 0.5

4.2 Imperfect Rankings

Next, we address the case where we have imperfect rankings. In this setting, we use auxiliary variables to estimate the probability of success through logistic regression. We then use these estimates to rank the units for RSS. Here we look at the proportions of males and females working. For males, this proportion is 0.743 and for females it is 0.544. We use three different auxiliary variables to rank the units and select units for inclusion in our sample. For both males and females we use the variables, "public assistance", "age", and "hours worked last week". The variable named public assistance is an indicator of whether or not the respondent's family received public assistance when the respondent was a child. We expect that the gain in precision in the RSS estimator over the SRS estimator 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 correlations between working (a 0 and 1 variable) and the auxiliary variables mentioned above for each of the populations are as follows:

Males:

	Correlation with Working
Public Assistance	0.0352
Age	0.319
Number of Hours Worked	0.5969

Females:

	Correlation with Working
Public Assistance	0.0394
Age	0.2378
Number of Hours Worked	0.75796

We will again vary the sample allocations away from the optimal allocation to each judgment order statistic to see what effect there is on the standard error of the estimate of p.

In the case of imperfect rankings, we first examine the women's data set. Here the proportion of interest is p = 0.544, the proportion of women reporting that they are working, and the effect of sample allocations on the relative precision of the RSS estimator is displayed in Figures 3 and 4 . The first case that we consider is where the "public assistance" (correlation of 0.03938 with working) is used for rankings. As we consider again what happens when we hold the "optimal" sample allocations associated with each of the judgment order statistics fixed. Since this ranking variable is virtually uncorrelated with the variable of interest, we can see that the ranking does not help improve the precision of the estimator. In this case, we always do worse with RSS than with SRS. This is the danger with using unbalanced RSS. Only balanced RSS guarantees us that we will not do worse than SRS. This leads us to the conclusion that it is important to consider the correlation that the ranking variable has with the variable of interest in the unbalanced RSS situation.

Consider the case where "age" (correlation of 0.2378 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. Since the ranking variable is not highly correlated with the variable of interest, it is not a surprise that the precision of the unbalanced RSS estimator is less than the precision of the SRS estimator unless the sample allocations are nearly optimal. As we vary the sample allocations for the judgment order statistics, the SRS estimator quickly outperforms the unbalanced RSS estimator with non-optimal sample allocations.

Next consider the ranking variable "hours worked last week", which has a correlation of 0.7579 with the variable of interest. When we varied the sample allocations to each judgment order statistic for this setting, we see that there was still an improvement in precision with the unbalanced RSS estimator. We had to miss the optimal allocation by more than 20 in a cell for the precision of the RSS estimator to be worse than that of the SRS estimator.

The second data set that we considered contains only males. In this situation the proportion working is p = 0.743. The same patterns occur with this data set as with the females. Again the situation occurs where unbalanced RSS never outperforms SRS when the ranking variable and the variable of interest are virtually uncorrelated. When we have a variable that is not highly correlated with the working variable, then we do not have flexibility to deviate from the optimal allocation if we want the RSS estimator to have better precision than the SRS estimator. As we approach correlations of 0.5 or higher we can once again deviate from the optimal allocation and still achieve greater precision with the unbalanced RSS estimator than with the SRS estimator. In a few of the 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 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.

There is further analysis on the male data set, varying p's, and numerous correlations. For more detailed information and graphical representations of these results, see Kohlschmidt *et al.* (2005).

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 RSS estimator has greater precision than the SRS estimator. 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 RSS estimator being worse than the SRS estimator. As the 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 RSS estimator still has greater precision than the SRS estimator.

In this paper, we were concerned with how much flexibility we have in varying the sample allocation and still improving on the SRS estimator. We showed that in most cases, unbalanced RSS outperforms SRS even if we are not under optimal allocation of the sample units. The purpose was not to show whether we did better or worse than unbalanced RSS with optimal allocation. It is obvious that 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 SRS. Here we wanted to show that if there was uncertainty in the proportion p we could still improve upon the SRS estimator. This gives us the freedom to use RSS even in situations in which there is uncertainty about the likely value of p.

References

Bohn, L. L., 1996. A Review of Nonparametric Ranked Set Sampling Methodology. Communications in Statistics, Theory and Methods 25, 2675-2685.

Chen, H., 2005. Alternative Ranked Set Sampling Estimators for the Variance of a Sample Proportion. to appear, Journal of Applied Statistical Science.

Chen, H., Stasny, E. A., and Wolfe, D. A., 2003. Improved Estimation of Disease Prevalence Using Ranked Set Sampling. Technical Report Number 718, Department of Statistics, The Ohio State University.

Chen, H., Stasny, E. A., and Wolfe, D. A., 2005a. Ranked Set Sampling for Efficient Estimation of a Population Proportion, to appear in Statistics in Medicine.

Chen, H., Stasny, E. A., and Wolfe, D. A., 2005b. Unbalanced Ranked Set Sampling for Estimating a Population Proportion, to appear in Biometrics.

Dell, T. R. and Clutter, J. L., 1972. Ranked Set Sampling Theory with Order Statistics Background. Biometrics 28, 545-555.

Kohlschmidt, J.K., Stasny, E. A., and Wolfe, D. A., 2005. Ranked Set Sampling: Allocation Of Sample Units To Each Judgment Order Statistic
Technical Report Number 764, Department of Statistics, The Ohio State University.

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.

McIntyre, G. A., 1952. A Method for Unbiased Selective Sampling, Using Ranked Sets. Australian Journal of Agricultural Research 3, 385-390.

Patil, G. P., 1995. Editorial: Ranked set Sampling. Environmental and Ecological Statistics 2, 271-285.

Patil, G. P., 2002. Ranked Set Sampling. Encyclopedia of Environmetrics 3, 1684-1690.

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.

Terpstra, J.T. and Liudahl, L.A., 2004. Concomitant-based Rank Set Sampling Proportion Estimates. Statistics in Medicine 23, 2061-2070.

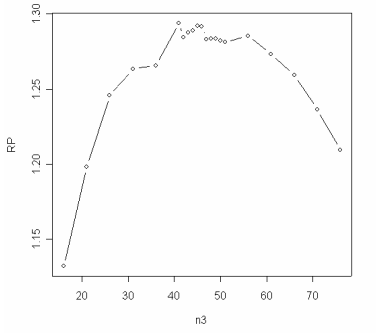
Wolfe, D. A., 2004. Ranked Set Sampling: An Approach to More Efficient Data Collection. Statistical Science 4, 636-643.

Appendix of Figures

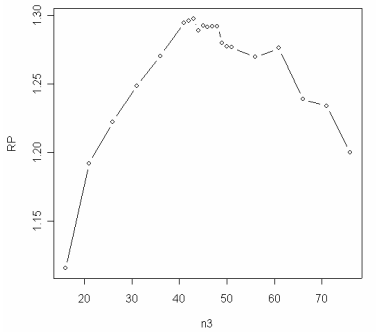
Figure 1: Relative Precision for Females with Age > 35, p = 0.575

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

a) $n_1 = 69$ fixed



b) $n_2 = 85$ fixed



c) $n_3 = 46$ fixed

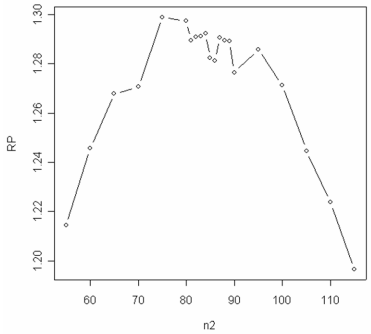
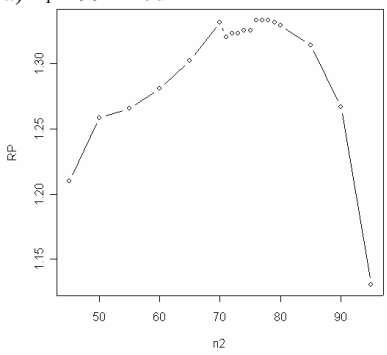


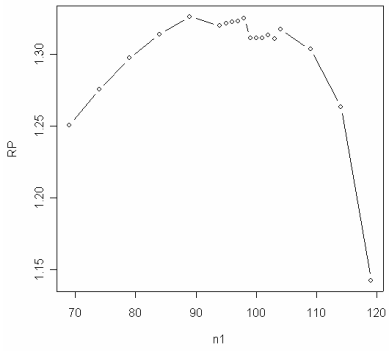
Figure 2: Relative Precision for Females with Age > 30, p = 0.743

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

a) $n_1 = 99$ fixed



b) $n_2 = 75$ fixed



c) $n_3 = 26$ fixed

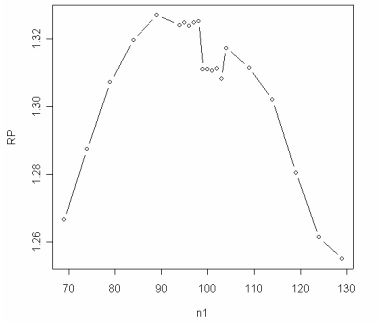
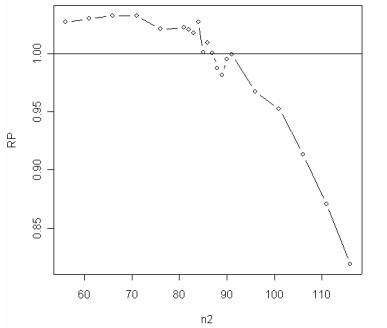


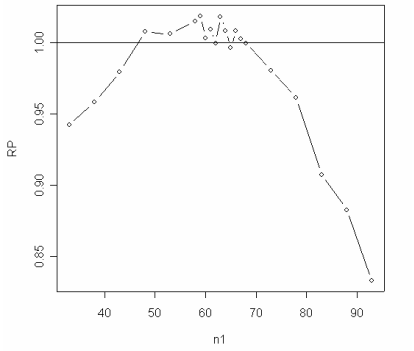
Figure 3: Relative Precision for Females Working, p = 0.544

Note: The ranking variable here is the probability of working from logistic regression on age. The optimal allocation assuming perfect rankings is $n_1 = 63$, $n_2 = 86$, and $n_3 = 51$.

a) $n_1 = 63$ fixed



b) $n_2 = 86$ fixed



c) $n_3 = 51$ fixed

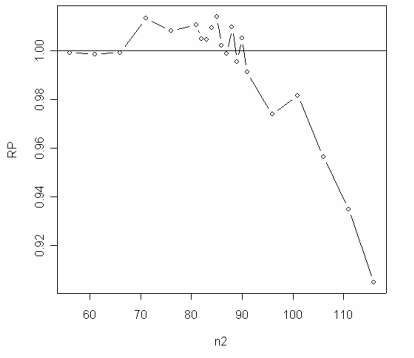
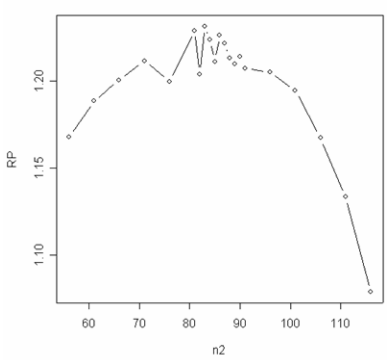


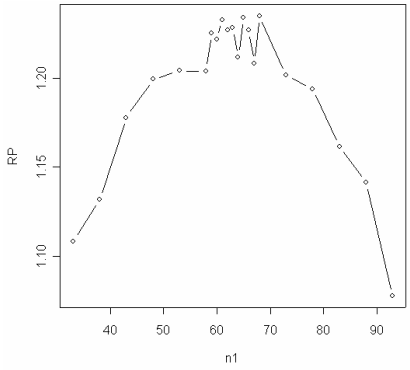
Figure 4: Relative Precision for Females Working, p = 0.544

Note: The ranking variable here is the variable "Number of Hours Worked". The optimal allocation assuming perfect rankings is $n_1 = 63$, $n_2 = 86$, and $n_3 = 51$.

a) $n_1 = 63$ fixed



b) $n_2 = 86$ fixed



c) $n_3 = 51$ fixed