COMPOSITE STRATUM VARIANCE ESTIMATION FOR AGRICULTURAL AREA FRAME SAMPLE ALLOCATION

Raj S. Chhikara[1], University of Houston-Clear Lake
Floyd M. Spears[1,3], Harding University
Charles R. Perry[4], USDA-NASS
Raj Chhikara, University of Houston-Clear Lake, 2700 Bay Area Blvd, Houston, TX 77058.

KEY WORDS: Variance Function, Weighted Model-Fit, Composite Estimation, NASS Sample Allocation

Abstract

In a previous study (JSM 2004) we proposed a method of improving upon the area frame sample allocation that NASS/USDA uses for the purpose of conducting annual agriculture surveys. It first entails modeling of stratum variance as a function of stratum agriculture item totals, and then estimating each stratum variance by a linear combination of the direct survey variance estimate and a model predicted variance estimate. The approach is shown to have superior properties for sample allocation compared to the current one used by NASS. Since the NASS sample allocation is derived from the past three or more years of survey data, the proposed approach is evaluated using three years data; hence, it parallels the current approach NASS uses for development of its sample allocation.

1 Introduction

A stratified design is employed in the NASS agricultural area frame surveys. A multivariate procedure is the basis of NASS area frame sample allocation. It requires input values for eight agricultural items, of which six are major crops and two are non-crop items. The six crops are corn, cotton, soybean, durum wheat, spring wheat and winter wheat. The two non-crop items are the number of farms and the number of cattle that are not covered by the list frames, designated as the not-on-list (NOL) cattle.

Determination of sample size and its allocation to different strata requires input of stratum variances for each of the eight agricultural items. NASS utilizes stratum variances determined from 3-5 years historical variances computed from survey data. However, sample survey variances can be unstable when a stratum has a small number of samples or when the percent crops or agricultural items are small. So the major objective of this study is to obtain more stable estimates of stratum variances for the eight items for their use in the multivariate sample allocation. An overall objective is to develop an alternative approach to utilizing NASS survey data in determination of sample allocation for the area frames sampling.

In the present study, six crop items and two non-crops items are considered for modeling of their stratum variances. Empirical model-fits for stratum variance (for each crop) and stratum standard deviation (for each non crop item) were obtained using the 2001, 2002, and 2003 survey estimates. A weighted model-fit is determined to be appropriate where the weight function is dependent upon the proportion of an agricultural item and the sample size used in estimating the historical variance for the stratum. The model fit details are described in Section 3.

For a stratum , the model predicted variance is combined with the survey estimated variance to obtain a composite variance estimate for each item. The multivariate allocation procedure is applied to obtain the new sample allocations in 2001, 2002 and 2003 using (i) model predicted standard deviations, (ii) survey estimated standard deviations, and (iii) composite standard deviations for all strata in the area frame. The three sample allocations are contrasted against the actual NASS survey design allocation in each of three years. Next, a sample allocation for 2004 is determined based on the three year stratum standard deviations as well as by combining the three year sample allocations. It is seen that

[1] These authors' work was supported in part under a cooperative research program of the United States Department of Agriculture at the University of Houston-Clear Lake.

[2] This author is with the University of Houston-Clear Lake, 2700 Bay Area Blvd, Houston, TX 77058.

[3] This author is with Harding University, Box 10764, Searcy, AR 72149.

[4] This author is with the National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030

the NASS sample allocation for 2004 is closer to the average of the sample allocations for the previous years (2001-2003) than that using the average of the three year stratum standard deviations.

2 Data for Modeling

The 2001, 2002, and 2003 survey data were used in developing the modeling data of stratum estimates and variance estimates for the agriculture items of six crops and two non-crops as listed in Section 1. Both the estimate and the variance estimate for an item are computed for each stratum.

NASS land use strata are grouped together by considering similarity in their major land use, depending upon the agricultural intensity or the location. Table 1 lists the characteristics of the five groups of land use strata, the area level at which the model-fits are made.

Table 1: Stratum Groups

Stratum Group	Land-use Strata	Description
1	10 - 19	Intense Cultivation
2	20 - 29	Moderate Cultivation
3	30 - 39	Urban Area
4	40 - 49	Low Cultivation
5	50 - 59	Non-Agricultural

Survey data were standardized on per acre basis. This was to eliminate the effect of segment size which varied from one group of land use strata to another group. This amounted to using the proportion of an ag item per acre in a stratum and the corresponding survey variance estimate. Furthermore, the direct survey estimates for each item was replaced by those determined using the Agriculture Statistics Board (ASB) estimates. This was done to have a covariate value obtained independent of the survey data used in stratum variance estimations. Since there are no Board estimates at the stratum level, state estimates were proportioned to strata within the state. Table 2 lists the agricultural items used for proportioning of ASB estimates for the nine items considered here. For example, the proportioning for all crop items was done based on the cultivated land in each stratum relative to the state cultivated land.

For land use stratum h, let x_h denote the ASB proportioned estimate of an item and s_h^2 the variance estimate computed from the survey data. The estimated crop proportion on a per acre basis in stratum h is

$$p_h = \left(\frac{c_h}{c_s} \cdot B_s \right) \ / \ A_h \qquad (1)$$

where c_h is the cultivated land in stratum h, c_s is the cultivated land in the state, B_s is the ASB estimate of crop acreage total in the state and A_h is the total acreage in stratum h. The corresponding variance is

$$s_h^2 = \frac{n_h \cdot s_{d,h}^2}{N_h \cdot a_h^2} \qquad (2)$$

where N_h is the total number of area frame segments in the stratum, n_h is the number of sample segments, $s_{d,h}^2$ is the direct survey estimated stratum variance for the crop acreage total and a_h is the segment size for stratum h. Here s_h^2 represents the estimate of the population variance in the stratum. For details on the computation of survey variance estimates $s_{d,h}^2$, refer to Kott (1990). Thus the data for modeling of stratum variance or standard deviation consists of (p_h,s_h^2) for stratum h.

Table 2: Proportioning Factors

Item Estimated	Proportioning Factor
Crop	Cultivated Land
Number of Farms	Land in Farms
NOL Cattle	Crop Land
Equine	Number of Farms

3 Modeling of Stratum Variance

3.1 Case of Crop Acreage

If the measurement unit is same as the sampling unit and it either is completely covered by the crop of interest or it is not, then sampling of a unit randomly in a stratum amounts to a Bernoulli trial. If p is the proportion of units being covered by having the crop of interest, then the stratum variance at the unit level is given by

$$\sigma^2 = p \cdot (1 - p).$$

Since the measurements are at the aggregate level of an area segment, the stratum variance can be expected to be less than $p(1-p)$ due to positive intra-cluster correlation. One can postulate it to be

$$\sigma^2 = \alpha \cdot y^\gamma$$

where $y = p(1-p)$ and $\gamma > 1$. .

As explained in Mahalanobis (1946), the between area segment variance for a crop acreage can be expected to be a power function of the variance under a binomial model. The power would depend upon the intra-class correlation of the measurement units devoted to the crop in area segment and may be empirically determined using the survey estimates. This formulation was the basis of an empirical modeling of stratum variance by Mahalanobis (1968). The following model for the survey estimated stratum variance is considered in the present study:

$$s^2 = \alpha y^\gamma + \epsilon \tag{3}$$

where s^2 is an estimate of stratum variance, y is as defined above and ϵ is the random error component. Initially a non-linear model-fit was made using the modeling data. In most model-fits, the estimated value of α did not differ significantly from 1. This lead to the assumption that $\alpha = 1$ in the model defined in Equation (3), and a linear model-fit was deemed appropriate. As such, the model considered in the case of crop acreage is

$$s^2 = \beta y + \epsilon. \tag{4}$$

3.2 Case of Non-Crop Items

A counting process seems applicable for the occurrence of the number of farms, cattle, or equines in an area segment. However, when the survey estimates of these items were examined, it was found that a fewer number of them occur much more frequently than do a higher number of them in a segment. It suggested that an exponential distribution can be assumed for the underlying probability distribution for each item. If so, the stratum standard deviation is equal to the stratum mean. Thus the following model is considered for the estimate of stratum standard deviation:

$$s = \beta p + \epsilon \tag{5}$$

where p is the item proportion estimate and s is the estimate of standard deviation in a stratum and ϵ is the random error component.

3.3 Modeling Error Variance

The model error was investigated to account for the error variance heterogeneity, reliability of survey variance estimates and outliers in the data. Scatter plots of modeling data, (p_h, s_h^2) in the case of crops and (p_h, s_h) in the case of non-crop items, were made showing each point by a bubble with size proportional to sample size used in computing the s_h^2. Since $0 < p_h < 0.5$ for any single crop item, the scatter plots can be done using p_h even though y_h is the regressor. This allowed us to examine the points in scatter plot according to their relative precision and hence, their importance. Since the s_h^2 (or s_h) are computed from sample survey data, these are not equally reliably estimated. Because a larger sample size leads to higher precision for s_h^2 (or s_h), its weight is considered proportional to n_h. This led us to assign weight to each data point based on the associated sample size in developing a model-fit.

For crops, scatter plots of s_h^2 vs. p_h showed that on the average the s_h^2 value increases as p_h increases, and so does the spread in the s_h^2 values. This implies the variance of s_h^2 increases as a function of p_h. This requires carrying out a weighted linear model-fit for s_h^2 where the model residuals are weighted by an inverse power of p_h. An estimation of variance of s_h^2 and hence, determination of its associated weight in model-fit is described in details in Chhikara, et al (2005).

In order to deal with potential outliers in data, Tukey's biweight procedure was used to assign lower weights to extreme observations. See Fox (2002) for its full description.

4 Model Fits

That data was fit using the model in Equation (4) for crop items and the model in Equation (5) for non-crop items. Scatter plots made for stratum variance s_h^2 in the case of crops, and standard deviation s_h in the case of non-crops, vs. covariate (p_h) for each item showed that the ranges in s_h^2 (as well as in p_h) differ considerably across the five groups of strata and thus separate model-fits were considered for the five stratum groups defined earlier in Table 1.

For each group stratum, an estimate of β, denoted as $\hat{\beta}$, was determined by minimizing the following quantities:

$$\sum_{h=1}^{H} (s_h^2 - \beta y_h)^2 w_h$$

in the case of crop acreage, and

$$\sum_{h=1}^{H} (s_h - \beta p_h)^2 w_h$$

in the case of a non-crop item. Here H represents the number of land use strata in a stratum group and y_h (or p_h) represents the covariate value as defined earlier. Separate model-fits were carried out

Table 3: Model fits of stratum variance for crop items and stratum standard deviation for non-crop items

	Strata	2001		2002		2003	
Item	Group	Obs	$\hat{\beta}$	Obs	$\hat{\beta}$	Obs	$\hat{\beta}$
Corn	1	61	0.124	61	0.159	61	0.159
Corn	2	54	0.100	50	0.080	50	0.080
Corn	3	100	0.225	100	0.227	100	0.227
Corn	4	46	0.084	46	0.099	46	0.099
Cotton	1	21	0.345	22	0.336	22	0.336
Cotton	2	24	0.100	25	0.122	25	0.122
Cotton	3	35	0.650	35	0.311	35	0.311
Soybeans	1	42	0.137	38	0.139	38	0.139
Soybeans	2	37	0.124	35	0.122	35	0.122
Soybeans	3	69	0.518	69	0.186	69	0.186
Soybeans	4	34	0.155	34	0.130	34	0.130
Durum Wheat	1	39	0.086	39	0.127	39	0.127
Spring Wheat	1	13	0.248	14	0.160	14	0.160
Spring Wheat	2	15	0.120	16	0.146	16	0.160
Spring Wheat	3	29	0.964	29	0.975	29	0.975
Spring Wheat	4	9	0.575	8	0.014	8	0.014
Winter Wheat	1	52	0.216	49	0.260	48	0.260
Winter Wheat	2	51	0.183	54	0.163	54	0.163
Winter Wheat	3	93	0.177	92	1.041	92	1.041
Winter Wheat	4	40	0.013	41	0.018	41	0.018
Farms	1	37	0.587	36	0.606	35	0.612
Farms	2	29	0.891	38	0.808	36	0.827
Farms	3	83	0.566	84	1.062	87	0.952
Farms	4	28	1.162	30	1.261	35	1.261
NOL Cattle	1	51	0.037	57	0.053	53	0.026
NOL Cattle	2	37	0.364	34	0.213	30	0.078
NOL Cattle	3	96	0.060	96	0.091	96	0.444
NOL Cattle	4	27	0.523	28	0.616	28	0.939

for various agricultural items using the 2001, 2002 and 2003 modeling data as outlined earlier.

Table 3 lists the estimates of β, listed as $\hat{\beta}$ in the tables, for model fits for variance for the crop items and standard deviation for the non-crop items. $\hat{\beta}$ has values less than 1 except in a couple of cases. Generally speaking, $\hat{\beta} \leq 0.30$ for all items other than the number of farms. This in turn implies a much smaller variances between area segments than obtained by considering the binomial model for crop acreage. Of course, it is expected due to a strong intracluster correlation for crop acreage within sample area segments.

A weighted residual plot was examined for any lack of fit or anomaly in each model fit. The model-fits were viewed to be good for stratum groups 1 and 2, but less so for stratum group 3 and 4. Since stratum groups 1 and 2 account for a substantially large amount of value for an item, the model-fits were judged to be useful for predicting stratum variances.

5 Stratum Variance Estimation

5.1 Composite Variance Estimate

The model predicted variances are expected to be most useful for strata when their survey estimated variances are unreliable due to a small number of sample segments used in computations. Otherwise, the survey estimated stratum variances are reliable and should be used in the estimation of stratum variances. Instead of making a choice between the two variance estimates, it is better to combine the two on the basis of an optimal criterion. This leads us to develop a composite estimate of stratum variance as a linear combination of its survey estimated variance S_D^2 and its model predicted variance S_M^2 given by:

$$S^2 = \alpha S_D^2 + (1 - \alpha) S_M^2.$$

Historically, a common practice among sample survey organizations is the strategy of using S_D^2 if it is based on a sample size of 30 or more, in which case $\alpha = 1$. When the sample size is less than 30, α is taken to be the proportion of sample size relative to 30. Since in the present case, S_D^2 has *df* degrees of freedom associated with it, this convention or approach leads to α defined to be:

$$\alpha = \begin{cases} 1 & \text{if } df \geq 30 \\ df/30 & \text{if } df < 30 \end{cases} \tag{6}$$

Here we use degrees of freedom instead of sample size since S_D^2 is a pooled variance for the stratum and so its reliability depends upon the degrees of freedom for the within sum of squares computed across substrata in the stratum.

5.2 Stratum Variance Comparisons

Composite variances for all strata were computed for each item using α from Equation 6. These are compared to the corresponding survey estimated variances, which showed that the composite estimates are in good agreement with the survey estimated variances except in the case of stratum group 3 and 4. In the later case, the composite estimates are mostly driven by the model-predicted variances. Overall, the model-predicted variances seem to be resistant to being outliers as compared to the survey estimated variances. However, the composite variances are derived more by the survey estimated variances than the model-predicted variances, and yet these are more robust than the survey estimated variances.

6 Sample Allocations

6.1 Prior Year 2001, 2002 and 2003

Sample allocations were performed using the NASS constraint inputs for 2001, 2002 and 2003 in the multivariate allocation procedure. Stratum standard deviations obtained for the eight agricultural items (six crop and 2 non-crop items) were used in carrying out the multivariate allocations. For comparisons purposes, we considered the use of three different stratum variances to determine the sample size and its allocation to land use strata:

1. Direct Survey Variance Estimate, S_D^2
2. Model-Fit Predicted Variance, S_M^2
3. Composite Variance Estimate, $S^2 = \alpha S_D^2 + (1 - \alpha) S_M^2$

Table 4 lists the total sample sizes for the three years 2001, 2002 and 2003 with allocations across the five stratum groups. Included here are also the actual design allocations used by NASS.

The use of model-predicted variances leads to the smallest sample size in stratum group1. It also tends to have a smaller total sample size, expect in 2003, than the other two variance cases. Moreover, the sample allocations using the model-predicted variances compares least favorably with the actual design allocation. The use of composite variances produces the sample allocation most comparable with the actual design allocation. As expected, the composite variance case has slightly more samples allocated in marginal cases than does the survey estimated variance case.

6.2 Sample Allocations for 2004

The NASS sample design allocation for a given year is based on the use of the stratum variances estimated from its survey data from the previous year and the historical sample allocations over the preceding few years. Any changes to the sample allocation based on the most recent survey data is predicate on the overall survey cost, trends reflected in the historical sample allocation, and revisions or updates in the area frame in one or more states. Thus the NASS sample design allocation technically is a product of 3-5 years of historical sample data, although there is no straightforward quantitative approach to determining it. It obviously raises a question, whether or not a quantitative approach can be used to combine the NASS historical sample data to determine its sample allocation for the following year. We currently explore this and use the NASS sample data from years 2001, 2002 and 2003 to determine a new sample allocation for year 2004.

Two approaches are considered for combining the three year data. The first approach is to obtain the stratum standard deviations by taking the mean or median of the stratum standard deviations obtained

Table 4: Sample Allocation by Stratum Group for (a) 2001, (b) 2002, and (c) 2003.

(a) 2001

	Actual	Allocation using Standard Deviations from		
Stratum Group	Design	Survey Estimated	Model Predicted	Composite
1	6037	6117	4430	6170
2	2605	2928	2291	2816
3	368	365	256	287
4	1523	2329	1940	2192
5	105	96	96	96
Total	10638	11835	9013	11561

(b) 2002

	Actual	Allocation using Standard Deviations from		
Stratum Group	Design	Survey Estimated	Model Predicted	Composite
1	6109	6268	5739	6466
2	2786	2400	2222	2496
3	376	285	273	271
4	1699	927	1364	1034
5	105	96	96	96
Total	11075	9976	9694	10363

(c) 2003

	Actual	Allocation using Standard Deviations from		
Stratum Group	Design	Survey Estimated	Model Predicted	Composite
1	4899	6919	6769	7088
2	2202	2521	2623	2621
3	351	266	326	276
4	1422	1137	1457	1319
5	104	96	96	96
Total	8978	10939	11271	11400

in survey years 2001, 2002 and 2003. The resulting sample allocations for the mean and median were very similar.

The second approach is to combine directly the sample allocations obtained for survey years 2001, 2002, and 2003. This would be pretty much in line with the NASS approach as described earlier. Considering the average of these three year sample allocations by stratum, the new sample allocations obtained for 2004 are as listed in Table 5.

In the three way comparisons between the survey estimated, actual design and composite cases, the results are similar to those obtained using the average of stratum standard deviations. However, almost a complete match in overall sample size is seen between the 2004 actual design and the composite estimated case. This seems to show that the NASS design allocation in a year is equivalent to taking the average of the historical sample allocations determined using the sample survey variances for the land use strata.

Figure 1 depicts a comparison of four different sample allocations by stratum group. The four allocations are referred to as: 2004 Design, Average of Standard Deviations, Median of Standard Deviations and Average 2001-2003. The last three sample allocations are those obtained using the composite variances. It shows that the 2004 Design allocation is mostly in agreement with the average 2001-2003 allocation with a slight exception of stratum group 1. Here again, it confirms that the NASS sample allocation can be mimicked by taking the average of historical allocation.

Since the samples across three survey years are not independent due to the use of a 5-year rotation design, the reliability of any of the two new allocations for 2004 using average or median standard deviations as described above may not be as same

Table 5: Average of Sample Allocations from 2001-03

		Allocation using Standard Deviations from		
Stratum Group	2004 Design	Survey Estimated	Model Predicted	Composite
1	6132	6431	5644	6571
2	2811	2615	2379	2644
3	371	296	285	276
4	1696	1453	1574	1502
5	107	96	96	96
Total	11117	10891	9978	11089

as one would expect had the samples been independently drawn from one year to another. Thus it may not be any more advantage to use the average or median standard deviations for strata in the multivariate allocations procedure than having simply to obtain the average of the 2001-2003 allocations. The average allocation obtained using the proposed composite variance estimates appear to be the most robust.

References

[1] Allen, Don (1999), "1997 Census Not on Mail List Survey". U.S. Department of Agriculture, National Agricultural Statistics Service Report (unpublished).

[2] Bethel, James (1989), "Sample Allocation in Multivariate Surveys," *Survey Methodology*, **15:1**, 47-57.

[3] Chhikara, Raj S. and Perry, Charles R. (1986), "Estimation of Stratum Variance in Planning of Crop Acreage Surveys"' *Journal of Statistical Planning and Inference*, **15**, 97-114.

[4] Chhikara, Raj S., Spears, Floyd M. and Perry, Charles R. (2002), "Sample Allocation for Estimation of the Number of "Not on Mail List" (NML) Farms for the 2002 Census of Agriculture," USDA-NASS, RDD Research Report Number RDD-02-01, January, 2002.

[5] Cramér, Harald (1946). *Mathematical Methods of Statistics*, Wiley, New York.

[6] Fox, John (2002), *Robust Regression,* Appendix to "An R and S-PLUS Companion to Applied Regression." January, 2002.

[7] Fuller, Wayne A. (1987), *Measurement Error Models,* Wiley, New York.

[8] Mahalonobis, P.C. (1946), "On Large Scale Sample Surveys," *Philosophical Transactions of Royal Society, Landon, Series B*, **231**, 329-451.

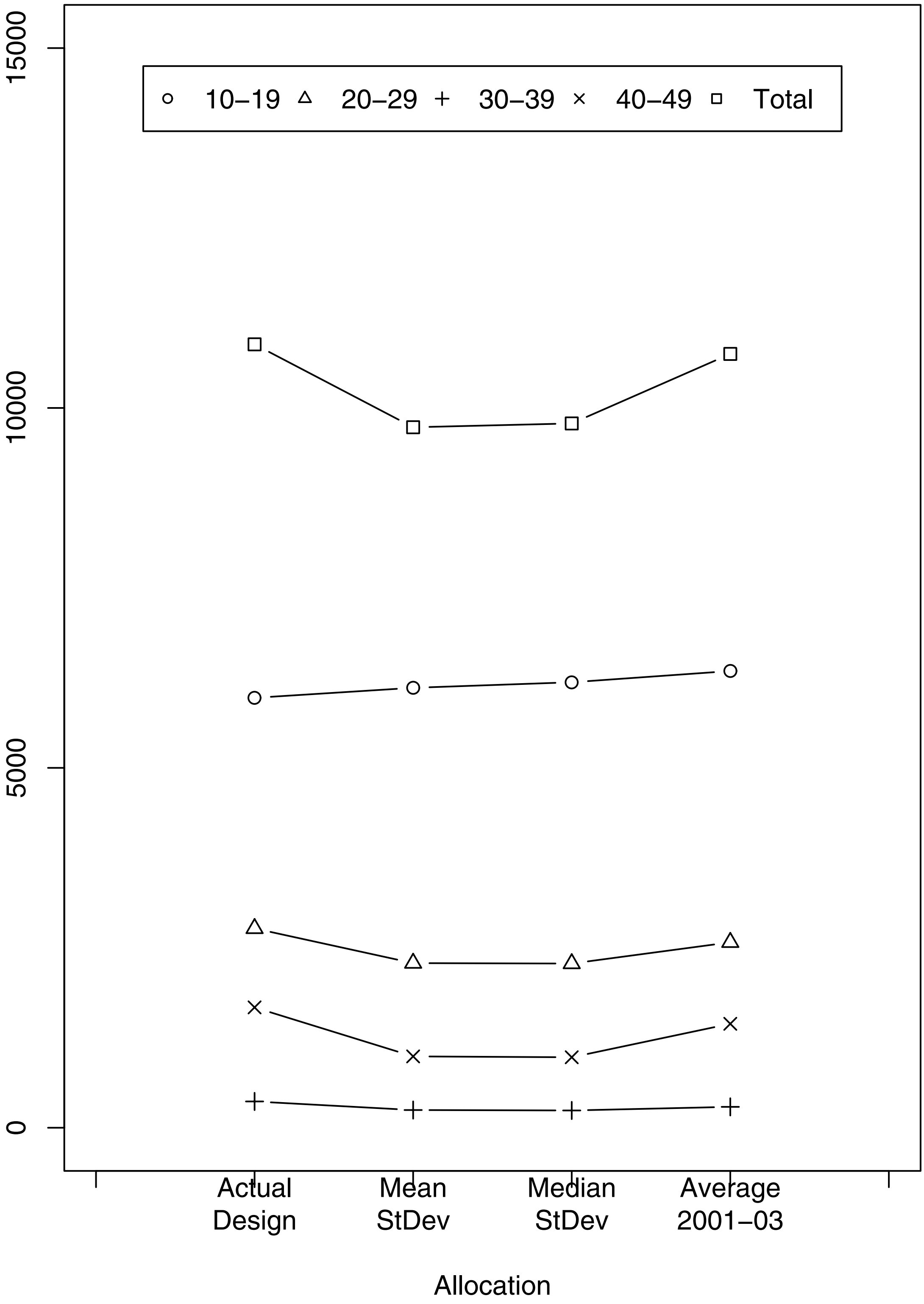
[9] Mahalanobis, P.C. (1968), "Sample Census of Area Under Jute in Bengal, 1940" *Statistical Publishing Society*, Calcutta.

[10] National Agricultural Statistics Service (1989), *Area Frame Design for Agricultural Surveys*, Washington, D.C.: NASS, U.S. Department of Agriculture.

[11] Perry, Charles R. (1992), "The Province Level Sample Size and Allocation for the Punjab, Sindh and Northwest Frontier Provinces," IPO Trip Report, NASS, USDA. Washington, D.C. December, 1992.

[12] Smith, H.F. (1936), "An Empirical Law Describing Heterogeneity in the Yield of Agricultural Crops," *Journal of Agricultural Science*: **28**.

Figure 1: Comparison of 2004 Sample Allocations


○ 10–19
△ 20–29
× 40–49
□ Total
Actual Design
Mean StDev
Median StDev
Average 2001–03
Allocation