# THE VARIANCE ESTIMATOR AND SAMPLE SIZE IN VARIABLES SURVEY SAMPLING

Vernon W. Stone and Russell W. Irvine, Georgia State University

## PROBLEM

While attributes sampling focuses on the question "How many?," variables sampling focuses on the question "How much?." In the objective determination of sample size, $\underline{pq}$ is the variability factor in the former, and $\underline{\sigma}$ is the variability factor in the latter. In variables sampling, depending upon the various and sundry estimators employed to estimate variance and/or standard deviation, a wide range of sample sizes can result. The formula application requires the direct use of the standard deviation. The tabular reading requires the use of the standard deviation as the denominator in the ratio of the sampling error to the standard deviation. The foregoing, then, suggests various estimates of variance and/or standard deviation, pursuant to a variety of estimators, and the subsequent objective determination of sample sizes employing such estimates.

## HYPOTHESIS

The basic hypothesis may be stated as follows: Depending upon the estimator(s) employed, marked variations can be observed among the variance estimates and, consequently, the objectively determined sample sizes.

## PROCEDURES

### Method

The descriptive methodology, making use of the survey technique was employed.

### Subjects

College professors, graduate students, and research practitioners, totaling 115 volunteers, in attendance at a 1976 sampling workshop, constituted the sample. The volunteers represented a variety of behavioral-science disciplines. Moreover, each participant possessed recognized preparation and expertise in the general sampling area.

### Instrumentation

A questionnaire, detailing the following contrived situation, was distributed to the 115 prospective participants:

1,000 college freshmen are annually enrolled in Mathematics 105, Mathematical Analysis, at City Junior College. During the last five academic years (1970-71 thru 1974-75), an annual mathematics achievement test has been administered to all of the 1,000 students. For the 1975-76 academic year, it is necessary, because of budgetary problems, to use a sample from the 1,000 students.

The following information was attached to the questionnaire: (1) Miscellaneous experimental data, including numerical indices of the achievement-test variances for the last five academic years; and (2) Population frame of the 1976 mathematics achievement scores for the 1,000 freshmen.

Also, the following sampling specifications were indicated: (1) Population is 1,000 ($N = 1{,}000$); (2) Confidence level is to be 95% (z (standard variate) $= 1.96$; and (3) Sampling error is to be 2 ($E = 2$).

The sampling task was to estimate the population variance--by any estimator desired--and report a numerical index for the variance estimate.

In addition to the foregoing, the respondent was requested to indicate, on the reverse side of the questionnaire, the estimator category, or categories, and the estimator sub-category, or sub-categories, employed in arriving at the reported numerical index of the variance. The checklist presented both the estimator categories and estimator sub-categories as follows:

1.___Empirical Bases
   a.___Past experience
   b.___Conventional practice
   c.___(Other)________________

2.___Ad Hoc Samples
   a.___One-sample statistics
   b.___Two-or-more-sample statistics
   c.___(Other)________________

3.___Central Limit Theorem Applications
   a.___Derived-formula basis
   b.___Normal distribution basis
   c.___(Other)________________

4.___Systematic Determinations
   a.___Computed values
   b.___Tabular values
   c.___(Other)________________

5.___Subjective Approaches
   a.___Intuition
   b.___Whim
   c.___(Other)________________

___(Employed Estimator Category Not Listed)

### Treatment

On the basis of the reported variances, sample sizes were determined by both formula and table.

(1) By formula. The starting point was the reliability, or precision, formula, based upon the standard error of the mean. Then, the sample-size (n) formula was derived, and the finite population correction was applied thereto, thus resulting in the following composite formula which is applicable to variables sampling:

$$n = \frac{N\,(z\,\sigma)^2}{(N\,E^2) + (z\,\sigma)^2}$$

(2) By table. The ratios of the sampling error to the standard deviation were computed. Then, each ratio was related to population size and confidence level. For purposes of this study, the appropriate sample-size table by Arkin was employed.

Except for minor variations, the formula and the table yield comparable sample sizes.

## RESULTS

Table 1, "The Frequency and Proportion of Subjects Employing Each of the Estimator Categories," indicates that the number of usable responses was 36, or approximately 31 per cent of the number of questionnaires distributed. The frequency of use of the five estimator categories ranged from 4 to 12, showing Systematic Determinations as the modal estimator category.

Table 1

THE FREQUENCY AND PROPORTION OF SUBJECTS EMPLOYING EACH OF THE ESTIMATOR CATEGORIES

| Estimator Category | Frequency | Proportion of Respondents (n=36) |
|---|---|---|
| Empirical Bases | 4 | .11 |
| Ad Hoc Samples | 7 | .20 |
| Central Limit Theorem Applications | 5 | .14 |
| Systematic Determinations | 12 | .33 |
| Subjective Approaches | 8 | .22 |
| Totals | 36 | 1.00 |

Table 2, "The Estimated Ranges of Variance and Standard Deviation Reported for Each of the Estimator Categories," shows that the Subjective Approaches are the most variable. Variance ranges from 36 to 361, and standard deviation ranges from 6 to 19. The Empirical Bases are observed to be the least variable. Hence, the lowest variance is 36, and the highest variance is 361, representing standard deviations of 6 and 19, respectively.

Table 2

THE ESTIMATED RANGES OF VARIANCE AND STANDARD DEVIATION REPORTED FOR EACH OF THE ESTIMATOR CATEGORIES

| Estimator Category | Estimated Range for Variance ($\sigma^2$) | Estimated Range for S.D. ($\sigma$) |
|---|---|---|
| Empirical Bases | 81-121 | 9-11 |
| Ad Hoc Samples | 64-225 | 8-15 |
| Central Limit Theorem Applications | 100-169 | 10-13 |
| Systematic Determinations | 81-196 | 9-14 |
| Subjective Approaches | 36-361 | 6-19 |

Pursuant to the five specified estimator categories, it was expected that Subjective Approaches would be the most variable since there is a discernible relationship between variability and subjectivity. It is unclear, however, as to why Empirical Bases are the least variable, insofar as a degree of subjectivity is involved therein. It was anticipated that Systematic Determinations would be the least variable because of the objective nature of this estimator category.

A comparison of Tables 1 and 2 shows that there is a relationship between the highest frequency (Table 1) and the most variable estimator category (Table 2). A Spearman calculation results in a coefficient of .70. Notwithstanding, the small number of estimator categories renders the index insignificant.

Table 3, "Sample Sizes on the Formula Basis (Standard Deviation) and the Tabular Basis (Ratio of Sampling Error to Standard Deviation) Determined for the Reported Standard Deviation Range (S.D. 6 thru S.D. 19)," makes use of the computational formula by using the standard deviation values and the table by using the ratio of the sampling error to standard deviation.

It will be noted that minor differences are to be observed in most instances with respect to the values of the two sample-size determinations. Observing that the standard deviations range from 6 thru 19 (variances 36 thru 361), the sample sizes, by formula, range from 34 thru 258 and, by table, from 41 thru 278.

Table 3

SAMPLE SIZES ON THE FORMULA BASIS (Standard Deviation) AND THE TABULAR BASIS (Ratio of Sampling Error to Standard Deviation) DETERMINED FOR THE REPORTED STANDARD DEVIATION RANGE (S.D. 6 thru S.D. 19)

| Standard Deviation ($\sigma$) | Sample Size By Formula (n) | Ratio of Sampling Error to Standard Deviation ($E/\sigma$) | Sample Size By Table (n) |
|---|---|---|---|
| 6 | 34 | .33 | 41 |
| 7 | 45 | .29 | 42 |
| 8 | 58 | .25 | 58 |
| 9 | 73 | .22 | 88 |
| 10 | 88 | .20 | 88 |
| 11 | 105 | .18 | 89 |
| 12 | 122 | .17 | 146 |
| 13 | 140 | .15 | 146 |
| 14 | 159 | .14 | 164 |
| 15 | 179 | .13 | 186 |
| 16 | 198 | .13 | 187 |
| 17 | 218 | .12 | 211 |
| 18 | 238 | .11 | 241 |
| 19 | 258 | .10 | 278 |

Table 4, "The Cost of Incremental Sample Units When Sample Size Is Based Upon Formula Determined for the Reported Standard Deviation Range (S.D. 6 thru S.D. 19)," underscores the direct, positive relationship between incremental sample units and cost, using the sample sizes, by formula, for the illustration. Pursuant to a request, a testing service representative suggested that the subject mathematics achievement test, under the conditions indicated, would cost approximately $1,180 for the first 75 students when group tested. Thereafter, a pro rata average cost of $30.00 for each incremental sample unit above 75 students would be reasonable on the basis of current educational costs. Even a degree of conservatism was discerned with respect to these cost figures. From a S.D. of 6 thru a S.D. of 9, a maximum sample of 73 is involved. Hence, $1,180 would represent the basic cost. From a S.D. of 10 thru a S.D. of 19, however, the incremental sample units at $30.00 each bring the total sample cost to $1,570 thru $6,670, representing sample sizes of 88 and 258, respectively.

Table 4

THE COST OF INCREMENTAL SAMPLE UNITS
WHEN SAMPLE SIZE IS BASED UPON FORMULA
DETERMINED FOR THE REPORTED STANDARD DEVIATION RANGE (S.D. 6 thru S.D. 19)

| Standard Deviation (6) | Sample Size (n) | Incremental Sample Unit Cost at $30.00 Each | Total Sample Cost Given Minimum Cost for 75 Units ($1,180) |
|---|---|---|---|
| 6 | 34 | $ --- | $ 1,180 |
| 7 | 45 | --- | 1,180 |
| 8 | 58 | --- | 1,180 |
| 9 | 73 | --- | 1,180 |
| 10 | 88 | 390 | 1,570 |
| 11 | 105 | 510 | 2,080 |
| 12 | 122 | 510 | 2,590 |
| 13 | 140 | 540 | 3,130 |
| 14 | 159 | 570 | 3,700 |
| 15 | 179 | 600 | 4,300 |
| 16 | 198 | 570 | 4,870 |
| 17 | 218 | 600 | 5,470 |
| 18 | 238 | 600 | 6,070 |
| 19 | 258 | 600 | 6,670 |

## CONCLUSIONS

The subject hypothesis was confirmed, clearly demonstrating that there are, depending upon the estimator(s), marked variations among variance estimates and, consequently, sample sizes.

The study constituted a pilot, designed to suggest the confirmation or disconfirmation of the hypothesis. To be sure, a tighter, more telescopic study may reveal different numerical results. Notwithstanding, there is ample reason to anticipate that such results will underscore the general direction of the findings reported herein.

Recognizing that samples which are too small and samples which are too large can adversely affect findings, it behooves the researcher to attend such with great care. At the threshold of the Third Century, researchers have an obligation to avoid both "overkill" and "underkill" via sample size.

## REFERENCES


Arkin, Herbert. Statistical Sampling for Auditors. N.Y.: Price Waterhouse & Co., 1960.

Arkin, Herbert. "Statistical Sampling in Auditing," The New York Certified Public Accountant, July, 1957.

Arkin, Herbert and Henry P. Hill. Sampling in Auditing. N.Y.: Price Waterhouse & Co., 1959.

Cochran, William G. Sampling Techniques, 2/e. N.Y.: John Wiley & Sons, Inc., 1963.

Deming, W. Edwards. Sample Design in Business Research. N.Y.: John Wiley & Sons, Inc., 1960.

Hansen, Morris H., William N. Hurwitz, and William G. Madow. Sample Survey Methods and Theory, Vol. I. N.Y.: John Wiley & Sons, Inc., 1953.

Slonim, Morris J. Sampling in a Nutshell. N.Y.: Simon and Schuster, 1960.

Stephan, Frederick F. "History of the Uses of Modern Sampling Procedures," Journal of the American Statistical Association, 43 (241), 1948.

Trueblood, Robert M. and R.M. Cyert. Sampling Techniques in Auditing. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957.

Vance, Lawrence L. and John Neter. Statistical Sampling for Auditors and Accountants. N.Y.: John Wiley & Sons, Inc., 1956.