DETERMINING THE SAMPLE SIZE FOR MULTIVARIATE ESTIMATES SATISFYING SIMULTANEOUS CONFIDENCE INTERVALS

Dhirendra N. Ghosh, Consultant, McLean, Virginia

The usual practice to determine the necessary sample size in a multivariate sample survey designed to estimate the means of several variables is as follows:

For each key variable i, we use an estimate for the variance (σ) and a specified precision requirement given in terms of an allowable error (d_i) and a precision level (say 95%.)

The following two equalities are to be satisfied

Prob $\left[\left|\bar{x}_i - M_i\right| \leq d_i\right] = .95$

Prob $\left[\left|x_i - M_i\right| \leq 1.96 \frac{\sigma_i}{\sqrt{n_i}}\right] = .95$

We equate $d_i = 1.96 \frac{\sigma}{\sqrt{n_i}}$

and solve for n_i

This process is repeated for all the key variables (k) and the largest of the n_i's is the required sample size $[n = \max (n_1, .. n_k)]$

For categorical data, the variance is calculated for the worst possible case (P=0.50) when the expected proportion is unknown.

If, however, we want to ensure that all the estimates satisfy a simultaneous confidence interval i.e., the estimates all fall within prescribed limits with a certain probability (say 95%), the following inequalities must be satisfied simultaneously:

Prob $\left[\left|(x_1 - M_1)\right| \leq d_1, \ldots \left|x_k - M_k\right| \leq d_k\right] = .95$

We first consider the special case when the key variables are pairwise independent, a highly unlikely case for a real life sample survey.

Then the above equation equals

$\prod_{i=1}^{k}$ Prob $\left(\left|x_i - M_i\right| \leq d_i\right) = .95$

For any possible n_i we can calculate each of the separate probabilities. Their product is the overall probability. We start with a plausible value of n and then keep on increasing or decreasing the value of n until the product equals .95.

For example let $\sigma_1^2 = 100$, $\sigma_2^2 = 200$

$d_1 = 2$, $d_2 = 3$

then n = 120

The following theorem due to Turkey provides a sample size which somewhat corresponds to the above case.

Theorem:

Subbase $M_1, M_2 \ldots M_k$ are unknown parameters and $(\underline{M}_1, \bar{M}_1), \ldots$ $(\underline{M}_k, \bar{M}_k)$ are $100\left[1 - \frac{1}{k}(1 - P)\right]\%$ confidence intervals for $M_1, \ldots M_k$ respectively. Then the probability is at least P that these confidence intervals simultaneously contain $M_1 \ldots M_k$ respectively.

For the general case when all the pairwise correlations are not zero, we need to know the pairwise covariances in addition to the individual variances. We need to calculate the multivariate normal probabilities where the integral for the ith variable has the limits $\left(\frac{-di}{\sigma_i/\sqrt{n}} \quad \frac{di}{\sigma_i/\sqrt{n}}\right)$

However, multivariate normal tables for more than 3 variables are not known. Thus this method does not work for the case with more than 3 key variables.

An approximation can be worked out in the following way.

We know

$$T^2 = n\left[\bar{x}_1 - M_1, \ldots \bar{x}_k - M_k\right] \Sigma^{-1} \left[\bar{x}_1 - M_1, \ldots \bar{x}_k - M_k\right]'$$

is distributed as X^2 with K degrees of freedom. If we substitute d_i for $X_i - M_i$ and equate T^2 to a specified tabulated value of X^2 we can solve for n.

For example

$\sigma_1^2 = 100$, $\sigma_2^2 = 200$

$d_1 = 2$, $d_2 = 3$

$\sigma_{12} = \sigma_{21} = 90$, $P = 0.64$

$$\Sigma = \begin{bmatrix} 100 & 90 \\ 90 & 200 \end{bmatrix}$$

$$n\left[d_1, d_2\right]\Sigma^{-1}\left[d_1, d_2\right]' = 5.99 = (X^2_{.95\,(2)})$$

n=113

However, this simultaneous confidence interval has the shape of an ellipse while the specified confidence interval was rectangular. Nevertheless, if the sizes of the confidence intervals are approximately equal, the above procedure provides a close approximation.

References:

1. Goodman, Leo: On Simultaneous Confidence Intervals for Multinomial Proportion. Technometrics, Vol 7, 1965.

2. Moonan, William J: On the Problem of Sample Size for Multivariate Simple Random Sampling. Journal of Experimental Education. Vol 22, 1952.

3. Roy, S.N. and Bose, R.C.: Simultaneous Confidence Interval Estimation, Annals of Mathematical Statistics. Vol 24, 1955

4. Tables of the Bivariate Normal Distribution Function and Relation Function. NBS, 1959

5. Steck, G.P.: A Table for Computive Trivariate Normal Population, Ann. Math. Stat. Vol 29, 1959