Analysis of Variance as a Basis for Sample Surveys

Dhiren Ghosh* Andrew Vogt†

Abstract

In sample surveys variance generally decomposes into the sum of two terms: variance among primary units and variance within primary units. If the secondary units are composite, variance within a primary unit can be decomposed into variance among secondary units and variance within secondary units. This can be extended as necessary. This decomposition is the central idea in stratified sampling, cluster sampling, regression and ratio estimation, and ANOVA. It can also be applied to cases where the primary units overlap.

Key Words: stratification, clustering, regression estimates, ratio estimates, between-cluster variance, within-cluster variance

1. Introduction

A population is divided into subpopulations or clusters, and we study a variable y on this population. The variable y has a mean μ and a variance σ^2. The variance σ^2 can be decomposed into two parts: the variance within clusters and the variance between clusters.

Our theme is how this decomposition takes place for the variable y on the population and for estimators of μ on random samples from the population. We consider multiple settings: clusters, strata, regression/ratio, and overlapping clusters.

Our subject matter is thus Analysis of Variance or ANOVA, and the role it plays in sample surveys. In ANOVA we apply multiple treatments and compare the resulting values of a variable y. Treatments can be likened to clusters or strata, and we undertake a comparison of treatments by ANOVA to determine whether the treatments are significantly different from one another. The basic equation is:

$$\Sigma_{i,j}(y_{ij} - \overline{y})^2 = \Sigma_i \Sigma_j (y_{ij} - \overline{y_i})^2 + \Sigma_i N_i (\overline{y_i} - \overline{y})^2$$

or

$$SST = SSW + SSB.$$

Here N_i is the size of the i-th treatment group. The above equation is the basis for ANOVA. Under assumptions of normality and independence we study the ratio SSB/SSW properly scaled, expecting it to behave as an F distribution if the treatments are indistinguishable.

We propose that analysis of variance also happens to be the fundamental theme in survey sampling (other than simple random sampling). In ordinary ANOVA we are testing hypotheses whereas in sample surveys we are estimating parameters, but the differences obscure the commonality. We make no claims for the originality of this observation. See, for example, Särndal et al. [2, p. 261 and p. 281]. Our presentation is merely a review and recasting of familiar ideas in this light. Standard formulas that appear below are to be found in Cochran [1].

*Synectics for Management Decisions, Inc., 1901 North Moore Street, Suite 900, Arlington, VA 22209

†Department of Mathematics and Statistics, Georgetown University, Washington, DC 20057-1233

2. Clusters and Strata

We consider a finite population on which a variable y is defined, the population being divided exhaustively into k mutually exclusive subpopulations, which for the moment we call "clusters."

The variable y has mean $E(y) = \mu$ and variance $\sigma^2 = E((y-\mu)^2)$. The deviation $y - \mu$ can be rewritten as $y - \mu = y - \mu_i + \mu_i - \mu$ where y is a value taken on a unit in the i-th cluster and μ_i is the mean of y on the i-th cluster (and σ_i^2 is the variance of y on the i-th cluster). Then it is easily seen that:

$$\sigma^2 = \frac{\Sigma_i N_i \sigma_i^2}{N} + \frac{\Sigma_i N_i (\mu_i - \mu)^2}{N}$$

where N_i is the size of the i-th cluster. This equation is the sample survey version of $SST = SSW + SSB$.

If we attempt to estimate the population mean μ by the sample mean $\overline{y}$ of a simple random sample without replacement from the population, we arrive at the familiar results:

$$E(\overline{y}) = \mu$$

and

$$\begin{aligned} V(\overline{y}) = (\frac{N-n}{N-1})\frac{\sigma^2}{n} &= (\frac{N-n}{N-1})\frac{1}{n}(\frac{\Sigma_i N_i \sigma_i^2}{N} + \frac{\Sigma_i N_i (\mu_i - \mu)^2}{N}) \\ \approx \frac{N-n}{Nn} s^2 &= \frac{N-n}{N}\frac{1}{n}(\Sigma_i \frac{n_i - 1}{n-1} s_i^2 + \Sigma_i \frac{n_i}{n-1}(\overline{y_i} - \overline{y})^2). \end{aligned}$$

We see from the above that when a sample mean is used to estimate the population mean, the variance of this estimator also divides into a portion that measures variance within clusters and a portion that measures variance between clusters. In the last line we have replaced the true variances by sample and subsample variances s^2, s_i^2, which are unbiased estimates of the former. Of course $\overline{y_i}$ and s_i^2 may not be available for some clusters unless they are represented by at least one or two elements in the sample.

Another possible estimator for the population mean μ is

$$\overline{y}_w = \frac{\Sigma_i N_i \overline{y_i}}{N}$$

provided that each cluster is represented in our random sample. In this case the familiar result is:

$$E(\overline{y}_w) = \mu$$

and

$$\begin{aligned} V(\overline{y}_w) &= \Sigma_i (\frac{N_i}{N})^2 (\frac{N_i - n_i}{N_i - 1})\frac{\sigma_i^2}{n_i} \\ &\approx \Sigma_i (\frac{N_i}{N})^2 (\frac{N_i - n_i}{N_i})\frac{s_i^2}{n_i}. \end{aligned}$$

Adjusting the weights of the subsample means eliminates the between component of the variance. Stratification has resulted in the disappearance of a component of variance!

Use of $\overline{y}_w$ has suppressed the between-cluster variance. The estimator works directly with the individual clusters assigning them weights proportional to the cluster sizes. The

only variability is that between the individual sample means and the individual cluster means.

Of great importance in this setting is that the clusters are exhaustive and all clusters are represented in the sample. In this situation we call the clusters strata, and we are engaging in stratified random sampling. If some clusters are not represented in the sample, a customary stratagem is to collapse clusters or strata so that each stratum is represented. The choice of clusters or strata, and any collapsing are based on a principle of homogeneity: we would like the within-cluster variance to be small so strata are chosen to be as homogeneous as possible with respect to the variable y or its surrogates.

Another indispensable requirement for the use of $\overline{y}_w$ is that the sizes N_i of each cluster are known.

3. Regression and Ratio

Now let us consider regression. In addition to y, we suppose that another variable x is available about which considerable information is available, and we consider a regression of y on x. The least squares model for such a relationship is:

$$\hat{y} = A + Bx$$

where $B = \frac{\sigma_{xy}}{\sigma_x^2}$ and $A = \mu_y - B\mu_x$ and $\hat{y}$ is the imagined component of y that depends linearly on x.

The variance of y on the population, alrady computed above, can be thought of in terms of clusters, one cluster corresponding to each value of x. Then the variance decomposes as follows:

$$V(y) = \frac{\Sigma_x N_x \sigma_{y/x}^2}{N} + \frac{\Sigma_x N_x (\hat{y} - \mu_y)^2}{N}$$

where $\sigma_{y/x}^2 = \frac{\Sigma_{\text{fixed } x}(y-\hat{y})^2}{N_x}$. The decomposed variance is here seen to be the mean within-cluster variance plus a between-cluster variance. In the mean within-cluster variance the quantity $\sigma_{y/x}^2$ can be thought of as the mean square error of y as an estimate of $\hat{y}$ within the cluster (fixed x value). The between-cluster variance is equal to

$$\frac{\Sigma_x N_x (\hat{y} - \mu_y)^2}{N} = B^2 \sigma_x^2 = \frac{\sigma_{xy}^2}{\sigma_x^2}.$$

We have decomposed the variance of y on the population into a component that represents variability for given x and a component that varies linearly with x. If the variability of y with x is far from linear, then the within-cluster variances will tend to be large and unstable.

Turning to estimation from a sample, we have two possible estimators based on the population mean of x, μ_x, assumed known:

$$\overline{y}_B = \overline{y} + B(\mu_x - \overline{x})$$

and

$$\overline{y}_b = \overline{y} + b(\mu_x - \overline{x}).$$

Here

$$B = \frac{\sigma_{xy}}{\sigma_x^2}$$

and is assumed known, while

$$b = \frac{\Sigma(y - \overline{y})(x - \overline{x})}{\Sigma(x - \overline{x})^2}$$

is determined by summations over a random sample of n points (x, y). Then

$$E(\overline{y}_B) = \mu_y,$$

and

$$E(\overline{y}_b) \approx \mu_y$$

if the sample size n is sufficiently large. In addition

$$V(\overline{y}_B) = (\frac{N-n}{N-1})\frac{\sigma_y^2 - B^2\sigma_x^2}{n}$$

and

$$V(\overline{y}_b) \approx (\frac{N-n}{N-1})\frac{\sigma_y^2 - B^2\sigma_x^2}{n} \approx (\frac{N-n}{N})(\frac{n-1}{n-2})\frac{s_y^2 - b^2 s_x^2}{n},$$

where the approximations are valid for n large.

In each of these cases we see that the variance of the regression estimate is adjusted downward by removing a term of the form $B^2\sigma_x^2$ (or $b^2 s_x^2$). This is the term previously identified with between-cluster variance. What is removed is variability due to monotonic variation of y with x. What is left is within-cluster variability together with the non-monotonic component of variation between y and x.

Turning to a ratio estimate of μ_y, we have many different estimators available. Examples ares $\overline{y}_{r1} = \frac{\overline{y}}{\overline{x}}\mu_x$, $\overline{y}_{r2} = \frac{\Sigma\frac{y}{x}}{n}\mu_x$ where the summation is over sample units, and $\overline{y}_{r3} = \frac{\Sigma_i N_i \frac{\overline{y}_i}{\overline{x}_i}}{N}\mu_i$. The third estimator can only be used when all clusters or strata are present in the sample, and their frequencies N_i and the mean of x on each are known. A special case of this is when the clusters are the different values of x.

All of these estimators are biased but for large sample sizes (and large enough in each stratum) they are approximately unbiased. The variance of $\overline{y}_{r1}$ is approximately the same as that $\overline{y}_B$ or $\overline{y}_b$, and the others have variances, too complex to report here, that attempt to eliminate variance due to differences in strata or x values.

4. Overlapping Clusters

In the case of overlapping clusters we assume that the population is contained in the union of the clusters. Suppose that the parameter of interest is $\hat{\mu} = \frac{\Sigma_i N_i \mu_i}{\Sigma_i N_i}$. This parameter is the mean of cluster means. In this case the variable y can be replaced by a new variable (y, i) representing the value of y on a unit in the i-th cluster. Then the variable y_o is defined to be the value of y on a unit of this new population, namely, the set of units-with-clusters. Furthermore

$$E(y_o) = \hat{\mu}$$

and

$$V(y_o) = \frac{\Sigma_i N_i \sigma_i^2}{\Sigma_i N_i} + \frac{\Sigma_i N_i (\mu_i - \hat{\mu})^2}{\Sigma_i N_i}.$$

Note further that if each element of the population is in exactly the same number of clusters, then $\hat{\mu} = \mu$ and $V(y_0) = V(y)$. In any case, the variance formula for $V(y_0)$ decomposes the variance into within-cluster variance SSW and between-cluster variance SSB.

The question is: what do we do if not all units are in the same number of clusters? This question is a significant one since many times populations are given by lists. The population is the union of the units in the lists but in general the overlap among lists may be pronounced. Within a single list there may be ways to eliminate repetitions but when we use many lists duplications in different lists can be a significant problem.

One example where this arises is when the population is intended to be all members of a given profession. In this case the clusters may consist of the different professional organizations that members of the profession belong to. If the population is households with schoolchildren, the clusters may be schools. From each school we obtain the names of parents, eliminating extra mentions of parents with more than one student in the same school (or just removing any redundancies in the subsample from that school). However, it is more difficult to account for overlaps between different schools. Another example is a population of consumers, say, of a specialized item, with the clusters being different shops that sell the item.

Here we do not give full details. However, the general approach in a population consisting of k clusters is to divide the population into subpopulations consisting of those units that belong to j clusters where $j = 1, 2, ..., k$. Then treat each subpopulation separately. Consider the two quantities

$$\begin{aligned} N_{ij} &= \text{the number of units} \\ &\quad \text{in the i-th cluster that belong to } j \text{ clusters} \\ \mu_{ij} &= \text{the mean of } y \text{ on the units} \\ &\quad \text{in the i-th cluster that belong to } j \text{ clusters.} \end{aligned}$$

From these two equations it is possible to compute both μ and $\hat{\mu}$, namely,

$$\hat{\mu} = \frac{\Sigma_{i,j} N_{ij} \mu_{ij}}{\Sigma_{i,j} N_{ij}} \text{ and } \mu = \frac{\Sigma_{i,j} \frac{N_{ij}}{j} \mu_{ij}}{\Sigma_{i,j} \frac{N_{ij}}{j}}.$$

The variance of y for each subpopulation can be decomposed into between and within components by cluster, and to some extent the analysis is like that for a double stratification or two-way ANOVA. The kind of issue that we must be concerned with is whether reliable estimates of N_{ij} are available, and what to do when our sample contains no elements in the i-th cluster that belong to j clusters. In fact it will often be the case that $N_{ij} = 0$ for large values of j and so our approach requires assumptions about values that can occur. In some cases we will need to assume that $N_{ij} = \frac{n_{ij}}{n} N_i$, and this is an extra source of variability.

We close by mentioning one extreme case that sheds light on the decomposition of variance that we have proposed. Suppose that each sample without replacement from a population is considered to be a cluster. Then the number of clusters in our population is:

$$k =_N C_n \text{ and } N_i = n \text{ for } i = 1, ..., k$$

and the decomposition SST = SSW + SSB takes the form:

$$\begin{aligned} V(y) &= V(y_0) = \frac{\Sigma_i N_i \sigma_i^2}{\Sigma_i N_i} + \frac{\Sigma_i N_i (\mu_i - \hat{\mu})^2}{\Sigma_i N_i} \\ &= \frac{\Sigma_i n \sigma_i^2}{n_N C_n} + \frac{\Sigma_i n (\overline{y} - \hat{\mu})^2}{n_N C_n} = \frac{\Sigma_i \sigma_i^2}{{}_N C_n} + \frac{\Sigma_i (\overline{y} - \hat{\mu})^2}{{}_N C_n} \\ &= \frac{\Sigma_i \frac{\Sigma (y - \overline{y})^2}{n}}{{}_N C_n} + V(\overline{y}) = E((\frac{n-1}{n}) s^2) + V(\overline{y}) \\ &= \frac{n-1}{n} \frac{N}{N-1} \sigma^2 + \frac{1}{n} (1 - \frac{n}{N}) \frac{N}{N-1} \sigma^2 \\ &= \sigma^2. \end{aligned}$$

If our clusters are themselves random samples, then most of the variance is within-cluster variance and only a small portion is between-cluster variance and that portion shrinks as the sample size grows!

REFERENCES

Cochran, W. G. (1977) , *Sampling Techniques* (3rd ed.), New York: John Wiley and Sons, Inc.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.