Journal of Statistical Theory and Applications, Vol. 14, No. 2 (June 2015), 192-203

Estimating Per Capita Rates Using Aggregate Measurements From Groups of Diverse Compositions

Donald N. Stengel

Department of Information Systems and Decision Sciences,
California State University, Fresno
5245 N. Backer Ave M/S PB07
Fresno, CA 93740-8001, USA

Priscilla Chaffe-Stengel

Department of Information Systems and Decision Sciences,
California State University, Fresno
5245 N. Backer Ave M/S PB07
Fresno, CA 93740-8001, USA

Received 16 November 2014
Accepted 25 January 2015

Abstract

This paper considers the problem of estimating a variable mean for a population of elements where data are only available as aggregate sums for groups of multiple elements. The proposed model addresses an additional complication created when the group measure includes the contribution of diverse elements that were only partially in operation or present as part of the group during the measurement period. The model also accounts for statistical dependency between the contributions of individuals belonging to the same group. The degree of statistical dependency is reflected in a correlation coefficient parameter, which, while not observable, can be adjusted to reduce heteroscedasticity in the group data. A simple example is provided to illustrate the model.

Keywords: Estimation, Sampling, Grouped data

2010 Mathematics Subject Classification: 62F10, 62D99, 62-07

1. Introduction

One of the authors was asked to design a study to estimate the utilization rate per year of an item by individual practitioners in a healthcare profession, where measurements of utilization are generally available only on the basis of total combined use by all partners in a practice because orders of supplies and inventories are maintained on combined basis. While units of the variable being measured are ultimately assignable to one individual member of the practice, the decomposition from aggregate group-level measurement to individual measurements is not observable.

Similar circumstances of sample measurements being available only for aggregates of individual elements occur in other settings. Examples are utilization of office supplies per worker where common supply closets are used by all employees in an administrative unit, consumption of food by individuals in a living unit with a shared kitchen, and water consumption per housing unit in neighborhood where single units do not have water meters.

Published by Atlantis Press

A common tool for addressing these situations is a ratio estimator. Cochran [1] provides a good presentation of ratio estimators. The sampling in this framework measures two variables: the variable of primary interest Y and an auxiliary variable X that counts or measures the size of the aggregated element. Ratio estimators focus on the ratio Y/X, which is a proxy for the rate of variable Y per unit of variable X. The theory provides equations for point and interval estimates of the ratio. In cases where the total number of elements is assumed to be known, the theory also provides point and interval estimates for the total of Y across the population.

The assumption for the classical case of ratio estimators is that Y is a linear multiple of X, the variance in Y is a linear multiple of X, and the residual in Y from the linear relationship between Y and X is normally distributed. For the problem studied here, the variable X would be the number of units included in the aggregate group measurement of variable Y. If the variance in Y increases linearly with the number of individual elements in the group being measured, the variance of Y would be equal to the variance for a group with a single element, multiplied by the number of units in the group.

Since the variance of the sum of a set of identical, statistically independent random variables is the variance of one of the variables multiplied by the number of random variables, the assumption above implies that the contribution of each individual element in a group aggregate is independent of other elements in their group. This assumption may be too strong. In the case of the usage of a medical supply by individual healthcare practitioners, if the usage by one professional in the group is higher than the mean per capita rate in the population, the per capita rates of others in the same group may be more likely to be high relative to the population. The variance within a group may be less than expected if group members really did perform independently of their group membership. This might occur because the individuals in a group have a particular type of practice or may influence one another as to the manner in which they execute their practice. Although some of this dependency can be removed by including covariates in the analysis, the drivers of intragroup dependency may not be readily obvious.

An alternative approach that does not presume that variance among elements within a group is equal to the variance among elements across the population is cluster sampling. Bhatti [2] provides a comprehensive treatment for analyzing the presence of reduced variation with group or cluster elements, called cluster effects. Typically in cluster sampling, clusters are sampled randomly from the population of clusters and then a random sample of elements within each selected cluster is selected and measured. A main difference between the problem studied in this paper and typical cluster sampling applications is that the second level of cluster sampling provides measurements for individual elements within the sample, while in this paper it is assumed there is a single measurement from each group: the sum of values across all elements in the group or cluster.

Another issue that needs to be addressed in our methodology is the impact of the composition of the group on intragroup variation. A healthcare practice consisting of a single full-time professional may not have the same variance as a healthcare practice of two half-time professionals. Although both practices have a similar size in terms of containing one full-time equivalent professional, the latter case of two half-time professionals may have a smaller variance than the single full-time professional.

This paper presents a model for estimating the ratio of a variable per unit that addresses the possibility that within-group variation of individuals in a group is less than the variation across all individuals in the population and considers the compositions of the sampled groups.

2. The Basic Model

Suppose a population is comprised of groups of one or more individual units. There is a measurable characteristic variable Y_{ij} for each individual unit j belonging to group i. However, only the sum of the Y_{ij} for a group is observable. A random sample of n groups will be drawn from the population and a measurement of $Z_i = \sum Y_{ij}$ for group i is recorded.

Individual units may only operate at a fraction of a full-time unit and the value of Y_{ij} is scaled to reflect the operating level. A parameter f_{ij} indicates the operating level, with $f_{ij} = 1$ for a unit that is operating full-time or 100%. Although the model assumes that individual values of Y_{ij} are not observable, the model assumes that the values of f_{ij} are observable for any group selected in the random sample.

The value of Y_{ij} is presumed to be normally distributed, with a mean of $f_{ij}\,\mu$ and a variance of $f_{ij}^2\,\sigma^2$. The values of μ and σ^2 are assumed to be unknown. The expected covariance in Y for individual units included in sample groups is presumed to be $\rho\, f_{ij}\, f_{ik}\, \sigma^2$ for two individual units j and k that belong to the same group i, where $0 \le \rho \le 1$, and zero for two individual units belonging to different groups. The parameter ρ is the coefficient of correlation in variable Y for any pair of individual units in the same group. The assumption of common intragroup correlation is employed widely in cluster sampling models, such as Scott and Holt [3]. In this section, we will consider the parameter ρ as a value set by the researcher; in a subsequent section, we will consider how to select the value of ρ.

Let m_i represent the number of individual units in group i. Based on known results for sums of random variables [4], we can conclude the expected value for the measured total Z_i for group i is

$$E(Z_i) = \sum_{1 \le j \le m_i} E(Y_{ij}) = \mu \sum_{1 \le j \le m_i} f_{ij}$$

And the variance for the measured total Z_i for group i is

$$Var(Z_i) = \sum_{1 \le j \le m_i} Var(Y_{ij}) + 2 \sum_{1 \le j < k \le m_i} Cov(Y_{ij}, Y_{ik}) = \sigma^2 \sum_{1 \le j \le m_i} f_{ij}^2 + 2\rho\sigma^2 \sum_{1 \le j < k \le m_i} f_{ij} f_{ik}$$

Splitting the first term and reorganizing leads to

$$Var(Z_i) = (1-\rho)\sigma^2 \sum_{1 \le j \le m_i} f_{ij}^2 + \rho\sigma^2 \Big(\sum_{i \le j \le m_i} f_{ij} \Big)^2$$

Since the number of full-time equivalent units in group i is known, the total observed value Z_i for group i can be transformed to a variable R_i for the per capita rate observed for the group, with the following expected value and variance:

$$R_i = Z_i / \sum_{1 \le j \le m_i} f_{ij}$$

$$E(R_i) = E(Z_i) / \sum_{1 \le j \le m_i} f_{ij} = \mu$$

$$Var(R_i) = Var(Z_i) / \Big(\sum_{i \le j \le m_i} f_{ij} \Big)^2 = (1-\rho)\sigma^2 \sum_{1 \le j \le m_i} f_{ij}^2 \Big/ \Big(\sum_{i \le j \le m_i} f_{ij} \Big)^2 + \rho\sigma^2 \qquad (1)$$

For the case where a group consists entirely of fully operating individual units (i.e., $f_{ij} = 1$ for all units), Eq. (1) simplifies to

$$Var(R_i) = (1-\rho)\sigma^2 / m_i + \rho\sigma^2 \tag{2}$$

When $\rho = 0$ and the value of each individual unit is independent of the values of other individual units in its group, the variance in Eq. (2) is the familiar expression for the average of identical but independent random variables. When $\rho = 1$, the values of individual units are perfectly correlated with other firms in the group, and the variance of the per capita rate is the same regardless of the number of units in the group.

Point estimates and confidence intervals for μ can be obtained by collecting the observed values of R_i from the random sample of groups and calculating the minimum variance estimate. However, since each group has a different variance for R_i due to different group compositions, heteroscedasticity is generally present, and it is necessary to use a weighted estimate of population parameters. The weights should be set such that the weighted variables have the same expected variation. Using Eq. (1), a suitable set of weights w_i would be

$$w_i = \frac{1}{Var(R_i)/\sigma^2} = \frac{1}{(1-\rho)\sum_{1 \le j \le m_i} f_{ij}^2 /(\sum_{i \le j \le m_i} f_{ij})^2 + \rho} \tag{3}$$

The goal of the model is to provide estimates of mean μ and variance σ^2 of the value of the variable being studied as it applies to a full-time single unit. Based on standard formulas used in statistical software [5], we calculate the estimates as follows:

The minimum variance point estimate for μ is weighted mean of the R_i values.

$$\overline{R}_w = \sum_{1 \le i \le n} w_i R_i / \sum_{1 \le i \le n} w_i$$

Since w_i is set so that $\sigma^2 = w_i \, Var(R_i)$, the squared deviation of R_i from the weighted mean is effectively a sampled value from a normal distribution with mean zero and variance σ^2 / w_i. By averaging the weighted squared differences and correcting for sample variation in the weighted mean, the unbiased estimate of variance σ^2 corresponding to a single, full-time unit is

$$s^2 = \frac{\sum_{1 \le i \le n} w_i (R_i - \overline{R}_w)^2}{n-1}$$

where n is the number of observed groups. The standard error for the estimate for μ is

$$s_e = \frac{s}{\sqrt{\sum_{1 \le i \le n} w_i}}$$

A 100(1-α)% confidence interval for μ is

$$\overline{R}_w - t_{1-\alpha/2,n-1} s_e \le \mu \le \overline{R}_w + t_{1-\alpha/2,n-1} s_e$$

3. Sensitivity of Group Weights to Intragroup Correlation, Group Size, and Group Composition

The group weights used to determine point estimates and confidence intervals on per capita levels are affected by the correlation ρ between individual units in each group, the number of units m_i in each group i, and the set of operating levels f_{ij} of the units in each group i. As shown in the Appendix, the formula for the group weights w_i in Eq. (3) has the following equivalent expression:

$$w_i = \frac{1}{1+(1-\rho)CV_i^2/m_i - (1-\rho)(m_i-1)/m_i} \tag{4}$$

where the CV_i^2 is the squared coefficient of variation in unit operating levels f_{ij}:

$$CV_i^2 = \frac{\sum_{1\le j\le m_i}(f_{ij}-\bar{f}_i)^2/m_i}{\bar{f}_i^2}$$

and the average of the f_{ij} values is

$$\bar{f}_i = \frac{\sum_{1\le j\le m_i} f_{ij}}{m_i}$$

A careful examination of Eq. (4) indicates that the weight will increase as the number of units m_i increases for fixed values of ρ and CV_i. However, the rate of increase diminishes as m_i gets large, asymptotically approaching 1/ ρ. The sensitivity of group size on the weight is greater when the correlation coefficient is low, as unit Y_{ij} values are only weakly dependent on other unit values in their group and the group per capita mean will have less variance for larger groups.

The examination of the effect of the coefficient of variation term in Eq. (4) offers two interesting insights. First, since the value in the denominator increases if CV_i increases, the weight for group i gets smaller as relative variation between the f_{ij} values becomes greater. Further, since Eq. (4) includes no other terms related to the participation levels of the units, the average of the f_{ij} values has no effect on the weight if the coefficient of variation remains the same. So, for example, a group involving two units both operating at f_{ij}=0.5 will have the same w_i weight as a group with two units operating at f_{ij}=1.0. Stated differently, when a group has units with uniform participation levels f_{ij}, the weight assigned to the group is the same for any average participation level and no additional information about the mean per capita rate is provided merely by higher average participation rates. This conclusion presumes that unit performance measurements are divisible, which is a reasonable assumption if the unit Y_{ij} values are fairly large or operating levels of the units f_{ij} are not fractions close to zero.

The value of the denominator in Eq. (3) will generally be less than one because for any group with multiple individual units operating at nonzero levels

$$\sum_{1 \le j \le m_i} f_{ij}^2 < (\sum_{i \le j \le m_i} f_{ij})^2$$

Therefore, for fixed values of CV_i and m_i, as the intragroup correlation coefficient increases, the denominator increases and the weight w_i for group i decreases, approaching $w_i = 1$ as ρ approaches one. This occurs because when $\rho = 1$, contributions to the group total from individual units in the group are perfectly correlated, so the group per capita rate effectively varies like a group comprised of a single unit. On the other hand, as ρ approaches zero, the weight w_i increases, approaching a value of

$$w_i = \frac{1}{1 + CV_i^2 / m_i - (m_i - 1) / m_i}$$

If $CV_i = 0$, weight w_i in Eq. (4) is m_i, reflecting the group per capita rate being the result of units operating independently of other units.

4. A Simple Example

Suppose a researcher is interested in the typical number of syringes used by a physician in a particular type of medical practice over one year. A random sample of six practices that focus on this specialty has been selected and queried for (1) total use of syringes over a specified 12-month period, (2) the number of physicians who worked in the practice office during the period, and (3) what fraction of a full-time, full year each individual physician worked during the period.

The results from the survey are in Table 1. Suppose we assume the intragroup correlation coefficient $\rho = 0.3$. Using the relationships from the model defined earlier, the calculated weights for the six groups (rounded to three decimal places) appear in the final column of Table 1.

Table 1. Sample results, use per full-time equivalent physician, and group weights when $\rho = 0.3$.

Practice #	Composition # physicians, participation levels	Total syringes used over year	Syringe use per FTE	Weight w_i for $\rho = 0.3$
1	1 physician, 100%	3150	3150	1
2	2 physicians, each 50%	2220	2220	1.538
3	2 physicians, each 100%	4520	2260	1.538
4	3 physicians, one 100%, two 40%	5880	3266.67	1.709
5	3 physicians, each 60%	6020	3344.44	1.875
6	10 physicians, each 100%	33330	3333	2.703

It is interesting to note that while Practice 1 and Practice 2 has one full-time, full year equivalent position, Practice 2 has a larger weight. This occurs because in the second case, there are two physicians contributing to the syringe use, and while there is only an expected correlation of $\rho = 0.3$, the count from Practice 2 provides more information about the mean per capita use.

Practice 3 has a similar organization to Practice 2, but the two physicians worked at 100% rather than 50%. Yet, the weights of the two practices are the same, because the use by each half-time physician can be doubled and provide an equivalent estimate of per capita use.

Practices 4 and 5 each have three physicians and a total of 1.8 full-time, full year equivalent physicians, yet the weight for Practice 5 is larger. This happens because there is no variation in the participation fractions in Practice 5, while Practice 4 has a positive coefficient of variation in participation levels, so as explained in the previous section, all other things equal, the weight is higher for the group with a lower coefficient of variation.

Practice 6 is at least five times as large as any of the other practices in the sample, yet the weight assigned to Practice 6 is less than twice as large as most of the other practice groups. This happens because the correlation coefficient of ρ=0.3 indicates that syringe use by one physician influences the use by other physicians in the group, thereby diminishing the value of the total syringe use in the practice toward estimating the population mean.

Using these group weights, the model yields the following sample statistics for annual per capita syringe use:

Point estimate: 2982 syringes per year
Sample standard deviation: 698.8 syringes per year
Standard error of the estimate of the population mean: 217.1 syringes per year
95% confidence interval for the population mean: between 2424 and 3540 syringes/year

Since the value of the correlation coefficient is hypothesized, the effect of the selection can be examined by recalculating the weights and statistics for other values of ρ. Table 2 shows the impact on group weights for different coefficient values. When ρ=0, practices are presumed to be comprised of statistically independent individual units. When the units in the group are uniform in participation level, the weight is equal to the number of units. These weights diminish as ρ increases, reflecting the higher correlation within the group and corresponding diminished value of each observation in estimating μ relative to the number of units. When ρ=1, every practice has a weight of one regardless of group composition, as perfect correlation between units in a group means that each group's per capita use reflects the equivalent to information about μ provided by the value from a practice with one full-time physician.

Table 2. Group weights for example using different values of correlation coefficient ρ.

ρ	0	0.2	0.4	0.6	0.8	1
w_1	1	1	1	1	1	1
w_2	2	1.667	1.429	1.25	1.111	1
w_3	2	1.667	1.429	1.25	1.111	1
w_4	2.455	1.901	1.552	1.311	1.134	1
w_5	3	2.143	1.667	1.364	1.154	1
w_6	10	3.571	2.174	1.563	1.220	1

Table 3 shows the effect of different hypothesized correlation coefficients on the sample statistics. The point estimate is somewhat sensitive to the selection, with greater sensitivity when ρ is closer to zero. The point estimate increases as ρ gets smaller due to the fact that the observed per capita levels for Practices 4, 5, and 6 were higher than for Practices 1, 2, and 3, and the weights for the larger practices decrease more as the coefficient of correlation increases. The 95%

confidence interval limits are tighter and more sensitive to changes in the coefficient when ρ is small, since the standard error of the estimate is inversely related to the square root of the sum of the group weights, and that sum increases at a growing rate as ρ decreases.

Table 3. Sample statistics for individual per capita usage using different values of correlation coefficient ρ.

ρ	0	0.2	0.4	0.6	0.8	1
Point Estimate	3104.0	3004.3	2966.5	2946.9	2935.7	2929.0
Sample Standard Dev	866.3	739.5	665.2	612.2	571.3	538.3
Standard Error of Estimate	191.5	213.9	218.7	220.1	220.2	219.8
LCL (95%)	2612	2454	2404	2381	2370	2364
UCL(95%)	3596	3554	3529	3513	3502	3494

5. Estimation of the Intragroup Correlation Coefficient

The model presented in the second section assumes that a value for the parameter ρ will be specified before being applied to a set of observations from groups of units. This parameter indicates the degree of correlation among pairs of units within groups, allowing dependency between the units for the variable of interest and higher variation in the aggregate group total than would be expected if individual units operated fully independently. However, the degree of correlation is not readily observable and thus is a practical concern in applying the model.

The ρ parameter affects the variability of the group aggregate variable and influences the weighting factors w_i used to compensate for heteroscedasticity in sample per capita rates due to diverse group compositions. In turn, the choice of ρ affects the point estimate of the mean operating level for a full-time unit and the standard error of that estimate.

One can examine the effect of the choice of ρ on point or interval estimates by doing sensitivity analysis on the parameter. The parameter could be tested at different levels, as was done in the example in the previous section. If the point estimate and confidence interval for μ do not change much, the selection of ρ is not a serious concern. If there are noticeable differences, a conservative approach would be to widen the confidence interval at the desired confidence level to include the confidence intervals for the range of correlation coefficients tested. For example, with the syringe use survey presented in the last section, if the researcher is uncertain of correlation value, but believes the correlation coefficient is no larger than 0.6, the interval from 2381 to 3596 would include the 95% confidence intervals for the per capita population mean for all values of ρ from 0 to 0.6.

If the researcher is uncertain about the degree of intragroup correlation, but has some prior belief about value of ρ, a probability distribution could be defined on that parameter and a Bayesian approach could be applied. Or, using Monte Carlo simulation and the model of the prior section, the simulation for randomly generated values of ρ would provide distributions for the point estimate of μ and the upper and lower limits of the confidence interval for μ.

When a fairly substantial number of groups are included in the sample, particularly for a sample that has considerable variation in the weighting factors w_i, the sample data can be used to inform the selection of ρ. Since the purpose of the weighting factors is to maintain equal variation in the weighted residuals from the population mean, scatterplots of the weighted residuals from the procedure can be examined to assess the effectiveness of the weighting factors. The squared residuals can be plotted against the values of weights w_i. If the spread of weighted residuals is fairly uniform across the weights, the selected correlation coefficient is probably sound. On the other hand, if the spread appears to either increase or decrease as the weights increase, this suggests that relative sizes of the weights are not appropriate for reducing heteroscedasticity and a different value for the correlation coefficient might be appropriate.

Since, as noted earlier, for any group with multiple individual units with a nonzero operating level,

$$\sum_{1\le j\le m_i} f_{ij}^2 < (\sum_{i\le j\le m_i} f_{ij})^2$$

there is a negative relationship between the selection of the correlation coefficient ρ and the weights w_i as calculated in Eq. (3). As such, if the larger weights need to be increased to balance the weighted squared residuals, decreasing the value of ρ would reduce the imbalance, whereas if the larger weights need to be reduced in relative magnitude, increasing the value of ρ should result in an improvement. A narrowing spread of the weighted squared residuals as the weights increase indicates that the weights are too aggressive and a larger value of ρ should be tried. Alternatively, a widening spread of weighted squared residuals suggests the higher weighted groups are relatively underweighted and ρ should be decreased.

Another approach to setting the correlation coefficient is to start with the assumption of no intragroup correlation effect (ρ =0) and test if the level of heteroscedasticity is significant. One option is to place the sampled groups into categories based on the w_i weights and apply the Levene test [6] to the weighted residuals to see if the differences in category variation are significant. Another option is to run a regression on the weighted squared residuals against the weight values (or group sizes) to see if there is a significant relationship. The White test [7] and Breusch-Pagan test [8] take this approach. If the test concludes that homoscedasticity must be rejected, the value of the correlation coefficient could be reset and retested for heteroscedasticity.

It should be noted that whenever a different correlation coefficient ρ is used, not only are the weights w_i affected, but in turn, the estimate of the mean per capita estimate will change as well. As a consequence, the residuals between the observed per capita rate and estimated population per capita mean need to recalculated and not merely reweighted.

There is no guarantee that heteroscedasticity can be removed sufficiently with any correlation coefficient ρ between zero and one. This situation may occur if the basic assumption of the model that groups have a similar degree of intragroup correlation is strongly violated.

6. Discussion

Most statistical software packages allow the entry of a weighting variable for calculating sample statistics for a single sample mean. To apply the methods in this paper, the weights would need to be calculated separately or the software would need to be augmented with a script in a language like R or Python.

While the formula for calculating the weighted mean is standard across statistical software tools, there are differences in the calculations of the weighted sample variance, standard error of the mean estimate, and confidence intervals. One source of the difference is the base used to calculate the sample variance. In this paper and in SAS [5], the base used to calculate the (unbiased) sample variance is the sum of classes minus one. However, in SPSS [9], the base used is the sum of the group weights minus one. The justification for using the number of groups minus one is that while groups vary in size and composition, each group in the sample provides a single per capita estimate, and, when weighted to reflect group composition, provides a similar contribution to other groups in estimating the population parameters.

Likewise, the designation of the number of degrees of freedom for the *t*-statistic used to create confidence intervals for the mean may differ based on the same issue. Since the sample variance in the model in this paper is an estimate based on the number of class observations, the confidence intervals on the estimate of the population mean employ a *t*-distribution using the number of groups minus one as the number of degrees of freedom. For algorithms where the sample variance is calculated using the total of the group weights in the denominator, the degrees of freedom will be the sum of the group weights minus one.

Obviously, variables other than intragroup correlation and group composition may be effective in explaining variation in per capita levels across groups. For the example in this paper, perhaps subspecialties of a group practice or geographical location would be significant. Weighted analysis of variance or weighted least-squares regression would allow the incorporation of the group weights based on hypothesized intragroup correlation and composition, as well at the consideration of other variables. However, as with the computation of weighted sample variance and confidence intervals, when using weighted algorithms in statistical software packages, care should be taken to see how the weights are used and degrees of freedom are determined.

In addition to providing inferences about the population from which the sample is drawn, the statistics generated from a sample can be used to determine whether the per capita use for a specific group is high or low relative to the population, given its composition. The group's per capita value can be evaluated based on its position with respect to the probability distribution created by a linear transformation of a *t*-distribution having (1) a mean equal to the weighted mean sample per capita rate, (2) a standard deviation equal to the sample standard deviation in per capita amount divided by the square root of the weight assigned to the group, and (3) the number of degrees of freedom based on the number of groups used in the calculation of the sample variance minus one. With computer tools, the per capita amount for the group in question can be converted to a quantile with respect to that distribution. This evaluation could be applied to groups whether or not they were not part of the sample used to generate the sample statistics.

Appendix: Derivation of the Alternate Formula for the Group Weights

In the presentation of the model in the second section, the formula Eq. (3) for the appropriate weights for each observed class was derived using the intragroup correlation coefficient ρ and participation levels of the individual units comprising the group. In the third section of the paper, alternative formula Eq. (4) was cited that relates the appropriate weight to correlation coefficient, the number of units in the group, and the relative variation in participation levels with the group. The alternative formula is derived here.

The first formula for weights (Eq. (3)) derived with the basic model was

$$w_i = \frac{1}{(1-\rho)\sum\limits_{1 \le j \le m_i} f_{ij}^2 / (\sum\limits_{i \le j \le m_i} f_{ij})^2 + \rho}$$

Inverting the equation and algebraic manipulation results in the following equation:

$$1/w_i = \frac{(\sum_{i\le j\le m_i} f_{ij})^2 + (1-\rho)(\sum_{1\le j\le m_i} f_{ij}^2 - (\sum_{i\le j\le m_i} f_{ij})^2)}{(\sum_{i\le j\le m_i} f_{ij})^2}$$

Since the denominator is the square of the sum of unit fractions, and the sum of the unit fractions is equal to the number of units m_i times the average unit fraction $\bar{f}_i$, using this substitution in the equation yields

$$1/w_i = \frac{m_i^2 \bar{f}_i^2 + (1-\rho)(\sum_{1\le j\le m_i} f_{ij}^2 - m_i^2 \bar{f}_i^2)}{m_i^2 \bar{f}_i^2}$$

Using the following

$$m_i^2 \bar{f}_i^2 = m_i \bar{f}_i^2 + (m_i^2 \bar{f}_i^2 - m_i \bar{f}_i^2)$$

the equation becomes

$$1/w_i = \frac{m_i^2 \bar{f}_i^2 + (1-\rho)(\sum_{1\le j\le m_i} f_{ij}^2 - m_i \bar{f}_i^2) - (1-\rho)(m_i^2 \bar{f}_i^2 - m_1 \bar{f}_i^2)}{m_i^2 \bar{f}_i^2}$$

Applying the computational formula for the variance of a set of data

$$Var(F_i) = \frac{\sum_{1\le j\le m_i} f_{ij}^2 - m_i \bar{f}_i^2}{m_i}$$

then extracting the squared coefficient of variation in the f_{ij} values using

$$CV_i^2 = \frac{Var(F_i)}{\bar{f}^2}$$

followed by cancellation of terms and re-inversion of the equation, results in Eq. (4):

$$w_i = \frac{1}{1 + (1-\rho) CV_i^2 / m_i - (1-\rho)(m_i - 1)/m_i}$$

References

[1] Cochran, W. G., *Sampling Techniques* (3rd ed.) (John Wiley & Sons, New York, NY, 1977).

[2] Bhatti, M., *Cluster Effects in Mining Complex Data*. (Nova Science Publisher's, New York, NY, 2012).

[3] Scott, A. J., and Holt, D., The Effect of Two-Stage Sampling on Ordinary Least Squares Methods, *Journal of the American Statistical Association*, **77**(380), (1982) 848-854.

[4] Feller, W., *An Introduction to Probability Theory and its Applications* (Vol. 1, 3rd ed.) (John Wiley & Sons, New York, NY, 1967).

[5] SAS Institute, Inc., *SAS/STAT 9.2 User's Guide*, Chapter 92 (SAS Institute, Cary, NC, 2009).

[6] Levene, H., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, eds. Olkin, Hotelling, et al (Stanford, CA: Stanford University Press, Stanford, CA, 1960), pp. 278–292.

[7] White, H., A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica*, **48**(4), (1980) 817–838.

[8] Breusch, T. S., and Pagan, A. R., A Simple Test for Heteroscedasticity and Random Coefficient Variation, *Econometrica*, **47**(5), (1979) 1287-1294.

[9] IBM Corporation, IBM SPSS Algorithms, t Test Algorithms, One Sample t Test, [online]. Available at: http://129.8.241.71:56773/help/index.jsp?topic=%2Fcom.ibm.spss.statistics.algorithms%2Falg_introduction.htm