Optimal Allocation in Multivariate Stratified Sampling

Rushil Sheth

May 11th 2017

Introduction

Presently, data sets have a very large population size as data is becoming increasingly easy to obtain. The buzz word, **"Big Data"**, is everywhere. Most data is very large because of the temporal and spatial aspects associated with it, mainly due to the digital age and the invention of new technologies, devices, and means for communication. Humans produced 5 billion gigabytes of data by 2003. This amount was matched every two days in 2011, and at an incredible rate of every 10 minutes in 2013.

Now that we have all this data, we must decide what to do with it. When dealing with these massive data sets, we are forced to use sampling methods for computational and validation purposes.

This raises 3 important questions:

1. How do we choose our samples?
2. What are we trying to maximize or minimize?
3. What is our optimization problem?

Before we derive our optimization problem, we shall define a few constraints involving the strata associated with sampling techniques. If a population U with N units is divided into H strata $E_1, E_2, ..., E_H$, formed by, respectively, $N_1, N_2, ..., N_H$ units, then:

$$N_1 + N_2 + ... + N_h + ... + N_H = N \tag{1}$$

$$E_i \cap E_k = \emptyset, k = 1, .., H, i = k + 1, ...H \tag{2}$$

$$\cup_{k=1}^{H} E_k = U \tag{3}$$

In equation (1), we can see that the sum of the strata sizes, must be equal to the population size. Equation (2) forces the intersection of any two distinct strata to be the empty set, i.e. each element of the population is only in **one** sample. The third condition ensures that each element of the population is in a sample.

The problem of optimal allocation of samples was first discussed by Jerzy Neyman, the founder of the Statistics department here at UC - Berkeley. He discussed how to maximize survey accuracy in univariate sampling. For Neyman allocation, the optimal sample size with a fixed sample size for stratum h is given by:

$$n_h = n * (N_h * S_h)/[\sum(N_i * S_i)]$$

where n_h is the sample size for stratum h, n is total sample size, N_h is the population size for stratum h, and S_h is the population standard deviation of stratum h.

This was proposed over 80 years ago. Since then researchers have hypothesized alternative solutions to sample allocation in univariate surveys. They have even gone further and derived methods for Optimal Allocation in Multivariate Stratified sampling.

Multivariate Stratified sampling

Overview

Multivariate Data is becoming increasingly prevalent and is relevant in many fields of study. For example, a dietician could perform a study on eating habits and have several predictors. Imagine, they have collected data on cholesterol, weight, and fat content of patients and their red meat, dairy, and carbohydrate consumption per week. The dietician wants to investigate the first three measures and how they affect eating habits(last three variables listed), which is a clear multivariate data set, since multiple variables are being predicted.

Optimization Problem formulation

From Neyman's allocation we can see that variance will play a part in our optimal allocation. Methods for multivariate stratified sampling are divided into two main classes. The first method attempts to discover the optimal allocation for the average variance by computing the weighted average of the stratum variances. The second method has to do with the response variables and assigns a coefficient of variation(cv) for each of them.

Unlike Neyman's allocation we do not always know the sample size per strata, a priori. In this case, we have different optimizations for the allocation of samples, $(n_h, h = 1, ..., M)$. The various optimizations will be explained below.

I will describe a nonlinear matrix optimization problem followed by a nonlinear formulation. I will conclude with a linear program which will be converted into a integer programming problem.

Nonlinear Matrix Optimization problem

First, we will look at a nonlinear matrix optimization problem which aims to minimize the variance of the estimated mean$(\bar{y}_{st})$ of the response variables and minimize the overall cost subject to a constraint on the variances.

We must define a few terms first:

Sample covariance in stratum h of the j^{th} and k^{th} response variable: $s_{h_{jk}} = \frac{\sum_{i=1}^{n_h}(y_{hi}^j - \bar{y}_h^j)(y_{hi}^k - \bar{y}_h^k)}{n_h - 1}$

Relative size of stratum h: $W_h = \frac{N_h}{N}$

Overall estimator of population(U) mean for j^{th} response variable: $\bar{y}_{ST}^j = \sum_{h=1}^H W_h \bar{y}_h^j$

Now we can define our Covariance matrix, but we wiil use an estimator of the covariance-variance matrix of $\bar{\mathbf{y}}_{ST}$ for G response variables:

$$\widehat{Cov}(\bar{\mathbf{y}}_{ST}) = \begin{bmatrix} \widehat{Var}(\bar{y}_{ST}^1) & \widehat{Cov}(\bar{y}_{ST}^1, \bar{y}_{ST}^2) & ... & \widehat{Cov}(\bar{y}_{ST}^1, \bar{y}_{ST}^G) \\ \widehat{Cov}(\bar{y}_{ST}^2, \bar{y}_{ST}^1) & \widehat{Var}(\bar{y}_{ST}^2) & ... & \widehat{Cov}(\bar{y}_{ST}^2, \bar{y}_{ST}^G) \\ . & & & \\ . & & & \\ . & & & \\ \widehat{Cov}(\bar{y}_{ST}^G, \bar{y}_{ST}^1) & \widehat{Cov}(\bar{y}_{ST}^1, \bar{y}_{ST}^2) & & ...\widehat{Var}(\bar{y}_{ST}^G) \end{bmatrix}$$

where $\widehat{Cov}(\bar{y}_{ST}^i, \bar{y}_{ST}^j) = \sum_{h=1}^H \frac{W_h^2 s_{h_{jk}}}{n_h} - \sum_{h=1}^H \frac{W_h^2 s_{h_{jk}}}{N}$, and $\widehat{Var}(\bar{y}_{ST}^j) = \sum_{h=1}^H \frac{W_h^2 s_{hj}^2}{n_h} - \sum_{h=1}^H \frac{W_h^2 s_{hj}^2}{N}$

Now we are ready to formulate our optimization problem:

$$\min_n \widehat{Cov}(\bar{\mathbf{y}}_{ST})$$

$$\text{subject to}$$

$$\mathbf{c}'\mathbf{n} + c_0 = C \tag{4}$$

$$2 \leq n_h \leq N_h, h = 1, 2, ..., H$$

$$n_h \in \mathbb{N}$$

The obvious difficulty in this formulation is how to make sense of minimizing a matrix, specifically a covariance matrix. From basic statistics, we know $\widehat{Cov}(\bar{\mathbf{y}}_{ST})$ is a positive semidefinite matrix, so we define a function $f : S- > \mathbb{R}$ where S is the set of positive semidefinite matrices.

This leads us to the following optimization problem, which looks very similar to the one above, but is more practical.

$$\min_n \mathbf{f}\left(\widehat{Cov(\mathbf{y}_{\bar{ST}})}\right) \tag{5}$$

$$\text{subject to}$$

$$\sum_{h=1}^{H} c_h n_h + c_0 = C \tag{6}$$

$$2 \leq n_h \leq N_h, h = 1, 2, ..., H \tag{7}$$

$$n_h \in \mathbb{N} \tag{8}$$

A few examples of f and how it could be useful are given below. Also notice that $\widehat{Cov}(\bar{\mathbf{y}}_{ST})$ is a function of $\mathbf{n}$, which is the stratum sample sizes we want to determine and optimize:

1. The Trace of the matrix, $\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))$:

$$\mathbf{f}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n})) = \text{trace}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n})))$$

2. The determinant of the matrix $\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))$:

$$\mathbf{f}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))) = |\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))|$$

3. The sum of the elements of the matrix, $\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))$:

$$\mathbf{f}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))) = \sum_{j,k=1}^{G} \widehat{Cov}(\bar{y}^j_{ST}, \bar{y}^k_{ST})$$

4. $$\mathbf{f}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))) = \lambda_{max}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n})))$$

   where λ_{max} is the maximum eigenvalue of the matrix $\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))$

5. $$\mathbf{f}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))) = \lambda_{min}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n})))$$

   where λ_{min} is the minimum eigenvalue of the matrix $\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))$

6. $$\mathbf{f}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))) = \lambda_{j}(\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n})))$$

   where λ_{max} is the j_{th} eigenvalue of the matrix $\widehat{Cov}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))$

7. Many others.

Our choice for $\mathbf{f}$ mainly depends on our choice or comparison of some experimental design model. For example, to determine regression estimators for multivariate models, we try to minimize the determinant or trace of $\widehat{Cov}(\bar{\mathbf{y}}_{ST})$. The second $\mathbf{f}$ aims to minimize the generalized variance, which is the 1 dimensional measure of multidimensional scattering. The max and min eigenvalue formulations can be used in dimension reduction techniques, while the j^{th} eigenvalue is used to look at a specific response variable. We will now propose a linear optimization problem, which is more familiar to the class.

Linear Programming Optimization problem

In this optimization problem, we will focus on minimizing the total sample size to be distributed among the strata.

First we will define a few variables: Assume $Y_1, Y_2, ..., Y_j, ..., Y_m$ are a set of m search variables. Then the population variance for each strata for each of these variables is

$$S_{hj}^2 = \sum_{\forall i \in E_h}^{H} (y_{ij} - \bar{y}_{hj})^2, j = 1, ..., m; h = 1, ..., H$$

The y_{ij} is the value of the i^{th} observation in stratum h associated with the j^{th} variable of research, and $\bar{y}_{hj}$ is the average for the h^{th} stratum for variable j.

The variance of the estimator of the total (t_y) for each of the m response variables is defined by:

$$V(t_{y_j}) = \sum_{k=1}^{H} \frac{N_h^2 S_{hj}^2}{n_h} \cdot (1 - \frac{n_h}{N_h}), j = 1, ..., m$$

We can easily calculate N_h and S_{hj}^2, so we observe $V(t_{y_i})$ depends solely on the sample size for stratum n_h.

The objective function is shown in equation (9) and minimizes the sum of sample sizes per strata. Equation (10) guarantees each strata has a sample size of at least 1 and no greater than the stratum size. We come to our reoccurring variance constraint in Equation (11), and it ensures the ratio between standard deviation of each search variable and its search total is less than or equal to a user defined cv.

$$\text{Minimize} \sum_{h=1}^{H} n_k \tag{9}$$

subject to

$$1 \leq n_h \leq N_h, h = 1, ..., H \tag{10}$$

$$\sqrt{V(t_{y_j})}/Y_j \leq cv_j, j = 1, ..., m \tag{11}$$

$$n_h \in Z_+, (h = 1, ..., H) \tag{12}$$

Now we will treat this as a binary integer programming formulation, where we want the sample sizes to be integers. Naturally, we introduce a binary variable x_{hk}.

The integer programming formulation is quite similar to the previous optimization problem. The constraint in equation (14) assures that there will be only one variable x_{hk} per strata, which says that there will be one and only one sample size per strata. Constraint (15) is the equivalent to constraint (12) above, detailing the coefficient of variation.

$$\text{Minimize} \sum_{h=1}^{H} \sum_{k=1}^{N_h} k x_{hk} \tag{13}$$

subject to

$$\sum_{k=1}^{N_h} x_{hk} = 1, h = 1, ..., H \tag{14}$$

$$\sum_{h=1}^{H} \sum_{k=1}^{N_h} \frac{1}{k} x_{hk \cdot hj} - p_{hj} \leq 1, j = 1, ...m \tag{15}$$

$$x_{hk} \in 0, 1, h = 1, ...H, k = 1, ..., N_h \tag{16}$$

References

Moura Brito, José André. Semaan, Gustavo Silva. Nascimento Silva, Pedro Luis. Maculan, Nelson. “An Integer Programming Formulation Applied to Optimum Allocation in Multivariate Stratified Sampling.” Component of Statistics Canada (Dec. 2008).

Díaz-García, José A. Cortez, Liliana U. “Multi-objective optimisation for optimum allocation in multivariate stratified sampling.”

Díaz-García, José A. Ramos-Quiroga, Rogelio. “Optimum allocation in multivariate stratified random sampling: Stochastic matrix optimisation.”

Sampling Techniques. Willian G. Cochran. Third Edition – Wiley, 1977.