The Bethel package

Michele De Meo
<< M-Data-Models >>
micheledemeo@gmail.com
http://micheledemeo.blogspot.com

March 27, 2009

The Bethel algorithm

Bethel's procedure [1] is a mathematical algorithm to achieve the optimum sample allocation in a multivariate sample survey, that is to say the study of several subject variables which are also stratified. The aim of Bethel's procedure is to ascertain the *minimum cost* of the sample, given the precision limits required for each stratum. The cost C is defined as:

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

where c_0 represents a fixed cost correlated with the survey, c_h represents the unit cost per interview within the stratum $h-th$ $(h = 1 \dots H)$, while n_h represents the number of units selected from within the *h-th* stratum. Given that the sampling is stratified, the constraints of precision levels of estimates can be expressed as follows [1]:

$$Var(\widehat{Y}_j) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{hj}^2}{n_h} \leq k_j^2 \qquad j = 1 \dots J$$

where $\widehat{Y}_j$ represents the total estimator for the *j-th* variable *(j=1... J)*, S_{hj}^2 represents an estimate (or a hypothetical value) of the variance of the *j-th* variable within the *h-th* stratum and k_j^2 represents the threshold level (in absolute terms) for the value of the variance of the total estimator (for the *j-th* variable). The previous formula can be equivalently expressed as follow:

$$\sum_{h=1}^{H} N_h^2 \frac{S_{hj}^2}{n_h} \leq k_j^2 + \sum_{h=1}^{H} N_h S_{hj}^2 \quad \Leftrightarrow \quad \frac{\sum_{h=1}^{H} N_h^2 \frac{S_{hj}^2}{n_h}}{k_j^2 + \sum_{h=1}^{H} N_h S_{hj}^2} \leq 1 \quad \Leftrightarrow$$

[1] The finite population correction was not considered in the original Bethel's article and the formula was: $Var(\widehat{Y}_j) = \sum_{h=1}^{H} N_h^2 \frac{S_{hj}^2}{n_h} \leq k_j^2$

$$\Leftrightarrow \quad \sum_{h=1}^{H} \underbrace{\frac{N_h^2 S_{hj}^2}{\epsilon_j^2 \widehat{Y}_j^2 + \sum_{r=1}^{R} N_r S_{rj}^2}}_{a_{hj}} \cdot \underbrace{\frac{1}{n_h}}_{x_h} \leq 1$$

where $\widehat{Y}_j$ represents the total estimator for the variable $j-th$, and ϵ_j represents the relative error (*coefficient of variation*) for the $j-th$ variable. The previuos formula can be more concisely expressed as follows:

$$a_{.j}\widetilde{x} \leq 1 \qquad j = 1 \ldots J \qquad \Leftrightarrow \qquad A'\widetilde{x} \leq \widetilde{1}$$

where $A = [a_{hj}]_{H \times J}$, $\widetilde{x} = [x_h]_{1 \times H} = \left[\frac{1}{n_h}\right]_{1 \times H}$ and $\widetilde{1} = [1]_{1 \times J}$.

The *minimum cost problem* can be summarized as follows[2]:

$$\min_{\widetilde{x}} g(\widetilde{x}) = \sum_{h=1}^{H} \frac{c_h}{x_h} \qquad \text{with} \qquad A'\widetilde{x} \leq \widetilde{1}.$$

Bethel demonstrated that this problem always has a solution, and that this corresponds to the following formula:

$$x_h^* = \frac{\sqrt{c_h}}{\sqrt{\sum_{j=1}^{J} \alpha_j^* a_{hj}} \cdot \sum_{k=1}^{H} \sqrt{c_k \sum_{j=1}^{J} \alpha_j^* a_{kj}}}$$

where α_j^* are the *normalized Lagrange multipliers*.

To calculate x_h^*, Bethel proposes an algorithm which is neither particularly efficient nor easy to apply. At that time, a better algorithm was already available, formulated by Chromy [2] and also available in the same publication of Bethel. The algorithm is composed of the following three steps. Step (2) and step (3) are repeated continually until reaching an acceptable criteria of convergence ($r = 1, 2, \ldots$):

$$\widetilde{\alpha}^{(0)} = \left[\alpha_j^{(0)}\right]_{1 \times J} = \left[\frac{1}{J}\right]_{1 \times J} \tag{1}$$

$$x_h(\widetilde{\alpha}^{(r-1)}) = \frac{\sqrt{c_h}}{\sqrt{\sum_{j=1}^{J} \alpha_j^{(r-1)} a_{hj}} \cdot \sum_{k=1}^{H} \sqrt{c_k \sum_{j=1}^{J} \alpha_j^{(r-1)} a_{kj}}} \tag{2}$$

$$\widetilde{\alpha}^{(r)} = \frac{\widetilde{\alpha}_j^{(r-1)} \left[a_{.j} \cdot \widetilde{x}(\widetilde{\alpha}^{(r-1)})\right]^2}{\sum_{k=1}^{J} \widetilde{\alpha}_k^{(r-1)} \left[a_{.k} \cdot \widetilde{x}(\widetilde{\alpha}^{(r-1)})\right]^2} \tag{3}$$

[2] c_0 is unnecessary to calculate the minimum.

Usage in R

The package consists of a single formula and a sample dataset. The following is the contents of the help pages of Bethel package.

The function *bth*

The input for the procedure consists of two dataframes and a constant:

S A dataframe or a matrix with strata, variances, population size and minumum sample size.

T A dataframe or a matrix with precision levels (coefficient of variation: CV) and totals.

eps The level of precision for the algorithm convergence. The default is 1e-10.

S is composed by a minimum of 6 columns ($ncol(S) \geq 6$), suppose $ncol(S) = k$. The first column shows the strata labels. The $k-th$ column shows the minimum sample rate for each strata, such as 0.04 if the sample will consist of at least 4% in each strata. Similarly, the $(k-1)-th$ column contains the absolute minimum sample size (for example 3 if each strata has to be composed at least of 3 sample units). The $(k-2)-th$ column shows c_h, that is the unit cost per interview (in each strata). Generally this value is equal to 1 to indicate the same cost in all strata. The $(k-3)-th$ column gives the size of the population in each strata. Finally, the estimated variances for the $k-5$ observed variables are shown in columns from the second to $(k-6)-th$. See the example below.

T is composed of 2 columns. The first column shows the coefficients of variation (CV) for $k-6$ analyzed variables (for example, $CV = 0.05$ for each variable). The second column shows the estimated totals for the same $k-6$ variables.

The output of the procedure is the dataframe with the Bethel sample size (*bethelNum*) and the minimum sample size (*bethelNum2*). If the sample size (in a generic strata) according to the Bethel algorithm is equal to 2, then *bethelNum* will be equal to 2, but *bethelNum2* will be 3 if, for example, 3 is the minimum sample size specified in the $(k-1)-th$ column of S. The distinction between *bethelNum* and *bethelNum2*is not considered in the original Bethel's article, but it is necessary from a practical point of view. The procedure of Bethel may indicate an optimum sample size (in a generic strata) of one unit, but generally you need a larger sample (at least 3 / 4 units) to estimate the variability.

The *pop* dataframe

A dataset with 1000 individuals (classified according to sex (M,F) and geographical area (area1 to area4)) in which we have collected yearly data on the following variables:

1. income;
2. number of books read;
3. total days of sporting activities.

Examples in R

To run a survey and to obtain the total estimates of these 3 variables (total income,total number of book, total number of days) we calculate the sample size to obtain, for example, a precision level (coefficient of variation) of 0.05:

```
>library(bethel)
>data(pop)
>attach(pop)

>#Calculate the dataframe with:
>##- strata labels
>##- estimated variances
>##- number of population units

>b1<-as.data.frame(cbind(var_Income=tapply(income,strata,var),
+ var_books=tapply(books,strata,var),
+ var_days=tapply(sportDays,strata,var),
+ num_units=tapply(sportDays,strata,length)))
>b1<-cbind(strata=row.names(b1),b1)
>row.names(b1)<-NULL

>#Add 3 columns:
>##- unit cost per interview
>##- minimum sample size n/N (where N is the population size)
>##- minimum sample size n

>b1<-cbind(b1, c=rep(1,8), n=rep(3,8), n_2=rep(0.04,8))

>#Dataframe with:
>##- precision levels (coefficients of variation)
>##- total estimates

>b2<-as.data.frame(cbind(CV=rep(0.05,3), tot=colSums(pop[,2:4])))

>#Bethel sample according to a precision level (CV) of 0.05

>bth(b1,b2)

 strata numBethel numBethel2
1 F_area1        25         25
2 F_area2        21         21
3 F_area3        19         19
4 F_area4         7          7
5 M_area1        21         21
6 M_area2        29         29
7 M_area3        30         30
8 M_area4        10         10

>#Bethel sample according to different precision level (CV)
```

```
>b2<-as.data.frame(cbind(CV=c(0.05,0.01,0.2), tot=colSums(pop[,2:4])))
>bth(b1,b2)

strata numBethel numBethel2
1 F_area1       120        120
2 F_area2        85         85
3 F_area3       105        105
4 F_area4        37         37
5 M_area1        96         96
6 M_area2       137        137
7 M_area3       149        149
8 M_area4        45         45
```

References

[1] Bethel, J.W. (1989), *Sample Allocation in Multivariate Surveys*. Survey Methodology, Vol. 15, pp. 47-57.

[2] Chromy, J. B. (1987), *Design Optimization With Multiple Objectives*. Proceedings of the Section on Survey Research Methods, 1987. American Statistical Association, pp. 194-199.