DETERMINING AN OPTIMAL SPLIT FOR A LENGTHY QUESTIONNAIRE

Dhiren Ghosh, Synectics for Management Decisions, Inc. and Andrew Vogt, Georgetown University
Andrew Vogt, Department of Mathematics, Georgetown University, Washington, DC 20057-1233

Keywords: Multiple regression, double sampling, Consumer Expenditure Survey

Introduction. Many surveys collect a large amount of data through repeated personal interviews. One way to reduce the response burden is to split a lengthy questionnaire into two or three parts and ask questions from only one or two parts to some interviewees. The issue then is how best to split the questionnaire.

Both theoretical issues such as how to utilize double sampling methodology and practical issues such as the desirability of topical continuity play a role in determining the optimal split. The variety of competing criteria leads us to propose an approach based on good judgment and subject matter knowledge. We illustrate this with data from the Bureau of Labor Statistics' Consumer Expenditure Survey.

Splitting. A split of a questionnaire is a decomposition of the questionnaire into possibly overlapping sets of questions, with each set to be asked to a subsample of the original sample of interviewees. The subsamples are disjoint and their union is the entire sample. Since we do not intend to separate questions that fall within the same section of the questionnaire (in our case), and since we will work directly with summary variables that combine responses to questions in a single section, we can regard a split as a decomposition of the summary variables into possibly overlapping sets of summary variables.

The most naive type of split is to decompose the summary variables into two nonoverlapping sets X and Y. Then questions in the sections corresponding to summary variables in X are asked to units in one subsample and like questions associated with Y are asked to units in another complementary subsample. The subsample mean of each summary variable in X or Y can be used as an estimate of the corresponding population mean. Such a split decreases respondent burden, improves the quality of the response, and decreases nonsampling error. However, it also reduces the sample size for each summary variable, makes no use of relationships between variables in X and variables in Y, and increases the sampling error. If, for example, the sample is divided into two equal subsamples, the standard error of each sample mean is multiplied by a factor of $\sqrt{2}$ ($\cong 1.414$).

In some situations frame variables (variables previously known) are available that are correlated with the variables in X and Y, or other variables that so correlate may be collected from every unit. These data can be used to lower the sampling error. Various double sampling methods are available for this purpose, including multivariate ratio estimates (Olkin [6]), difference estimates (Des Raj [3]), multiple regression estimates (Khan and Tripathi [4]), and other regression estimates (those of B. Ghosh described in [5]). Here we utilize multiple regression.

Suppose Z is a set of variables whose values are known or collected on the entire sample. Then the split is $X \cup Z$ and $Y \cup Z$ where X, Y, and Z are nonoverlapping sets of variables that encompass the entire questionnaire (and possibly some additional frame variables). In the first subsample for every X variable a multiple regression is fitted with the Z variables as covariates. The multiple regression yields an improved estimate of the population means of the X variables, one that takes into account the values of the Z variables in both the first and second subsamples. This is the standard double sampling methodology. Similarly the population means of the Y variables can be estimated by double sampling.

The above methods do not use correlations between X and Y variables. Another simple split that does use this feature is $X \cup Y$, X, Y In the first subsample for every X variable a multiple regression is fitted with the Y variables as covariates. The multiple regression obtained is used to estimate the population means of the X variables, taking into account the values of the Y variables in both the first and third subsamples. These double sample estimates are combined optimally with the means obtained from the second subsample to obtain overall estimates for the population means of the X variables. A similar method is applied to obtain overall estimates for the population means of the Y variables.

A method that includes the features of both of the last two methods is to use a split of the form $X \cup Y \cup Z$, $X \cup Z$, $Y \cup Z$. In the first subsample for every X variable a multiple regression is fitted with Y and Z variables as covariates. The multiple regression obtained is used to estimate the population means of the X variables, taking into account the values of the Y and Z variables in both the first and third subsamples. These double sample estimates are combined optimally with the means obtained from the second subsample to obtain overall estimates for the population means of the X variables. A similar method is applied to obtain overall estimate for the population means of the Y variables. The Z variables are measured in each subsample and their population means can be estimated by the ordinary sample mean.

The Estimators and their Estimated Precisions. Hereafter we suppose that the X∪Y∪Z, X∪Z, Y∪Z subsamples have respective sizes n_1, n_2, and n_3.

For the population mean of each X variable we propose to use an estimator of the form:

$$\bar{x}_c = \frac{V\bar{x}_{ds}}{V_{ds}+V} + \frac{V_{ds}\bar{x}}{V_{ds}+V}$$

where $\bar{x}_{ds}$ is a double sampling estimate based on the X,Y, Z subsample and the Y, Z subsample and $\bar{x}$ is the regular mean on the X, Z subsample. V_{ds} and V are the respective sampling variances of the two estimators. The combined estimator $\bar{x}_c$ is the well-know optimal combination[1] of two independent estimates. Its sampling variance is given by:

$$S^2 = V_c = \frac{1}{\dfrac{1}{V_{ds}} + \dfrac{1}{V}}$$

The sampling variances V and V_{ds} are given by:

$$V = \frac{S_x^2}{n_2}$$

and

$$V_{ds} =$$

$$\frac{S_x^2}{n_1}(1 - R^2)(1 + \frac{n_3}{n_1 + n_3}\frac{p}{n_1 - p - 2})$$

$$+ \frac{R^2 S_x^2}{n_1 + n_3}$$

In these formulas S_x^2 denotes the population variance of one of the X variables x. The number p denotes the number of variables in Y∪Z, and R is the multiple correlation coefficient of x with the variables in Y∪Z as covariates. The formula for V_{ds} is to be found in Cochran [2, p. 340] and in [4]. It assumes a multiple linear regression model for x in terms of the variables from Y∪Z, with a joint multivariate normal distribution for all variables and the residuals. The proof of this formula uses the theory of double sampling, as well as of multivariate normality properties considered in [1].

The overall sample size is taken to be $n_1 + n_2 + n_3 = N$ with N fixed for various choices of n_1, n_2, n_3. The estimated precision $S = \sqrt{V_c}$ is compared with

$$S_0 = \frac{S_x}{\sqrt{N}}$$

The relative precision is the ratio of S to S_0. Note that this ratio does not depend on S_x.

Choosing X, Y, and Z. Multiple criteria govern the splitting of a questionnaire, and thus the decision for how to split involves judgment rather than rigid application of simplistic quantitative measures.

One criterion is that thematically closely related questions should remain in the same part. Thus, for example, it is not advisable to split automobile expenses so that one person is asked how much is spent on gasoline but not how much is spent on oil and another is asked how much is spent on oil but not how much is spent on gasoline. It is easier to answer questions that center upon the same topic: the thought process used to formulate one answer can assist in the formulation of the other; joint consideration of both can correct errors of proportion; records and/or memories can be consulted that bear on both answers. Conversely, when one does not ask thematically related questions, the alternative is to ask unrelated ones. Jumping from one theme to another can decrease attention, concentration, responsiveness, and cooperation.

Opposed to this criterion, and a source of tension, is the criterion that closely correlated answers are somewhat redundant. If one answer is known, the other can be predicted. That being the case, why ask both? If questions leading to correlated responses are asked to different subsamples, the correlation can be used to supply the missing answer. However, thematically related questions often have answers that are statistically correlated.

Another criteria is to keep questions that have low correlation together since the answers tend to be independent and the combined information accordingly much greater.

[1] This combined estimate is not the only one possible for a given split. More elaborate estimators formed by linear combinations of all subsample means of all variables could be developed, but we do not pursue this development here.

Variables that have many zero values are usually sampled more often to increase the likelihood of acquiring the non-zero values. Variables with larger variances or larger coefficients of variation are usually sampled more often.

The Consumer Expenditure Survey. The Consumer Expenditure Survey, conducted by the Bureau of Labor Statistics, is a survey of expenditures by American households (consumer units). It consists of two parts: a quarterly interview survey and a weekly diary. The quarterly interview survey is used as an example here.

Each month a panel of consumer units is interviewed. An interviewer visits the same consumer unit every three months for five successive quarters and administers a questionnaire to a member of the unit that concerns expenses in the preceding three months. New panels are introduced into the interview sample on a monthly basis, as other panels complete their participation.

The first interview gathers general information about the unit. In subsequent interviews the full questionnaire is administered. This questionnaire has 24 sections, some of which include very precise and detailed subsections.

The sample design is a multi-stage cluster sample of households, a standard design for household surveys. Approximately 1800 usable interviews on average took place each month in the years 1996 and 1997. The respondent provides information about spending in the preceding three months. Data from earlier interviews, including the initial interview, are used to assist accurate recall.

However, interviews take about 1.5 hours on average. This is a high respondent burden, adversely affecting the quality of the data and increasing the drop-out rate of units.

To alleviate the response burden, it is desirable to split the lengthy questionnaire so as to reduce the burden for some or all respondents. Such a split should also attempt to minimize the loss of precision. Below we outline the development of such a split and offer consequent estimates of precision.

Assumptions. The data on which the analysis was based are the 1997 Interview Survey Public Use Microdata. We adopted as a starting point the principle that sections of the questionnaire will remain intact in any split and that the rotating panel design will not be substantially altered. To represent sections we used summary variables, i.e., for each section a single numerical variable that sums the chief expenses reported in the section. For example, section 2 concerns rented living quarters. Our summary variable for this section is the sum of all rental payments made in the reference period, adjusted for business and rooms rented to others, and summed over all members of the consumer unit.

A few sections were omitted since they contain general information or non-expense data (e.g. credit information). A total of 19 sections were studied by means of the corresponding 19 summary variables.

Statistics about these 19 variables were gathered from the consumer units in the 1997 microdata. The number of interviews in these data was 22,183, an average of 1849 per month. For simplicity, these 22,183 interviews were treated as if they were a simple random sample representing an average quarter in a year.

Based on the statistics of the summary variables, subsamples of a hypothetical monthly sample of size N = 1800 are to be proposed and splits (discussed below) of the questionnaire sections are to be derived to be administered to these subsamples. The monthly sample is treated as if it were a simple random sample, and the subsamples as if they were random subsamples from the sample.

The subsamples may also be regarded as independent of each other since the population is large.

To assess the effect of the splitting, we developed estimators that estimate the population mean of each summary variable. These estimators were based on the sample data, and estimates of their standard errors were developed for a range of subsample sizes. These standard errors are indications of how the splitting will affect the reliability of individual variables within the corresponding section.

The subsample sizes and questionnaire splits should offer a balance among various competing criteria discussed below.

Our analysis was cross-sectional. We did not investigate the relationships between monthly samples in successive months or quarters, or between interviews of the same consumer unit over time. Although temporal correlations between variables can be used to improve estimators, they necessitate significant adjustments in the panel design and they were not considered in the basic splitting decision described in this paper.

Relative Precisions. The summary variables were divided into three groups X, Y, and Z. Table 2 below shows the decomposition that gave the most favorable result. The variables in Z are measured on all consumer units and hence the standard error of their sample means is unchanged. (Recall that the overall sample size is taken to be 1800.) For each of the X and Y variables we computed the value S_0 of the estimated standard error of the sample mean in case the variable was measured on the full sample of size 1800. When the variables in X and Y were measured respectively on subsamples of size $n_1 + n_2$ and $n_1 + n_3$, where $n_1 + n_2 + n_3 = 1800$ and $n_2 = n_3$, we computed the ratio of the new standard error S to S_0. This ration is called the DEFT (the square root of the design effect). The ratio is larger than 1 and reflects the loss of precision that results from the splitting of the questionnaire.

Table 1 below summarizes the results for the split in Table 2. It shows the range of relative precisions, i.e., the DEFT, for the X and Y variables for different choices of n_1, n_2, and n_3, with $n_2 = n_3$. It is apparent that these ranges are fairly narrow and that the use of double sampling to compensate for the splitting gives only a somewhat modest improvement over 1.414 (corresponding to the half-sample split mentioned above).

Table 1
Relative precision (i.e. DEFT) when double sampling is used.

n_1 X, Y, Z	n_2 X, Z	n_3 Y, Z	range of S/S_0 for X and Y variables
200	800	800	1.30-1.34
400	700	700	1.22-1.27
600	600	600	1.16-1.22
800	500	500	1.12-1.17
1000	400	400	1.09-1.13
1200	300	300	1.06-1.09

References

[1] Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*, Second Edition, John Wiley & Sons, New York, 1984.

[2] Cochran, W. *Sampling Techniques*, Third Edition, John Wiley & Sons, New York, 1977.

[3] Des Raj. On a method of using multiauxiliary information in sample surveys, JASA 60, pp. 270-277, 1965.

[4] Khan, S. and T. P. Tripathi. The use of multivariate auxiliary information in double sampling, J. Indian Statistical Association 5, pp. 42-48, 1967.

[5] Murthy, M. N. *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta, India. 1967

[6] Olkin, I. Multivariate ratio estimation for finite populations, Biometrika 45, pp. 154-165, 1958.

Table 2
List of Subject Areas by Split

Subject Area	Split
Rented Living Quarters	Z
Owned Living Quarters and Other Owned Real Estate	Z
Utilities and Fuels for Owned and Rented Properties	Y
Construction, Repairs, Alterations, and Maintenance of Property	Z
Appliances, Household Equipment, and Other Selected Items	Y
Household Equipment Repairs, Service Contracts, and Furniture Repair and Reupholstering	Z
Home Furnishings and Related Household Items	Y
Clothing and Sewing Materials	X
Rented and Leased Vehicles	Z
Owned Vehicles	Z
Vehicle Operating Expenses	X
Insurance Other than Health	X
Hospitalization and Health Insurance	Z
Medical and Health Expenditures	Z
Educational Expenses	Z
Subscriptions, Memberships, Books, and Entertainment Expenses	Y
Trips and Vacations	X
Miscellaneous Expenses	Y
Expense Pattern for Food, Beverages, and Other Selected Items	X
Family Size	Z
Work Experience and Income	Z