Martin R Frankel
City University of New York*
USA

6.2 Ascription in magazine audience research

INTRODUCTION

The *Oxford English Dictionary* defines ascription as "The action of ascribing, attributing, imputing or declaring that something belongs to a person or thing."

In the most recent edition of *A Dictionary of Statistical Terms*, by Kendall and Buckland, one finds neither ascription, nor imputation nor attribution. This omission provides appropriate testimony to the status of the subject in the formal statistical literature.

The body of mathematics which constitutes the basic theory of probability sampling assumes a data collection process which produces complete and unbiased information for all population elements that are selected into a sample. The words non-response, non-cooperation, missing data, incomplete data or partial data are themselves missing from the standard vocabulary of theoretical probability sampling.

In the world of media research, and particularly in the world of magazine research as practised in the US, the situation is somewhat different. Some of the respondents selected into our samples simply refuse to cooperate at all. Others will cooperate, but only in a partial fashion. Unlike the users of syndicated studies which measure television and radio, users of syndicated magazine studies have come to expect that magazine audience estimates will be accompanied by information about other media exposure and general product consumption. This makes it necessary to ask for respondent cooperation on more than one occasion. Unfortunately not all of our respondents who cooperate the first time will cooperate at subsequent times.

As a survey sampling statistician who was first exposed to the particular problems of magazine audience research only four years ago, I found it somewhat ironic that there seemed to be a number of commonly accepted and practised methods for dealing with the problem of total non-response, but the methods and procedures for handling partial non-response or incomplete response were, at best, *ad hoc* algorithms which could only be deciphered by examining computer codes and were unaccompanied by anything approaching a well-defined statistical justification. It seemed to me at the time that the task of inference in the absence of any data should be more complex than the task of inferences in the case of partial data. I subsequently discovered that this state of affairs was not a function of the complexity of the problem but more a function of the complexity of the solutions.

In the US it is standard practice to compensate for the absolute and relative impacts of total non-cooperation or total non-response by the application of final ratio weighting factors which bring certain weighted totals into agreement with externally available population estimates. This may be accomplished by explicit ratio adjustment on a cell by cell basis or by sample balancing, a term suggested by Jerry Greene to describe the successive ratio adjustment of marginal totals which was first described by Deming and Stephan in the 1940s. (*1*)

Methods used to adjust for partial non-response are much less uniform. When respondent level micro-data files are not made available to users, the research organisation may make use of a number of adjustments including question-specific weighting, intermediate ratio adjustment or item ascription. When respondent level micro-files are to be made available to users, adjustment for partial or item level missing data takes the form of ascription. By ascription I mean the assignment of numerical values to specific items within respondents for whom actual responses should exist, but do not.

Having felt the same way at one time, I am not surprised that researchers who are quite comfortable with the use of other adjustments for partial non-response or missing data items, may become somewhat sceptical when the word ascription is mentioned in the context of a survey. With this in mind I would first like to review the alternative approaches that may be applied to the problem of missing data at the within respondent level. Considering the available alternatives I expect that you will accept by conclusion that in many situations ascription is the most practical and the most statistically appropriate alternative for dealing with this problem. I will then briefly review some of the presently available techniques for ascription noting their strengths and limitations. Finally I will be describing a method of ascription recently developed at SMRB which represents a significant improvement over presently available methods from both a statistical and practical perspective.

ADJUSTMENTS FOR ITEM NON-RESPONSE

As a simple means of exploring the issue of why some type of adjustment might be appropriate under conditions of

* *Professor Frankel is Consultant to the Simmons Market Research Bureau.*

item level missing values, let us consider a somewhat oversimplified case. Suppose we are undertaking a simple readership survey involving three magazines, called A, B and C. Further, let us assume that our average readership measure is based upon a single interview from a cross-sectional sample of 100 respondents. After data entry, cleaning and global weighting we must deal with some respondents who have not answered all of the readership questions. Let us suppose that three of our respondents did not answer the readership questions for magazine A, three of the respondents did not answer the readership questions for magazine B, and one of our respondents did not answer the readership questions for either magazine A or magazine B. We have several basic options:

(**1**) if any readership questions are unanswered, throw out the entire case.

(**2**) retain all the cases and allow a non-response category in data tabulations.

(**3**) retain all the cases, but restrict the readership tabulations to responding individuals. Apply a ratio adjustment on a magazine-by-magazine basis by determining a rating and then multiplying by the appropriate population projection.

(**4**) apply a separate weighting procedure for each of the specific readership questions.

(**5**) ascribe or impute a response to the readership questions which were not answered by the respondent.

Throw out the entire case if any items are missing

Let us consider the first option, if any missing data is present, throw out the case. Under the assumption that missing data is independent of both readership and demographics or anything else, the procedure is not biased but it is rather wasteful. Under the best of assumptions we are simply lowering reliability (ie increasing standard error). In surveys which ask a large number of questions, the probability of at least one item specific non-response may become quite large and we could find ourselves throwing out a significant portion of our sample. Under the more realistic assumption that item non-response may be related to either background or outcome variables, the removal of an entire case can introduce statistical bias into our estimates in addition to lowering reliability. For example, if the probability of skipping readership questions was higher for downscale persons and if magazines A and B tended to be broad based while C was upscale the use of this strategy would tend to inflate the audience of magazine C at the expense of magazines A and B. (*2*)

Allow and report a non-response category

The strategy of retaining all cases and exhibiting a non-response or no-answer is often thought of as the statistically appropriate way to deal with non-response. Basically we pass the buck and transform our problem into the users' problem. Very sophisticated users may apply some type of simple *ad hoc* correction similar to our third or fourth option, but for the most part users will simply ignore the non-response and effectively penalise those magazines which have the misfortune of being affected. As the use of magazine and other media audience data becomes more and more computer assisted, this penalisation of magazines which happened to have higher item non-response becomes more probable. It is just too complicated to expect the consumer of audience estimates to program the full set of contingency corrections for non-response and the question level. Again, under the assumption that item non-response is random and uncorrelated with any variable of interest the impact of this type of strategy and the subsequent non-use of any correction will be a downward bias that is similar for all magazines. However, under the more realistic assumption that item specific non-response may be related to background and outcome variables the use of a non-response reporting category and its probable non-use by research consumers will result in downward bias that is variable across different magazines.

Under the assumptions that non-response tabulation categories are simply ignored by users the assumptions mentioned for the first strategy alternative will lead to a larger downward bias for magazines A and B than for C. (*3*)

Ratio adjust individual questions

The third option which specifies that we restrict readership tabulations to responding individuals, compute a rating and then multiply this rating by the appropriate population projection, provides a reasonable solution for estimating audiences under the assumption of independence between item non-response and any background or outcome variable. But, to the extent that item non-response is not independent across all respondents and magazines, we may be exchanging some of the downward bias associated with strategy two for differential bias, both upward and downward among different magazines.

Even if we accept this strategy as statistically appropriate, which is in fact not the case, we are describing a procedure which is cumbersome in its implementation. Most tabulation software presently in use accumulates weighted counts within a predefined cell structure. Total population projections are simply the sum of the weights for all sample cases. Total audience projections are sums of weights restricted to respondents who are readers. Estimation of a rating value for each magazine would require an additional vector of cells (one

for each magazine) to accumulate weighted counts of respondents who responded in some way to each specific magazine question. This would essentially double the number of storage cells required for any tabulation.

While this requirement would not, and indeed on its own, should not, cause us to reject such a strategy, there are a number of other more practical problems. First, some type of tabulation strategy would have to be adopted for the estimation of audiences on a sub-class basis (eg males and females). Assuming that this strategy were extended to sub-classes, users of our printed reports would find that audiences for mutually exclusive and exhaustive sub-classes would not add to total audiences. By appropriate subsetting of the data, audience estimates derived by summing across sub-groups could be artificially increased or decreased. This multiplicity of estimates available to users of micro-data files would, no doubt, make the survey organisation somewhat hesitant about release of micro-data tapes.

Weight each question on an individual basis

Option four, the development of weights on a magazine specific basis provides a solution to the problem of missing item level data which removes some of the basic theoretical objections to options one through three above. When properly implemented this weighting strategy can come to grips with both overall bias inherent in the use of a non-response tabulation category and the differential bias that arises due to the differential interaction among demographics and readership across magazines. If the item non-response for a particular magazine is more concentrated among younger males, the use of weights defined on the basis of sex and age will yield projected audiences that add across age by sex sub-classes and remove some of the differential bias across the age by sex sub-classes. Decisions concerning the depth of the weighting sub-classes (ie the number of dimensions to be used in weighting) must, of course be made with recognition of the potential increase in standard error that occurs when the degree of weighting is increased. It is my belief, however, that a producer of syndicated audience estimates must adopt a utility function that minimises as much as possible differential bias across the magazines measured, even when the result of this minimisation may be an increase in random statistical error.

While the use of separate weights for each magazine provides a theoretically appropriate method of dealing with item non-response, the use of such a procedure is from a practical standpoint, impossible. Practical and existing tabulation software is not set up to make use of different question-specific weights in a simple and straightforward fashion. While it can be modified to make use of these separate weights, we essentially double the storage requirements for any tabulation. Of a more limiting nature however, is the fact that estimates of more complex statistics are not well-defined. A very simple example of these more complex statistics is the pairwise inter-title duplication rate required for scheduling models. Even the apparently straightforward strategy of restricting the estimation of a duplication rate to respondents who have provided answers to both magazine questions becomes somewhat problematic when we realise the sum of weights for qualifying cases will not match any other projections for total populations or sub-classes. Thus, if we are to develop separate weights for individual magazines we should also be developing weights for all possible pairs of magazines, all possible triples of magazines, etc.

Ascribe or impute the missing value

The final option is that of ascription or imputation. Under this strategy we retain all cases, but attach data values to those items for which values are not present. In this way tabulations may be accomplished in straightforward manner from all cases in the sample, without the need to perform separate ratio adjustments, or use question-specific weights. From a mathematical standpoint, ascription may be viewed as a form of data weighting of the type described in some of the options above. This realisation that properly executed ascription is an alternate form of data weighting provides a mechanism for its statistical justification. If, for example, we were to impute missing item values by a random selection among all respondents for which a response existed we would be following strategy 3 at an overall sample level in much the same way that weighting was accomplished in the days of counter sorters. If we subdivide the sample cells and use stratified random selection of actual responses to stand in for item non-responses we are using a variant of strategy 3 or 4. Given that we can specify a question-specific weighting process that satisfies our statistical concerns about bias correction due to missing data, we can transform this weighting process into an ascription algorithm. The only real theoretical difference between the resulting ascription and the use of separate question weights is the potential for a small increase in standard error associated with the use of integer rather than fractional weight values. Even if this increase in standard error is present, it represents a small theoretical price to pay for gaining a practical solution to the missing data problem.

STANDARD METHODS OF ASCRIPTION

Although the techniques of ascription, or as it is sometimes called, imputation, have not found their way

6.2 Ascription in magazine audience research

into the mainstream of the survey sampling literature, the practice of these methods has existed (sometimes in out of the way and darkened computer rooms) for at least four decades. One of the more prestigious organisations to make use of these methods is the US Bureau of the Census. It is perhaps indicative of a trend toward new found respectability that a recently released 171 page document describing the Current Population Survey devotes an entire paragraph to the fact that 16% of all persons do not report complete income information, and that a new procedure was instituted in 1968 to improve the imputation of this missing information. (5)

In general, ascription and imputation methods are usually classified as being either model based or empirical. Some of the more commonly used techniques include:

(**a**) random sampling from the total sample of responses.
(**b**) random sampling of responses within cells.
(**c**) cell mean.
(**d**) hot deck – man (person) next door.
(**e**) external data – cold deck.
(**f**) linear model – regression with/without random error.

Random selection of responses

We have already mentioned the first method which is, at an overall sample level, similar in impact to an overall ratio correction on an item-specific basis. The major disadvantage of this method arises in the development of projections for sub-classes. To the extent that item non-response is correlated with background variables (eg demographics) our sub-class estimates will suffer from some degree of differential bias.

Random selection within cells (strata)

It is possible to eliminate some of this differential bias by restricting the pool from which an imputed value will be selected to individuals within the same predefined sub-set of the sample. This random sampling from responses within cell is quite similar in impact to post-stratification weighting on an item-specific basis. The major problem associated with this method is the limitation on the number of cells that may be defined. We must be certain that there is at least one valid response within each cell which will be used. One of the desirable features of this method with respect to magazine research audience surveys is implicit control over inter-title duplication rates. It is well known that duplication between pairs of magazines is not a random phenomenon. Yet, when imputation is based upon random sampling from the total sample of responses, resulting duplication rates may be shifted toward independence in much the same way that the addition of random terms induces attenuation in correlation. By selecting imputed values within sub-groups we will still produce effective random duplication within sub-groups, but if sub-groups are chosen in a way that is correlated with readership, the overall sample impact is the maintenance of the non-independent duplication rates on an inter-title basis.

Cell mean

A variation on this general cell approach is known as the cell mean method. Rather than randomly selecting a single response within a cell for each imputation, we first compute the mean for each question within the cell and use this mean value when an imputation is required. Recent work in the area of James-Stein estimation by Effron and Morris suggests that this method of ascription may lead to somewhat more precise estimates at the univariate level. From a practical standpoint, this approach has several basic drawbacks. First, most tabulation programmes used in magazine research are set up to deal with zero-one variables for readership computations. The cell mean approach would force us to impute proportion values somewhere between zero and one. This would basically change the format of the required data files. In addition to producing imputed values which were not in the same form as data from the questionnaire, the use of cell means would unduly complicate the computation of turnover rates, duplication rates and various cross-tabulations in which the entry criterion was readership for a particular magazine.

Hot deck

The hot deck approach is still another variation of the general method of selecting imputed values from sub-sets or cells of the total sample. The popularity of this method is probably due to its extensive use by the US Census Bureau. Its use at the Census Bureau was most likely due to its simplicity of implementation in a sequential data processing environment.

The method requires the definition of a mutually exclusive, mutually exhaustive cell structure. The data cases are sorted by ultimate sampling unit (cluster or segment) nested within PSU and nested within broad geographic region. Starting with the beginning of the file, a determination is made of the cell to which the case belongs. All response data for the case is stored on a temporary basis, replacing any response data that might have been stored from a preceding case within the cell. When a case is found with a missing response item, the item value held in the temporary storage area is used for ascription.

One of the basic appeals of this method would seem to be related to its implicit geographic control over and above the control brought about by the basic cell structure. In practice, however, the distribution of specific item non-response will be such that 'donor' values do not come from either the same cluster or even the same PSU. Anderson, Kasper and Frankel found that in sparse

imputation situations, with an extensive cell structure, the geographic ordering may have a significant impact on the resulting survey estimates. (*6*) This situation is simply the result of the fact that the geographically nearest non-missing data value may be quite far removed from the case requiring imputation. In addition, when sparse imputation is attempted with an extensive cell structure, the basic geographic sorting of the data file must be modified in order to assure that at least one response item is available prior to the arrival, in the sequence, of an item for which ascription is required. Unless this modification is introduced, certain items will not have required donors.

Cold deck

A further variation of the within cell imputation model does not use data from the same survey, but instead makes use of item values from either past surveys or other data sources. There are several appealing features of this method, but it has not seen widespread use in the United States.

With the exception of the cell mean technique, the methods described above have the common feature of developing ascribed values from individual respondent observations. The cell mean technique makes use of actual observations in aggregate form. For this reason these models are often described as empirical methods for ascription.

Regression-linear model

Another class of techniques which is sometimes employed for ascription may be generally described as regression methods. Based upon least squares fitting, models are developed which can predict each of the data items that may be subject to missing values. These predictive equations are then used directly (with their actual predicted y values) or with the addition of a random error term. This random error term is introduced when it is necessary to preserve the same variation that exists among respondents who provided responses to the actual items. If this error term is not included, estimators which are descriptive of relationships (eg correlation coefficients) may be subject to either attenuation or ecological effects.

Model based ascription procedures permit a much larger number of variables to be used in the development of imputed values than is possible with traditional cell approaches since it does not involve the specification of a mutually exclusive and exhaustive cell structure. This very desirable feature has its price, however, since we are forced to accept a predetermined data generation structure that is most often linear and does not include interaction. Thus, for example, if on a marginal basis having the attribute male increases the probability of reading a certain magazine by one unit and having at least a college degree increases reading levels by two units, we will add three units whenever a missing readership question occurs for a male with a college degree. In reality however if we examine the male by college graduate cell we might find that the increase in readership level is only 2.1 units.

EXTENDED CELL ASCRIPTION

Extended cell ascription is a method which retains the more realistic data structure assumptions of the cell-specific imputation methods while significantly increasing the number of variables on which the ascription is based. As was mentioned earlier, cell-specific imputation procedures are severely limited in their number of dimensions and number of cells due to the necessity of having at least one actual response to each item within each of the imputation cells. Typical standard cell ascription methods are usually limited to four or perhaps five dimensions, which result in between 100 to 300 implied cells. Through the use of extended cell ascription, SMRB has, for the past two years, been able to base its market book ascription process on 23 dimensions involving in excess of 1,000,000 implied cells. These dimensions include sex, census division, household income, presence and number of children, employment status, number of adults in household, home ownership, locality, race, marital status, person in household, age, education, occupation, dwelling type, readership of four separate magazine types, hours of TV viewing, county size and sampling stratum. Those of you who are familiar with the SMRB sample will recognise that the number of implied cells greatly exceeds 15,000, the approximate number of SMRB respondents in a given survey year.

The vast majority of cells contain no respondents and thus it is never necessary to select a response to ascribe to an item non-response (in this case a product diary). For those cells which contain both respondents with responses and respondents requiring ascribed responses the ascription is accomplished by random selection. It is only those cells which contain respondents requiring ascription but no respondents with actual response values that has been the limiting boundary of standard cell approaches. With extended cell ascription techniques it is possible to specify a series of predetermined searches that are undertaken in order to collapse neighbouring cells until at least k potential donors are found. This search is carried out in a way that allows various forms of priority ordering in the basic dimensions and their categories.

It is not surprising that once the extensive software required for implementing extended cell ascription had been developed we quickly found a number of more general applications. For a number of years SMRB has

collected information about radio listening and television viewing from a stratified sub-sample of its basic sample. Over the years, the size of this sub-sample has varied from 4500 to 6000 relative to a total magazine sample size of 15,000. The traditional approach to tabulation and data release was to apply a separate weighting process to the sub-sample and to base printed tabulations on the sub-sample alone. Data release was also restricted to this weighted sub-sample. The basic problem with this approach was that overall magazine audience levels derived from this data base were different from the magazine audience levels shown in the basic Simmons reports and available from the total magazine data base. While these differences were essentially within expected sampling limits, the fact was that magazine levels contained in television by magazine and radio by magazine cross-tabulations were significantly less reliable than magazine levels available elsewhere in the study. We found that users were not, in general, willing to apply appropriate correction factors. The solution to this problem was to treat the respondents not in the radio television sub-sample as respondents for which this data was missing and apply extended cell ascription. In this way, intermedia relationships between radio and magazines as well as television and magazine are preserved but marginal magazine levels are based upon the full sample of 15,000 respondents.

Once we became convinced that extended cell ascription was both practical and statistically appropriate when the presence of missing data was not a data collection phenomenon but rather a sampling phenomenon, the next logical step was its use as a method of merging external samples into the general SMRB data base. We are presently involved in a number of pilot experiments which should permit organisations that work with in-depth brand-specific studies to integrate these studies with the SMRB media and consumption data. This will allow the development of extensive media profiles for specific brands and subsegments within brand.

REFERENCES

1 Deming, W Edwards and Stephan, F F (1942). 'On a least squares adjustment of a sample frequency table when the expected marginal totals are known *Annals of Mathematical Statistics* **XIII** 2, June, pp 166–178.

2 Sample individuals who are downscale have a higher probability of being removed from the sample. This will tend to make the sample contain a disproportionately high number of upscale persons.

3 Downscale respondents have higher probability of missing questions and relatively higher probability of being readers of magazines A and/or B, rather than C.

4 For example, if the rate of reading and item non-response is different for males and females, the total sample rating will be computed in a way that assumes that the reading rate for both sexes is the same.

5 US Bureau of the Census (1977). *The Current Population Survey: Design and Methodology* p 87.

6 Andersen, R, Kasper, J, Frankel, M R *et al* (1979). *Total Survey Error* Jossey-Bass.