

LI Analysis Training Series

Data Cleaning and Replacement of Missing Values

(Last Revised: 6/24/99)

Melissa McDermeit, Rodney Funk and Michael Dennis

Chestnut Health Systems

Bloomington IL 61701

309-827-6026

www.chestnut.org

Acknowledgement: This document was developed under grant No. TI11320 from the Center for Substance Abuse Treatment (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA). The opinions expressed here are solely those of the authors and do not reflect official positions of the U.S. Government.

Purpose: To provide methods of handling missing and inconsistent data. Consistency codes that tell the analyst the type of missing data, will be covered first. Then the replacement of those codes based on type of question and other cleaning issues will be discussed. The methods described herein are based on data cleaning done with the Global Appraisal of Individual Needs (GAIN; Dennis, 1998).

Background: Missing data is a common problem, and the best approach to minimize the problem is through careful administration and/or quality assurance. Rates of less than 1% missing data are generally considered trivial, 1-5% manageable. However, 5-15% require sophisticated methods to handle, and more than 15% severely impact any kind of interpretation. Rates of missing data are usually higher when data are from records, self-reports, or are collected by largely unmonitored staff. Regardless of why there is missing data, it is a problem because most analytic software procedures require observations on all individuals-variables and will use listwise deletion (i.e., dropping all variables for a case where any single variable listed in the procedure is missing) by default. Different people missing data on different items can amount to the loss of a fifth or more of the total sample, significantly reducing statistical power (Dennis, Lennox, & Foss, 1997). Recent work has shown that listwise deletion is less hazardous if it involves minimal loss of sample size (minimal missing data or a sufficiently large sample size) and there is no structure or pattern to the missing data (Figueredo, 1999). For other situations where the sample size is insufficient or some structure exists in the missing data, listwise deletion has been shown to produce more biased estimates than alternative methods (Little & Rubin, 1987).

If only a few percent (<5%) are missing, the data can be replaced using the mean (if normal), median (if skewed) or mode (if categorical). Where the goal is to compare several groups (e.g. gender or treatment conditions), it is often desirable to do this replacement within each group. As the percentage of missing data approaches or exceeds 5% a new problem arises. Replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. It is therefore recommended that data be replaced using one or more of the advanced methods based on hot-deck imputation (used here), multiple imputation (modeling uncertainty due to missing data, while using the existing data (Rubin,

1987)) or a regression model (predicting the missing value based on the other available data). Multiple imputation and regression models are more elegant, but much more difficult because each variable requires a different equation and in many cases multiple equations per variable because some predictors may also be missing. Regression models are also dependent on the order in which variables are replaced. In this paper, we will focus on the less complicated method of hot-deck imputation.

Consistency Codes: Consistency codes are needed to help explain to the analyst why a specific data element is missing. A question could be unanswered for several reasons including: 1) the subject refused to answer, 2) it was legitimately skipped due to a prior response, 3) it was skipped because the subject did not know the answer, or 4) it was not asked in the current version of the questionnaire. Each reason can have different indications on how to replace them. We have used negative numbers for consistency codes, so they will not be confused with legitimate answers. These negative values are then defined as User Missing in SPSS. The values used are: -7 for refused to answer, -8 for don't know, -9 for legitimately skipped and -3 for question was never asked (i.e., due to different instrument versions). If the item response was not in the valid range of responses for a question or a missing value is not already coded using one of the above missing codes, it is assigned a value of -8 (don't know/unknown/missing).

Cleaning the Data: Data is always checked first for illogical or out of range responses. If the correct response can be determined, incorrect values are replaced immediately. (For example, if there were a series of yes/no questions, sometimes the entered value is the number of the item within the series rather than 0 or 1; or year of birth is indicated as 1997, but age is given.) If the correct response can not be determined, the response was set to missing (given a value of -8). For most variables, missing values were later replaced using methods discussed in more detail below.

Missing Data That Should NOT be Replaced With This Procedure . It is important to distinguish between legitimate skips and questions that were never asked or were not applicable (-3). There are two situations where this happens. First, there are some situations where a legitimately skipped question does not have any logical value (e.g, if you never used marijuana, age of first use is not meaningful). Second, if an item was not asked in the particular version of the survey, the value should not be "assumed". Another situation in which the procedures below are inappropriate is when an entire wave of data is missing (e.g, you have intake but no three-month follow-up). Replacement of waves of data (based on other waves of data) will be the subject of a different memo. The following paragraphs discuss the additional replacement procedures.

Recoding Legitimate Skips: For analysis purposes, legitimately skipped questions are set to the logical value of zero (0) except in cases listed above where replacement would not be meaningful. For example, if someone reported never using alcohol (recency question), then 'days of alcohol use in the past 90 days' (frequency question) was legitimately skipped (-9) and should be recoded to 0. Note that it is still possible to consider only days of use for those who had ever used by selecting on the "ever used" (recency) variable.

Logically Replacing Missing Values: Responses to several items depended upon responses to earlier items. If the earlier item was left missing or indicated lower use or frequency than the later item, the earlier item was 'coded up' to reflect the later response. For example, if recency of drug use was left blank, but the client reported 30 days of drug use, recency was coded to having used in the past 30 days. Similarly, if a client reported 10 days of any alcohol use, but 30 days of drinking 5 or more times per day, the value for any alcohol use was coded up to 30. If the responses indicated that there was no recent use (in the past 90 days) and the recency question was missing, recency was randomly replaced with one of the following: never, 4-12 months ago or more than 1 year ago. An exception to random replacement occurred for a few of the recency questions where there was an additional series of questions indicating the last use of a particular substance. These questions were used to logically determine if a recency question should be coded to 4-12 months ago, or more than one year. Then remaining missing values were randomly recoded as indicated above. An example of the code used follows:

```

Compute mp9=p9
do if (mp9u gt 0 and missing(mp9)).
  if (mp9u ge 88) mp9=6.
  if (mp9u ge 83 and mp9u le 87) mp9=5.
  if (mp9u ge 61 and mp9u le 82) mp9=4.
  if (mp9u ge 1 and mp9u le 60) mp9=3.
end if.
do if (missing(mp9) and mp9u=0).
  compute pick=uniform(1).
end if.
end if.

if (pick gt 0 and pick le .33) mp9=2.
if (pick gt .33 and pick le .66) mp9=1.
if (pick gt .66) mp9=0.

```

The variable p9 is the recency question and mp9u is a past 90-day question that has already been median-replaced (explained below). The first 'Do if' statement replaces the recency based on the past 90-day report where: 6 is 1-2 days ago, 5 is 3-7 days ago, 4 is 1-4 weeks ago and 3 is 1-3 months ago. If there are no days reported in mps9u, the recency gets randomly assigned. The second 'Do if' creates a random variable 'pick' that is uniformly distributed between 0 and 1, whenever the 90 days question is 0 and the recency is missing. The last series of 'If' statements then assigns values to mp9 of: 2 being 4-12 months ago, 1 being over a year ago and 0 being never. These are based on the values randomly assigned to pick.

Random Missing Value Replacement. Once the values for items had been logically replaced or coded up from other responses, data were sorted by level of care, gender, race and year of birth. The remaining missing variables were then replaced in one of two ways. For interval data, missing values were replaced with the (rounded) median of the four surrounding values for interval level data. Missing categorical values were replaced with the mode of the four surrounding values. This is accomplished by using the RMV command in SPSS (Version 9.0, 8.0.1 or 7.5). An overview on this command can be found on pages 784-787 in the *SPSS Base 7.5 Syntax Reference Guide (1997)*. Following is an example of the syntax used:

```

missing values s2w (lo thru -3).
sort cases by loc xchk1.

```

```

rmv ms2w=median(s2w,2).
compute ms2w=rnd(ms2w).
missing values s2w (-8).
if (s2w=-9) ms2w=0.
if (s2w=-3) ms2w=-3.
missing values s2w (10 thru -3).

```

The first ‘missing values’ command defines -3, -7, -8 and -9 as missing. The ‘sort’ command sorts the data by level of care and xchk1, which is a variable made up of the clients’ gender, race and age. This puts clients’ records near the records of other similar clients. The ‘RMV’ command creates a variable ‘ms2w’ that has the valid answers from s2w and replaces the missing values with the median of the two cases before and the two cases following the missing value. SPSS will label the new variable with the part after the equal sign in the RMV command. In this case, the label would be ‘median(s2w,2)’. The ‘compute’ statement rounds ms2w into a whole number (.5 becomes 1). The second ‘missing values’ command is so that -9s can be replaced with 0s. This is done with the ‘if’ command. If -9s are still defined as missing, this command will not work. The second ‘if’ replaces the median replaced, if the question was not asked. The -3s needed to be defined as missing or else they will be used in computing the median for replacement. The final ‘missing values’ command simply resets the missing values. The median was used in this particular example due to the skewness of the data. If your data is closer to a normal distribution, the mean can be used instead.

One exception to this format was replacement of missing values for sexual risk variables. For these variables, data were first coded up or set to zero based on logical progression. For example, if a client reported no sex with a man in the past year, missing values for number of male partners in the past 90 days were set to zero. Clients were categorized by current and yearly sexual pattern. Sexual patterns combined gender with sexual orientation based on the gender of the client’s partners (e.g., men having sex with women, men with men, men with both men and women). Missing values for specific sexual behaviors (e.g., frequency of being the penetrating partner in intercourse, receiving end of oral sex) were replaced within gender and recency or within sexual pattern.

Composite Score Replacement Within Individual: Scales often have a lot of missing data because the software will use listwise deletion by default (i.e. everything must be answered). As long as three or more valid answers have been given, missing answers are typically replaced within individual by multiplying the average of the valid answers times the expected number of items. This can be easily accomplished by using the Mean function in a compute statement:

```

compute scale=mean.3(var1 to var4).

```

The mean function computes the scale score based on the average answers to variables 1 through 4, based on those with valid answers. The ‘.3’ indicates the minimum number of valid values accepted. Therefore, scores will only be calculated for those with 3 or more valid answers. These scores can be seen in Table 1.

Table 1.

var1	var2	var3	var4	Scale
1	0	0	1	0.5
1	1	1	1	1
1	0	0	.	0.33
1	1	1	.	1

If all the answers to items in a scale are 0 or 1, and there is one item missing due to a change in the version of the questionnaire, these items can be imputed from the average answer to the other items in the scale. For example, if the last two clients in the above sample data were not asked var4, but we want a scale based on 4 items. We would impute this variable based on the average of the other variables then compute the scale. Note that this is ONLY appropriate if the items form an internally consistent scale (alpha of .7 or more). The syntax is:

```
compute replace=var4.

do if (missing(replace)).
  compute replace=rnd(mean(var1 to var4)).
end if.
compute scale2=sum(var1,var2,var3,replace).
```

So when var4 is missing, the value of 'replace' will be the rounded average response to the other items in the scale. The resulting data are presented in Table 2.

Table 2.

var1	var2	var3	var4	replace	Scale2
1	0	0	1	1	2
1	1	1	1	1	4
1	0	0	.	0	1
1	1	1	.	1	4

Alternatively, you can simply multiply the mean score times the expected number of items (in this case, four) to get the scale score and skip actually replacing the individual item level variables. The value of this alternative is that the table can reflect the items based on actual respondents but then give the best estimate of the scale score for each person.

Comments:

Good consistency codes that indicate the type of missing data will make the replacement process much more valid. The examples given above demonstrate ways to clean and replace missing data. These examples can be used extrapolate to more complicated situations that may occur in your data, and make replacement of missing data less of a problem.

Describing These Procedures. These procedures would normally be described in a report or paper as follows:

On the key items used in this analysis, the average percent missing was X, with only Y missing 5% or more. All skipped items were coded to their implied values, and other logical imputations were made (e.g, if someone reports using a substance every day of the last 90 days, a missing recency question can be inferred as “in the past two days”). Since listwise deletion is the most biased method for analysis, we replaced the remaining data. Despite the simplicity of simple mean replacement, it can artificially deflate the variance and consequently inflate statistical tests. It is, therefore, generally recommended that missing data be replaced using some form of hot deck imputation, multiple imputation or regression as we have done here (Dennis, Lennox, & Foss, 1997; Rubin, 1996, Little & Rubin, 1987). For individual items, missing data were replaced using SPSS (1997) Replace Missing Value (RMV) procedure. To do this, individuals were sorted based on type of treatment, gender, race and age, then the missing value was replaced by the mean (normally distributed), median (skewed) or mode (categorical) of the four nearest valid answers in the ordered list. For scales with high internal consistency (Cronbach’s alpha =.7+), an alternative procedure was used. As long as data were available on three or more items, the missing value was replaced with the average of the valid answers to other scale items for the same individual.

References

- Dennis, M.L. (1998). Global Appraisal of Individual Needs (GAIN). Bloomington, IL: Chestnut Health Systems (www.chestnut.org/li/cyt/gain).
- Dennis, M.L., Lennox, R.I., & Foss, M. (1997). Practical power analysis for substance abuse health services research. In K.J Bryant, M Windle, and S.G. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research, (pp. 367-405). Washington, DC: American Psychological Association.
- Figueredo, A.J., McKnight, P.E., McNight, K.M., Sidani, S. (1999 manuscript) Multivariate Modeling of Missing Data Within and Across Assessment Waves.
- Little, R., & Rubin, D. B. (1989). The analysis of social science data with missing values. Sociological Methods and Research, 18(2), 292-326.
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley.
- Rubin, D.B. (1996). Multiple imputation after 18 years. Journal of the American Statistical Association, 91, 473-489.
- Statistical Program for the Social Sciences. (1997). SPSS Base 7.5 syntax reference guide. Chicago: Author (www.spss.com).