

# Missing values – what they are and why they are important

Ray Hoare, HRS Ltd, copyright 2003.

## Introduction

A numeric code to represent missing values is vital to nearly all data analysis. It is needed when you have more than one variable in a dataset, and where you have in some cases (rows or records) data for some variables but not for others.

For example, if you are a chemist analysing water samples, initially collected in several kinds of bottles to properly preserve the water characteristics, one of the types of sample bottles may have been lost or destroyed. You may have data on dissolved phosphorus in the water, but nothing on the microbiology. You will want to put as much information as you can in your data file, but the microbiology data is recorded as missing.

All statistics programs provide a code to perform this function. In some the actual value is never seen, but takes the value of the largest possible negative integer. In others the value is chosen by the user.

In some situations you may wish to record several different reasons for there being missing data. For instance, you may carry out a survey of people, asking about various health matters. A question like “How many pregnancies have you had” would be inapplicable to men, and may be seen by some woman as a type of question they will not answer. So you need to record a code for “not applicable” and another for “no answer”. You may also find there is a group who truly don't know, for medical reasons, so you could have a third group for whom a simple number is not appropriate.

In most statistics programs you would store these as numbers, such as -1 or 99, that are not part of the likely range, and temporarily change them to missing values when performing summaries or analysis. The precise way to make this change depends on the software. In some programs you can define one or more "user-defined" missing values in addition to the system-defined one.

A related problem is "censored" data, such as <0.01 or >10000. The < or > symbol can't be stored with numeric data in any of the usual statistics programs, but a common work-around is to store the data as -.01 etc. It is vital to note this in the header of the dataset where one is provided, and to ensure that the coded values are removed before conducting numerical summaries.

## Actions controlled by missing values

### *Missing values when summarising variables*

When you count or calculate the average of a numeric data column with missing values in it, the missing cases are ignored. Note that this is different from treating them as zero.

### ***Missing values when computing new variables***

When you write a compute expression to make a new variable, say to compute density from weight of a sample divided by its volume, the resulting value will always be missing if any of the values are missing for a particular case.

### ***Missing values when combining variables***

Sometimes you want to combine variables using rules which are different from the normal rules for missing values. For instance, you may wish to take the average of several variables and put the result in a new one. If you say  $\text{varnew} = \text{var1} + \text{var2} + \text{var3}$  then a missing value in any of the 3 variables will give you a missing value in  $\text{varnew}$ . You can circumvent this by using functions that are present in most statistics programs.

For instance, many programs have a function SUM, and in this case  $\text{varnew} = \text{SUM}(\text{var1}, \text{var2}, \text{var3})$  will give you a non-missing value in  $\text{varnew}$  unless values in all 3 of the variables are missing. Other column-oriented functions will usually behave in a similar way.

### ***Missing values when comparing variables***

If you have a set of numeric variables and make a correlation between them, your program will probably by default remove every case that has a missing value in any of the columns specified, and then use only that data for the calculation. This is fine if you have only a small proportion of your data as missing, but it can lead to unexpected results. I did a correlation on a set of data once, after just gaining access to a statistics program, and got correlation coefficients of 1.000 for all of my relationships. It turned out that the missing value pattern had left me with only two cases for my correlation.

You can eliminate this problem by choosing an option for “pairwise deletion” of missing values. So when the program correlates  $\text{var1}$  with  $\text{var2}$ , it just deletes the cases with missing values in either of those columns, and ignores missing values in other variables that will be dealt with later in the correlation. This lets you use more of your data, but has its own problem. In extreme cases, the subset of data used for correlating  $\text{var1}$  and  $\text{var2}$  can have little in common with that used with correlating  $\text{var2}$  and  $\text{var3}$ , to the extent that  $\text{var1}$  and  $\text{var2}$  are highly correlated, as are  $\text{var2}$  and  $\text{var3}$ , but there is little correlation between  $\text{var1}$  and  $\text{var3}$ .

### ***Stepwise multiple regression***

When you do stepwise multiple regression the software will normally discard cases with missing values in any of the variables you include. (“Listwise deletion”) At the end of the regression it will normally turn out that you initially included variables that were not included in the final regression.

You should check whether your software will automatically run the regression again, using only the cases with non-missing values in the variables that remain. This is a different data set from the initially-specified one, and the coefficients of the regression

are not necessarily still significant in the final model. Stepwise regression is not the automatic process it seems, when you have data sets with patterns of missing values.

### ***Missing values in multiple response analysis***

In a multiple response analysis SPSS deletes any case that has no response for any of the set of variable. This can reduce the apparent number of cases in your dataset, and give you puzzling values of proportions. Check the behaviour of your software so you understand what you are getting.

### ***Missing values in time series***

Most kinds of time series analysis require a continuous sequence of equal interval data. When your data series is broken you have a number of things you can do.

- Ignore data before the gap.
- Ignore data after the gap.
- Interpolate across the gap, using one of several algorithms.
- Substitute a value (perhaps zero) for the missing value.
- Delete the case with a missing value.

You must be aware of what the default behaviour of your program is, and if necessary choose a different option.

### ***Systematic patterns of missing data***

In some situations you get systematic patterns of missing data that are themselves capable of interpretation. You may wish to recode the missing values and then use conventional methods to interpret them.

You can purchase software that is designed to look for and to correct problems caused by patterns of missing data, but in most cases this is not necessary.