Missing Data Coding

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised September 20, 2024

NOTE: This is a slightly tweaked repeat of material contained in the Basic Data Exploration handout. I include it here to make sure the most critical points are now clear to you if they weren't before.

Before you do any extensive analysis with your data, you should make sure missing data is coded correctly. The Stata missing value codes are ., .a, .b, .c,…, .z (i.e. . and .a to .z). Even if you downloaded your data in Stata format, the missing data codes may not be correct. For example,

```
. use https://www3.nd.edu/~rwilliam/statafiles/fixcoding, clear
. fre var1

var1
-------------------------------------------------------------------------------
                              |      Freq.    Percent      Valid       Cum.
------------------------------+------------------------------------------------
Valid   1  Strongly Disagree  |         54      24.11      24.11      24.11
        2  Disagree           |         75      33.48      33.48      57.59
        3  Agree              |         29      12.95      12.95      70.54
        4  Strongly Agree     |         42      18.75      18.75      89.29
        97 Don't Know         |          8       3.57       3.57      92.86
        98 Refused            |          5       2.23       2.23      95.09
        99 Not Applicable     |         11       4.91       4.91     100.00
        Total                 |        224     100.00     100.00
-------------------------------------------------------------------------------

. sum var1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        var1 |        224     12.5625    29.72513          1         99
```

The values 97, 98, and 99 are missing data codes. That might be correct coding for a program like SPSS, but in Stata those codes are treated as legitimate values, which totally distorts statistics involving the variable, e.g. the mean and standard deviation are wrong here. OLS or logistic regression results could also be way off if you don't fix the MD coding.

The `mvdecode` command is one of the many ways to solve the problem (the `recode` command is another) :

```
. mvdecode var1, mv(97=.a\ 98 = .b\ 99=.c)
        var1: 24 missing values generated
. fre var1
```

```
var1
-------------------------------------------------------------------------------
                            |      Freq.    Percent      Valid       Cum.
----------------------------+--------------------------------------------------
Valid   1  Strongly Disagree |         54      24.11      27.00      27.00
        2  Disagree          |         75      33.48      37.50      64.50
        3  Agree             |         29      12.95      14.50      79.00
        4  Strongly Agree    |         42      18.75      21.00     100.00
        Total                |        200      89.29     100.00
Missing .a                   |          8       3.57
        .b                   |          5       2.23
        .c                   |         11       4.91
        Total                |         24      10.71
Total                        |        224     100.00
-------------------------------------------------------------------------------
```

```
. sum var1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        var1 |        200       2.295    1.083441          1          4
```

Much better! Further, suppose var1 thru var20 are consecutive variables in the data set and are all coded the same way. We might then be able to say

```
mvdecode var1-var20, mv(97=.a\ 98 = .b\ 99=.c)
```

Or, better yet, suppose all variables in the data set use the same missing value codes. You could then say

```
mvdecode _all, mv(97=.a\ 98 = .b\ 99=.c)
```

If we want, we can also tidy up the value labels a bit. `var1` uses a value label called `agreement` (using the same value label for several variables that share the same values is often convenient). We can get rid of the old labels and add the new with the commands

```
. label define agreement 97 "" 98 "" 99 "", modify
. label define agreement .a "Don't Know" .b "Refused" .c "Not Applicable", add
. fre var1
```

```
var1
-------------------------------------------------------------------------------
                            |      Freq.    Percent      Valid       Cum.
----------------------------+--------------------------------------------------
Valid   1  Strongly Disagree |         54      24.11      27.00      27.00
        2  Disagree          |         75      33.48      37.50      64.50
        3  Agree             |         29      12.95      14.50      79.00
        4  Strongly Agree    |         42      18.75      21.00     100.00
        Total                |        200      89.29     100.00
Missing .a Don't Know        |          8       3.57
        .b Refused           |          5       2.23
        .c Not Applicable    |         11       4.91
        Total                |         24      10.71
Total                        |        224     100.00
-------------------------------------------------------------------------------
```

Other notes:

- Never just assume you did things right! Check things out before and after like I did.
- The missing data codes were pretty obvious in this case. Other times they won't be. Try to check the dataset documentation if you can.
- It is nice when every variable uses the same MD codes, but that doesn't have to be the case. For example, 99 may be a missing value for one variable and a valid value for another.
- Sometimes all missing data are just coded ., the system missing value. That is often fine, but at other times it is helpful to know why data are missing. If you use Stata's multiple imputation commands it is very important that you use different MD codes for different types of MD. For example, it can make a difference if a respondent refuses to answer or if they were never asked the question in the first place. Eventually, you may decide that different types of missing data will be treated differently in your analysis.
- See `help mvdecode` for more information and examples.
- Chuck Huber has a nice 2-minute video on "How to convert missing value codes to missing values". I prefer to directly write out code when I can, but sometimes the menu-driven approach he shows is better or easier. See https://www.youtube.com/watch?v=6HV2773-dVM.