Missing value hell

Steve Simon

2022-07-04

There’s no mincing words here. Missing values are hell for every statistician. Here are a few examples.

### The null license plate

Joseph Tartaro thought it would be funny to get the following California vanity license plate.



Figure 1. Null license plate

The joke is that NULL is the missing value code used by many databases. But, as described in Wired Magazine, his life turned into a living hell. Police officers will always write the license plate number of any car that was caught speeding or parking illegally. Well, almost always. In the very rare cases that the officer forgets, this is converted to a NULL value in the database. This is then matched to Joseph Tartaro and the ticket was sent his way. He got tickets from all over California–Corona, Fresno, Rancho Cucamonga–mostly to places he’d never been. These tickets were written for all sorts of different automobiles that he had never driven–Mercedes, Toyotas. He got hundreds of tickets worth tens of thousands of dollars. It was not an easy problem to fix.

This is caused by a programming error. The four letters N-U-L-L are not the same thing as the internal code of Null in a database. But a programmer writes a statement like

if var="NULL" then ...

and everything goes to hell.

### The $9,700 bargain

A while back I was browsing through the geekier parts of Office Max and noticed the following tag.



Figure 2. Printer pricing tage

The computer database used the common practice of using all 9’s to denote a missing value. That’s fine, but then whoever programmed the calculation of how much saving you would get by buying the printer at $299.99 dollars instead of the actual price, figured through a simple subtraction that you would save a whopping $9,700!

### How to avoid missing value problems

Whenever I see stories like this, I think of the quote from John Bradford, a sixteenth century preacher–“There but for the grace of God go I.”

How easy it is to forget to account for missing values. To avoid this, you must develop an obsession about missing values. You must track their number through every version of every dataset. This number can grow, if you are not careful, so monitor it with a religious fervor.

Direct comparisons though can lead to trouble, so it is always best to use an explicit function that detects missing value. This is “missing” in SAS, SPSS, and Stata. In R, use the “is.na” function.

Best practice is to account for missing values first and last. So start off any recodes with the case where the original variable is missing and end with a catch-all code.

Consider a variable for gender with 1 for male, 2 for female that might or might not have missing values. You want to recode from number codes to a string. Then a recode in SAS might look like

if missing(gender) then gender\_label="Unknown";  
 else if gender=1 then gender\_label="Male";  
 else if gender=2 then gender\_label="Female";  
 else gender\_label="Coding error";

The syntax changes slightly with other statistical packages, but the general concept applies. Account for missing value codes first and then catch anything that fails to match at the end.

### Conclusions

Missing values are one of the most troublesome aspects of data management. Every statistical package stores missing values differently, so be sure to use the function that they provide for detecting missingness. Account for missing values at the start and at the end of any recodes.