Avoiding Careless Errors: Know Your Data

Kristin L. Sainani, PhD

When authors and reviewers worry about errors in statistics, they tend to worry about high-level issues: Did I choose the right statistical model? Did I violate a statistical assumption? But it's the far simpler errors—missing data, sloppy data handling, transcribing errors, and mistakes in arithmetic—that worry me more. These types of errors are often far more devastating to an analysis than fancier statistical issues. Take, for example, a recent scandal at Duke University [1] in which a large number of medical papers were retracted; the errors that undermined these analyses included accidental shifting of data cells in Excel and the reversal of data labels (such that drug-sensitive cells became drug resistant and vice versa). These types of careless errors are hard to detect or quantify systematically, but anecdotal evidence suggests that they are shockingly common [2,3].

The primary problem is that researchers become so focused on choosing the correct statistical test and trying to get the computer to spit out *P* values that they overlook more fundamental steps, such as plotting and checking the data. The good news is that avoiding careless errors doesn't require an advanced degree in statistics. Here are a few easy actions that researchers can take to minimize avoidable errors.

KEEP THE DATA IN ONE PLACE

Researchers sometimes create or save multiple copies of a dataset during the data-entry process, which is problematic because it is easy to lose track of which is the "current" copy of the data. For example, I recently reviewed a paper in which the main outcome variable was presented in both table and figure form. However, the values in the table and figure didn't match—likely the result of different authors working from different versions of the dataset. The solution is to always enter the data into a single database (or 2 independent databases if the data are being double-entered). If the researcher wishes to alter or parse the dataset (such as making separate datasets for men and women), this should be done from within a data-analysis program (see the section Don't Manually Alter Data).

DON'T MANUALLY ALTER DATA

Manual data entry is often unavoidable and is an obvious source of error. Unfortunately, researchers compound the problem by also altering the data—that is, fixing values or calculating new variables—from within data-entry programs such as Excel. Instead, researchers should always load the dataset into a statistical analysis program (such as R, SAS, or SPSS) and make changes within this environment. Unlike data-entry programs, these programs leave a clear record of exactly how the data were altered. They also avoid the kind of human error that might occur if, for example, the data-entry specialist calculated age from birth date manually. These kinds of derivative variables should always be created automatically from within the statistical analysis program.

DON'T ANALYZE DATA IN EXCEL

In my experience, many medical researchers are quite comfortable using Excel and thus prefer to analyze their data in that program. Unfortunately, analyzing data in Excel often involves cutting, pasting, and rearranging rows and columns of data—a surefire way to introduce errors. Researchers should always import data from Excel into a program that was designed expressly for statistical analysis.

K.L.S. Division of Epidemiology, Department of Health Research and Policy, Stanford University, HRP Redwood Bldg, Stanford, CA 94305. Address correspondence to: K.L.S.; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

1934-1482/13/$36.00
Printed in U.S.A.

Vol. 5, 228-229, March 2013
http://dx.doi.org/10.1016/j.pmrj.2013.01.012

PLOT EACH VARIABLE FIRST

When researchers come to me for statistical advice, they often don't have a good sense of their data. For example, they can't answer simple questions, such as: What was the highest value of the outcome? How many people were missing the outcome? This inevitably leads to confusion later. Thus researchers should spend considerable time and energy getting to know their data before they run any statistical tests. The first step is to make a distributional plot (such as a histogram) of each variable. Plots make it easy to spot wild data points (which need to be resolved) and outliers (which need to be considered in all analyses). Researchers should also generate simple statistics for each variable, such as the N (because there may be missing data), mean, standard deviation, maximum, and minimum. They should also get a feel for simple relationships, such as which pairs of variables are correlated. Researchers who understand their data at this basic level will avoid careless errors and have a better framework for interpreting results.

CHECK YOUR Ns

Most regression analyses automatically throw out incomplete observations, so if a subject is missing the value for just one of the variables in the model, that subject will be excluded. When the multivariate model includes many variables, this can add up to a lot of omissions, even when the number of missing data points for any individual variable is low. For example, a researcher might feed 400 observations into the regression procedure, but the final model might include only 200. Unfortunately, many researchers forget to check the Ns of their multivariate analyses. I recently reviewed a paper in which, unbeknownst to the authors, 40% of their subjects had been excluded from the multivariate analysis. Obviously, throwing out nearly half the data may drastically change the results of the analysis.

DOUBLE-CHECK SIMPLE MATH AND USE COMMON SENSE

Many careless errors go unnoticed because people become so intimidated by statistics that they stop using simple math skills and common sense. I once reviewed a paper that described a study of 10 subjects and concluded impossibly (through a slip of averaging) that "78% of subjects improved." In another example, I reviewed a paper in which I calculated a follow-up rate of 20% (by simply dividing the number of outcomes by the number of people who started the study), whereas the authors reported a follow-up rate of 65%. Catching such blatant errors doesn't require an advanced degree in statistics—and these errors usually portend deeper problems in the analysis, such as sloppy data handling, poor research quality, and a lack of statistical consultation.

For more of these types of tips, I refer readers to an excellent article by Andrew Vickers [2].

REFERENCES

1. Kolata G. How Bright Promise in Cancer Testing Fell Apart. New York Times, July 7, 2011. Available at: http://www.nytimes.com/2011/07/08/health/research/08genes.html?_r=0. Accessed January 25, 2013.
2. Vickers A. Look at Your Garbage Bin: It May Be the Only Thing You Need to Know About Statistics. Nov 3, 2006; www.Medscape.com. Available at: http://www.medscape.com/viewarticle/546515. Accessed February 7, 2012.
3. Must Try Harder. Editorial. Nature 2012;483:509.