

A Guide to Bulletproofing Your Data

by Jennifer LaFleur, ProPublica

Integrity Checks for Every Data Set

- Make sure you know how many records you should have and that you have them all. Certain versions of Excel, especially ones used in gov't agencies, have row limits; be concerned if you or someone in your newsroom tells you they have 65,536. Database programs like MySQL have limits on the size of integers and on the length of string fields. These limits tend to be powers of two -- 2^n or 2^{n-1} . or 1,048,576 records, so be skeptical if your strings are 255 characters long or if a sum appears to be 2,097,152.
- Double-check totals or counts. Check for studies or summary reports.
- Consistency-check all fields. Are all city names spelled the same way? How about other important fields? Check by running a GROUP BY and sorting alphabetically by every important field. Check it for spelling inconsistencies. For example, if you're analyzing a database of highway accidents, GROUP BY and sort ascending on the road name to check for inconsistencies.
- Other basic checks: make sure all states/cities/counties are included. Check the range of fields. (For example, check for DOBs that would make people too old or too young.)
- Check for missing data or blank fields. Are they real values, or did something happen with an import or append query?
- Check your methodology (if necessary) against other similar research.

Beyond Basic Checks

- Keep a data notebook (or computer file) and write down everything you do.
- Know the source of the data.
- Get similar data from another source.
- Create a back-up copy of the database.
- Check against reports.
- Make sure you're using the right tool. You may need to do more than counting and sorting.
- Check with experts from different sides of the issue.
- Find similar stories and study what they did.
- Look at it. If you can actually physically go spot check records, do it.
- Don't forget the gut check. If someone just doesn't seem right, it probably isn't.
- Beware of lurking variables
- Come up with a standard naming convention for your files so you can stay organized. I don't recommend naming files "final" or even "super final."
- If you think you're in over your head, call on an expert to help.

Some Notes About Others' Studies

- Get the questionnaire and the methodology. If they won't give it to you, that's a red flag.
- Beware of nonscientific methods: Web surveys, man on the street or other self-selection.
- Know the sample size, which will give you the sampling error.
- Again, know the source.
- Account for margin of error and non-response or "don't know" when drawing conclusions.
- If possible, run statistical tests on the data. What may look significant to you, may not be.
- When reporting, avoid false precision. Saying 52.18 percent of people think "blah, blah, blah" is portraying an impossible accuracy to readers.
- Put your numbers in perspective

Find the Right Methodology

- Read research reports.
- Finding an existing data model - There are some accepted methodologies for dealing with certain types of data.
- Find an expert to bounce your methodology off during the process.
- Show findings to the targets of the story.
- Duplicate your work. To make sure you didn't mess something up along the way.
- Maintain a consistent universe of cases. If you have to filter or redefine your universe, be able to explain why you isolated certain records or cases.
- Give yourself enough time to follow through on collecting information for your database before you start writing. If you've built an organic database, where information may need to be updated or will change after additional reporting, set a cut off date and don't make any more changes to the database unless the data is inaccurate or the new information will change the meaning of the story.

Other Tips From Our Colleagues

Sarah Cohen of the New York Times on homemade databases:

- Number the pages of your documents to keep them in order and include the number when you enter the data. It helps you stay organized and with double-checking later on.
- Add fields that relate to how "publishable" information is. I usually create fields that anyone can fill in that say 1) is the name spelling checked.

Ron Campbell of the Orange County Register on documenting your work:

"I try to document every stage of my work using three tools:"

- A work log (in Word or a text file); I describe what I'm trying to do at each stage and paste in queries.
- Query files. Easy to do in SQL Server. Just be sure to put in a comment above the query explaining what the hell you were trying to do.
- The Comment tool in Excel. Again useful for documenting what you're trying to do.

Russell Clemmings of the Fresno Bee on rechecking your data:

- Write a different query that should yield the same results and see if it does so.
- Pull a random sample from your results and check them against the raw data.
- Have someone who knows the data check your results before publication -- even the target of the story, if possible.
- Double-check surprising results -- if citations spiked by 50 percent in one year, it could be a story or it could (more likely) be an error.

For More Information

Numbers in the Newsroom: Using Math and Statistics in News by Sarah Cohen for Investigative Reporters and Editors, Inc.

Precision Journalism by Philip Meyer. Indiana University Press, Bloomington. 4th Edition. 2002.

News and Numbers by Victor Cohn. Iowa State University Press, Ames. 1989.

How to Lie with Statistics by Darrell Huff. W. W. Norton & Company, New York. 1954 (renewed 1984)

Innumeracy: Mathematical Illiteracy and Its Consequences by John Allen Paulos. Vintage Books, New York. 1990.

A Mathematician Reads the Newspaper by John Allen Paulos. Anchor Books, New York. 1995. (Also, check out the tape from Paulos keynote address at NICAR 2002 in Philadelphia)

IRE Resource Center: www.ire.org

Danielle Cervantes from The San Diego-Union Tribune contributed to this handout.