

ROBUST DATA

Ashish Sen

Director, Bureau of Transportation Statistics
Department of Transportation

1. ROBUST STATISTICS

When data quality is poor, statisticians have traditionally used robust estimators.

For example, suppose one wants to estimate the average of cars involved in accidents. But all the missing values have been recorded as 9999 (an old Cobol convention), and thus a naïve analyst will get a number that is much too large.

To solve this problem, one could use a robust estimate of the mean, such as the 10% trimmed mean. Here one throws away the largest and smallest 5% of the data, then averages the rest. Thus one could be missing 5% of the data, but the missing value codes would not mislead the analyst.

It is possible to have almost half of the data (actually, $\frac{p}{p+1} \epsilon$, for p the number of variables) be of poor quality, but still obtain good estimates.

2. ROBUST DATA

BTS encourages the use of robust statistics. But this is not enough.

One reason is the problem of subgroups. People want to look at traffic Safety data by age, gender, type of automobile, city, and weather. This level of detail ensures small samples in each cross-tabulation, and robust methods lose too much efficiency.

Another reason is legislative. Congress has made BTS responsible for DOT data quality. We believe it is more cost-effective to build quality into DOT processes than to retroactively enforce it.

The third, and most important, reason is public trust. Our data drive decisions that save lives. Managers should never delay action because of questions about our data quality.

For these reasons, BTS is trying to develop robust data. These have high quality and make good inference easy and make poor inference hard.

2.1 DATA QUALITY PROGRAMS

In April, congress asked BTS to report on DOT quality. We did a quality audit on four randomly chosen databases. One was excellent, two were adequate, and one was poor.

Based on those results, we are planning to extend this review to all major DOT databases—about 80 especially crucial. This is a major effort, and we plan to proceed in two stages. The first stage is a screener that flags problematic databases. The second stage is an in-depth audit that diagnoses difficulties in order to catalyze change.

Using feedback from the audit teams, input from participants in the Safety In Numbers workshops, joint work with the new databases quality initiative at the National Institute of Statistical Sciences, and relevant research literature, we have developed a preliminary design for the database screener that should enable rapid and accurate assessment of eight key dimensions of database quality.

The screener plan currently aims at capturing information on the following quality features:

1. Is the sampling procedure used in collecting the database statistically adequate?
2. Is there a formal procedure for auditing/verifying data entries?
3. Is there appropriate documentation for the data definitions, the collection procedures, and other meta-information.
4. Is there a standing committee to review and improve the database collection procedures, instruments, and documentation?
5. Does the data collected in the database adequately and directly the ostensible purposes?
6. Are the data accessible?
7. Is the data collection program reasonable and well-coordinated with other DOT efforts?
8. Is the data collection program more burdensome than necessary?

2.2 DENOMINATOR DATA

Statisticians know that good denominator data re crucial. The uncertainty in a ration is much more sensitive to denominator error than to numerator error.

As an example, consider estimates of the traffic fatality rate. We have such good data on the numerator (the number killed, available from FARS) that DOT feels obliged to sue the corresponding denominator (the total number of miles driven on U.S. roads, available from the HPMS).

But it is very difficult to estimate the number of miles driven. Different approaches have been tried, and experts feel the standard error in those estimates could be as large as 10^8 miles. A simple calculation (the delta method) shows that this means our confidence interval on the fatality rate is very wide.

To ensure robust data, we need to help design capture processes that especially focus on the accuracy of denominator data. This may entail reallocating resources away from numerator data.

2.3 INTERMODAL TRANSPORTATION DATABASE

BTS is building the Intermodal Transportation Database (ITDB). This will eventually assume all major DOT databases, and provide a unified gateway for transportation information.

The ITDB is a virtual wrapper around databases that are and will continue to be maintained by different DOT administrations. BTS will help these administrations improve their data quality, and the ITDB will improve accessibility and usability. The ITDB has statistics front-end to support analysis and graphics.

ITDB users can easily combine information from multiple databases, without even needing to know that they are doing so. BTS believes this new connectivity will improve data quality by encouraging each database to be measured against the best.

The ITDB will contain a Source & Accuracy Statement for each major database. These were pioneered by the bureau of Labor Statistics, and we hope these will become standard government practice for documenting the content and limitations of databases.

3. STATISTICAL GRAPHICS

One way to improve data quality is to visualize it. This highlights defects. BTS is pushing all of DOT to graph their data.

As an example of the power of graphics consider the figure on pedestrian fatalities in 1998. It shows: (Graphic below)

1. The cline from north to south, probably reflecting the fact that the south has more days of good walking weather.
2. Hotspots in Florida, New Mexico, and Arizona. These have large populations of elderly pedestrians, many elderly drivers, and snow bird drivers unfamiliar with local streets. They also have immigrant populations from rural areas of Central America, where street-crossing skills are not drilled in childhood.
3. A hotspot in Washington D.C., the most intensely urbanized area.

4. A hotspot in Nevada, perhaps due to intoxicated pedestrian traffic (alcohol is involved in most pedestrian fatalities it may be even more conspicuous in resort cities).

4. CONCLUSIONS

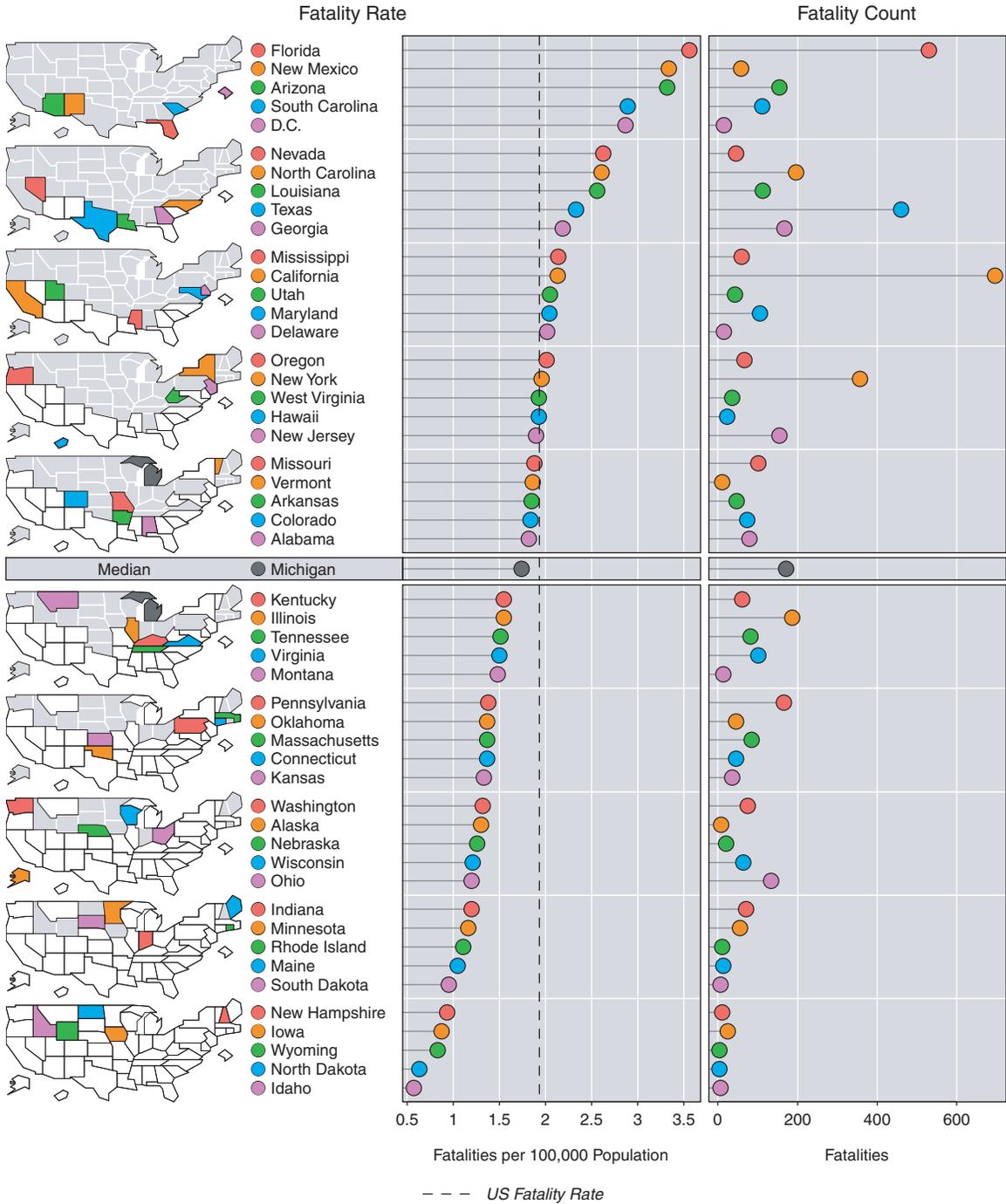
BTS is committed to improving data quality.

Our strategy for improvement is rooted in process control methods, statistical theory, and graphical tools.

Our tactics for improvement are tailored to meet the practical and political demands of our environment--we don't want "the best to be the enemy of the good."

The next two years will tell the tale. Success is contingent on our ability to provide rapid and constructive help that will inspire others to make quality a priority.

Pedestrian Traffic Fatalities by State: 1998



SOURCE: U.S. Department of Transportation, National Highway Traffic Safety Administration, Traffic Safety Facts 1998 (Washington DC: October 1999), p. 150, Table 109.