## TECHNICAL AIDS

*by*
***Lloyd S. Nelson***

# Display Tables and Significant Digits

In an effort to give readers a rest between discussions of tests and critical values, I would like to review the generally sad state of tabular data. Further I would like to make some recommendations that can ameliorate the situation. Let us direct our attention to the two-way table as a communication device, *not* as an array awaiting statistical processing.

Many years ago I took a course in data analysis in which the professor stated flatly that "data are good to no more than two significant figures". At that time I doubted this; I doubt it much less now. For the moment let us assume that this statement is true. We will evaluate it quantitatively later.

Consider the following table which contains the hypothetical hours of down-time per month for each of four machines over a period of five months. That is, for example, the various down times for Machine A in January added to 7 hours, 45 minutes and 7 seconds (7.752 hours). The engineer reported the full precision at his command.

| | MACHINE | | | |
|---|---|---|---|---|
| | A | B | C | D |
| JAN | 7.752 | 8.114 | 6.048 | 7.173 |
| FEB | 8.016 | 7.241 | 8.312 | 6.853 |
| MAR | 7.685 | 7.819 | 7.504 | 7.147 |
| APR | 8.211 | 7.568 | 7.301 | 7.912 |
| MAY | 8.013 | 7.310 | 7.552 | 8.069 |

This tables gives only a fuzzy impression of what is going on. There are too many digits that are of no use. Consider the following version edited to two significant figures.

| | MACHINE | | | |
|---|---|---|---|---|
| | A | B | C | D |
| JAN | 7.8 | 8.1 | 6.0 | 7.2 |
| FEB | 8.0 | 7.2 | 8.3 | 6.9 |
| MAR | 7.7 | 7.8 | 7.5 | 7.1 |
| APR | 8.2 | 7.6 | 7.3 | 7.9 |
| MAY | 8.0 | 7.3 | 7.6 | 8.1 |

**KEY WORDS:** Distribution of First Digits, Significant Digits, Tabular Display

Notice how the situation has come into focus. But there is still more that we can do. A third arrangement with the columns of data ordered by the column averages makes the suggestively inconsistent January value for Machine C stand out a little more.

| | MACHINE | | | | |
|---|---|---|---|---|---|
| | C | D | B | A | AVG |
| JAN | 6.0 | 7.2 | 8.1 | 7.8 | 7.3 |
| FEB | 8.3 | 6.9 | 7.2 | 8.0 | 7.6 |
| MAR | 7.5 | 7.1 | 7.8 | 7.7 | 7.5 |
| APR | 7.3 | 7.9 | 7.6 | 8.2 | 7.8 |
| MAY | 7.6 | 8.1 | 7.3 | 8.0 | 7.8 |
| AVG | 7.3 | 7.4 | 7.6 | 7.9 | 7.6 |

The re-ordering of the columns is reasonable because the machines have no numerical relationship with each other. (They are on a nominal scale.) Generally the levels of the factor of greatest interest should be the columns because it seems that columns are a little easier to compare than rows.

By the way, a plot of the data in this table produces a criss-cross of lines that is no easier to interpret than the original numbers. The moral is: rounding to two significant figures and re-ordering columns and/or rows when appropriate can help the reader understand the message. For further examples and expert discussion of this and related matters I highly recommend the book by Ehrenberg (1978).

How much information is lost when numbers are rounded to two significant figures? We first note that the *maximum* loss in accuracy results from rounding 1.05 to 1.0. (Remember the rule for rounding when the digit to be rounded is five: the preceding digit is made even or left even.) The decimal place is of no consequence; it can be anywhere. The loss experienced is $100(0.05/1.05) = 4.8$ percent.

The *maximum* loss for numbers starting with the digit two is $100(0.05/2.05) = 2.4$ percent, and for the rest of the digits (starting digit, percent loss in accuracy) is (3, 1.6%), (4, 1.2%), (5, 1.0%), (6, 0.83%), (7, 0.71%), (8, 0.62%) and (9, 0.55%). One might be tempted to find the average maximum

loss by simply averaging the nine percent losses. However this uses the assumption that numbers in general are equally likely to start with any digit (excluding zero, of course). Most people find it astonishing that this assumption is known to be wrong.

It was shown empirically by Benford (1938) and theoretically by Pinkham (1961) that the cumulative distribution of first digits is $\log_{10}(n + 1)$, where $n$ equals the digit in question. For our purposes we need individual probabilities. First differences yield these probabilities (digit, probability of being the first digit) as (1, 0.301), (2, 0.176), (3, 0.125), (4, 0.097), (5, 0.079), (6, 0.067), (7, 0.058), (8, 0.051) and (9, 0.046). This is to say, for numbers in general about 30 percent will begin with the digit one, almost 18 percent will begin with the digit two, and so on. The doubting reader can verify these probabilities using a sampling experiment with any large body of physical data such as can be found in *The World Almanac.* When these probabilities are used as the weights for the corresponding maximum losses, we find that the average *maximum* loss in accuracy attendant upon rounding numbers in general is 2.4 percent.

If we can assume (as has been shown reasonable) that the second and subsequent digits of a number are essentially uniformly distributed, then the average loss will be about one percent. Thus, with such obvious exceptions as a 10.6 percent interest rate or $1.04 earnings/share, it appears that little would be lost in accuracy and much might be gained in clarity of communication through rounding to two significant figures.

## References

BENFORD, F. (1938). "The Law of Anomalous Numbers". *Proceedings of the American Philosophical Society* 78, pp. 551–572.

EHRENBERG, A. S. C. (1978). *Data Reduction. Analysing and Interpreting Statistical Data*, John Wiley and Sons, New York, NY.

PINKHAM R. S. (1961). "On the Distribution of First Significant Digits". *The Annals of Mathematical Statistics* 32, pp. 1223–1230.