Statistical Science
1999, Vol. 14, No. 2, 149

Introduction to "Solving the Bible Code Puzzle" by Brendan McKay, Dror Bar-Natan, Maya Bar-Hillel and Gil Kalai

Robert E. Kass

One of the fundamental teachings in statistical training is that probability distributions can generate seemingly surprising outcomes much more frequently than naive intuition might suggest. For good reason, experienced statisticians have long been skeptical of claims based on human perception of extraordinary occurrences. Now that computer programs are widely available to help nearly anyone "mine" available data, there are wonderful new possibilities for discovering misleading patterns.

In this context, when the article "Equidistant Letter Sequences in the Book of Genesis," by Witztum, Rips and Rosenberg, was examined by reviewers and editorial board members for *Statistical Science*, none was convinced that the authors had found something genuinely amazing. Instead, what remained intriguing was the difficulty of pinpointing the cause, presumed to be some flaw in their procedure, that produced such apparently remarkable findings. Thus, in introducing that paper, I wrote that it was offered to readers "as a challenging puzzle."

Unfortunately, though perhaps not so surprisingly, many people outside our own profession interpreted publication of the paper as a stamp of scientific approval on the work. However, even though the referees had thought carefully about possible sources of error, no one we asked was willing to spend the time and effort required to reanalyze the data carefully and independently. Rather, we published the paper in the hope that someone would be motivated to devote substantial energy to figuring out what was going on and that the discipline of statistics would be advanced through the identification of subtle problems that can arise in this kind of pattern recognition.

In this issue, Brendan McKay, Dror Bar-Natan, Maya Bar-Hillel and Gil Kalai report their careful dissection and analysis of the equidistant letter sequence phenomenon. Their explanations are very convincing and, in broad stroke, familiar. They find that the specifications of the search (for hidden words) were, in fact, inadequately specific: just as in clinical trials, it is essential to have a strict protocol; deviations from it produce very many more opportunities for surprising patterns, which will no longer be taken into account in the statistical evaluation of the evidence. Choices for the words to be discovered may seem innocuous yet be very consequential. Because minor variations in data definitions and the procedure used by Witztum et al. produce much less striking results, there is good reason to think that the particular forms of words those authors chose effectively "tuned" their method to their data, thus invalidating their statistical test.

Considering the work of McKay, Bar-Natan, Bar-Hillel, and Kalai as a whole it indeed appears, as they conclude, that the puzzle has been solved.

Robert E. Kass is Professor and Head, Department of Statistics, Carnegie Mellon University, Pittsburgh Pennsylvania 15213-3890 (*e-mail: kass@stat.cmu.edu*).