

Statistics and *Homeland Defense* David Banks Asks

How Are We Meeting the New Challenge?

by David Banks
Duke University

The statistics profession is being sidelined on national defense and homeland security research. This marginalizes our discipline, diminishes our influence, and—most crucially—provides inferior solutions to problems of enormous practical importance.

Currently, there are no statisticians in the Department of Homeland Security. And there are essentially no statisticians in any of the established DHS research centers. To the limited extent that our allied sciences are involved in counterterrorism, most of the quantitative work is being done by computer scientists and operations researchers.

Our concern should be larger than the parochial interest of our profession. The main issue is that the work be done in the best possible way. Certainly, this requires interactions among many kinds of professionals, but it is shortsighted to overlook the areas to which statisticians have contributed in the past and to which we can contribute in the future. Some of these key areas include:

- **Probabilistic Risk Assessment.** Statisticians have done this for decades, starting with the analysis of nuclear reactor safety and moving on to achieve widely-recognized success with drug approval, consumer product safety, and options pricing. There is an ASA Section devoted wholly to this topic, and its members are expert in exactly the kind of balancing of cost versus risk reduction that should be the main criterion for responsible defense investment. Many of the members also cope specifically with risk analysis in complex and uncertain situations, such as public health.

- **Syndromic Surveillance.** This is a relatively new area, in which statistics from hospital emergency rooms, transit ridership, or over-the-counter sales of medicine are used to give early warning of emerging epidemics. It combines time series and spatial modeling, and has benefits for both counter-bioterrorism and the early detection of natural diseases, such as the flu or a chickenpox outbreak. Some of the key statisticians working in this area include Henry Rolka, Ken Kleinman, Galit Shmueli, and Martin Kulldorff.

- **Cybersecurity.** The DOD is worried that terrorist hackers might try to shut-down, flood, corrupt, or even hijack crucial defense networks. And the financial industry has similar concerns and greater vulnerabilities. Preventing such sabotage depends in part on automated statistical algorithms that identify attack signatures or detect system anomalies quickly and adopt appropriate countermeasures before damage is done. Dave Marchette, Alan Karr, and Martin Theus have worked in this area.

- **Biometric Identification.** The Holy Grail in this area is the ability for pattern recognition software to scan airport photographs and automatically discover terrorists in disguise—but this is really hard to do. Andrew Rukhin is working to improve identification performance in the DARPA Ferret database of faces by combining results from multiple algorithms; Sinjini Mitra is using model-based techniques to a similar end. Other aspects of biometric identification include improvements in fingerprint matching, iris scanning, and perhaps continual in-session authentication of computer users through idiosyncratic keystroke patterns.

- **Record Linkage.** This field started with the work of Fellegi and Sunter in 1969; since then, the area has expanded rapidly, driv-

en in large part by statisticians at the U.S. Census Bureau and especially the research of Bill Winkler. One application is forensic investigation; here, the purpose is to match records across multiple databases (e.g., one might have a terrorist suspect's name and an address, which can be partially matched to a driver's license application, which might be partially matched to a student enrolled in a flight school.) If the confidence in the match is sufficiently great, law enforcement can act. The major concern, obviously, is the false alarm rate; all statisticians understand this, but it can be overlooked by contractors who build such linking systems and the agencies that want to use them.

- **Data Mining.** There are oceans of data, and for counterterrorism, the signals one wants to find are concealed often deliberately. Cluster analysis and classification methods that can perform in high dimensions are essential for separating out the salient features; then, one often needs to represent those findings in visualizations or other summaries that highlight the important parts. Carey Priebe, Ed Wegman, David Madigan, and David Scott all have worked on national security problems from this perspective—this should be one of our core strengths in counterterrorism research.

- **Privacy Protection.** This is the flip-side of both data mining and record linkage. The war on terrorism is tightly connected with issues of individual autonomy and privacy. Statistical methods help government agencies decide how much and what kind of noise public releases of federal data should contain in order to ensure a prescribed level of confidentiality (See the work of George Duncan, Steve Fienberg, and Laura Zayatz). This also relates to problems in secure computation, which develops protocols for combining information across separate, confidential databases without pooling them or otherwise violating individual privacy. This relates to work by Jerry Reiter and Alan Karr.

- **False Discovery Rates.** Statisticians know about the problems with repeated significance tests. In counterterrorism applications such as syndromic surveillance, record linkage, and data mining, one makes enormous numbers of inferences. The Benjamini-Hochberg approach, and more recent variations by Chris Genovese, Larry Wasserman, and John Storey, all point to the kinds of corrections that need to be made to avoid unacceptably high numbers of false positives.

These examples only scratch the surface. The point is to whet the imaginations of statisticians interested in such work and to lay out a smorgasbord of programs that illustrate to policymakers the largely untapped potential that statistics has in this area.

Obviously, the major challenge for statisticians is to get into the game. Our profession has been, and is being, proactive in addressing counterterrorism problems. The Board on Mathematical Sciences held an open meeting on the role of statistics in homeland defense in 2002; the ASA and the Washington Statistical Society cosponsored a conference in May 2003; the ASA approved a new Section on Statistics in National Defense and Security in August 2004; and this new Section, partnering with the Risk Analysis Section and the National Institute of Statistical Sciences, sponsored a meeting in New York in November, 2004. More directly, there has been significant investment of time and energy by statisticians at the national laboratories, by ASA Executive William Smith, and by individual researchers. And Sallie Keller-McNulty has made national defense the theme of her ASA presidency.

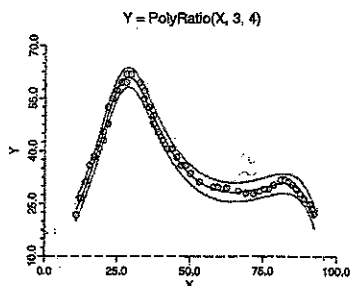
Regrettably, none of this activity seems to have connected with those who frame research policy. From conversations with program directors in statistics at the NSF, the Office of Naval Research, and the Army Research Office, it appears that few statistics proposals have been funded for counterterrorism. Similarly, the Department of Homeland Security research centers have not involved

NCSS - Statistical Software - PASS

PASS 2005 Now Available

Data Analysis

NCSS 2004 is an easy-to-use statistical analysis package that includes over 220 statistics and graphics procedures. The latest edition contains new procedures for proportions, binary diagnostic tests, meta-analysis, ROC curves, tolerance intervals, curve fitting, histograms, and much more.



Power Analysis

PASS 2005. We are pleased to announce the release of a new edition of our industry-leading power analysis and sample size program. **PASS 2005** adds 44 new procedures and 21 enhanced procedures. New procedures include: inequality of proportions (17), non-inferiority of proportions (13), equivalence of proportions (13), non-inferiority of means (7), equivalence of means (8), cross-over designs (10), simulation routines (10), and coefficient alpha (2). No other program calculates power for more procedures than does **PASS!**

Download free trial copies at

www.ncss.com

NCSS Statistical Software • 329 North 1000 East • Kaysville, Utah 84037

Internet: <http://www.ncss.com> • Email: Sales@ncss.com

Toll Free: (800) 898-6109 • Tel: (801) 546-0445 • Fax: (801) 546-3907



statisticians as part of their (otherwise large and diverse) teams. Some of our national laboratories are in slightly better shape, due to the aggressive leadership of people such as Keller-McNulty, Brent Pulsifer, and Dale Anderson, but even here the work that gets discussed in the open can appear to be a bit specialized, peripheral, and artificial. Perhaps their outputs are eagerly discussed in the corridors of power, but my sense is that the statisticians at the national labs are not receiving the attention they deserve, nor are they being allowed to fully play the role defined in their ostensible mission.

For the long-run health of our profession, we need to understand why there is so much indifference to the value we can add. It may be a perception problem—computers are sexy and statistics is not. It may be that we have been too timid and acquiesced in our own diminution; perhaps our careful expressions of uncertainty, limitation, and professional responsibility are not the kind of message policymakers want to hear. It may be that we have not trained up a generation of communicators, or that we have not worked to embed enough visionary statisticians at leadership levels in the military, the government, and industry. It also may be that we work too slowly. One major cultural difference between traditional statisticians and many other technical fields is that we do not use a laboratory model in which staffs of post-docs and graduate students collaborate to produce a steady stream of results, with junior hands doing the spadework while seniors focus on strategy.

The new hope in this arena is the Statistical and Applied Mathematical Sciences Institute. It is sponsoring a year-long research program on national defense and homeland security under the leadership of Nell Sedransk and Larry Cox. The kickoff meeting for this is September 11–14 and the topics to be covered include all those listed above and whatever other areas the attendees determine should be on the docket. For the good of the profession, and for the good of the nation, I hope statisticians interested in this area will lend their strength and support to its success. ■

Average Love Songs

This lyric may be sung to the tune of Paul McCartney's 1976 #1 hit "Silly Love Songs" (italicized lyrics can be overlaid by a second vocalist while first vocalist sings "I love MU" refrain).

You'd think people would've had enough of average love songs
On my TV, I see... it isn't so—oh no.
Is it mean to fill the world with average love songs?
Is it Greek to you? I'd like to know
'cause here I go... again

I love MU
I love MU
Add the values up and divide by their number
I love MU
The mean need not be a data value

Is it Greek to you? I'd like to know 'cause here I go... again
I love MU
I love MU

Song length has a central limit
Just to get played at all!
120 beats per minute:
It's expected, it's expected,
Expected value for all.....

I love MU
The mean is in between the maximum and minimum
I love MU
The mean is routine with symmetry and no outliers
I love MU
I love MU

You'd think people would've had enough of average love songs.
I follow my "mus" and see it isn't so—oh no.
I play... a mean guitar... on average love songs.
With X-barre chords.....

Copyright 2002, Lawrence Mark Lesser
(see the Winter 2002 issue of *STATS: the Magazine for Students of Statistics* for more songs)

