Response to Jon Bentley’s
“Little Experiments on Algorithms”

Jim Landwehr
(with input from Colin Mallows and Jean Meloche)

Data Analysis Research Dept.
Avaya Labs
March 6, 2008
Jon’s examples show value of data analysis in the context of small, careful experiments

For these processes to be successful, need to:
- Use substantive knowledge of the problem, along with good ideas about possible models
- Look at data in different ways, use EDA tools
- Explore different variables
- Iterate

I’m a proponent for this approach to statistics, but there are also other key ideas in statistics that have proven useful

My Discussion
- Challenger example
- Related (?) concepts
- Network experiments
The Challenger data and scatter plot discussed in a conference call the night before launch.
The mistake was that the flights with zero incidents of damaged O-rings were left off the plot because it was felt that these flights did not contribute any information about the temperature effect. The scatter plot including all the data is shown below.
Randomized Trials vis-à-vis Observational Studies and Modeling

- Randomization – great successes in agriculture, medicine, clinical trials
- Currently, economics of development aid … “The basic idea behind the lab is to rely on randomized trials … to study antipoverty programs” (*NY Times*, Feb. 20, 2008)
- Interesting editorial by David Freedman (*Chance*, Winter 2008)
- “Hiccups in data” – variables you didn’t know or think about, address via randomization?
- Question: To what extent do, or do not, the concepts of randomized trials apply to the types of experiments Jon considers?
Further Related Statistical Topics

- **Blocking**
  - Arrange experimental units into groups (blocks) so that the units in a block are similar to one another
  - Try for treatment comparisons within blocks

- **Factorial experiments vis-à-vis one-factor-at-a-time experiments**
  - Both types have roles
  - Are there situations where there might be interactions, leading to factorial experiments?
  - In order to reduce time to run experiment, do benefits from partial factorials apply?
  - Are purposes related to exploring a response surface, or more to establish/confirm/fit a theory?

**Question:** To what extent do, or do not, these concepts apply?
Experiments and Networks

- I will mention some issues and relationships to Jon’s talk
- Problem Context: VoIP, video over corporate networks
- Goals:
  - Trouble shooting for intermittent problems
  - Provide ability to adapt to changing network conditions in real time, e.g. provide very high QoS and reliability for these applications
  - Evaluating network changes (today)
  - Evaluating possible changes to network equipment and structures (for tomorrow’s uncertain applications and traffic)
Topology of Corporate Network

Colorado

New Jersey
Testing and Experimentation Process

- End-to-end (E2E) tests across network
  - Use phones or special devices for testing
- Path across network for an E2E test does matter, it can change, and it needs to be discovered
- Analyst might or might not be able to select endpoint locations for the E2E tests
  - Sometimes can insert special tests
  - Sometimes use data from phone calls occurring naturally
- Implies that the analysis inherently involves tomography
Mathematical Formulation

Notation

\[ Y_i = \begin{cases} 
1 & \text{if test fails} \\
0 & \text{if test succeeds} 
\end{cases} \]

\[ R = \text{routing matrix} \]

\[ R_{ij} = \begin{cases} 
1 & \text{if the } j^{\text{th}} \text{ node (or edge) is involved in the } i^{\text{th}} \text{ test} \\
0 & \text{otherwise} 
\end{cases} \]

for \( j = 1, \ldots, K \) and \( K \) is total number of nodes (or edges)

\[ \pi_i = \text{probability the } i^{\text{th}} \text{ test fails} \]

\[ p_j = \text{probability that performance on this node fails} \]

\[ p = \text{probability test fails although none of the nodes directly "causes" failure} \]

Then

\[ Y_i \sim \text{Bernoulli}(\pi_i) \]
Probability of Success

- End-to-end test success implies success on each traversed node and also for the background

\[(1 - \pi_i) = (1 - p) \times \prod_{\{j:R_{ij}=1\}} (1 - p_j)\]

- Log likelihood

\[C + \sum_{\{i:y_i=0\}} [\log(1 - p) + \sum_{\{j:R_{ij}=1\}} \log(1 - p_j)] + \sum_{\{i:y_i=1\}} \log(1 - [(1 - p) \times \prod_{\{j:R_{ij}=1\}} (1 - p_j)])\]
Problems

- Likelihood a complicated non-linear expression
  - Could be intractable
  - Heuristics useful
- More fundamentally, truth can be on boundary of parameter space so standard asymptotic theory does not apply
- Hence, is likelihood analysis a good approach?
- Real networks are large and raise complicated issues such as these in spades, so perhaps we should study the topic using Jon’s ideas with careful, appropriately selected “small network experiments”?
Summary

• Congratulations to Jon for
  – His series of careful “small experiments”
  – For using data analysis creatively and successfully to gain insights
  – And for promoting the approach

• I’ve tried to raise questions on how, and whether or not, other statistical topics fit into experimentation in this domain
  – Randomized trials
  – Blocking variables
  – One-variable-at-a-time vis-à-vis factorial experimentation
  – Observational data and modeling vis-à-vis randomized controlled experiments

• Networks offer a fruitful area for pursuing these topics