Big Data, Statistical Engineering, and the Future

Geoff Vining
Professor of Statistics
Virginia Tech
Outline

• Understanding Big Data
• Big Data: Opportunities and Cautions
• Large, Unstructured, Complex Problems
• Statistical Engineering
• The Science for Solving Problems with Data
• The Future
Understanding Big Data

• Several Talks on Big Data:
  – Roger Hoerl
  – Alberto Ferrer
  – Lluis Marco-Almagro
Understanding Big Data

• Current Abilities:
  – 100% Inspection via Sensors
  – Ability to Sample at Very High Rates
  – Image Data
  – Scrapping Data Bases
  – Internet of Things
Understanding Big Data

• Challenges:
  – Extreme oversampling (sometimes for a reason!)
  – Looking for a needle in a haystack
  – Large n, large p often produces sparse data
  – Complex variance structures

• The Issue: How to Distill the Information?
Understanding Big Data

• Must Consider the Fundamental Nature of the Data:
  – Information rich
  – Information poor.
    • Very few, very small needles in a massive haystack.
    • Often suffer from significant oversampling
    • Modeling Noise!
Understanding Big Data

• Most informative data points: outliers!
• An important issue: What is an outlier?
• Big Data Dominated by Computer Scientists
  – Data Quality
  – Data Cleansing
  – Populations versus Samples
Understanding Big Data

• Standard Statistical Approaches Viewed as Obstructions
  – What Do Power and ARLs Mean with Very Frequent Sampling?
  – Standard Statistical Approaches Require too Much Computation
  – Ultimately, Not Valid, Not Informative.
Understanding Big Data

• Irony: “Needle in the Haystack” Is a Statistical Process Monitoring Problem!
• Proper Solution Requires Going Back to Basics
Understanding Big Data

- NASA Example:
  - Sampling Rate: 100 Times a Second!
  - Test Duration: 168 Hours (One Week)
  - More than 60,000,000 Observations!
  - Actually, 32 Test Stands (Almost 2,000,000,000 Observations a Week!)
Understanding Big Data

• Key insight: Data Buffering
• Can Calculate Simple Statistics in Real Time over Specified Period of Time
• Challenge Remains: How to Use the Data?
Understanding Big Data

- 5 serial numbers were selected
- Approximately 1 hour, 100 Hz data
- Minor excursions: deviation from target outside approx. +/- 0.085 lbs.
- Major excursions outside +/- 2 lbs.
Understanding Big Data

- Counted minor and major excursions for each period of test data
- Specific Goal: Determine Reasonable Engineering Definition of Minor
- Final Definition of Minor: .0075 lbs.
<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Excursions +/- Control Limits</td>
<td>16</td>
</tr>
<tr>
<td>Major Excursions</td>
<td>0</td>
</tr>
<tr>
<td>Set Point</td>
<td>0.075</td>
</tr>
<tr>
<td>Upper Control Limit Minor</td>
<td>0.075</td>
</tr>
<tr>
<td>Lower Control Limit Minor</td>
<td>0.075</td>
</tr>
<tr>
<td>Upper Control Limit Major</td>
<td>2</td>
</tr>
<tr>
<td>Lower Control Limit Major</td>
<td>2</td>
</tr>
</tbody>
</table>

**SN 00604**
Proposed Phase I I-Chart

I-MR Chart of ln_adj

Data points omitted from the calculations: 1-32, 136-168
One Hour Problem

One Hour from SN 02430

- dev
- Upper
- Lower
Approximately 200 Seconds
Phase I Limits Applied to SN 00604

At least one estimated historical parameter is used in the calculations.
Opportunities and Cautions

• Massive data are an important opportunity.
• Challenge to explain to management:
  – Management believes that the data set must contain all the answers to its questions.
  – Management rarely understands the fundamental issues with observational studies.
Opportunities and Cautions

• Massive data are vulnerable to misuse.
• Must Manage Expectations!
• As a community,
  – Must address this issue in greater detail
  – Warn our practitioners about making improper conclusions.
Large, Unstructured, Complex Problems

• Problems that Keep Your CEO Awake at Night
• Solutions Are Extremely Valuable!
• These Problems Are:
  – Large
  – Unstructured
  – Complex
Large, Unstructured, Complex Problems

• Require the Proper Data!
  – Properly Collected
  – Properly Analyzed
  – Properly Interpreted

• Proper Solutions Must Be Sustainable
Large, Unstructured, Complex Problems

- Examples:
  - Roger Hoerl’s Talk
  - Pete Parker’s Talk
- Solutions Combine
  - Academic – Basic Tools
  - Practical Experience – How to Apply the Tools
Large, Unstructured, Complex Problems

- Currently, No Text Discusses Large, Unstructured, Complex Problems!
- Emerging New Discipline: Statistical Engineering
Solutions

• Interdisciplinary Teams
  – Technical – Engineering/Scientific
  – Statistical/Analytical
  – Organizational Psychology/Anthropology
  – Project Management
• Strong Leadership
Statistical Engineering

• The heart of Statistical Engineering is the scientific method.
• Most theories underlying statistical engineering involve the scientific method.
  — Deming-Shewhart PDCA
  — Six Sigma’s DMAIC
Statistical Engineering

• The Scientific Method Is
  – Fundamental Approach for Discovery and Problem Solving
  – Inductive/deductive problem solving process
Strategy of Statistical Engineering

• Identify Problem
• Provide Structure
• Understand Context
• Develop the Solution Strategy
• Develop and Execute Tactics
• Deploy Final Solution
Tactics of Statistical Engineering

• Data Acquisition
• Data Exploration
• Analysis
  – Traditional Statistical Methodologies
  – Modern Analytics (Big Data)
Tactics of Statistical Engineering

• Inference to the Process/Problem
• Deployment of Tentative Solution
  – Does It Work?
  – Is It Sustainable?
Overarching Methodologies

• Data Visualization
• Project Management
• Teamwork
• Organizational Culture
The Science for Solutions

• The Basic Tactical Tools Are Taught in Statistics/Industrial Engineering

• Data Acquisition:
  – Planning the Data Collection
  – Designing Experiments
  – Data Cleansing
The Science for Solutions

• Data Exploration:
  – Exploratory Data Analysis
  – Analytics!
  – Preliminary Models
  – Follow Up
The Science for Solutions

• Analysis
  – Statistical Inference: Estimation and Testing
  – Formal Model Building
  – Predictive Analytics
  – Critical Point:
    • Right Tool
    • Properly Applied
The Science for Solutions

• Inference Back to the Problem – What Did We Discover?
• Propose, Evaluate, Deploy Solutions
• Statistical Engineering: The Science for Deploying the Tools!
NASA Statistical Engineering Example
The Future

• Success Requires Providing Value to Organizations
• Solving Large, Unstructured, Complex Problems Provides Great Value
• Current State: People Know Tools, Often Limited Number of Tools
The Future

• Future State: How Do We Use the Tools Most Effectively
  – Most Understand the Full Range of Tools Required
  – Most Understand Best Practices for Using the Tools
  – Must Understand Leadership in the Broad Sense

• The Future Is a Journey!