The Use, Misuse, and Abuse of Statistics in Information Security Research

Julie J.C.H. Ryan, D.Sc.

The George Washington University

presented to

ASEM 2003

St Louis, MO

The Challenge

- You, the Engineering Manager, must allocate resources for security purposes
  - Options range from zero to 100% of available resources
  - How do you decide?
    - Annualized loss expectation based on probability distribution of threat activities
    - This approach requires data
      - Who are the threats, what are they doing, how often do they do it, how often are they successful, etc

Data Sources

- Plenty of data sources out there
  - Computer Security Institute (CSI) conducts an annual survey and reports the results widely
  - Large international consulting firms conduct surveys
    - Ernst & Young
    - PriceWaterhouseCoopers
  - And others
- Methods of collecting the data range from web-based collection to mailed surveys to professional organization members

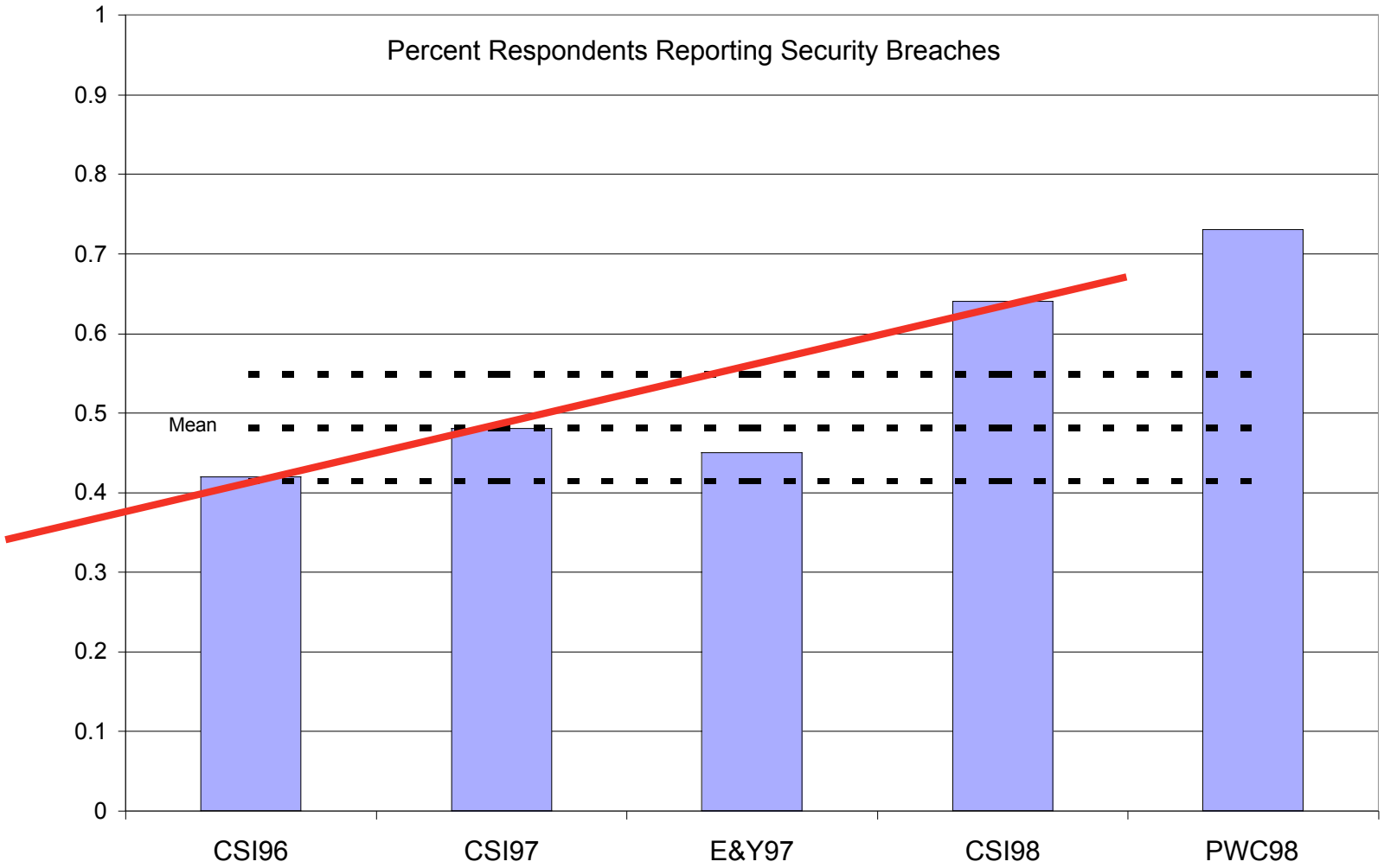
The Problem

- Data is only as good as it's meant to be
  - Many surveys are not designed to be valid
    - And say so in the body of the report
  - Most surveys include multiple responses from a single organization
    - The data is not generalizable to the organization level
- But the data is reported and represented as reliable, valid, and generalizable to the company level by many
  - Ranging from students to advisors to the US Government

Design Problems

- Design & Methodology
  - Validity of questionnaire design
    - Question design
    - Answer options
  - Selection of respondents
    - Not random, not even stratified random
    - Professionals working in the area
      - Selected through professional mailing lists
    - Same respondents year after year -- learning problem
    - Many working in the same company

Example


Percent Respondents Reporting Security Breaches
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean
CSI96
CSI97
E&Y97
CSI98
PWC98

Reporting Problems

- Descriptive statistics generally used
  - Widespread use of the term “average”
    - Mean? Median? Mode?
  - No information on distribution of data or variance
- Inferences incorrectly implied
  - Generalized responses to company level
  - No caveats given
    - Confidence interval, confidence level, alpha, distribution…

Example 1

- From a student paper:

> The Computer Security Institute recently published the 2001 CSI/FBI Computer Crime and Security Survey and it contained some very interesting statistics:
>
> -- Ninety-one percent of surveyed organizations detected employee abuse of Internet access privileges (for example, downloading pornography or pirated software, or inappropriate use of e-mail systems). Only 79% detected net abuse in 2000.
>
> -- Ninety-four percent detected computer viruses (only 85% detected them in 2000).
>
> Just in these two findings, companies must realize that they need to do everything they can to not only require security awareness training but also require the testing of those employees to determine if they have actually retained the information they were taught and to MAKE SURE they have a basic understanding of information security.

Sustaita, David, "Security Awareness Training Quiz - Finding the WEAKEST link!" August 13, 2001; http://www.sans.org/rr/papers/47/396.pdf, accessed May 31, 2003.

Example 2

- Another student paper:
  - "According to the FBI/CSI 2002 study, even though 89% of the companies surveyed have firewalls and 60% use intrusion detection systems (IDS), an alarming 40% of those surveyed still detected intrusion from the outside…"
    - Tan, Ding, "Quantitative Risk Analysis Step-By-Step'" December 2002, http://www.sans.org/rr/papers/5/ 849.pdf, accessed May 31, 2003

What's Scary

- These papers were produced in the course of the students becoming certified as information security professionals
  - The folks that others look to as experts
- The same pattern of data abuse is found in government documents
  - Example:
    - U.S. General Accounting Office, "GAO Report to the Committee on Government Affairs, U.S. Senate: Information Security Serious Weaknesses Place Critical Federal Operations and Assets at Risk", September 1998, http://www.gao.gov/archive/ 1998/ai98092.pdf, accessed May 31, 2003.

What's Needed

- Obviously, good data
  - Valid research and experimentation on infosec solutions and resource allocations
    - Beginning to occur in various research organizations
- More importantly, appreciation for statistical meaning
  - In all professions, including journalism
  - Especially for engineering managers
    - Those who will be called upon to make the hard decisions about how to manage the enterprise

Contact Information

Julie J.C.H. Ryan, D.Sc.

1776 G. Street NW #110

Washington DC, 20052

jjchryan@gwu.edu

http://www.seas.gwu.edu/~infosec/

The George Washington University is an NSA Certified Center of Academic Excellence in Information Assurance Education and meets the Federal Training Standards for Information Systems Security Professionals (NSTISSI 4011). We offer Graduate Certificate, Master's, and Doctoral level education in Information Security Management for professionals from all educational backgrounds. GWU is located in the heart of Washington DC very near the White House and other government offices.