

**Making Sense of
Statistics in
Information Security**

ISSA-Hartford Meeting
Tuesday 16 June 2009

M. E. Kabay, PhD, CISSP-ISSMP

CTO, Adaptive Cyber Security Instruments, Inc.
Assoc Prof Information Assurance
School of Business & Management
Norwich University
<http://www.mekabay.com>

1

Copyright © 2009 M. E. Kabay. All rights reserved.

Topics

- Introduction
- Fundamentals of Statistical Design and Analysis
- Resources for Further Study

2

Copyright © 2009 M. E. Kabay. All rights reserved.

Introduction

- Professional Background in Applied Statistics
- Value of Statistical Knowledge Base
- Limitations on Our Knowledge of Computer Crime
- Limitations on Applicability of Computer-Crime Statistics

3

Copyright © 2009 M. E. Kabay. All rights reserved.

Professional Background in Applied Statistics

- Studied biology, genetics at McGill 1966-1970
- Fascinated by biometrics (statistics applied to biological research) taught by Prof Hugh Tyson 1969 using Sokal & Rohlf's *Biometry* text
- Continued study independently during MSc at McGill in teratology 1970-1972
- Took PhD Dartmouth in invertebrate zoology & applied statistics 1972-1976;
 - One of PhD examiners was Dr Thomas E. Kurtz, co-inventor of BASIC (and a statistician)
- Have taught applied statistics at universities since 1975 & served as statistical consultant to scientists and industry

4

Copyright © 2009 M. E. Kabay. All rights reserved.

Value of Statistical Knowledge Base

- Security professionals often asked about
 - Frequency and security breaches
 - Severity of damage
- Bear upon risk management
 - Quantitative
 - Qualitative
- Competitive analysis
- Litigation
 - Standards of due care and diligence
 - Commonly-accepted or best practices

5

Copyright © 2009 M. E. Kabay. All rights reserved.

Limitations on Knowledge of Computer Crime: Detection

- AKA problem of ascertainment
- Not always possible to detect breach of security
- E.g., data leakage using covert channel has no record and no evidence (until competitor steals the market)
- But DoD DISA research 1995-1996 showed experimental evidence of non-detection
 - 68,000 non-classified DoD systems
 - Penetration tests broke into 2/3 of them
 - Only 4% of sysadmins noticed penetrations

6

Copyright © 2009 M. E. Kabay. All rights reserved.

Limitations on Knowledge of Computer Crime: Reporting

- Few reported in systematic way
- Unquantified, anecdotal reports of information assurance specialists
 - ❑ Only ~10% of all breaches known publicly
- DoD DISA studies support this view
 - ❑ Only ~½% of all *detected* breaches were properly reported as required by procedures
- "... COMPUTER CRIME STATISTICS SHOULD GENERALLY BE TREATED WITH SKEPTICISM."

7 Copyright © 2009 M. E. Kabay. All rights reserved.

Limitations on Applicability of Computer-Crime Statistics

- Enormous variability in computer systems and networks
 - ❑ Processors
 - ❑ Operating systems
 - ❑ Topologies
 - ❑ Firewalls
 - ❑ Encryption
 - ❑ Applications
 - ❑ ...
- How do we generalize from specific cases?
- How do we build database of usable statistics?

8 Copyright © 2009 M. E. Kabay. All rights reserved.

Fundamentals of Statistical Design and Analysis

- Descriptive Statistics
- Inference
- Hypothesis Testing
- Random Sampling
- Confidence Limits
- Contingency Tables
- Association vs Causality
- Control Groups
- Confounded Variables

9 Copyright © 2009 M. E. Kabay. All rights reserved.

Descriptive Statistics (1)

- Presentation of data can greatly influence perception of reality
- Amateurs (e.g., some reporters and PR personnel) can inadvertently or deliberately distort information through elementary mistakes
- E.g., consider 3 companies who report following losses from security breaches:
 - ❑ \$1M
 - ❑ \$2M
 - ❑ \$6M

Next page shows different ways of representing these data

10 Copyright © 2009 M. E. Kabay. All rights reserved.

Descriptive Statistics (2)

Class	Frequency
≤ \$2M	2
> \$2M	1

Left-hand table:

- Wrong impression of where the data lie
- No sense of lower or upper bounds
- No idea of gap between 1, 2 & 6
- Cannot compute mean, median at all

Class	Frequency
< \$1M	0
≥ \$1M & < \$2M	1
≥ \$2M & < \$3M	1
≥ \$3M & < \$4M	0
≥ \$4M & < \$5M	0
≥ \$5M & < \$6M	0
≥ \$6M & < \$7M	1
≥ \$7M	0

Right-hand table:

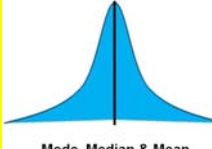
- Still wrong mean

11 Copyright © 2009 M. E. Kabay. All rights reserved.


Descriptive Statistics (3)

Measures of central tendency

- Mean (computed) – sum / total number
- Median (counted) – value of middle of sorted list
- May differ if distribution is *skewed* (asymmetric)



Mode, Median & Mean



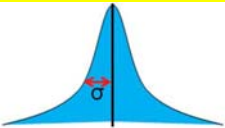
Mode Median Mean

12 Copyright © 2009 M. E. Kabay. All rights reserved.

Descriptive Statistics (4)

Measures of dispersion (variability)

- Range – largest value – smallest value
- Variance – average of squared deviations from mean (σ^2)
- Standard deviation – square root of variance (σ)
- In a Gaussian (“Normal”) frequency distribution, standard deviation is distance between mean & inflection point of curve (where slope stops increasing)



Copyright © 2009 M. E. Kabay. All rights reserved.

Inference (1)

- **Population** is entire set of all possible members
 - ❑ E.g., population of residents of USA is all people residing in USA at a specific time
 - ❑ Sample statistic is known as *parametric value*
- **Sample** is enumerated or measured set of observations
 - ❑ E.g., 100,000 people selected from US population is a sample
 - ❑ Statistic computed on sample is *sample statistic or estimator of parametric value*

Copyright © 2009 M. E. Kabay. All rights reserved.

Inference (2)

- Statisticians try to infer population statistics from sample statistics
 - ❑ Called *statistical inference*
 - ❑ E.g., population mean is μ and sample mean is \bar{Y} ; parametric variance is σ^2 and sample is s^2
- Sample statistics sometimes have different formula from parametric statistic
 - ❑ E.g., \bar{Y} estimates μ
 - ❑ But estimator s^2 of σ^2 is sum of squared deviations from mean divided by $(n-1)$ instead of by n [where n is sample size]

Copyright © 2009 M. E. Kabay. All rights reserved.

Hypothesis Testing (1)

- Often need to test an idea (hypothesis) about populations based on sample statistics; e.g.,
 - ❑ Testing idea that μ lies between 1.3 & 4.3 based on a sample mean $\bar{Y} = 2.8$
 - ❑ Testing idea that $\sigma \leq 35.6$ based on $s = 52.8$
- Can also test hypotheses about relationships
 - ❑ E.g., given observed data in table, test idea that firewalls and penetration

Firewalls	Penetration		Totals
	No	Yes	
No	25	75	100
Yes	70	130	200
Totals	95	205	300

Copyright © 2009 M. E. Kabay. All rights reserved.

Hypothesis Testing (2)

- Null hypothesis (H_0) is that there is no relationship
- Testing for relation between two independent variables
 - ❑ Presence of firewall
 - ❑ Detection of penetration
- Various calculations available to test for *independence*; e.g.,
 - ❑ Chi-square χ^2
 - ❑ Log-likelihood ratio G
- Both are 0 in a population where there is no relationship between variables
 - ❑ Compute probability that sample statistic would occur by chance alone if really 0 in population

Firewalls	Penetration		Totals
	No	Yes	
No	25	75	100
Yes	70	130	200
Totals	95	205	300

Copyright © 2009 M. E. Kabay. All rights reserved.

Hypothesis Testing (3)

Probability that the null hypothesis is true

- $p(H_0) > 0.05$: *not statistically significant (symbols ns)*
- $0.05 \geq p(H_0) > 0.01$: *statistically significant (*)*
- $0.01 \geq p(H_0) > 0.001$: *highly statistically significant (**)*
- $p(H_0) \leq 0.001$: *extremely statistically significant (***)*.

Copyright © 2009 M. E. Kabay. All rights reserved.

Random Sampling (1)

- Randomization essential to all of statistical inference
- Sample is *random* when every member of population has equal likelihood of being selected for sample
- Non-random sample is *biased*
 - ❑ E.g., population is all members of multinational company BUT most employees picked are disproportionately from US subsidiaries – biased toward US sub-group
 - ❑ E.g., population is all adult US residents but 2x as many men are selected as women – gender bias

19 Copyright © 2009 M. E. Kabay. All rights reserved.

Random Sampling (2)

- *Surveys can suffer from response bias*
 - ❑ What if survey is known only to a subset of desired population?
 - ❑ What if results report only those who respond?
 - ❑ What if those who respond are different from those who do not respond?
- The response bias can *confound* variables:
 - ❑ Subjects of the questions are confounded with
 - ✓ Awareness of the survey
 - ✓ Tendency to respond

20 Copyright © 2009 M. E. Kabay. All rights reserved.

Confidence Limits (1)

- Point estimates not generally useful
 - ❑ The average salary was \$38,232
 - ❑ The cost of gasoline rose \$0.12 per week last quarter
- Generally prefer to have a sense of reliability
 - ❑ Often report mean ± standard deviation
 - ✓ The average salary was \$38,232 ± \$1955
 - ✓ The cost of gasoline rose \$0.12 ± \$0.035 per week last quarter
- Should specify sample size to give *intuitive* sense of reliability
 - ❑ The average salary was \$38,232 ± \$1955 (n = 12)
 - ❑ The average salary was \$38,232 ± \$1955 (n = 12,000)

21 Copyright © 2009 M. E. Kabay. All rights reserved.

Confidence Limits (2)

- Can compute ranges that have a known probability of including the parametric value being estimated:
 - ❑ The probability that the average salary was between \$36,277 & \$40,187 based on the sample statistics is 95%.
 - ❑ The 95% confidence limits of the average salary were \$36,277 & \$40,187
- *Confidence limit* computations depend on
 - ❑ Random sampling
 - ❑ Known error distribution (e.g., Normal/Gaussian)
 - ❑ Equal variances at all values
 - ✓ Larger values no more variable than smaller values

22 Copyright © 2009 M. E. Kabay. All rights reserved.

Contingency Tables

- Contingency tables present counted (enumerated) data for two or more variables
- Common error: Presenting only part of contingency table
 - ❑ “Over 70% of systems without firewalls were penetrated last year”
 - ❑ Yes, but what % of systems with firewalls were penetrated?

FIREWALLS AND PENETRATION	FIREWALLS AND PENETRATION	
	Without Firewalls	With Firewalls in Default Config
Penetrated	70	70
Not Penetrated	30	30

FIREWALLS AND PENETRATION	FIREWALLS AND PENETRATION	
	Without Firewalls	With Firewalls Properly Config
Penetrated	70	10
Not Penetrated	30	90

23 Copyright © 2009 M. E. Kabay. All rights reserved.

Association vs Causality

- Don't mistake *association* for *causality*
 - ❑ Error of logic known as *post hoc, ergo propter hoc* – after the fact, thus because of the fact
 - ❑ E.g., suppose study shows that organizations with lots of fire extinguishers have lower rate of computer network penetration than those with few fire extinguishers
 - ❑ Do we conclude that presence of fire extinguishers causes better resistance to penetration?
- Many possible explanations for association other than causality

24 Copyright © 2009 M. E. Kabay. All rights reserved.

Control Groups

- When associated variables may be confounded, one can *control* for the variables
- E.g., in fire-extinguisher case
 - ❑ Measure state of security awareness
 - ❑ Compare groups with similar level of awareness
- Statistical techniques exist to control for independent variables and their interactions
 - ❑ Analysis of variance with regression
 - ❑ Multivariate analysis of contingency tables

25 Copyright © 2009 M. E. Kabay. All rights reserved.

More about Confounded Variables

- "One in 10 employees admitted stealing data or corporate devices, selling them for a profit, or knowing fellow employees who did."
- Confounds
 - ❑ Theft of data
 - ❑ Theft of devices
 - ❑ Selling things for profit
 - ❑ Knowing of others who did such criminal acts
- Cannot tease out the individual contributions
- "Knowing" particularly bad: confounds occurrence with social networking
 - ❑ If everyone knows everyone's business, could have 100% +ve response even if only 1% were criminals

26 Copyright © 2009 M. E. Kabay. All rights reserved.

For Further Reading

- Kabay, M. E. (2009). Understanding Studies and Surveys of Computer Crime:
http://www.mekabay.com/methodology/crime_stats_methods.pdf
(the apparent blanks are the underscore character, _)
http://www.mekabay.com/methodology/crime_stats_methods.htm
- Any introductory text for applied statistics in the social sciences
- Any introductory text on survey design and analysis

27 Copyright © 2009 M. E. Kabay. All rights reserved.

Sample Textbooks

- Babbie, E. R., F. S. Halley & J. Zaino (2003). *Adventures in Social Research : Data Analysis Using SPSS 11.0/11.5 for Windows, 5th Ed.* Pine Science Press (ISBN 0-761-98758-4).
- Sirkin, R. M. (2005). *Statistics for the Social Sciences, 3rd Ed.* Sage Publications (ISBN 1-412-90546-X).
- Schutt, R. K. (2003). *Investigating the Social World: The Process and Practice of Research, Fourth Edition.* Pine Science Press (0-761-92928-2).

28 Copyright © 2009 M. E. Kabay. All rights reserved.

Sample Web Sites

- Creative Research Systems "Survey Design"
<http://www.survevsystem.com/sdesign.htm>
- New York University "Statistics & Social Science"
<http://www.nyu.edu/its/socsci/statistics.html>
- StatPac "Survey & Questionnaire Design"
<http://www.statpac.com/surveys/>
- University of Miami Libraries "Research Methods in the Social Sciences: An Internet Resource List"
<http://www.library.miami.edu/netguides/psymeth.html>

29 Copyright © 2009 M. E. Kabay. All rights reserved.

Discussion

30 Copyright © 2009 M. E. Kabay. All rights reserved.