

Understanding Studies and Surveys of Computer Crime*

by M. E. Kabay, PhD, CISSP-ISSMP

***Assoc Prof of Information Assurance
School of Business and Management,
Norwich University, Northfield, VT***

Version 6 – June 2009

* The original 2001 version of this paper eventually became the basis for Chapter 4 of Bosworth, S. & M. E. Kabay (2002), eds. *Computer Security Handbook, 4th Edition*. Wiley (New York). ISBN 0-471-41258-9. 1184 pp. Index. An updated version was published as Chapter 10 of the 5th Edition of the *Computer Security Handbook* edited by Bosworth, S., M. E. Kabay & E. Whyne (2009). Wiley (New York) ISBN 0-471-71652-9. Two volumes; 1856 pp. Index.

Contents

1	Introduction	3
1.1	Value of statistical knowledge base.....	3
1.2	Limitations on our knowledge of computer crime	3
1.2.1	Detection.....	3
1.2.2	Reporting.....	3
1.3	Limitations on the applicability of computer-crime statistics.....	4
2	Basic research methodology	4
2.1	Some fundamentals of statistical design and analysis	4
2.1.1	Descriptive statistics	4
2.1.2	Inference: sample statistics versus population statistics	6
2.1.3	Hypothesis testing.....	7
2.1.4	Random sampling, bias and confounded variables	8
2.1.5	Confidence limits	8
2.1.6	Contingency tables.....	9
2.1.7	Association versus causality.....	10
2.1.8	Control groups	10
2.1.9	Confounded variables.....	10
2.2	Research methods applicable to computer crime.....	12
2.2.1	Interviews	12
2.2.2	Focus groups.....	13
2.2.3	Surveys	13
2.2.4	Instrument validation	13
3	Summary	14
4	For Further Reading	15

1 Introduction

This review is intended to provide guidance for critical reading of research results about computer crime. It will also alert designers of research instruments to the need for professional support in developing questionnaires and analyzing results.

1.1 Value of statistical knowledge base

Security specialists are often asked about computer crime; for example, customers want know who is attacking which systems how often using what methods. These questions are perceived as important because they bear upon the strategies of risk management; in theory, in order to estimate the appropriate level of investment in security, it would be helpful to have a sound grasp of the probability of different levels of damage. Ideally, one would want to evaluate an organization's level of risk by evaluating the experiences of other organizations with similar system and business characteristics. Such comparisons would be useful in competitive analysis and in litigation over standards of due care and diligence in protecting corporate assets.

1.2 Limitations on our knowledge of computer crime

Unfortunately, in the current state of information security, no one can give reliable answers to such questions. There are two fundamental difficulties preventing us from developing accurate statistics of this kind. These difficulties are known as the problems of ascertainment.

1.2.1 Detection

The first problem is that an unknown number of crimes of all kinds are undetected. For example, even outside the computer crime field, we don't know how many financial frauds are being perpetrated. We don't know because some of them are not detected. How do we know they're not detected? Because some frauds are discovered long after they have occurred. Similarly, computer crimes may not be detected by their victims but may be reported by the perpetrators.

In a landmark series of tests at the Department of Defense, the Defense Information Systems Agency found that very few of the penetrations it engineered against unclassified systems within the DoD seem to have been detected by system managers. These studies were carried out from 1994 through 1996 and attacked 68,000 systems. About two-thirds of the attacks succeeded; however, only 4% of these attacks were detected.

A commonly-held view within the information security community is that only one-tenth or so of all the crimes committed against and using computer systems are detected.

1.2.2 Reporting

The second problem of ascertainment is that even if attacks are detected, it seems that few are reported in a way that allows systematic data collection. This belief is based in part on the unquantified experience of information security professionals who have conducted interviews of their clients; it turns out that only about ten percent of the attacks against computer systems revealed in such interviews were ever reported to any kind of authority or to the public. The Department of Defense studies mentioned above were consistent with this belief; of the few penetrations detected, only a fraction of one percent were reported to appropriate authorities.

Given these problems of ascertainment, computer crime statistics should generally be treated with skepticism.

1.3 Limitations on the applicability of computer-crime statistics

Generalizations in this field are difficult to justify; even if we knew more about types of criminals and the methods they use, it would still be difficult to have the kind of actuarial statistic that is commonplace in the insurance field. For example, the establishment of uniform building codes in the 1930s in the United States led to the growth in fire insurance as a viable business. With official records of fires in buildings that could be described using a standard typology, statistical information began to provide an actuarial basis for using probabilities of fires and associated costs to calculate reasonable insurance rates.

In contrast, even if we had access to accurate reports, it would be difficult to make meaningful generalizations about vulnerabilities and incidence of successful attack for the information technology field. We use a bewildering variety and versions of processors, operating systems, firewalls, encryption, application software, backup methods and media, communications channels, identification, authentication, authorization, compartmentalization and operations.

How would we generalize from data about the risks at (say) a mainframe-based network running MVS in a military installation to the kinds of risks faced by a UNIX-based intranet in an industrial corporation or to a Windows NT-based Web server in a university setting? There are so many differences among systems that if we were to establish a multidimensional analytical table where every variable was an axis, many cells would likely contain no or only a few examples. Such sparse matrices are notoriously difficult to use in building statistical models for predictive purposes.

2 Basic research methodology

This is not an article about social sciences research. However, many discussions of computer crime seem to take published reports as gospel, even though these studies being discussed may have no validity whatsoever. In this short section, we will look at some fundamentals of research design so that readers will be able to judge how much faith to put in computer crime research results.

2.1 Some fundamentals of statistical design and analysis

The way a scientist or reporter represents data can make enormous differences in the readers' impressions.

2.1.1 Descriptive statistics

Suppose three companies reported the following losses from penetration of their computer systems: \$1M, \$2M and \$6M. We can describe these results in many ways. For example, we can simply list the raw data; however, such lists could become unacceptably long as the number of reports increased and it is hard to make sense of the raw data.

We could define classes such as "2 million or less" and "more than 2 million" and count how many occurrences there were in each class, as shown in Table 1.

Table 1. Crude representation of frequency data.

Class	Freq
$\leq \$2M$	2
$> \$2M$	1

Understanding Studies and Surveys of Computer Crime

Alternatively, we might define the classes with finer granularity as $< \$1M$, $\geq \$1M$ but $< \$2M$, and so on; such a table might look like Table 2:

Table 2. More granular frequency distribution.

Class	Freq
$< \$1M$	0
$\geq \$1M$ & $< \$2M$	1
$\geq \$2M$ & $< \$3M$	1
$\geq \$3M$ & $< \$4M$	0
$\geq \$4M$ & $< \$5M$	0
$\geq \$5M$ & $< \$6M$	0
$\geq \$6M$ & $< \$7M$	1
$\geq \$7$	0

Notice how the definition of the classes affects perception of the results: the first table gives the impression that the results are clustered around \$2M and gives no information about the upper or lower bounds. The second table provides more information at the level of million-dollar granularity.

2.1.1.1 Location

One of the most obvious ways we describe data is to say where they lie in a particular dimension. The *central tendency* of our three data can be represented in various ways; for example, three popular measures are

the *arithmetic mean* or *average* = $\$(1+2+6)M/3 = \$3M$

the *median* (the middle of the sorted list of losses) = \$2M

the *mode* (the most frequent category in a frequency distribution (not meaningful in our *multimodal* example because there are three equally-frequent classes)

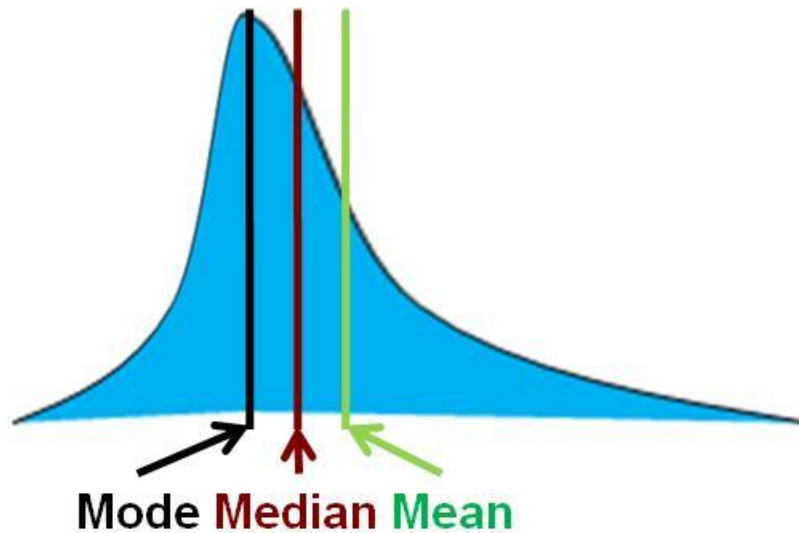
Note that it is not possible to compute the mean or the median from the first table, with its crude approximations. Attempting to compute the mean from Table 2 gives $\$(1.5+2.5+6.5)M/3 = \$3.5M$ and the approximated median is \$2.5M.

Such statistics should be computed from the original data if possible, not from summary tables; for summary tables that are the only source of data, the more granular the better.

Note that the mode, arithmetic mean and median are the same for symmetric distributions but may be very different for asymmetric (*skewed*) distributions, as shown in Figure 1 below.

Understanding Studies and Surveys of Computer Crime

Figure 1. Measures of central tendency in an asymmetric frequency distribution.



2.1.1.2 Dispersion

Another aspect of our data that we frequently need is *dispersion* – i.e., variability. The simplest measure of dispersion is the range – the difference between the smallest and the largest value we found; in our example, we could say that the range was from \$1M to \$6M or that it was \$5M. Sometimes the range is expressed as a percentage of the mean; then we would say that the range was $5/3 = 1.6\dots$ or $\sim 167\%$.

The *variance* (σ^2) of these particular data is the average of the squared deviations from the arithmetic mean; the variance of the three numbers would be $\sigma^2 = (1-3)^2 + (2-3)^2 + (6-3)^2 / 3 = (4+1+9)/3 \approx 4.67$.

The square root of the variance (σ) is called the *standard deviation* and is often used to describe dispersion. In our example, $\sigma = \sqrt{4.67} \approx 2.16$. The standard deviation divided by the mean is the *coefficient of variation*, $c_v = \sigma/\mu$. In our example, $c_v = 2.16/3 = 72\%$.

Dispersion is particularly important when we compare estimates about information from different groups. The greater the variance of a measure, the more difficult it is to form reliable generalizations about an underlying phenomenon, as I'll describe in the next section.

2.1.2 Inference: sample statistics versus population statistics

We can accurately describe any data using descriptive statistics; the question is what we then do with those measures.

Usually we expect to extend the findings in a *sample* or subset of a *population* to make generalizations about the population. For example, we might be trying to estimate the losses from computer crime in commercial organizations with offices in the United States and with more than 30,000 employees. Or perhaps our sample would represent commercial organizations with offices in the United States and with more than 30,000 employees and whose network security staff were willing to respond to a survey questionnaire.

Understanding Studies and Surveys of Computer Crime

In such cases, we try to infer the characteristics of the population from the characteristics of the sample. Statisticians say that we try to estimate the *parametric* statistics by using the *sample* statistics.

For example, we estimate the parametric (population) variance (usually designated σ^2) by multiplying the variance of the sample by $n/(n-1)$. Thus we would say that the estimate of the parametric variance (s^2) in our sample above would be $s^2 = 4.67 * 3/2 = 7$. The estimate of the parametric standard deviation (s) would be $s = \sqrt{7} \approx 2.65$.

2.1.3 Hypothesis testing

Another kind of inference that we try to make from data is *hypothesis testing*. We test ideas based on the information collected in samples. For example, one can test the idea that a population mean is no larger than a certain value or that a parametric variance lies within a specified range.

We can also test for the *existence* of relationships. Suppose we were interested in whether there was any *association* between the presence or absence of firewalls and the occurrence of system penetration. We can imagine collecting the following data about penetrations into systems with or without firewalls as shown in Table 3:

Table 3. Contingency table for firewalls vs penetration.

		Penetration		<i>Totals</i>
		<i>No</i>	<i>Yes</i>	
<i>Firewalls</i>	<i>No</i>	25	75	100
	<i>Yes</i>	70	130	200
<i>Totals</i>		95	205	300

We would frame the hypothesis (the *null hypothesis*, sometimes represented as H_0) that there was *no* relationship between the two independent variables, penetration and firewalls and *test* that hypothesis by performing a test of independence of these variables. In our example, a simple *chi-square test of independence* would give a *test statistic* of $\chi^2_{[1]} = 2.636$ with what is referred to as a single degree of freedom (the “[1]” symbol). If there really were no association between penetration and firewalls in the population of systems under examination, the parametric value of this statistic would be zero. In our imaginary example, we can show that such a large value (or larger) of $\chi^2_{[1]}$ would occur in only 10.4% of the samples taken from a population where firewalls had no effect on penetration. Put another way, if we took lots of samples from a population where the presence of firewalls was not associated with any change in the rate of penetration, we’d see around 10.4% of those samples producing $\chi^2_{[1]}$ statistics as large as or larger than 2.636.

Statisticians have agreed on some conventions for deciding whether a test statistic deviates enough from the value expected under the null hypothesis to warrant inferring that the null hypothesis is wrong. Generally we describe the likelihood that the null hypothesis is true – often shown as $p(H_0)$ – as follows:

- When $p(H_0) > 0.05$ we say the results are *not statistically significant* (often designated with the symbols *ns*);

Understanding Studies and Surveys of Computer Crime

- When $0.05 \geq p(H_0) > 0.01$ the results are described as *statistically significant* (often designated with the symbol *);
- When $0.01 \geq p(H_0) > 0.001$ the results are described as *highly statistically significant* (often designated with the symbols **);
- When $p(H_0) \leq 0.001$ the results are described as *extremely statistically significant* (often designated with the symbols ***).

2.1.4 Random sampling, bias and confounded variables

The most important element of sampling is randomness. We say that a sample is *random* or *randomized* when every member of the population we are studying has an equal probability of being selected. When a population is defined one way but the sample is drawn non-randomly, the sample is described as *biased*. For example, if the population we are studying were designed to be, say, all companies worldwide with more than 30,000 full-time employees but we sampled mostly from such companies in the United States, the sample would be biased towards US companies and their characteristics. Similarly, if we were supposed to be studying security in all companies in the United States with more than 30,000 full-time employees but we sampled only from those companies who were willing to respond to a security survey, we would be at risk of having a biased sample.

In this last example involving studying only those who respond to a survey, we say that we are potentially *confounding* variables: we are looking at people-who-respond-to-surveys and hoping they are representative of the larger population of people from all companies in the desired population. But what if the people who are *willing* to respond are those who have better security and those who don't respond have terrible security? Then *responding to the survey* is confounded with *quality of security* and our biased sample could easily mislead us into overestimating the level of security in the desired population.

Another example of how variables can be confounded is comparisons of results from surveys carried out in different years. Unless exactly the same people are interviewed in both years, we may be confounding individual variations in responses with changes over time; unless exactly the same companies are represented, we may be confounding differences among companies with changes over time; if external events have led people to be more or less willing to respond truthfully to questions, we may be confounding willingness to respond with changes over time. If the surveys are carried out with different questions or used by different research groups, we may be confounding changes in methodology with changes over time.

2.1.5 Confidence limits

Because random samples naturally vary around the parametric (population) statistics, it is not usually enough for practical purposes to report a *point estimate* of the parametric value. For example, if we read that the mean damage from computer crimes in a survey was \$180,000 per incident, what does that imply about the population mean?

To express our confidence in the sample statistic, we calculate the likelihood of being right if we give an *interval estimate* of the population value. For example, we might find that we would have a 95% likelihood of being right in asserting that the mean damage was between \$160,000 and \$200,000. In another sample, we might be able to narrow these *95% confidence limits* to \$175,000 and \$185,000.

In general, the larger the sample size, the narrower the confidence limits will be for particular statistics.

Understanding Studies and Surveys of Computer Crime

The calculation of confidence limits for statistics depends on some necessary *assumptions*:

- random sampling;
- a *known error distribution* (usually the *Normal* distribution – sometimes called a *Gaussian* distribution);
- *equal variance* at all values of the measurements (e.g., larger values are no more variable than smaller values).

If any of these assumptions is wrong, the calculated confidence limits for our estimates will be wrong; i.e., they will be misleading. There are tests of these assumptions that analysts should carry out before reporting results; if the data do not follow Normal error distributions, sometimes one can apply *normalizing transformations*.

In particular, percentages do not follow a Normal distribution. Here is a reference table of confidence limits for various percentages in a few representative sample sizes.

Table 4. 95% Confidence Limits for Percentages.

Sample size			
Percentage	100	500	1000
0	0-3.0%	0-0.6%	0-0.3%
10	4.9-17.6%	7.5-13.0%	8.2-12.0%
20	12.7-29.1%	16.6-23.8%	17.6-22.6%
50	40.0-60.1%	45.5-54.5%	46.9-53.1%
80	70.9-87.3%	76.2-83.4%	77.4-82.4%
90	82.4-95.1%	87.0-92.5%	88.0-91.8%
100	97.0-100%	99.4-100%	99.7-100%

2.1.6 Contingency tables

One of the most frequent errors in reporting results of studies is to provide only part of the story. For example, one can read statements such as “Over 70% of the systems without firewalls were penetrated last year.” Such a statement may be true, but it cannot be correctly be interpreted as meaning that systems with firewalls were necessarily more or less vulnerable to penetration than systems without firewalls. The statement is incomplete; to make sense of it, we need the other part of the implied *contingency table* – the percentage of systems *with* firewalls that were penetrated last year – before making any assertions about the relationship between firewalls and penetrations. Compare, for example these two hypothetical tables:

Understanding Studies and Surveys of Computer Crime

Table 5a & b. Firewalls and penetration.

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

In both cases, someone could say that “70% of the systems without firewalls were penetrated,” but the implications would be radically different in the two data sets. Without knowing the right-hand column of each table, the original assertion would be meaningless. This example also illustrates confounding of variables, which is discussed in more detail below in section 2.1.9.

2.1.7 Association versus causality

Continuing our example with rates of penetration, another error that untrained people often make when studying statistical information is to mistake *association* for *causality*. Imagine that a study showed that a lower percentage of systems with fire extinguishers were penetrated than systems without fire extinguishers and that this difference were statistically highly significant. Would such a result necessarily mean that fire extinguishers *cause* the reduction in penetration? No, we know that it’s far more reasonable to suppose guess that the fire extinguishers might be installed in organizations whose security awareness and security policies were more highly developed than in the organizations where no fire extinguishers were installed. In this imaginary example, the fire extinguishers might actually have *no causal effect* whatever on resistance to penetration. This result would illustrate the effect of confounding variables – *presence of a fire extinguisher with state of security awareness and policies*.

2.1.8 Control groups

Finally, to finish our penetration example, one way to distinguish between association and causality is to *control* for variables. For example, one could measure the state of security awareness and policy as well as the presence/absence of fire extinguishers and make comparisons only groups with the same level of awareness and policy. There are also statistical techniques for mathematically controlling for differences in such *independent variables*. These are part of multivariate statistical analysis such as analysis of variance (ANOVA), multivariate ANOVA (MANOVA), regression and multiple regression, and multivariate contingency-table analysis (e.g., using log-likelihood ratios, G).

2.1.9 Confounded variables*

Sometimes one reads statements such as, “One in 10 employees surveyed admitted stealing data or corporate devices, selling them for a profit, or knowing fellow employees who did. This finding was most prevalent in France, where 21% of employees admitted knowledge of this behavior.”

* This section is drawn from a column I wrote for *Network World Security Strategies* published in January 2009 entitled “Confounded Nonsense.”

Understanding Studies and Surveys of Computer Crime

The problem with the statement about admitting stealing data, admitting stealing devices, selling stolen data for a profit, selling stolen devices for a profit, or knowing employees who stole is that it combines many different causative factors that can result in the response. For example, suppose we are studying the effect of a new series of security-awareness cartoons on employees. One could form two groups, the cartoonified group (C+) and the uncartoonified group (C-) and then study their susceptibility to, say, phishing attacks sent to them via e-mail. Sounds great! We do the test and end up with the following table:

Table 6. Contingency table for cartoons and trickery.

	Tricked	Not Tricked
C-	72	128
C+	52	148

For statistics aficionados, we compute a chi-square statistical test of independence with a value of 4.219 (with 1 degree of freedom) for a probability of 0.04 that there is no relationship between cartoon exposure and resistance to phishing. So obviously exposure to the cartoons increased resistance to phishing messages, at least at the 0.05 level of significance, right?

Ah, but suppose that, without reporting the fact, we actually have an additional orthogonal (independent) factor defining two groups of employees: those who have previously been given a full-day security-awareness workshop (W+) and those who have not (W-). Well, that means that there are actually four test groups: W-C-, W-C+, W+C- and W+C+. And then we find out belatedly that the results, when classified with the additional information about security training, are as follows:

Table 7. Contingency table for workshop, cartoons and trickery.

		Tricked	Not Tricked
W-	C-	48	52
W-	C+	44	56
W+	C-	24	76
W+	C+	8	92

So the results with both variables displayed indicate quite a different story: the cartoons have very little effect on people who had no security-awareness training but there was a noteworthy improvement after exposure to the cartoons among those who had been trained. In statistical terms, we call this phenomenon an *interaction* between the independent variables (workshops and cartoons); there are tests for decomposing the effects precisely (the log-likelihood ratio, G, is my favorite). For readers who have studied analysis of variance (ANOVA), the G-test is the non-parametric equivalent of a multivariate ANOVA. But enough of this airy persiflage.

Understanding Studies and Surveys of Computer Crime

In statistical analysis, we refer to *confounded* variables when an analysis varies more than one attribute in a measured variable (an independent variable) and then ascribes fluctuations in a result (a dependent variable) to only one of the variables. In our example, the study confounds exposure to cartoons (what the survey claims accounts for resistance to phishing) with the unreported variable, exposure to security-awareness training.

You can see that because the original study confounded the two variables (cartoon exposure and awareness training), the analysis of the pooled data was misleading: it falsely ascribed the difference in response to phishing to the cartoons alone. So now let's go back to the issue of people who admitted to having stolen data *or to knowing someone who did*.

The general principle here is as follows: Statements of the form "X% of respondents admitted doing Y or knowing of coworkers who did Y" don't mean *anything*. They confound a number of factors into a meaningless jumble:

- 1) How many people do Y?
- 2) How many people who do Y are willing to admit it to an interviewer or on a survey?
- 3) How secretive are people who do Y about letting coworkers know about their actions?
- 4) How many people learn about a single person's transgressions?

I won't even discuss the possibility that some people will report personally-held beliefs or rumors they have heard.

The issue of responsiveness (factor 2) is inherent in all studies and surveys, but factors 3 and 4 are at the heart of the problem here. If criminals are blabbermouths, those "knowing of coworkers who do" will rise; similarly, if the social networking of criminals is high, more people will know about the crimes than if the criminals are relatively private people. In any case, the confounding of *doing* the crimes and *knowing about* the crimes makes the statistic useless.

As a simple example of how misleading the garbled statistic can be, imagine a reduction to absurdity. Suppose a single person in a company of 10,000 steals trade information and gets arrested and convicted. The security department releases an alert about the case as part of the security-awareness program and all 9,999 other employees therefore know about the case. An interviewer arrives some time later and interviews a hundred employees, all of whom say that they have NEVER stolen trade secrets but ALL of whom say they know of someone who did. The report would state that "100% of the employees admitted stealing data or knowing fellow employees who did."

Nonsense.

2.2 Research methods applicable to computer crime

2.2.1 Interviews

Interviewing individuals can be illuminating. In general, interviews provide a wealth of data that are unavailable through any other method. For example, one can learn details of computer crime cases or motivations and techniques used by computer criminals. Interviews can be structured (using precise lists of questions) or unstructured (allowing the interviewer to respond to new information by asking additional questions at will).

Interviewers can take notes or record the interviews for later word-for-word transcription. In unstructured interviews, skilled interviewers can probe responses to elucidate nuances of meaning

Understanding Studies and Surveys of Computer Crime

that might be lost using cruder techniques such as surveys. Techniques such as thematic analysis can reveal patterns of responses that can then be examined using exploratory data analysis.

2.2.2 Focus groups

Focus groups are like group interviews. Generally the facilitator uses a list of predetermined questions and encourages the participants to respond freely and to interact with each other. Often the proceedings are filmed from behind a one-way mirror for later detailed analysis. Such analysis can include non-verbal communications such as facial expressions and other body language as the participants speak or listen to others speak about specific topics.

2.2.3 Surveys

Surveys consist of asking people to answer a fixed series of questions with lists of allowable answers. They can be carried out face-to-face, or by distributing and retrieving questionnaires by telephone, mail, fax, and e-mail. Some questionnaires have been posted on the Web.

The critical issue when considering the reliability of surveys is *self-selection bias* – the obvious problem that survey results include only the respondents of people who agreed to participate. Before basing critical decisions on survey data, it is useful to find out what the response rate was; although there are no absolutes, in general we tend to trust survey results more when the response rate is high. Unfortunately, response rates for telephone surveys are often less than 10%; response rates for mail and e-mail surveys can be less than 1%. It is very difficult to make any case for random sampling under such circumstances, and all results from such low-response-rate surveys should be viewed as indicating the range of problems or experiences of the respondents rather than as indicators of population statistics.

As for Web-based surveys, there are two types from a statistical point of view: those using strong identification and authentication and those that don't. Those that do not are vulnerable to fraud such as repeated voting by the same individuals. Those that provide individual URLs to limit voting to one per person nonetheless suffer from the same problems of self-selection bias as any other survey.

2.2.4 Instrument validation

Interviews and other social-sciences research methodologies can suffer from a systematic tendency for respondents to shape their answers to please the interviewer or to express opinions that may be closer to the norm in whatever group they see themselves belonging to. Thus if it is well known that every organization ought to have a business continuity plan, some respondents may misrepresent the state of their business continuity planning to look better than they really are.

In addition, survey instruments may distort responses by phrasing questions in a biased way; for example, the question “Does your business have a completed business continuity plan?” may have a more accurate response rate than the question, “Does your business comply with industry standards for having a completed business continuity plan?” The latter question is not neutral and is likely to increase the proportion of “yes” answers.

The sequence of answers may bias responses; exposure to the first possible answers can inadvertently establish a baseline for the respondent. For example, a question about the magnitude of virus infections might ask

“In the last 12 months, has your organization experienced total losses from virus infections of

Understanding Studies and Surveys of Computer Crime

- a) \$1M or greater;
- b) less than \$1M but greater than or equal to \$100,000;
- c) less than \$100,000;
- d) none at all?"

To test for bias, the designer can create versions of the instrument in which the same information is obtained using the opposite sequence of answers:

“In the last 12 months, has your organization experienced total losses from virus infections of

- a) none at all;
- b) less than \$100,000;
- c) less than \$1M but greater than or equal to \$100,000;
- d) \$1M or greater?"

The sequence of questions can bias responses; having provided a particular response to a question, the respondent will tend to make answers to subsequent questions about the same topic conform to the first answer in the series. To test for this kind of bias, the designer can create versions of the instrument with questions in different sequences.

Another instrument validation technique inserts questions with no valid answers or with meaningless jargon to see if respondents are thinking critically about each question or merely providing any answer that pops into their heads. For example, one might insert the nonsensical question, “Does your company use steady-state quantum interference methodologies for intrusion detection?” into a questionnaire about security and invalidate the results of respondents who answer “Yes” to this and other diagnostic questions.

Finally, independent verification of answers provides strong evidence of whether respondents are answering truthfully. However, such intrusive investigations are rare.

3 Summary

In summary, all studies about computer crime should be studied carefully before we place reliance on their results. Some basic take-home messages about such research:

- What is the population we are sampling?
- Keep in mind the self-selection bias: how representative of the wider population are the respondents who agreed to participate in the study or survey?
- How large is the sample?
- Are the authors testing for the assumptions of randomness, normality and equality of variance before reporting statistical measures?
- What are the confidence intervals for the statistics being reported?
- Are comparisons confounding variables?
- Are correlations being misinterpreted as causal relations?
- Were the test instruments validated?

4 For Further Reading

- Textbooks:
 - If you are interested in learning more about survey design and statistical methods, you can study any elementary textbook on the social sciences statistics. Here are some sample titles.
 - 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).
- Web sites:
 - Creative Research Systems “Survey Design” <http://www.surveysystem.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>

