Appendix
Methodology and Statistics

Introduction

Science and Engineering Indicators (SEI) contains data compiled from a variety of sources. The purpose of this appendix is to explain the methodological and statistical criteria used to assess possible data sources for inclusion in SEI and to develop statements about the data. It also provides some basic information about how statistical procedures and reasoning are applied.

The first section describes the statistical considerations that are part of the selection process for data sets to be included in SEI. The next section discusses the different types of data (e.g., sample surveys, censuses, and administrative records) used in the report and provides some information about each type. A section on data accuracy follows, discussing factors that can affect accuracy at all stages of the survey process. The last section discusses the statistical testing employed to determine whether differences between sample survey-based estimates are *statistically significant*, i.e., greater than could be expected by chance. The appendix concludes with a glossary of statistical terms commonly used or referred to in the text. Selected key terms appear in bold in the text.

Selection of Data Sources

Four criteria guide the selection of data for SEI:

- Representativeness. Data should represent national or international populations of interest.
- Relevance. Data sources should include indicators central to the functioning of the science and technology enterprise.
- Timeliness. Data that are not part of a time series should be timely, i.e., substantial and unmeasured changes in the population under study should not have occurred since the data were collected.
- Statistical and methodological quality. Survey methods used to acquire data should provide sufficient assurance that statements based on statistical analysis of the data are valid and reliable.

Data that are collected by U.S. government agencies and that are products of the federal statistical system meet rigorous statistical and methodological criteria as described below. Unless otherwise indicated, these data are representative of the nation as a whole and of the demographic, organizational, or geographic subgroups that comprise it.

For data collected by governments in other countries and nongovernment sources, including private survey firms and academic researchers, methodological information is examined to assess conformity with the criteria U.S. federal agencies typically use. Government statistical agencies in the developed world cooperate extensively in developing data quality standards and improving international comparability for key data, and methodological information about the data generated by this international statistical system is relatively complete.

Methodological information about data from nongovernmental sources and from governmental agencies outside the international statistical system is often less well documented. These data are evaluated and must meet basic scientific standards for representative sampling of survey respondents and adequate and unbiased coverage of the population under study, and the resulting measurements must be sufficiently relevant and meaningful to warrant publication despite methodological uncertainties that remain after the documentation has been scrutinized. The most important statistical criteria are described in general terms below and in greater detail in the following sections.

Many data sources that contain pertinent information about some segment of the S&E enterprise are not cited in SEI because their coverage of the United States as a nation is partial in terms of geography, incomplete in terms of segments of the population, or otherwise not representative. For example, data may be available only for a limited number of states or studies may be based on populations not representative of the United States as a whole. Similarly, data for other countries should cover and be representative of the entire country. (In some cases, data that have limited coverage or are otherwise insufficiently representative are referenced in sidebars.)

Data included in SEI must be of high quality. Data quality can be measured in a variety of ways, some of which are described in the following sections. Some key dimensions of quality include:

- **Validity.** Data have *validity* to the degree that they accurately measure the phenomenon they are supposed to represent.

- **Reliability.** Data have *reliability* to the degree that the same results would be produced if the same measurement or procedure were performed multiple times on the same population.
- **Lack of bias.** Data are *unbiased* to the degree that estimates from the data do not deviate from the population value of a phenomenon in a systematic fashion.

Data Sources

Much of the data cited in SEI come from surveys. Surveys strive to measure characteristics of target populations. To generalize survey results correctly to the population of interest, a survey's **target population** must be rigorously defined and the criteria determining membership in the population must be applied consistently in determining which units to include in the survey.

Some surveys are censuses (also known as **universe surveys**), in which the survey attempts to obtain data for all population units. The decennial census, in which the target population is all U.S. residents, is the most familiar census survey. SEI uses data from the Survey of Earned Doctorates, an annual census of individuals who earn doctorates from accredited U.S. institutions, for information about the numbers and characteristics of new U.S. doctorate holders.

Other surveys are **sample surveys**, in which data are obtained for only a representative portion of the population units. The Survey of Recent College Graduates, which gathers data on individuals who recently received bachelor's or master's degrees in science, engineering, and health fields from U.S. institutions, is an example of a sample survey.

A sample is a **probability sample** if each unit in the sampling frame has a known, nonzero probability of being selected for the sample. Probability samples are necessary for inferences about a population to be evaluated statistically. Except for some Asian surveys referenced in chapter 7, sample surveys included in SEI use probability sampling. In **nonprobability sampling**, a sample is selected haphazardly, purposively, or conveniently, and inferences about the population cannot be evaluated statistically. Internet surveys and phone-in polls that elicit responses from self-selected individuals are examples of nonprobability sample surveys.

In sample surveys, once a survey's target population has been defined, the next step is to establish a list of all members of that target population (i.e., a **sampling frame**). Members of the population must be selected from this list in a scientific manner so that it will be possible to generalize from the sample to the population as a whole. Surveys frequently sample from lists that to varying extents omit members of the target population, because complete lists are typically unavailable.

Surveys may be conducted of individuals or of organizations, such as businesses, universities, or government agencies. Surveys of organizations are often referred to as *establishment surveys*. An example of an establishment survey used in SEI is the Survey of Research and Development Expenditures at Universities and Colleges.

Surveys may be longitudinal or cross-sectional. In a **longitudinal survey**, the same individuals (or organizations) are surveyed repeatedly. The primary purpose of longitudinal surveys is to investigate how individuals or organizations change over time. The Survey of Doctorate Recipients is a longitudinal sample survey of individuals who received research doctorates from U.S. institutions. SEI uses results from this survey to analyze the careers of doctorate holders.

Cross-sectional surveys provide a "snapshot" at a given point of time. When conducted periodically, cross-sectional surveys produce repeated snapshots of a population, enabling analysis of how the population changes over time. However, because the same individuals or organizations are not included in each survey cycle, cross-sectional surveys cannot, in general, track changes for specific individuals or organizations. National and international assessments of student achievement in K–12 education, such as those discussed in chapter 1, are examples of repeated cross-sectional surveys. Most of the surveys cited in SEI are conducted periodically, although the frequency with which they are conducted varies.

Some of the data in SEI come from **administrative records** (data previously collected for the purpose of administering various programs). Examples of data drawn directly from administrative records in SEI include patent data from the records of government patent offices; bibliometric data on publications in S&E journals, compiled from information collected and published by the journals themselves; and data on foreign S&E workers temporarily in the United States, drawn from the administrative records of immigration agencies.

Many of the establishment surveys that SEI uses depend heavily, although indirectly, on administrative records. Universities and corporations that respond to surveys about their R&D activities often use administrative records developed for internal management or income tax reporting purposes to respond to these surveys.

Surveys are conducted using a variety of modes (e.g., mail, telephone, the Internet, or in person). They can be self- or interviewer administered. Many surveys are conducted in more than one mode. For example, the Survey of Graduate Students and Postdoctorates in Science and Engineering, a census of establishments (university departments) from which students earn S&E graduate degrees, collects most of its data via a Web-based questionnaire but also allows respondents to answer a paper questionnaire. The National Survey of College Graduates, a longitudinal sample survey that collects data on individuals with S&E-related degrees and/or occupations, is initially conducted by sending a paper questionnaire by mail. Later, potential participants who did not respond to the questionnaire are contacted via telephone or in person.

Data Accuracy

Accurate information is a primary goal of censuses and sample surveys. Accuracy can be defined as the extent to which results deviate from the true values of the characteristics in the target population. Statisticians use the term "error" to refer to this deviation. Good survey design seeks to minimize survey error.

Statisticians usually classify the factors affecting the accuracy of survey data into two categories: nonsampling and sampling errors. **Nonsampling error** applies to all surveys, including censuses, whereas **sampling error** applies only to sample surveys. The sources of nonsampling error in surveys have analogues for administrative records: the processes through which such records are created affect the degree to which the records accurately indicate the characteristics of relevant populations (e.g., patents, journal articles, immigrant scientists and engineers).

Nonsampling Error

Nonsampling error refers to error related to survey design, data collection, and processing procedures. Each stage of the survey process is a potential source of nonsampling error. For most types, there is no practical method of measuring the extent of nonsampling error. A brief description of five sources of nonsampling error follows. Although for convenience the descriptions occasionally refer to samples, they apply equally to censuses.

Specification Error. Survey questions often do not perfectly measure the concept for which they are intended as indicators. For example, the number of patents is not the same as the amount of invention.

Frame Error. The sampling frame, the list of the target population members used for selecting survey respondents, is often inaccurate. If the frame has omissions or other flaws, the survey is less representative because coverage of the target population is incomplete. Frame errors often require extensive effort to correct.

Nonresponse Error. Nonresponse errors occur because not all members of the sample respond to the survey. *Response rates* indicate what proportion of sample members respond to the survey. Other things being equal, lower response rates create a greater possibility that, had nonrespondents supplied answers to the questionnaire, the survey estimates would have been different.

Nonresponse can cause *nonresponse bias*, which occurs when the people or establishments that respond to a question, or to the survey as a whole, differ in systematic ways from those who do not respond. For example, in surveys of national populations, complete or partial nonresponse is often more likely among lower-income or less-educated respondents. Evidence of nonresponse bias is an important factor in decisions about whether survey data should be included in SEI.

Managers of high-quality surveys, such as those in the U.S. federal statistical system, do research on nonresponse patterns to assess whether and how nonresponse might bias survey estimates. SEI notes instances where reported data may be subject to substantial nonresponse bias.

The response rate does not indicate whether a survey has a problem of nonresponse bias. Surveys with high response rates sometimes have substantial nonresponse bias, and surveys with relatively low response rates, if nonrespondents do not differ from respondents on important variables, may have relatively little.

Measurement Error. There are many sources of measurement error, but respondents, interviewers, and survey questionnaires are the most important. Knowingly or unintentionally, respondents may provide incorrect information. Interviewers may inappropriately influence respondents' answers or record their answers incorrectly. The questionnaire can be a source of error if there are ambiguous, poorly worded, or confusing questions, instructions, or terms, or if the questionnaire layout is confusing.

In addition, the records or systems of information that a respondent may refer to, the mode of data collection, and the setting for the survey administration may contribute to measurement error. Perceptions about whether data will be treated as confidential may affect the accuracy of survey responses to sensitive questions about business profits or personal incomes.

Processing Error. Processing errors include errors in recording, checking, coding, and preparing survey data to make them ready for analysis.

Sampling Error

Sampling error is probably the best-known source of survey error and the most commonly reported measure of a survey's precision or accuracy. Unlike nonsampling error, sampling error can be quantitatively estimated in most scientific sample surveys.

Chance is involved in selecting the members of a sample. If the same, random procedures were used repeatedly to select samples from the population, numerous samples would be selected, each containing different members of the population with different characteristics. Each sample would produce different population estimates. When there is great variation among the samples drawn from a given population, the sampling error is high and there is a large chance that the survey estimate is far from the true population value. In a census, because the entire population is surveyed, there is no sampling error.

Sampling error is reduced when samples are large, and most of the surveys used in SEI have large samples. Sampling error is not a function of the percentage of the population in the sample (when the population is large) or the population size but is a function of the sample size, the variability of the measure of interest, and the methods used to produce estimates from the sample data.

Sampling error is measured by the standard error of the estimate, sometimes called the "margin of error." The standard error of an estimate measures how closely the estimate from a particular sample approximates the average result of all possible samples. The standard error of the estimate is expressed as a range in the size of the difference (e.g., ±2%) between the sample estimate and the average result of all possible samples.

Statistical Testing for Data From Sample Surveys

Statistical tests determine whether differences observed in sample survey data could have happened by chance, i.e., as the result of random variation in which people or establishments in the population were sampled. Differences that are very unlikely to have been produced by chance variations in sample selection are termed **statistically significant**. When SEI reports statements about differences on the basis of sample surveys, the differences are statistically significant at the .05 level. This means that, if there were no true difference in the population, the chance of drawing a sample with the observed difference would be no more than 5%.

A statistically significant difference is not necessarily large, important, or significant in the usual sense of the word. It is simply a difference that cannot be attributed to chance variation in sampling. With the large samples common in SEI data, extremely small differences can be found to be statistically significant. Conversely, quite large differences may not be statistically significant if the sample or population sizes of the groups being compared are small. Occasionally, apparently large differences are noted in the text as not being statistically significant to alert the reader that these differences may have occurred by chance.

Numerous differences are apparent in every table in SEI that reports sample data. The tables permit comparisons between different groups in the survey population and in the same population in different years. It would be impractical to test and indicate the statistical significance of all possible comparisons in tables involving sample data.

As explained in "About Science and Engineering Indicators" at the beginning of this volume, SEI presents indicators. It does not model the dynamics of the S&E enterprise, although analysts could construct models using the data in SEI. Accordingly, SEI does not make use of statistical procedures suitable for causal modeling and does not compute effect sizes for models that might be constructed using these data.

Glossary

Most glossary definitions are drawn from U.S. Office of Management and Budget, Office of Statistical Policy (2006), "Standards and Guidelines for Statistical Surveys" and U.S. Bureau of the Census (2006), "Organization of Metadata, Census Bureau Standard Definitions for Surveys and Census Metadata." In some cases, glossary definitions are somewhat more technical and precise than those in the text, where fine distinctions are omitted to improve readability.

Administrative records: Data collected for the purpose of carrying out various programs (e.g., tax collection).

Bias: Systematic deviation of the survey estimated value from the true population value. Refers to systematic errors that can occur with any sample under a specific design.

Coverage: Extent to which all elements on a frame list are members of the population and to which every element in a population appears on the frame list once and only once.

Coverage error: Discrepancy between statistics calculated on the frame population and the same statistics calculated on the target population. *Undercoverage* errors occur when target population units are missed during frame construction, and *overcoverage* errors occur when units are duplicated or enumerated in error.

Cross-sectional sample survey: Based on a representative sample of respondents drawn from a population at a particular point in time.

Estimate: A numerical value for a population parameter derived from information collected from a survey and/or other sources.

Estimation error: Difference between a survey estimate and the true value of the parameter in the target population.

Frame: A mapping of the universe elements (i.e., sampling units) onto a finite list (e.g., the population of schools on the day of the survey).

Item nonresponse: Occurs when a respondent fails to respond to one or more relevant item(s) on a survey.

Longitudinal sample survey: Follows the experiences and outcomes over time of a representative sample of respondents (i.e., a cohort).

Measurement error: Difference between observed values of a variable recorded under similar conditions and some fixed true value (e.g., errors in reporting, reading, calculating, or recording a numerical value).

Nonresponse bias: Occurs when the observed value deviates from the population parameter due to differences between respondents and nonrespondents. Nonresponse bias may occur as a result of not obtaining 100% response from the selected units.

Nonresponse error: Overall error observed in estimates caused by differences between respondents and nonrespondents. Consists of a variance component and nonresponse bias.

Nonsampling error: Includes measurement errors due to interviewers, respondents, instruments, and mode; nonresponse error; coverage error; and processing error.

Population: See "target population."

Precision of survey results: How closely results from a sample can reproduce the results that would be obtained from a complete count (i.e., census) conducted using the same techniques. The difference between a sample result and the result from a complete census taken under the same conditions is an indication of the precision of the sample result.

Probabilistic methods: Any of a variety of methods for survey sampling that give a known, nonzero probability of selection to each member of a target population. The advantage of probabilistic sampling methods is that sampling error can be calculated. Such methods include random sampling, systematic sampling, and stratified sampling. They do not include convenience sampling, judgment sampling, quota sampling, and snowball sampling.

Reliability: Degree to which a measurement technique would yield the same result each time it is applied. A measurement can be both reliable and inaccurate.

Response bias: Deviation of the survey estimate from the true population value due to measurement error from the data collection. Potential sources of response bias include the respondent, the instrument, and the interviewer.

Response rates: Measure the proportion of the sample frame represented by the responding units in each study.

Sample design: Sampling plan and estimation procedures.

Sampling error: Error that occurs because all members of the frame population are not measured. It is associated with the variation in samples drawn from the same frame population. The sampling error equals the square root of the variance.

Standard error: Standard deviation of the sampling distribution of a statistic. Although the standard error is used to estimate sampling error, it includes some nonsampling error.

Statistical significance: Attained when a statistical procedure applied to a set of observations yields a *p* value that exceeds the level of probability at which it is agreed that the null hypothesis will be rejected.

Target population: Any group of potential sample units or individuals, businesses, or other entities of interest.

Unit nonresponse: Occurs when a respondent fails to respond to all required response items (i.e., fails to fill out or return a data collection instrument).

Universe survey: Involves the collection of data covering all known units in a population (i.e., a census).

Validity: Degree to which an estimate is likely to be true and free of bias (systematic errors).