

Generic Guide to Statistical Aspects of Developing an Environmental Results Program

April 25, 2003

Prepared on behalf of EPA OPEI by Michael Crow, Tracy Dyke Redmond, Rachel Cleetus, & Jeanne Herb of Tellus Institute,

With assistance from Ray Bienert, TetraTech, EM Inc.

Contact: Tracy Dyke Redmond, Tellus Institute • Tel: (617) 266-5400 • Fax: (617) 266-8303 • tdykeredmond@tellus.org

TABLE OF CONTENTS

Table of Contents 2

Executive Summary 3

 Overview of ERP 3

 Role of Statistics in ERP 3

 Purpose and Audience for this Statistical Guide 4

 Key Steps in Developing a Statistical Methodology for ERP 4

 Limitations of Statistics in ERP 6

 Benefits of Using Statistics in ERP 6

1. Introduction 9

 1.1. Purpose of this Document 9

 1.2. Audience 9

 1.3. Background on ERP 9

 1.4. Assumptions Used in this Document 13

 1.5. Additional Resources on Statistics 14

 1.6. Purpose and Goals for Statistical Analysis as part of ERP 14

 1.7. How to Use This Document 16

 1.8. Questions ERP Can Answer 16

2. Steps in Designing Your ERP Methodology 19

 2.1. STEP 1: Identify the Goals of ERP and the Questions You Want the Statistical Methodology to Answer 20

 2.2. STEP 2: Identify the Population of Facilities 21

 2.2.1. Define the Characteristics of the Target Population 22

 2.2.2. Identify and Locate Facilities in the Sampling Population 22

 2.3. STEP 3: Identify Environmental Business Practice Indicators (EBPIs) 24

 2.3.1. Purpose of EBPIs 24

 2.3.2. Two Primary Types of EBPIs 24

 2.3.3. How to Decide on EBPIs 26

 2.4. STEP 4: Develop Data Collection Instruments 27

 2.4.1. Principles for Constructing High Quality Data Collection Instruments 28

 2.5. STEP 5: Develop a Results Database 30

 2.6. STEP 6: Plan for Sample Selection 30

2.6.1.	A Few Key Facts about Margin of Error and Confidence Level	32
2.6.2.	How to Decide Upon a Your Sample Size.....	38
2.6.3.	Available Resources and Project Timeframe.....	44
2.6.4.	Mandatory v. Voluntary Participation in the Program.....	44
2.7.	STEP 7: Deciding on a Drop-Out Protocol.....	45
3.	Steps in Implementing Your ERP Statistical Methodology	47
3.1.	Generating the Sample	49
3.2.	Training Inspectors and Data Entry Personnel	49
3.3.	Entering Data and Conducting Needed Follow-Up	49
3.4.	Conducting Statistical Analysis	50
3.4.1.	Testing the Validity of Environmental Business Practice Indicators (EBPIs)	50
3.4.2.	Evaluation of Facility Performance	51
3.4.3.	Verification of the Accuracy of Self-Certification Data.....	51
3.4.4.	Impact of Questionnaire Design on Analyses.....	52
3.5.	Mid-Course Corrections.....	53
APPENDIX 1:	Considerations for Pilot Projects.....	54
Targeting Larger, Statewide Populations May Be More Resource Efficient		54
Keep an Eye Toward the Future		54
Consider Collecting Information on Non-Environmental Facility Characteristics		54
Consider Compiling Data on the Population for the Full Program		55
APPENDIX 2:	Expanding the Population: Florida DEP's Experience with Identifying New Facilities	56
APPENDIX 3:	Review List for Statistically Sound Data Collection	59
Question Format.....		59
General Question Format.....		59
Qualitative Questions.....		62
Questionnaire Content and Format		65
APPENDIX 4:	Verifying Accuracy of Self-Certification Data.....	68
Four Types of Observations.....		68
Observation #1: Facility and Inspector Declare Compliance		68
Observation #2: Facility and Inspector Declare Non-Compliance		69
Observation #3: Facility Reports Non-Compliance and Inspector Reports Compliance ...		69

Observation #4: Facility Reports Compliance and Inspector Reports Non-Compliance ...	69
Impact of the Timing of Follow-up Inspections	70
APPENDIX 5: Common ERP Statistical Calculations and Tests	71
Introduction.....	71
Computing Confidence Intervals	72
Constructing Hypothesis Tests	74
Calculation of the Minimum Sample-Size Required for the One- and Two-Sample Tests of Proportions.....	75
Calculation of Sample Size for the One-Sample Test of Proportions	76
Calculation of Sample Size for the Two-Sample Test of Proportions.....	77
Conducting a One-sample Test of Proportions	78
Computational Details for the One-sample Tests of Proportions	78
Conducting a Two-sample Test of Proportions	79
Computational Details for the Two-sample Test of Proportions.....	81
Software Tools for Determining Sample Size and Performing Tests of Proportions.....	81
Correlation Analysis	82
APPENDIX 6: Alternative Approaches for Analyzing Survey Data - Analysis of Categorical Data Using Contingency Tables	83
APPENDIX 7: Additional Resources on Survey Methods.....	87
Books and Miscellaneous Documents	87
Software Packages	88
Resources Available on the Internet	89
GLOSSARY	90

LIST OF BOXES AND FIGURES

Box 1: Abbreviated Illustrative Inspector Checklist	25
Figure 1: The ERP Cycle	12
Figure 2: Seven Steps in Designing Your ERP Statistical Methodology.....	19
Figure 3: The Target Population and the Sampling Population.....	22
Figure 4: EBPI Connection to Questions.....	26
Figure 5: Four Random Samples.....	31
Figure 6: Margin of Error and Confidence Level.....	32
Figure 7: Margin of Error Decreases as Sample Size Increases, for Two Confidence Levels.....	36
Figure 8: Relationship of Sample Size to Population Size.....	37
Figure 9: Ten Steps in Implementing Your ERP Statistical Methodology.....	48

This page is intentionally blank.

Executive Summary

EXECUTIVE SUMMARY

Overview of ERP

The Environmental Results Program (ERP) is an innovative approach to solving high-priority environmental problems in industry sectors largely comprised of large numbers of smaller sources of environmental pollution, usually smaller businesses. ERP relies in part upon a statistical approach to enhance confidence in the reliability and accuracy of performance measurement data. The data can then be used for many purposes, including strategically targeting agency resources to problem areas and, ultimately, providing accurate and compelling evidence of performance improvements and the success of the program.

Role of Statistics in ERP

ERP uses statistics to evaluate environmental performance for a group of facilities without requiring that every facility be inspected. The fundamental role of statistics is to draw inferences about the environmental performance of a group of facilities (e.g., an entire sector) based on collection of data at a smaller, statistically valid sample of facilities. Based on this information, agencies can identify problem areas and more efficiently target resources to improve environmental performance. Agencies also can measure whether the environmental performance changes over time.

ERP combines inspections, compliance assistance, and self-certification in an integrated approach to performance measurement. In particular, inspections are conducted at a randomly selected sample of facilities before self-certification. In addition to targeted inspections, another set of inspections, conducted at randomly selected facilities, takes place after compliance assistance and self-certification has occurred. Statistical analysis of inspection data and self-certification data allows agencies to:

- 1) **Assess Baseline Performance**, illuminating the nature, scope, and seriousness of the environmental problems presented by the targeted group of facilities before the agency intervenes with compliance assistance and self-certification requirements.
- 2) **Assess Changes in Performance** by comparing performance between randomly selected samples of facilities before and after the compliance assistance/self-certification period. This analysis enables agencies to draw inferences about the degree to which the overall group of facilities is following compliance requirements and/or best management practices, and the extent to which the performance of facilities has changed over time.
- 3) **Strategically Target Resources** by focusing the agency's attention on problem areas, such as groups of facilities that are not in compliance with certain requirements. At the same time, agencies may spend fewer resources where a higher percentage of those facilities are meeting key requirements.
- 4) **Increase Public Accountability** by publicly reporting on specific elements of environmental performance for entire sectors and for individual facilities. The performance

data that results from ERP statistical analysis promotes public accountability for businesses and for regulators.

Purpose and Audience for this Statistical Guide

This statistical guide is intended to help states that are developing their own Environmental Results Programs understand the range of statistical issues that ERP presents and how program decisions can affect the statistical validity of performance measurement. While this guide cannot take the place of a qualified statistician, it will help readers understand the key concepts so they can more effectively work with a statistician. This statistical guide assumes a general familiarity with ERP, such as is provided by “The Massachusetts Environmental Results Program: User’s Guide for Government Agencies” which is available at www.epa.gov/permits.

The first several sections of this document focus on high-level concepts and key decisions points related to statistics and ERP. These sections are intended for government officials who are familiar with ERP but do not have a background in statistics, as well as any program manager seeking an understanding of the role of statistics in ERP. This Executive Summary provides a brief, high-level overview of key statistical aspects of ERP. Section 1, the Introduction, provides further background on ERP and how statistics supports the program as a whole. Section 2, Steps in Designing Your ERP Methodology, discusses how elements of ERP program design affect the statistical analysis of ERP data, and consequently how program design affects the conclusions that can be drawn about facilities’ environmental performance. Both of these sections are intended to be helpful to readers thinking about statistical aspects of ERP *before* they begin ERP planning and design.

The remaining sections of the document are designed for readers that are involved in the daily decisions of developing and implementing an ERP. Section 3, Steps in Implementing Your ERP Statistical Methodology, discusses specific tasks in ERP implementation that involve statistics. Finally, the Appendices provide more detailed information for staff involved in implementing ERP. Appendix 5, in particular, is intended for readers with some technical background. Once readers have reviewed Appendix 5, they can use the accompanying Excel spreadsheet to easily conduct some of the basic statistical calculations in ERP.

Key Steps in Developing a Statistical Methodology for ERP

This statistical guide provides an orientation to all the major steps in designing and implementing a statistical analysis as part of ERP. The document is designed to enable readers to work with statisticians to develop program-specific statistical methodologies for their ERPs. Program-specific statistical methodologies are necessary to account for the unique circumstances in their states and the sectors in which they are implementing ERP. The main text of this document outlines key steps to be taken and critical questions to be answered in designing a statistical methodology for your ERP. While this Executive Summary does not review each step discussed in the main text, a few of the key steps in designing a statistical methodology include:

- **Clarifying the goals of ERP and how to measure progress towards those goals:** The first step of designing a statistical methodology for ERP is to clarify overall program goals and the questions that the ERP statistical methodology should answer. For example, if a key program goal is to reduce leaks from underground storage tanks (USTs), then a key question that the statistical methodology could be designed to answer is: “What percentage of USTs are complying with technology requirements that are designed to eliminate UST leaks?” Being clear about the goals of the program at the outset will help ensure that the statistical methodology is capable of measuring progress towards those goals. In order to make these goals and questions operational, it is important to select a relatively short list of key metrics (typically called Environmental Business Practice Indicators) used to evaluate environmental performance at individual facilities and for groups of facilities.
- **Defining the types of facilities to target through ERP, and identifying specific facilities:** The group of facilities targeted in ERP may constitute a sector or another group of facilities that have certain characteristics. For example, some ERPs have focused on the auto repair sector, while others have focused on certain types of auto repair facilities (e.g., auto body repairs). It is also possible to design a cross-sector ERP (e.g., an ERP for new industrial boilers below a certain size). Whatever the target group of facilities, it is important to identify individual facilities in that group as completely and accurately as possible, otherwise the statistical analysis could be biased.
- **Developing sound data collection instruments:** Data collection instruments used in ERP include the inspector checklist and self-certification form, both of which are designed to record facility performance data. Data collection instruments should be designed so that they collect data accurately and consistently (e.g., so that two different people filling out the form would interpret the questions in the same way and provide comparable data). To further ensure consistency and accuracy, well-designed data collection instruments are complemented by providing information to inspectors and facilities through workshops and workbooks. It is important to carefully consider how data is collected for ERP, because without good data collection instruments, even perfect statistical calculations will produce inaccurate and misleading results.
- **Selecting a representative sample:** In the event that resource limitations prevent agencies from collecting data from every facility in the target group, the best method to estimate facilities’ performance is to collect data from a representative sample of facilities. The size of the sample is critical, because it has a substantial impact on how performance measurement results can be interpreted. In general, the larger the sample size, the more confident you can be in your estimates of the overall performance. Deciding on an appropriate sample size depends on many factors, such as the total number of facilities in the ERP, what kinds of questions will be answered using the data, and how much uncertainty about the data can be tolerated.

Once the statistical methodology has been developed, there are a number of steps needed to implement it which are too detailed to discuss here. These steps are discussed in detail in section 3.

Limitations of Statistics in ERP

In order to effectively use statistics in ERP, it is important to understand the limitations of statistics and the impact of those limitations on ERP program decisions. Also, it is important to understand the limitations of this statistical guide in providing statistical guidance for ERP.

- **Estimates based on samples are not as accurate as a census of all facilities.** If resources allowed, or if the group of facilities were small, inspecting every facility would be preferable to using statistics to estimate characteristics of facilities. Keep in mind that all estimates you develop based on a sample are just that – *estimates* – and they will be subject to some uncertainty.
- **Both the design and implementation of a statistical methodology for ERP affect how performance can be measured and what conclusions can be drawn about program effectiveness.** If there are flaws in the statistical methodology it will undermine the reliability of performance measurement results for the program, and may appreciably reduce the credibility of the program as a whole.
- **EPA recommends that ALL ERP programs be implemented under the guidance of an experienced statistician.** This generic document cannot take the place of a qualified statistician in ensuring that a statistical methodology is sound. However, this document will help readers become familiar with the statistical concepts and issues embedded in ERP so that they may use a statistician's time more effectively.

Benefits of Using Statistics in ERP

Using statistical sampling to measure performance is a good way to more accurately measure performance and effectively use limited resources. Where resources are an issue, using statistical analysis as part of ERP can help states better estimate characteristics of ERP facilities and describe the uncertainty associated with those estimates. Without a statistical approach to performance measurement, it is not possible to make inferences about a whole population based on a smaller sample. Consequently, without a statistical approach, program managers are very limited in the conclusions they can draw about changes in performance or program effectiveness.

A statistically valid approach to performance measurement gives program managers the tools to:

- Answer key programmatic questions about certain compliance levels and extent to which facilities have adopted best practices;
- Differentiate performance levels among different types of facilities;
- Determine how well facilities are performing in the absence of ERP;
- Estimate the impact ERP over time;
- Learn how to improve ERP program design to motivate better environmental

performance;

- Hold facilities and sectors accountable for their performance; and
- Demonstrate program results to key stakeholders and the public clearly and decisively.

In summary, the statistical approach to performance measurement is an integral part of ERP because it allows agencies to better understand the effectiveness of ERP in generating real environmental results.

Introduction

1. INTRODUCTION

1.1. Purpose of this Document

This document is designed to help regulatory agency staff understand the key statistical concepts that are important to the Environmental Results Program (ERP). ERP is an innovative approach to solving high-priority environmental problems in industry sectors largely comprised of under-regulated small businesses. The ERP approach combines technical assistance, self-certification, inspections, and statistically based performance measurement in order to reduce the environmental impacts of business. ERP relies upon a statistical approach to enhance confidence in the reliability and accuracy of performance measurement data where states do not have the resources to inspect every facility. This statistical methodology is intended to help states that are developing their own Environmental Results Programs understand the range of statistical issues that ERP presents and how program decisions can affect the statistical validity of performance measurement. This document provides more detailed guidance on developing an ERP statistical methodology than that contained in a previously available document, "The Massachusetts Environmental Results Program: User's Guide for Government Agencies."¹

The document describes generic statistical issues and steps that agencies should take in designing and implementing an ERP. Once regulatory staff have become familiar with this document, they should develop a specific statistical methodology they will employ in their own ERP. The specific statistical methodology should account for the unique circumstances of the state and sector in which ERP is being implemented.

1.2. Audience

This generic statistical guide should be useful to government officials who are familiar with the ERP framework but who do not have a background in statistics. This guide also includes appendices that provide greater depth readers with a more technical background who will be developing, implementing, or reviewing the program-specific statistical approach. EPA recommends that all ERP programs be implemented under the guidance of an experienced statistician.

1.3. Background on ERP

ERP is an innovative environmental management approach pioneered and originally implemented by the Massachusetts Department of Environmental Protection. Massachusetts and other states are adapting ERP to address continually emerging environmental priorities. ERP is an on-going initiative that offers states the opportunity to cost-effectively improve environmental performance through a less burdensome, more transparent regulatory system. The ERP approach integrates several compliance assurance approaches into an effective, synergistic package that can supplement a state's traditional compliance inspection program, including:

- *Randomly selected inspections* that provide statistically valid performance measurement and enable the regulatory agency to make inferences about a sector's compliance with

¹ The ERP Users Guide, along with other ERP background material, is available on line at www.epa.gov/permits/masserp.htm, or by calling Gregory Ondich at EPA (202) 556-2215.

specific environmental requirements. Randomly selected inspections, like any other inspections, also deter facilities from failing to comply with the law;

- *Targeted inspections* enhance the deterrence effect specifically among facilities that show indications of non-compliance on their certification forms;
- *An annual self-certification of compliance* by companies to increase self-evaluation and accountability and to provide additional performance data;
- *Compliance assistance* from the agency through outreach and innovative workbooks.

Regulatory agencies implementing ERP typically combine these tools in several steps described here and illustrated in Figure 1.

- **Baseline Inspections:** First, the agency specifically identifies a group or sector of facilities it is targeting and conducts baseline inspections for a randomly selected *sample*² of facilities among that group.
- **Self-Certification Period:** Next, the agency provides compliance assistance through workbooks and workshops. Facilities use the compliance assistance materials to complete and submit self-certification forms and, if necessary, provide information on how they will return to compliance.
- **Targeted Follow-Up:** Following baseline inspections and self-certification, agencies may also conduct targeted follow-up inspections with facilities that are not in compliance to resolve compliance issues. Data from these targeted follow-up inspections are not used in the statistical analysis to characterize performance of the sector overall, although they are important to overall ERP effectiveness.
- **Post-Self-Certification Inspections and Analysis:** Following the self-certification period, the agency conducts another set of inspections for a randomly selected sample of facilities. The agency then compares this round of inspection data with the baseline inspection data and with the data provided on self-certification forms to better understand changes in environmental performance over time and as a result of the ERP program. Additionally, comparing self-certification forms with the random inspections enable the agency to judge how reliable the self-certification forms are in terms of indicating the sector's compliance with specific requirements.

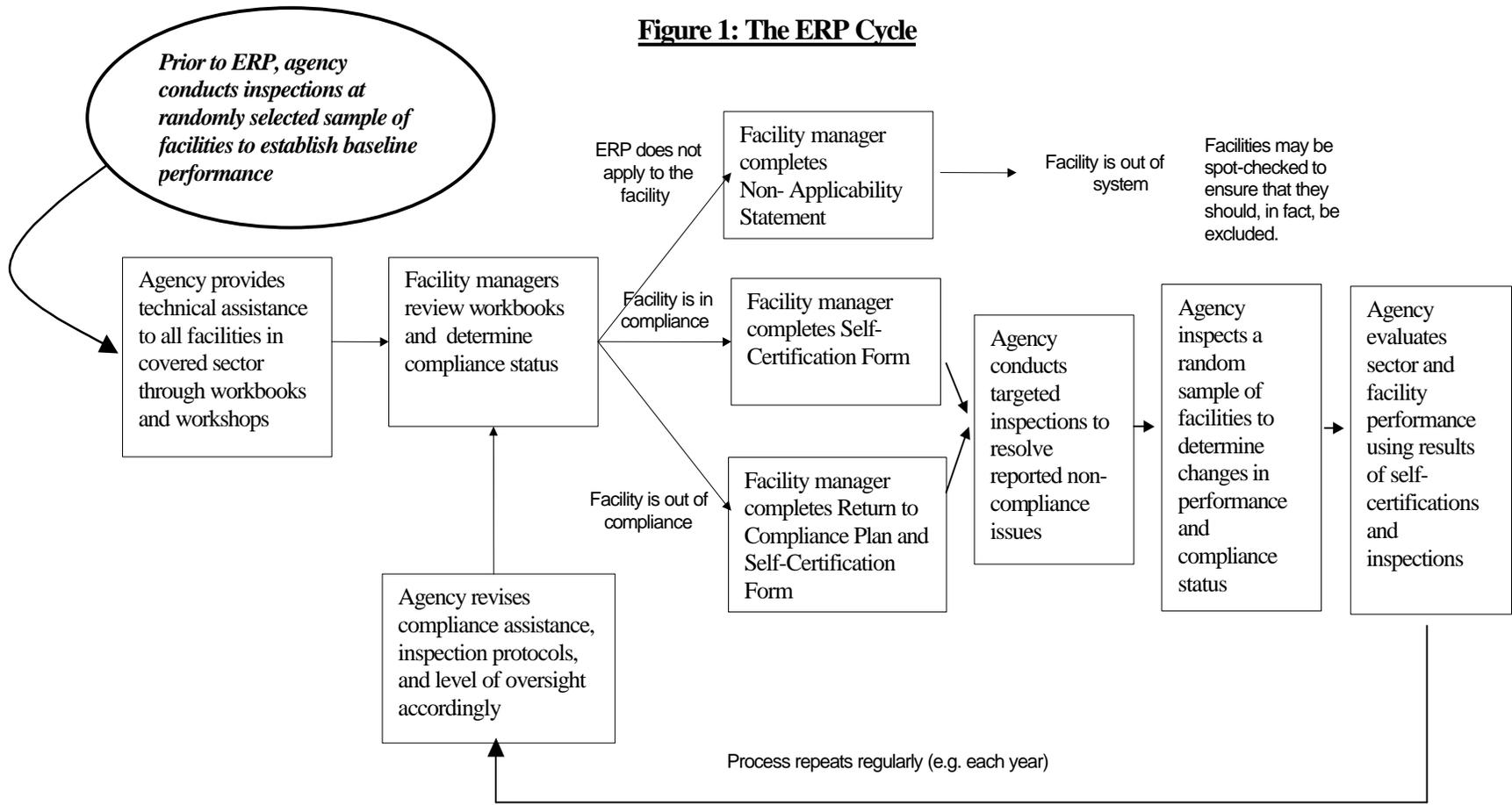
The self-certification and inspection process is repeated at regular intervals to maintain facility performance levels and confidence in an agency's understanding of those levels. Each step of the ERP process described here (except for the targeted follow-up inspections) has statistical aspects

² Statistical terms printed in italics in the main text are defined in a glossary at the end of this document.

April 25, 2003

that relate to the agency's ability to conduct statistically valid performance measurement. This document will identify and explain these statistical aspects of each stage in the ERP process.

Figure 1: The ERP Cycle



1.4. Assumptions Used in this Document

In order to make this document accessible to readers who are not fluent in statistics, this guide describes only simple ERP statistical designs and approaches. In particular, this statistical guide assumes that, unless otherwise noted:

- **You can identify the facilities in the sector you are trying to address.** Although you may not know all of the facilities in your target sector at the outset of designing your ERP, many states have used regulatory or commercially available databases to identify relevant facilities. For more information on how to identify facilities in your target sector, see section 2.2 in this document. If you cannot find data sources that identify facilities in your target sector, you should consult with a qualified statistician to help you design a sound statistical approach.³
- **Your main research questions can be reduced to “yes/no” or binary questions** (e.g., whether facilities are in or out of compliance, or whether facilities are or are not following a particular best practice). Although this statistical guide does briefly discuss some other types of questions (e.g., finding the average number of pounds of hazardous air emissions produced by facilities in your sector), this document provides guidance for analyzing data derived only from yes/no questions. This is because it is more straightforward, from a statistical perspective, to analyze data derived from yes/no questions.

There are many factors that may require you to take a more complex statistical approach to ERP. This document highlights some of those situations, but it is not possible to be comprehensive due to the limited scope of this document and to the likelihood that unanticipated situations will emerge as more states adopt ERP. EPA suggests that ALL states designing an ERP, at a minimum, consult with a qualified statistician once they have drafted their ERP statistical methodologies. If you have more complex situations, such as those identified in Appendix 5, you are encouraged to work closely with a qualified statistician to ensure that your statistical approach is appropriate for your situation. Even though this generic document cannot take the place of a qualified statistician in ensuring that a statistical methodology is sound, this document is intended to help you become familiar with the statistical concepts and issues embedded in ERP so you may use a statistician's time more effectively. It may be helpful to think of the role of a statistician as analogous to that of a lawyer. Like a lawyer, a statistician need not be involved in every step of program design or implementation. However, expert advice should be sought out at critical junctions. Also, like a lawyer, a statistician should not tell program staff what programmatic decisions they should make, but rather should inform program staff about their options and the implications of important decisions.

Note that this guide has been designed specifically for application within the standard ERP model described in the ERP Users Guide cited above. The guidelines presented in this document may require substantial revision if fundamental elements of the standard ERP model are changed during implementation.

³ For more information on statistical approaches for an unknown population, see Appendix G in EPA's Guide for Measuring Compliance Assistance Outcomes, found at www.epa.gov/compliance/planning/results/tools.html.

Several states that have adopted ERP are experimenting with voluntary approaches (i.e., those where facilities have the option to self-certify, but are not required to do so). While some portions of this document are relevant only to mandatory ERP programs where facilities are required to self-certify, states that are using a voluntary ERP approach will also find this document useful. Section 2.6.4 specifically highlights considerations for voluntary programs.

1.5. Additional Resources on Statistics

Readers with a technical background, and those charged with implementing elements of a statistical methodology for ERP, may wish to consult additional resources for more detailed discussions of the concepts presented here. Resources that may be particularly helpful are listed in Appendix 7. In addition, EPA has developed a separate manual for applying statistics in the context of compliance assistance programs. That document, the “Guide for Measuring Compliance Assistance Outcomes,” provides further explanation for some aspects of statistical design and analysis. This Generic Guide to Statistical Aspects of Developing an Environmental Results Program and the Guide for Measuring Compliance Assistance Outcomes are generally consistent in their recommendations. However, the latter document was designed to help states gauge the effectiveness of voluntary compliance assistance programs, not ERP. Given the different intent of the two documents, they differ somewhat in their assumptions and emphases. Despite this difference in goals and focus, the Guide for Measuring Compliance Assistance Outcomes does provide a good additional resource for states implementing ERP. It can be found online at: <http://www.epa.gov/compliance/resources/reports/planning/results/comeasuring.pdf>.

1.6. Purpose and Goals for Statistical Analysis as part of ERP

ERP uses statistics to estimate performance (e.g., compliance with select requirements and adherence to best practices) for a large group of facilities where inspection resources are limited. If resources were unlimited, or if the group of facilities were small, inspecting every facility would be preferable to using statistics to estimate characteristics of facilities. However, where there are not enough resources to inspect every facility, using statistical analysis as part of ERP can help states better estimate characteristics of ERP facilities and describe the uncertainty associated with those estimates. The statistical approach to performance measurement is an integral part of ERP because it allows agencies to better understand the effectiveness of ERP in generating real environmental results. Such an approach to performance measurement allows an agency to monitor environmental performance of a large group of facilities (e.g., an entire sector) based on collection of data at a considerably smaller, statistically valid sample of facilities. There are several roles that statistics plays in the implementation of ERP, as follows:

- **Baseline Performance Measurement:** An initial statistical analysis is typically conducted prior to ERP implementation, after an agency has conducted a baseline round of inspections at randomly selected facilities. This baseline analysis helps the agency understand the level of compliance in the absence of ERP and illuminates the nature, scope, and seriousness of the environmental problems presented by the targeted group of facilities. The results of the baseline analysis can help agencies design more effective

and targeted compliance assistance and self-certification materials that focus on key compliance problems identified in the sector.

- **Analysis of Changes in Performance:** After an agency has distributed compliance assistance and self-certification materials to all facilities in the program, and conducted a follow-up round of performance inspections at another set of randomly selected facilities, another statistical analysis is conducted. This analysis enables an agency to gauge changes in facility performance over time. Facility performance is measured in two ways: (1) by comparing overall results of the post-implementation round of inspections to the performance baseline data from the pre-implementation round of inspections; and (2) by analyzing the data provided by all facilities in their self-certification forms.
- **Strategically Targeting Inspection Resources:** Statistical analysis allows an agency to use limited government resources to inspect a sample of facilities, and based on that sample, to draw inferences for a much larger group of facilities (i.e., the target sector or population). Based on the analysis of performance improvement, an agency can focus its attention on problem areas, e.g., groups of facilities that are not in compliance for certain requirements, or specific requirements that a large percentage of facilities are failing to meet. At the same time, agencies can spend less resources inspecting groups of facilities where a high percentage of those facilities are meeting key requirements.
- **Public Accountability:** Statistically-based performance measurement enables agencies to publicly report on specific elements of environmental performance for entire sectors and for individual facilities. The performance data that results from ERP statistical analysis promotes public accountability for businesses and for regulators.

In order to achieve the above goals of statistical analysis as a part of ERP, it is important to verify the performance measurement tools that are part of ERP to ensure that they accurately reflect performance. There are two primary types of verification that are important to ERP:

- **Analysis of Accuracy of Self-Certification Forms:** Data collected from random inspections after self-certification enable agencies to gauge the accuracy and reliability of self-certification forms. This analysis compares results from post-self-certification compliance inspections to the data reported by inspected facilities on their self-certification forms. The self-certification verification step provides an agency with an understanding of how much credence to give to the information contained in self-certification forms, which in turn allows the agency to make well informed decisions about allocating inspection resources toward the targeted sector or other sectors. If it appears that self-certification forms accurately represent environmental performance, agencies may, over time, begin to rely on the data collected by self-certification forms to gauge environmental performance. On the other hand, if it appears that facilities are not accurately and honestly filling out the self-certification forms, this represents a serious concern, and facilities should be targeted for follow-up inspections, and, where necessary, enforcement action. However, even if an analysis reveals that self-certification forms are not very reliable in providing accurate data, they may still serve an effective purpose in leading the managers of many facilities through the process of learning about and implementing compliance procedures and best management practices.

- **Verification of Environmental Business Practice Indicators:** Environmental Business Practice Indicators (EBPIs) are a short set of data points used in ERP to indicate environmental performance for individual facilities and the entire group of facilities. It is important to verify that EBPIs are, in fact, good measures of the most important aspects of performance, since the EBPIs are the metrics by which performance is measured.

This statistical guide provides more detail on how to conduct the statistical tests and verifications that are integral to ERP. In particular, section 3.4 provides more information on how statistics is used in ERP.

1.7. How to Use This Document

This document describes selected introductory concepts in statistics along with steps that agencies must take in designing and implementing an ERP. After reading this document, you should prepare a specific statistical methodology for your ERP which accounts for the unique circumstances in your state and the sector in which you are implementing ERP. This document identifies the key steps to be taken and critical questions to be answered in designing a statistical methodology for your ERP.

1.8. Questions ERP Can Answer

ERP is designed to answer regulator's questions about environmental performance among facilities in an ERP target sector. There are many questions that regulators can answer by analyzing ERP data. Depending on the type of question you want to answer, you may rely either on descriptive statistics or inferential statistics to come up with an answer. *Descriptive statistics*, which are used to organize and summarize information for an entire *population* of entities.⁴ *Inferential statistics* are used to draw inferences, or conclusions, about the whole population of facilities, based on the random sample of facilities.

In ERP, data collected on self-certification forms can be summarized using descriptive statistics, since self-certification forms are collected from all facilities in a sector.⁵ However, self-certification forms may not be completely reliable. Therefore, agencies also conduct compliance inspections, which are expected to provide more reliable data. These inspection data are expected to be more reliable because inspectors are likely to be better trained than facility managers regarding environmental requirements, and inspectors have no incentive to report inaccurate data. (Facilities, on the other hand, may have an incentive to falsely report that they are in compliance in order to avoid enforcement action.)

Conducting compliance inspections requires significant resources on the part of regulatory agencies. One benefit of the ERP measurement approach is that it uses random sampling to conduct compliance inspections for a relatively small sample of facilities. In ERP, regulators primarily rely on inspection data for samples of facilities to draw inferences about levels of environmental performance within the entire population. Whenever you use inferential statistics,

⁴ Note: the term "universe" is often used as a synonym for "population."

⁵ This assumes that self-certification is mandatory.

you will be dealing with uncertainty, since you are only studying a sample of facilities and not the entire population. Even though using inferential statistics means that there will be some uncertainty in the conclusions, inferential statistics is useful when you do not have the resources to study the entire population. The focus of this document is on the inferential statistics involved in ERP.

Based on the ERP statistical approach to performance measurement, you can answer questions such as a) – d) below using inferential statistics. You can answer question e) using descriptive statistics:

- a. Based on a random sample, what percentage of facilities in this sector are in compliance or are following certain best practices **prior to** ERP implementation?
- b. Based on a random sample, what percentage of facilities in this sector are in compliance or are following certain best practices **after** ERP implementation?
- c. Based on random samples, is there a statistically significant difference in the percentage of facilities in compliance or following best practices before and after ERP implementation?
- d. Based on a random sample, how does compliance or implementation of best practices vary with certain key facility characteristics (e.g., business size, type of ownership, location, etc.)?
- e. According to self-certification forms, what percentage of facilities are independently run? What percentage of facilities are part of a franchise?

Note that all of the questions above deal with percentages of the population. In statistical terms, the percentage of the population that has a specified attribute is known as the *population proportion*. Population proportions are a way of summarizing information that can be ascertained by asking yes/no questions (e.g., “Is the facility in compliance for this requirement?”). You may also be interested in other types of questions, such as the average number of pounds of waste generated by facilities or the average number of gallons of effluent emitted by facilities. These values can be described by the *population mean* (i.e., the average value for a variable with a population). Since ERP statistical analysis generally deals with population proportions, and since this type of analysis is generally more straightforward, this guide focuses on population proportions. Estimating a population mean based on a sample is more complicated, and should be undertaken with the help of a qualified statistician.

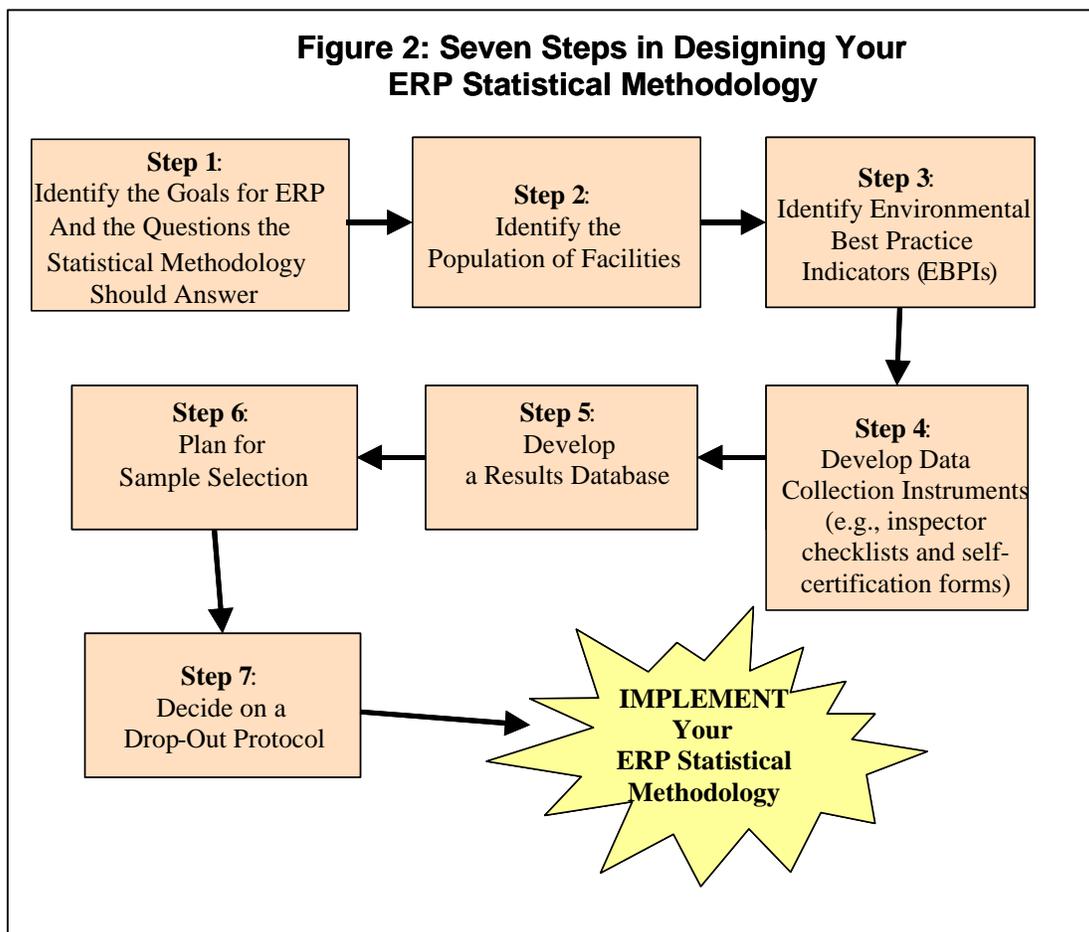
Steps in Designing Your ERP Methodology

2. STEPS IN DESIGNING YOUR ERP METHODOLOGY

There are seven key steps in designing an ERP statistical methodology. These steps are illustrated in Figure 2 below. Each step is described in detail in the following sections.

Keep in mind that in practice, you may implement these steps iteratively, and in a slightly different order, depending on your state's specific circumstances. For example, you may begin identifying the population of facilities before you clearly identify the questions as you want your ERP statistical methodology to answer. Similarly, since Environmental Business Practice Indicators are directly related to the questions your ERP methodology can answer, as well as to the data collection instruments you develop, there are iterative feedback loops between these steps. For the purpose of simplicity, this document suggests the following order of steps as one logical possibility.

Also, keep in mind that issues related to automation should be considered from the very earliest stages of ERP design, including design of the statistical methodology. For example, if you want self-certification forms to be electronically scanned into a database, or if you want data to be submitted on-line, it may influence the way you design the questions on your data collection instruments, as well as your results database. It is advisable to begin consulting with information technology staff from at the outset of ERP planning.



2.1. STEP 1: Identify the Goals of ERP and the Questions You Want the Statistical Methodology to Answer

The first step of designing a statistical methodology for ERP is to reflect on what you want to accomplish through ERP. What are your agency's environmental performance and compliance goals? One way to think about these goals is to ask yourself "If this ERP is successful, five years from now, what will have changed about the sector in which we are conducting ERP?" Once you have clarified the goals in your mind, you can begin to think about how you will measure progress. A key step in thinking about how to measure progress is to clarify what you are most interested in measuring, or put another way, what questions you want the ERP statistical methodology to answer.

As noted above, there are many types of questions that can be answered through an analysis of ERP data. For example, you may be interested in simply understanding what percentage of facilities in your ERP sector are in compliance. Alternatively, you may be interested in determining patterns of non-compliance. For example, are certain types of facilities (e.g., smaller facilities, or those affiliated with a parent corporation) more likely to be out of compliance? You may want to focus purely on compliance, or you may want to look at best practices that are not required. Most likely, you will want to consider all of these types of questions. Going through the process of clarifying what it is you hope to find out through your ERP statistical analysis, and how this relates to your overall ERP strategy, will help you design an appropriate methodology that will address your highest priority concerns.

Depending on the types of questions you want to answer, you may need to alter the design of your ERP statistical methodology. Brainstorming about potential questions with program staff and inspectors can help ensure that you identify the most important questions. Note that in some cases, field inspectors may have a good sense of certain answers from their experience with facilities. You may wish to confirm their experience through ERP, or you may wish to focus on other aspects of performance that are less well known. In either case, it is best to get agreement on the questions you want to answer before proceeding. In addition to writing down the questions, you may wish to note the following characteristics of each question:

- Does this question refer to yes/no questions (e.g., whether or not a facility is meeting a certain requirement or following a certain practice), or does the question relate to quantitative variables (e.g., data measured in time, volume, mass, or counts)?
- Do you want to look at a snapshot of performance, or do you want analyze trends over time (e.g., before and after ERP implementation)?
- Do you want to understand the performance of a whole sector, or do you want to compare subgroups within a sector?

Note the limitations of this guide in answering certain types of questions described in section 1.4. You may want to revisit and revise these questions as you proceed through the subsequent steps of designing your statistical methodology.

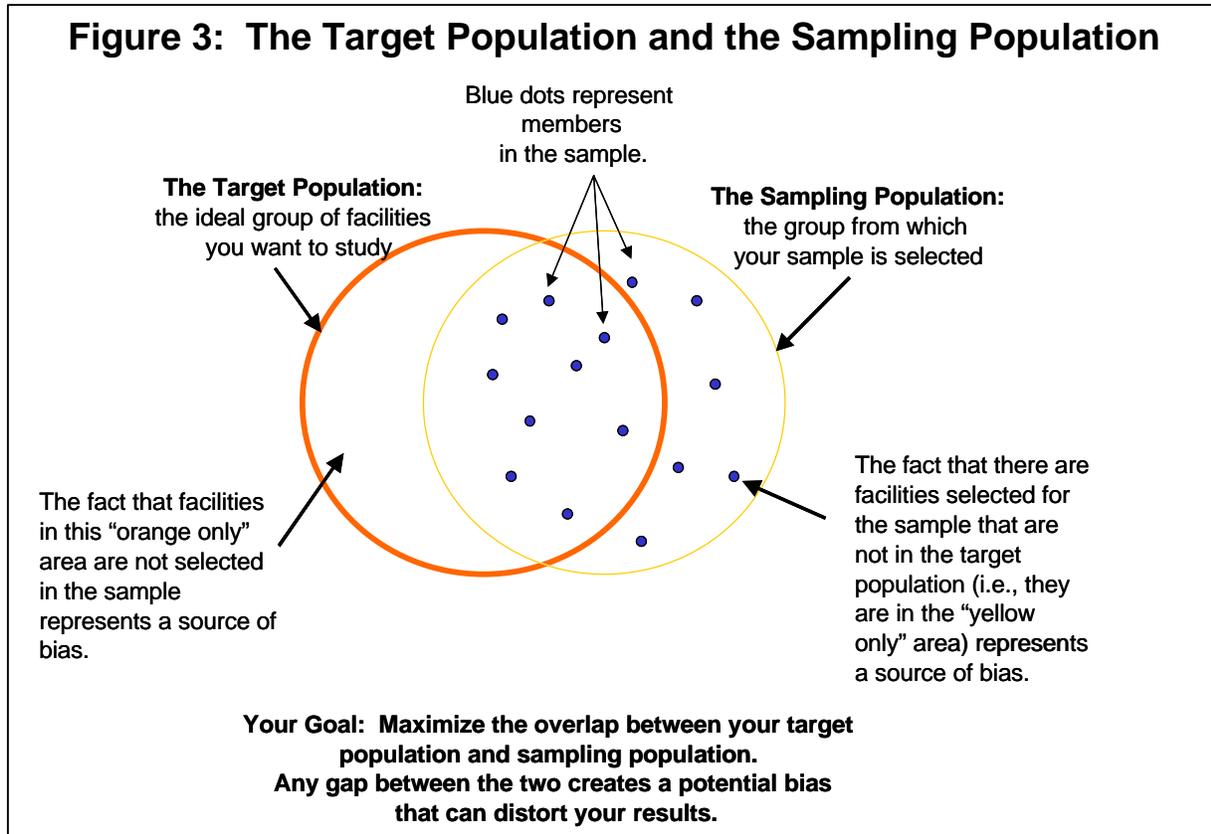
2.2. STEP 2: Identify the Population of Facilities

A very important initial process in ERP is careful identification of the population of facilities to be targeted by ERP. Identifying the population of facilities includes both defining characteristics of the population (i.e., what types of facilities you want to include), as well as finding individual facilities that have these characteristics (e.g., identifying their business name and address). In practice, it is useful to think about the *target population* and the *sampling population*.⁶ The target population is the group of entities in which you are interested. The target population is an ideal – it represents the theoretical list of all facilities that perfectly match the characteristics you have defined. However, you will probably not be able to perfectly identify all of the entities in the target population. Thus, the sampling population is the set of entities within the target population that you can identify. The sampling population is an actual list of facilities that serves as the pool from which you will draw your sample. The sampling population may differ from your target population in two ways:

- First, you may not be able to find all of the individual facilities in your target population. For example, if your ERP focuses on drycleaners, you probably would like your ERP to offer insights about all drycleaners in your study area. However, you may only be able to identify dry cleaners that are listed in the phone book and in regulatory databases. If the phone book or databases are incomplete and miss some drycleaners, then your sampling population will not include all facilities in the target population.
- Second, you may not be able to perfectly identify facilities that have the characteristics of your target population. For example, if your target population is limited to dry cleaners that have certain types of equipment or processes, but you do not have data to correctly identify those types of dry cleaners, your sampling population may either erroneously include facilities that are not in the target population or exclude facilities that are in, in fact, in the target population. For example, suppose you want to include only drycleaners that actually do the cleaning on site (i.e., not drop-off facilities). It may be difficult to select only those drycleaners that do on site dry cleaning based on available databases or information in the phone book.

A key task in identifying your population of facilities is make sure that your sampling population is as close as possible to your target population. This is because, in the end, you can only draw statistically valid conclusions about facilities that are in your sampling population. **Any gap between your target population and the sampling population is a potential source of bias that can seriously undermine your ability to accurately interpret your results.** For example in Figure 3 below, facilities that appear in the “yellow only” area not in the target population, but they may be included in the sample because they are listed in the sampling population. Yet facilities in the “orange only” area are in the target population, but they cannot be included in the sample because they are not in the sampling population. Either situation creates a source of bias. Therefore, the goal in selecting the target and sampling population is to minimize the difference between the two. Ideally, there would be complete overlap between the two circles in Figure 3.

⁶ The sampling population is sometimes referred to as the “sampling frame.”



2.2.1. Define the Characteristics of the Target Population

Identifying the target population involves defining the characteristics of those facilities to be targeted and/or those facilities to be excluded. For example, Florida DEP has designed an ERP pilot project that targets the automotive repair sector, but it excludes shops that do collision repair (i.e., autobody shops), even if they also do mechanical repairs. The pilot project also excludes facilities located outside the boundaries of the two Florida DEP districts that will initially test the program.⁷ In order to define the characteristics of the population of facilities for your ERP, think about the goals for the ERP, the questions you wish to answer, and which types of facilities would be relevant to those questions. Keep in mind that your ability to accurately identify facilities in the target population may be limited by available data.

2.2.2. Identify and Locate Facilities in the Sampling Population

After defining the characteristics of the target population, you should accurately identify and locate as many facilities as possible within the target population. This step is defining the sampling population. A well-defined sampling population will enable you to generate appropriate and relevant random samples for measuring performance and efficiently target compliance assistance and self-certification materials.

⁷ For the sake of brevity, the main text of this methodology does not discuss any of the issues related to developing a smaller pilot project as part of the path toward a full ERP in the chosen sector. Appendix 1 highlights a few key issues for states to consider when taking this approach.

For small target populations, you may be able to identify facilities individually, without needing to consult databases or registries. For example, agencies whose ERP programs have a distinct geographic, or neighborhood focus may wish to identify facilities by walking or driving around the area and/or seeking help from interested community groups in identifying facilities.

In most cases, target populations are too large to allow agencies to identify facilities individually. If this is the case, you can identify facilities by combining data from a variety of available data sources, including the following:

- environmental agency compliance databases;
- state business registries;
- trade associations;
- privately managed national facility databases, such as InfoUSA or Dun & Bradstreet, which compile data from multiple sources; and
- specialized databases that may be available for certain sectors. For example, state insurance regulatory agencies may require automotive collision repair facilities to be registered with the state in order to receive insurance payments.

When combining data from different sources, keep in mind the strengths, weaknesses and practicality of using each database. Factors to consider include the following:

- cost of acquiring and updating the data set over time;
- restrictions on usage of the data set (such as with third-party data sets);
- geographic identification of facilities (e.g., by ZIP code, county, etc.);
- level of specificity of facility classification (e.g., sector, subsector, etc.);
- extent of cross-referencing for facilities engaged in multiple activities;
- frequency with which database records are updated;
- methods of data collection, considering what percentage of the targeted population is likely to be in a database;
- compatibility of different data sources; and
- extent to which failure to develop a good advance database will create systematic sampling bias or will lead to high numbers of visits to non-applicable facilities.

You should combine multiple data sources if your best data source does not capture all facilities in the target universe. You will then probably need to exclude from the combined list those facilities that do not have the characteristics of the targeted population. Remember that if the data on which you base your exclusion decisions is unreliable, it can create a *bias* in your sampling population. Also, keep in mind the risks of being over- vs. under-inclusive. If you inadvertently exclude facilities that should be in your sampling population, you will not be able to add them back in to a sample later, and they will not receive ERP materials such as self-certification forms and workbooks. Conversely, if you inadvertently include facilities outside

your target population, you can always exclude them from the sampling population later. However, the risk of being over-inclusive is that you will waste inspection resources on irrelevant facilities. Appendix 2 discusses one state's experience in developing a sampling population and in addressing data gaps.

2.3. STEP 3: Identify Environmental Business Practice Indicators (EBPIs)

There are many aspects of environmental performance that could be measured through analysis of ERP data. Since analysis requires time and resources, it is advisable to pick a few key indicators of performance. Although you may choose to collect more comprehensive data in inspections and on self-certification forms, analyzing data can be resource intensive, and so it is important to select a relatively small set of performance indicators, known as "Environmental Business Performance Indicators" (EBPIs). For example, in Massachusetts, the ERP inspector checklist and the self-certification forms included approximately 10-15 EBPIs, as well as longer subsets of questions that led up to, and supported, these EBPIs. Not all EBPIs need to appear on self-certification forms, but they should be reflected (at least indirectly) on inspector checklists.⁸

2.3.1. Purpose of EBPIs

As their name suggests, the EBPIs serve as indicators for drawing inferences as to both compliance status and overall environmental performance of facilities. EBPIs should relate directly to the questions you want your ERP statistical methodology to answer – they are the metrics that make it possible to answer your ERP questions. Each EBPI should represent a key performance concern and/or indicate the degree of performance for a larger subset of questions. Instead of analyzing all data collected, the EBPIs enable agencies to streamline the analytical process by focusing on the most important indicators of performance. Thus, analyzing EBPIs represents a much more practical approach than conducting a full statistical test on each of 100-200 questions that may appear on an inspector checklist. Individual EBPIs or aggregate EBPI performance scores can be compared across time or across facilities.

It is strongly recommended that you verify the validity of EBPIs after conducting the baseline data collection to ensure that the EBPIs are actually indicating broader and deeper performance information. You should also re-assess the EBPIs on a regular basis throughout the long-term implementation of the program, to ensure that they are keeping pace with changes in the target population. This issue is discussed in section 3.4.1.

2.3.2. Two Primary Types of EBPIs

EBPIs can be classified into two primary categories: roll-up EBPIs and stand-alone EBPIs. Box 1, Abbreviated Illustrative Inspector Checklist, shows sample EBPIs of both types.

A roll-up EBPI is one that explicitly represents or summarizes a number of sub-questions. For example, in Box 1, Question #5 is a roll-up EBPI intended to measure whether hazardous wastes are being managed to prevent releases. It is a roll-up question because it can only be answered

⁸ If space is limited on self-certification forms, EBPIs may be excluded, since measurements of performance will primarily be based on an analysis of inspector checklists. However, since EBPIs are selected as the best indicators of performance, if space allows it usually makes sense to have facilities certify as to their performance on EBPIs. This should help facilities better understand the most important aspects of their performance. EBPIs should be reflected on inspector checklists, but they do not need to be explicitly included if they can be unambiguously derived from data that inspectors do collect.

“yes” if all sub-questions are also answered "yes." Question #2 is a quantitative roll-up question. The average monthly hazardous waste generation rate should be the additive amount of the amounts provided for the separate waste streams in Question #3.

Box 1: Abbreviated Illustrative Inspector Checklist (EBPIs in italics)

Hazardous Waste Generation

- 1) Facility's Hazardous Waste Generator Status: LQG SQG CESQG Non-handler
- 2) *Average monthly generation rate (gallons) of hazardous waste (over most recent 12-month period)* _____
- 3) *Average monthly generation rate (gallons) of individual hazardous waste streams (over most recent 12-month period):¹*

Waste Stream & Code	Generation Rate
_____	_____
_____	_____
_____	_____

Hazardous Waste Container Management

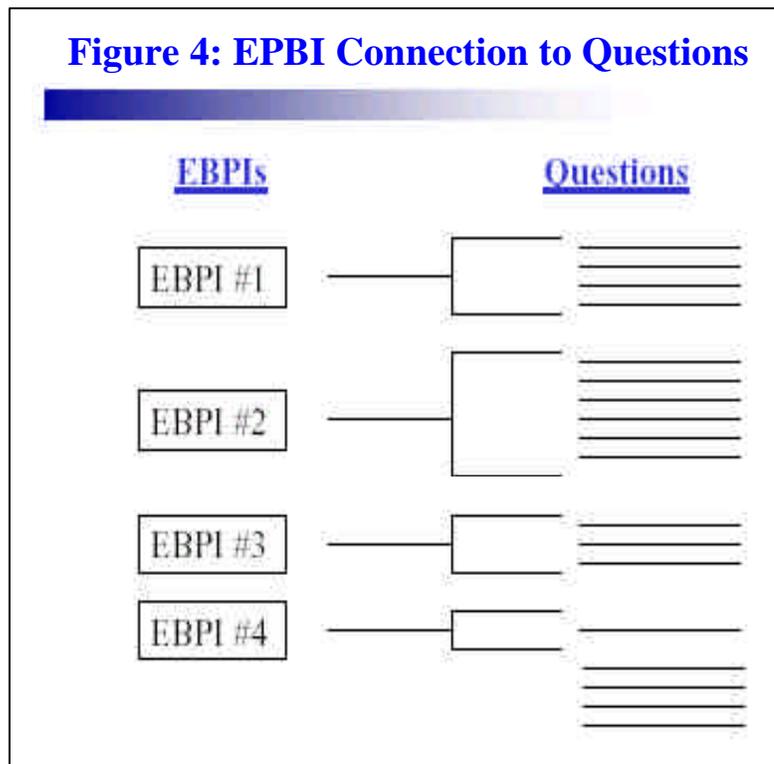
- 4) Does the facility accumulate hazardous waste on-site prior to treatment or disposal?
- 5) *Are containers holding hazardous waste managed to prevent releases? (*All of the following must be affirmative in order for this to be marked "yes"):*
 - a) Are the containers closed?
 - b) Are the containers in good condition? (Check for leaks, corrosion, bulges, etc.)
 - c) Are the containers handled in a manner to prevent the container from rupturing or leaking?
 - d) Is the hazardous waste compatible with the container and/or its liner?
 - e) Are containers holding incompatible waste kept apart by physical barriers or sufficient distance?

Hazardous Waste Training

- 6) *Have facility personnel received hazardous waste training?*
- 7) *Are employees familiar with proper waste handling and emergency procedures as relevant to their job duties?*
- 8) Was the training timely? (*All of the following must be affirmative in order for this to be marked "yes")
 - a) Was the training within the last year?
 - b) Are people trained within 6 months of hiring?
 - c) Are new workers supervised prior to training?
 - d) Is training reviewed annually?
- 9) Does the training cover the required areas? (*All of the following must be affirmative in order for this to be marked "yes")
 - a) Does the training cover emergency response procedures, including equipment handling and inspection?
 - b) Does the training cover safety?
 - c) Does the training cover hazardous waste identification and handling procedures?
- 10) *Did facility personnel receive Pollution Prevention training within the last year?*

A stand-alone EBPI can serve several purposes. It can indicate performance on an issue of particular interest, or it can be a leading indicator, i.e., something that indicates performance in other areas. Alternatively, a stand-alone EBPI may be one that does not necessarily capture the answers of questions that would arguably be related to it (e.g., questions in the same section that are not associated with a roll-up EBPI). This kind of the EBPI is typically used when it is difficult to create a roll-up question, when it is not considered vital to capture the numerous "minor" questions with a roll-up EBPI, and/or when a question is a key agency concern not easily rolled into another EBPI (e.g., a best management practice). Question #10 in Box 1, regarding pollution prevention training, is an example of such an EBPI.

Figure 4 helps visualize these two different kinds of EBPIs and their connections to ERP questions (which can appear both in inspector checklists and self-certification forms). In Figure 4, EBPIs #1, #2 and #3 are all roll-up EBPIs -- with different numbers of underlying questions. EBPI #4, rather, is a stand-alone question. Also note the remaining individual questions below EBPI #4 that are not explicitly linked to any EBPI.



2.3.3. How to Decide on EBPIs

Each of the agency's key concerns related to environmental performance should be captured in an EBPI. In formulating EBPIs, you may wish to consider what motivated you to select the sector you are focusing on through ERP. In other words, what are the critical compliance and/or performance issues you expect facilities in this sector to have? It is recommended that EBPIs include a mix of compliance and best management practice questions. In addition, you should consider formulating EBPIs in such a way that they will be relevant in the long term (e.g., 5-10 years from ERP initiation). For example, you may wish to include an EBPI regarding an emerging significant issue. Doing so will provide you with data to develop a better understanding of the problem and how to target solutions, and in the future this will enable you to examine the problem historically.

You may wish to consider existing agency-wide metrics when determining EBPIs, so that ERP results can feed into agency-wide results and be compared to other agency programs. Similarly, you may wish to consider developing EBPIs that are similar to those of ERPs developed in other states, as this will allow you to benchmark your ERP results with those in other states.

2.4. STEP 4: Develop Data Collection Instruments

Appropriately generating a random sample is extremely important in this regard, and will be discussed below. Just as fundamental, however, is the accuracy and precision of the data collection methods. Imprecise or poorly-thought-through data collection methods can ruin perfectly good random samples. Consequently, you should be attentive to developing solid data collection instruments and training agency personnel in the use of those instruments. For the purposes of statistical analysis, ERP has three primary data collection instruments:⁹

- **Inspector checklists** are used by agency personnel in inspecting facilities. Inspectors take these checklists to the field to record the compliance and performance status of a random sample of facilities in the population. Because inspectors are expected to collect the most reliable data, the data collected using the inspector checklist serve as the foundation for analyzing facilities' environmental compliance and performance. The inspector checklist should collect all of the data related to performance that an agency later wants to analyze.
- **Self-certification forms** are filled out by all facilities in the population. Facilities complete these forms on their own, with the help of the self-certification workbook. Self-certification forms should focus on the key indicators of a facility's performance, e.g., EBPIs. The primary role of the self-certification forms is to educate facility managers about their environmental requirements and to encourage them to take responsibility for keeping their facilities in compliance. The data from self-certification forms may also be used to measure environmental performance for the population of facilities. However, the data from these forms are generally considered less reliable than the data collected with inspector checklists.
- **Non-applicability forms** are submitted by facilities to certify that they are not valid members of the target population. Supporting documentation for the self-certification program should guide all targeted facilities through a process of determining whether they are included in, or excluded from, the program. This language should precisely match those criteria inspectors use to make such determinations, which in turn should exactly match the criteria used in developing the database containing the population of facilities. Non-applicability forms may also collect data on facility characteristics such as number of employees, ownership classification, and location. Collecting such

⁹ In addition to the data collection instruments above that are used for statistical analyses, ERP also collects data from Return-to-Compliance (RTC) plans. Facilities must submit and act upon RTC plans if, through the self-certification process, they identify themselves as being out of compliance. However, these data are not used for direct statistical inferences. For example, in Massachusetts, RTC plans are handled in a relatively traditional regulatory fashion. All RTC plans are reviewed by Massachusetts DEP to determine if they are reasonable. Facilities that submit RTC plans frequently receive targeted, non-random inspections to determine if they meet the commitments in their RTC plans.

information may serve several purposes, including: 1) providing additional information to judge the veracity of the non-applicability statement and/or to target investigations of potential falsifications, and 2) enabling a better understanding of why a facility may have been mistakenly targeted, or if there are particular kinds of facilities that are unexpectedly being excluded.

The next section explains principles that will be helpful in designing these data collection instruments.

2.4.1. Principles for Constructing High Quality Data Collection Instruments

In order to allow statistically valid comparisons, ERP data collection instruments should follow several principles. These principles are described briefly below, and are further detailed in Appendix 3:

Consistency between Inspector Checklist and Self-Certification Form: The inspector checklist is often more detailed and comprehensive than the self-certification form. Where both forms collect the same data, phrase questions in a consistent manner so that the data can be compared between the forms. It is important to have some overlap between inspector checklists and self-certification forms, in order to be able to verify whether facilities are accurately filling out their self-certification forms.

Internal Consistency Checks: Data collection instruments may be designed with internal consistency checks in order to ensure overall data quality. Two types of errors may occur in data collection. The first type of error is simple oversight, error, or misunderstanding in recording data on forms or in transferring data into an electronic database. The second type of error is intentional falsification on self-certification forms or non-applicability forms filled out by facility managers. Internal consistency checks on forms can be useful in detecting both types of errors. These internal checks may be created by developing questions that have a necessary logical relationship between the answers (e.g., questions that ask for the same information in different ways, or questions where the response to one question should rule out responses to other questions). Questions that are designed specifically to detect falsifications on ERP self-certification forms have been called “red-flag questions.” For example, some ERP self-certification forms have been designed so that regulators can compare reported volumes of waste to reported quantities of chemicals purchased. If a facility reports waste volumes that are much higher than purchase volumes, this sends up a red-flag that there may be a problem with the facility’s certification data, and regulators may target that facility for a follow-up inspection. ERP data collection instruments may be automatically scanned for internal consistency once the data are entered into an electronic database. This type of automatic quality control procedure, called “rules-based processing” is discussed in section 2.5 below.

Comparability Over Time: Consider what data you may want to analyze in the future, as changes in an inspector checklist or self-certification form over time can make comparison to earlier years difficult or impossible. Similarly, during implementation, be very conservative when making changes to questions or forms, and pay particular attention to EBPIs.

Precision: Carefully consider whether questions will necessarily elicit the expected response. Avoid vague language, particularly on self-certification forms. Plain, easily understood and

direct language is always preferable. Again, pay particular attention to EBPIs. Testing questions with potential respondents is an important way to make sure that questions will be interpreted as intended.

Limiting Open-ended Responses: Design questions to elicit yes/no answers wherever possible, even if it takes several yes/no questions to get to the ultimate answer of interest.

Collecting Accurate Quantitative Data: Quantitative questions ideally should be preceded by yes/no questions that indicate whether a facility uses/generates a constituent of interest. Quantitative questions themselves should elicit answers expressed in pre-specified units that provide consistent and comparable volume and time (such as pounds of waste generated per month). If possible, it is also desirable to collect data that can be used to normalize quantitative data in relation to the level of production. Again, pay particular attention to EBPIs.

Providing Decision Rules for Respondents: Provide facilities and inspectors with clear rules for determining non-applicability of a question. For example, if a question asked whether a facility was providing adequate hazardous waste management training to its employees, the question should define the term "adequate." For example, a decision rule might inform respondents that they are allowed to indicate that the training is "adequate" only if the respondent is able to answer yes to several sub-questions related to the frequency and content of the training.

Reasonable Length: To the extent possible, limit the length of forms, because increased length increases the likelihood of error. This is particularly true of self-certification forms, because facilities will fill out the form with varying levels of experience, training, and technical expertise. You can shorten the self-certification form in part by focusing upon the EBPIs. You may also make the form modular, so that facilities do not have to complete the entire form at once. For example, a self-certification form might have separate, clearly distinct sections for different facility processes.

Providing Data on Other Relevant Characteristics: Consider including questions on inspector checklists, self-certification forms, and non-applicability forms to better understand the characteristics of the facilities being targeted. For example, questions could provide information on differentiating sub-groups of facilities that may be hypothesized to have different responses to ERP -- such as by number of employees or by ownership type.

Because creation of good data collection instruments and their supporting documentation (such as compliance workbooks) can be resource-intensive, it is strongly recommended that you consider adapting relevant, tested materials previously created by EPA, your agency, or by other states, particularly ERP-specific materials.

To ensure that you can fulfill the principles described above, you should allow sufficient preparation time to draft several iterations of all data collection instruments to be used in the program. In addition, it is recommended that you involve relevant internal and external stakeholders in reviewing the documents to ensure that questions are easily understood and are likely to elicit the expected response. If reviewers have limited time, you should suggest that they focus upon the EBPIs. In the case of inspector checklists, agency field inspectors will likely provide very valuable input. In the case of self-certification forms, trade associations and

individual facilities in the target population can be expected to provide useful feedback. Be cautious when sharing inspector checklists and EBPIs with so facilities do not gain the false impression that they only need to comply with EBPIs, rather than with all applicable requirements.

2.5. STEP 5: Develop a Results Database

As the inspector checklist and self-certification forms are being developed, work should begin on the database that will store the information collected in those forms. It is very helpful to start talking with Information Technology (IT) staff people at this time to ensure that databases to store information collected are appropriately structured. IT staff can provide a good perspective on ensuring that your data will be appropriately collected and managed.

Take care to ensure that data can be easily and seamlessly entered from inspector checklist and self-certification forms into the database, in order to minimize errors in translating the data into an electronic format. One way to efficiently and accurately enter data is to create forms that can be electronically scanned into the database. The database should be compatible with the database developed to house the complete list of the population of facilities. You may also wish to design the database so that it can be dynamically integrated with other, pre-existing, stand-alone state databases (e.g., a compliance databases or facility registries).

One important function a results database can serve is to assist in quality assurance procedures. If you have designed internal consistency checks into your data collection instruments, you can set up rules in your database to automatically highlight inconsistencies in the data. Such rules-based processing systems can allow you to pinpoint needed follow up activities, such as potential compliance problems highlighted by red-flag questions. You can also design the database in order to minimize errors in entering data. For example, if a response to one question indicates that a particular section on the form should be skipped because the questions are not applicable, you can design the database so that it will not accept responses for the “skipped” questions. You should work with IT staff to create a user-friendly and error-minimizing interface for data submission. This is particularly important if you are considering allowing facilities to submit data electronically. Also, if facilities are submitting certifications electronically, you will need to develop an effective electronic signature protocol that mimics the ERP mechanism of requiring the signature of a responsible corporate officer.

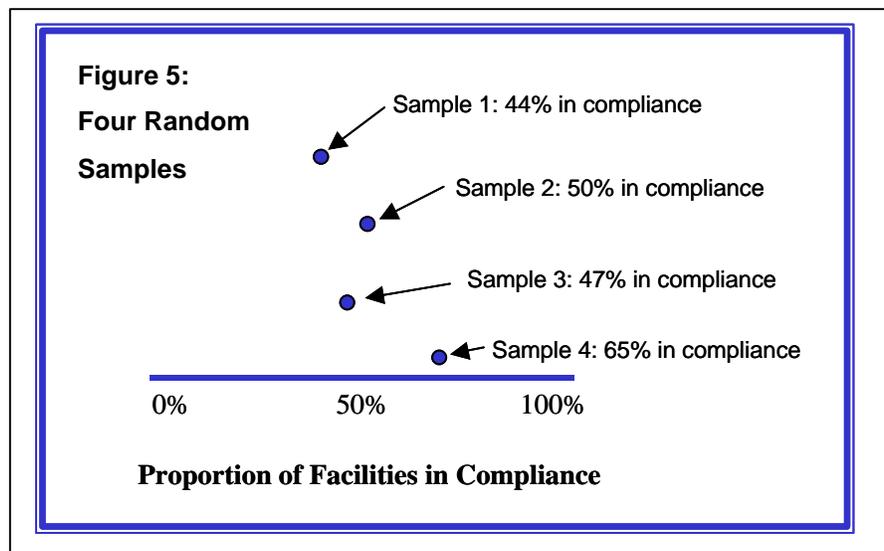
2.6. STEP 6: Plan for Sample Selection

As noted in the Introduction, ERP performance measurement relies upon collecting information for a sample of facilities in order to draw conclusions about the target population of facilities. Section 2.2 above discussed developing the sampling population (i.e., the list of facilities from which the sample will be drawn). In order to draw a statistically valid sample from the sampling population, you must take care to select a *random sample* so that you can use the sample to make conclusions about the total population. Moreover, you must draw a sample of sufficient size so that your results are meaningful. Statistics tells us that if: (1) the sampling size is large enough, (2) the sample is randomly selected, and (3) the data collection instruments are not biased or the data otherwise flawed (as discussed in section 2.4), then one can develop a meaningful estimate

of the population proportion based on the sample proportion.¹⁰ In other words, one can estimate the percentage of facilities with a certain characteristic in the total population based on the percentage of facilities with that characteristic in the sample.

In order to help clarify these concepts, it is useful to think about how sampling works. Suppose you wanted to understand what percentage of auto repair facilities in your state comply with a certain hazardous waste handling requirement. Since you do not have the resources to go out and inspect all of the relevant auto repair facilities, you could inspect a sample of facilities to estimate the percentage of all facilities that are in compliance. Of course, any random sample that you take would likely come up with a slightly different proportion of facilities that are in compliance (see Figure 5). Suppose you sampled 100 randomly selected facilities, and you found that 44% of them were in compliance. If you took another sample of 100 randomly selected facilities, you might find that 50% of them were in compliance. A third sample of 100 randomly selected facilities might find 47% compliance, and so on.

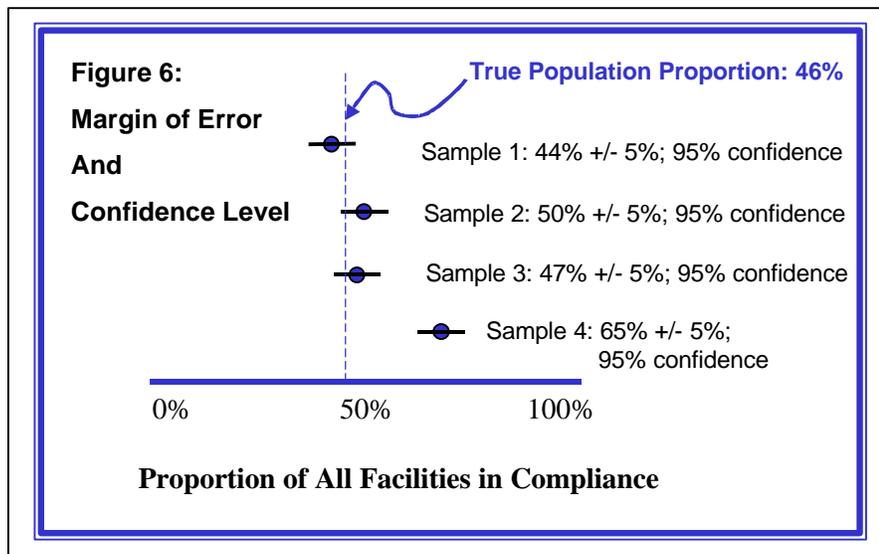
Any individual random sample you might take is unlikely to have the exact same percentage of facilities in compliance as the total population of facilities. However, statistics gives us a way to estimate the population proportion from the sample proportion. Two key concepts in understanding this estimate are *margin of error* and the *confidence level*. The margin of error for a



particular sample reflects a range of values in which the true population proportion is likely to lie. The margin of error is often expressed as an interval of X percentage points above/below the sample observation (e.g., +/- 5%). The confidence level is a measure of the confidence we have that the true population proportion is within the interval described by the margin of error. The confidence level is typically expressed in terms of a percentage (e.g., 90% or 95%). Thus, for our first sample of auto repair facilities where we found 44% of facilities in compliance with the hazardous waste requirement, we could estimate the population proportion as 44% +/- X% with a Y % confidence level, where X is the margin of error and Y is the confidence level. For instance, if the finding were 44% +/- 5% with a 95% confidence level, this would mean that we are 95% confident that the true population proportion in compliance with the hazardous waste requirement is between 39% and 49%. Note that even if we are 95% confident that the true population proportion is within the range defined by the margin of error, in 1 out of 20 cases it

¹⁰ Some states implementing ERP have been concerned that the inspections that are part of ERP could bias the sample or the statistical analysis. This is not the case, as long as results are properly interpreted. Inspections should have a deterrence effect for facilities, and this effect is part of the ERP approach.

will turn out not to be in that range (Sample 4 in Figure 6 illustrates such an occurrence).¹¹ Since we are only taking samples and not a census of all facilities, we will never know the true population proportion. However, by using the margin of error and confidence level, we can quantify the degree of uncertainty we have about our estimate of the true population proportion. This approach is substantially more desirable than collecting data that are not statistically valid, and not understanding the degree of uncertainty in your estimates. Note that margin of error and confidence level do not reflect any uncertainty and error that would result from biased data collection, non-random samples, or other flaws in the design of the statistical methodology.



2.6.1. A Few Key Facts about Margin of Error and Confidence Level

Understanding how the margin of error and confidence level work is very important to the performance measurement aspect of ERP, since they determine how precise your measurements can be and how confident you can be in the results. Since these concepts are so important, this section explains a few key facts about how margin of error and confidence level are related to each other and to the size of the sample and the population.

- The first key fact is that sample size has a big impact on the margin of error and confidence level. In general, the larger your sample size, the more precise your estimate can be (i.e., the smaller your margin of error) and the more confident you can be in that estimate (i.e., the higher your confidence level). Of course, obtaining a larger sample size requires conducting more inspections, which takes more agency resources. Therefore, states implementing ERP need to think about how they want to balance resource expenditures on sample inspections vs. uncertainty in performance measurement results.
- For a fixed sample size, if you decrease the margin of error, you also decrease the confidence level. Thus, for a given sample size, by making your estimate more precise, you sacrifice some of your confidence in that estimate.

¹¹ If the confidence level is 95%, the likelihood that the range of values described by any individual sample's margin of error will not contain the true population proportion is 5%.

- You can go about selecting a sample size in one of two ways. One approach is to let your resources determine how many inspections you can conduct, and live with the margin of error and confidence level that the sample size produces. While this approach may seem pragmatic, the downside is that if your sample size is too small, your estimates will be so imprecise and/or you will have so little confidence in the results that they may be of little or no value. Therefore, this guide focuses on another approach, which is to set the desired confidence level and stipulate the largest margin of error you are willing to accept at the outset, and let these figures determine your sample size. This latter approach (or a combination of the two approaches) is more typically used because it helps ensure that your results will be meaningful. Researchers typically require a margin of error of 5% or less and a confidence level of 90% or 95%, depending on the types of statistical analysis they are performing. Confidence levels lower than 90% are not typically considered acceptable. A margin of error of 5% may be acceptable if you expect the proportion to be in the range of 40% to 60%, but not if you expect the proportion to be, say 10% to 20%. In the latter case, you may only be comfortable with a margin of error of 2% for the results to be meaningful. A method typically used by researchers is to set the maximum allowable margin of error at 10% of the expected proportion. Thus, the margin of error would be 5% if the expected proportion is around 50% (since 10% of 50% equals 5%), but only 1% if the expected proportion is 10% (since 10% of 10% equals 1%).
- The margin of error established prior to sampling should be understood as an estimated figure. The exact margin of error can only be calculated after the sample has been taken, and the exact number of responses is known. The number of responses received may be smaller than anticipated for two reasons. First, some facilities may be closed or otherwise not relevant for the survey because they are no longer involved in the activity under study. Such facilities do not belong to the target population so they should not be included in the sample. Ideally, in this situation, these “drop out facilities” would be replaced with other randomly selected facilities from the sampling population (see section 2.7 for more information on establishing a drop-out protocol). Second, while a facility may be a valid member of the target population, some questions may not be relevant for that facility (e.g., if it does not have a particular type of regulated process or issue being studied). For any questions that receive one or more legitimate “N/A” (not applicable) responses, the effective sample size is reduced by the number of N/A responses. Thus, when analyzing these questions, the exact margin of error may be greater than the preset maximum. Of course, you can always decrease the margin of error if you are willing to decrease the confidence level, but the latter should not be less than 90%, and you should check with a qualified statistician before doing so. This issue of exact vs. estimated margin of error can become a problem when sampling from a somewhat heterogeneous population. For example, consider a population of facilities that includes conditionally exempt, small quantity, and large quantity generators of hazardous waste. In this case, questions relevant only to large quantity generators will receive fewer responses than questions relevant to all generators. Therefore, you will be able to make statistical inferences regarding large quantity generator questions with less confidence than the overall number of facilities sampled would suggest. This reduced confidence may be acceptable if the questions you are analyzing are not critical to determining facility performance. If, however, you anticipate that many questions that are

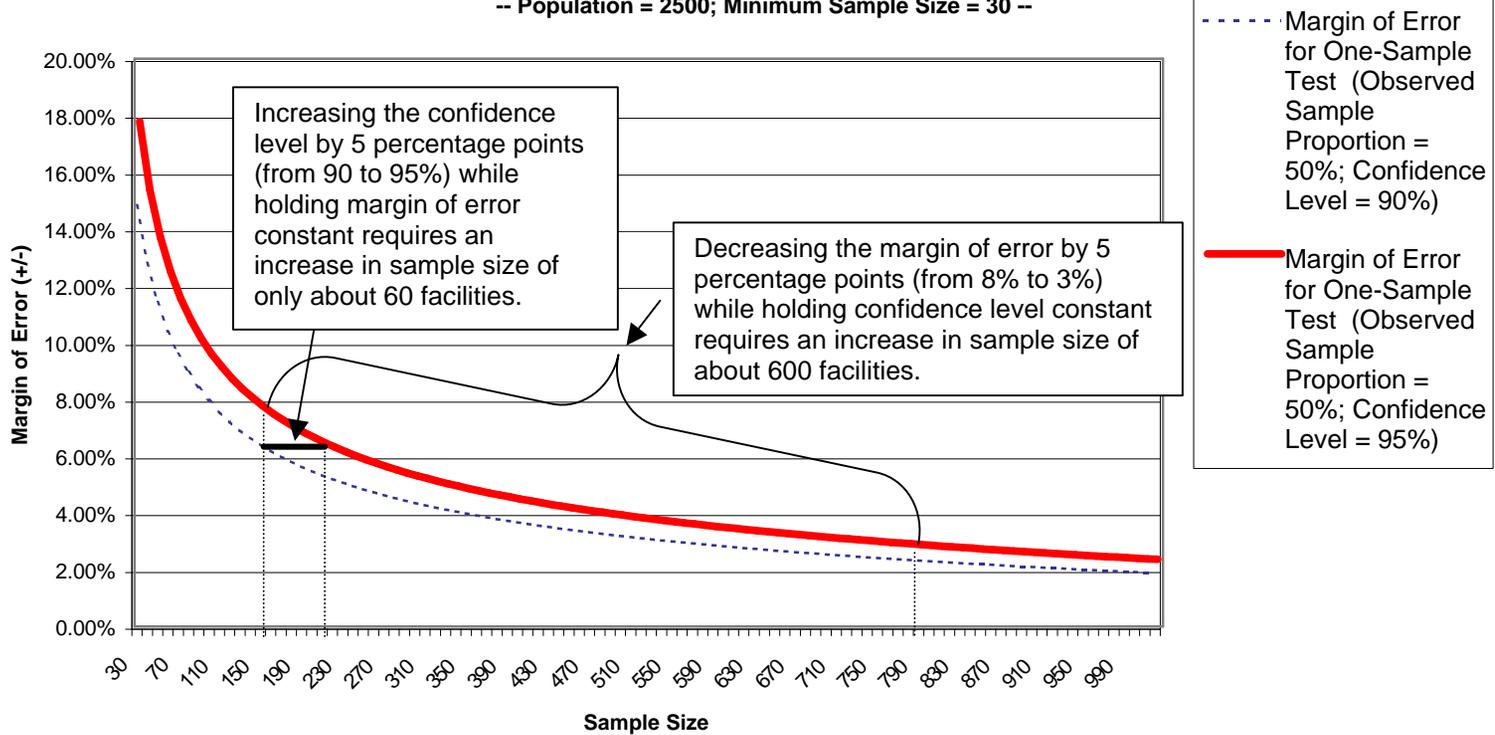
important to determining facility performance will result in a substantial number of N/A responses, you should consider stratifying the sample (for more information on stratifying the sample, see section 2.6.2, Question 2). Alternatively, if you do not have the data about facility characteristics that would be needed to stratify the sample, or if the questions that may generate N/A responses are not critical to evaluating ERP success, you may choose simply to increase the sample size to account for likely N/A responses.

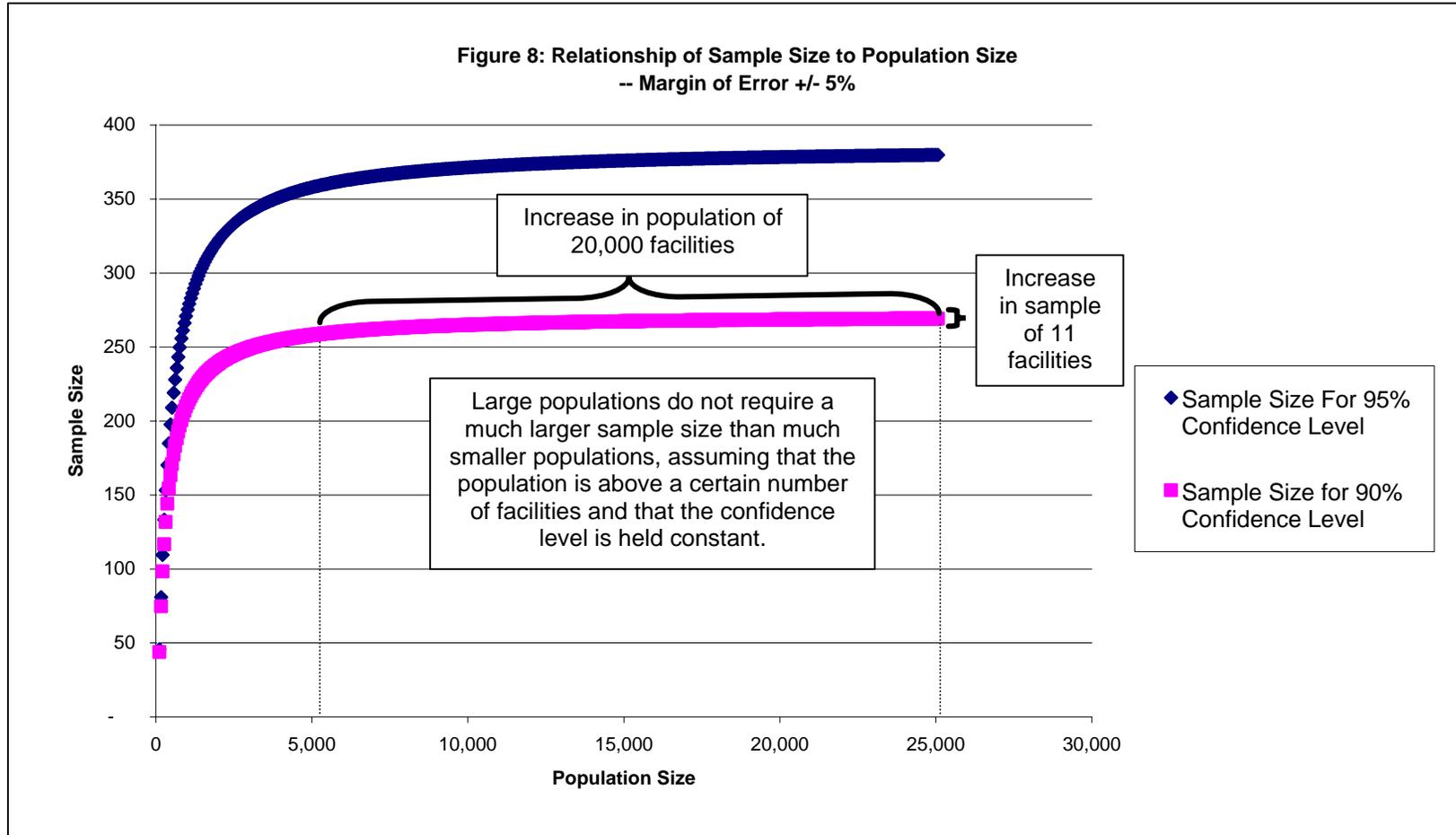
- While sample size and margin of error are related, the relationship is not linear. When a sample size is very small (e.g., less than 30 inspections), margin of error is very large. As the sample size increases above 30, margin of error drops rapidly, but it begins to level off as sample size reaches above 200. To cut the margin of error in half, you must quadruple the sample size. In Figure 7, the margin of error is about 6% with a sample size of 200, but you need a sample of about 800 to cut the error to 3%. Even with very large sample sizes there is still some margin of error. For the purposes of ERP, this means that it is important to conduct a minimum level of inspections, but that beyond a certain point the value of continuing to increase the sample size diminishes. Figure 6 illustrates this point. It shows that, for a given confidence interval (either 90% or 95%) and population size (in this case 2500), the margin of error decreases as sample size increases. Note that it takes a far smaller increase in sample size to improve one's confidence level by five percentage points than it does to decrease the margin of error by five percentage points.
- Figure 8 illustrates another very important point: the positive, but non-linear relationship between population size and sample size. For the given margin of error of +/- 5.0 %, note how the sample size necessary for a given population grows rapidly at small population sizes, but levels off quickly. This fact is important for staff designing ERP programs, since it means that an ERP designed for a small target population may not be as resource-efficient as an ERP designed for a large target population, for a given margin of error and confidence level.
- A final important fact to know about margin of error relates to the common situation where you want to compare results between two different samples (e.g., one sample before self-certification and one sample after certification). In this case, the margin of error associated with the comparison is larger than the margin of error associated with either of the individual samples.¹² For example, if the margin of error for each sample is 5%, the margin of error for the comparison between the samples is 7%. A good way of thinking about the margin of error associated with a comparison between two samples is the “minimal detectable difference.” In other words, when comparing two samples that each had a margin of error of 5%, a minimal detectable difference of 7% in the sample proportions would be needed to conclude the population proportions were actually different. So, an observed difference in the sample proportions greater than 7 percentage points would be statistically significant, but

¹² Specifically, the margin of error for the comparison of two samples is the square root of the sum of the squares of the error for each sample (i.e., $\sqrt{(\text{error for sample 1})^2 + (\text{error for sample 2})^2}$). For two samples of equal size and equal margin of error, the margin of error for the comparison between the two samples will always be the margin of error for one sample multiplied by 1.4.

smaller observed differences in the sample proportions would not be statistically significant. For example, if you observed a difference in the sample proportions of 6%, you could not say with any statistical confidence that there is an actual difference in the population proportions.

Figure 7:
Margin of Error Decreases As Sample Size Increases, for Two Confidence Levels
-- Population = 2500; Minimum Sample Size = 30 --





2.6.2. *How to Decide Upon a Your Sample Size*

As noted above, the sample size you choose is very important in determining how you can interpret your results. This section describes seven questions that you need to answer in order to select your sample size. Once you have answered these questions (and assuming your answers match the simplifying assumptions this guide uses), then you can use formulas listed in Appendix 5 in order to determine your sample size. Note that if your situation does not match the simplifying assumptions, or if you are unsure of your answers to the questions outlined in this section, you should consult a qualified statistician to ensure that you design your sample appropriately. Also, be aware that the list of facilities from which you draw your sample of facilities to be inspected should include *all* relevant facilities, even if they do not respond to ERP mailings. This issue is discussed in more detail in section 3.1.

Question 1: Are you measuring a proportion or an average value?

As noted in the introduction, this statistical guide only discusses how to estimate population proportions (e.g., percentage of facilities in compliance with a select requirement). If your primary interest is in estimating a population mean (e.g., average amount of emissions produced), you should consult a statistician.¹³ If your primary interest is in measuring a proportion, but you also decide to collect non-binary data (e.g., pounds of emissions, quantity of hazardous waste produced, etc.), you may be able to use such data, but this document does not review how to set up a statistical analysis for estimating population means.

Question 2: Do you expect that the population of facilities will be naturally divided into subgroups?

If you expect that the characteristic(s) you are estimating for the population of facilities will vary based on some objective and relevant criterion, then you should create a *stratified sample* according to that criterion. For example, if you expect auto repair shops owned by dealerships to have a consistently higher rate of compliance than independently owned auto repair shops, you should stratify your sample based on facility ownership. If your interest is simply in drawing conclusions about the whole population (and not in characterizing or comparing the subgroups in the population), the simplest way of creating a stratified sample is to use a technique called *proportional allocation*. First, decide on a sample size based on the size of the total population. Then, separate the sampling population into the relevant subgroups (i.e., the strata). Finally, draw a proportional sample from each subgroup that adds up to the overall sample size. Thus, each subgroup's proportion in the sample should be equal to each subgroup's proportion in the sampling population.¹⁴ This method of proportional allocation allows you to make statistically valid conclusions for the whole population.

¹³ A key reason why estimating population means is more complicated is that in order to do so you must estimate the sample variance (a measure of variation in values) before you can estimate sample size.

¹⁴ For example, suppose an agency has a sampling population of 2500 auto repair facilities, 60% of which are owned by dealerships and 40% of which are owned independently. The agency wants to draw a proportional sample, so that dealerships and independent shops are proportionally represented in the sample. Suppose that based on the size of the population and other considerations discussed in this section, the agency determines it needs a total sample size of 250 facilities in order to draw statistically valid conclusions about the entire population. To draw a proper proportional sample, it should select a random sample of 150 facilities owned by dealerships, and a random sample

Proportional allocation can also be used to control the number of inspections that any individual district or region within an agency's jurisdiction might have to conduct, even if it is not believed that geographic location of facilities is correlated to their compliance response. Many agencies divide inspection resources among different geographic jurisdictions within the state. Without proportional allocation, there is a possibility that one or more geographic jurisdictions would face a disproportionate burden of inspections.

Note that another sampling approach -- *cluster sampling* -- may be used if facilities are expected to be so widely dispersed geographically that it would be impractical to spend the resources needed to select a simple random sample. Cluster sampling can also be important in ERP where each facility may have several different "units" that would be individually evaluated for compliance. For example, gas stations with multiple underground storage tanks, or facilities with multiple degreasers or boilers, may be appropriately evaluated using cluster sampling. If you wish to consider cluster sampling, you should consult with a qualified statistician for help with designing the sample.

Question 3: Is it important to you to make statistically valid statements about subgroups of the population or to compare subgroups in a population?

If your population falls naturally into subgroups and you want to draw statistically significant conclusions for any one subgroup or to make comparisons between subgroups, you should stratify your population by these subgroups. However, rather than using proportional allocation, you should use *equal allocation*. This essentially means taking a statistically valid sample for each subgroup individually. This approach will enable you to make statistically valid statements about the subgroups and about comparisons between subgroups. However, equal allocation requires a somewhat larger total sample size than proportional allocation to achieve the same level of precision for the estimate of the overall population proportion.¹⁵

Question 4: Are you interested in making comparisons with your ERP data?

Depending on whether or not you will be interested in making comparisons with ERP data, and what kinds of comparisons you will want to make, the minimum sample size needed may vary. Keep in mind that you should choose a sample size that is large enough to allow you to make any of the comparisons that you may want to make over the course of ERP implementation. Appendix 5 contains specific instructions for how to choose your sample size once you know the types of analysis you want to conduct.

of 100 independently-owned facilities. The result is that 60% of the total sample will be owned by dealerships, and 40% of the total sample will be independently owned, thus reflecting the proportion of ownership categories in the total population.

¹⁵ For example, suppose a population of 1,000 facilities is divided into two strata, with 250 facilities in stratum 1 and 750 in stratum 2. To obtain an overall estimate of the population proportion with margin of error plus or minus 5% with 90% confidence, a sample of 270 facilities is required if proportional allocation is used (68 from stratum 1 and 202 from stratum 2). If the primary concern is to compare the strata, you want to sample 135 from each stratum. However, the margin of error would be plus or minus 5.6%. In order to get that error back down to 5.0% you would need to increase the sample from 135 to 169 in each stratum, or to 338 overall from the original 270 (a 25% increase).

In some cases, you may not need to make any comparisons with your data. For example, if you are analyzing baseline inspection data, you may simply want to estimate the proportion of facilities in compliance with a particular requirement for a single group of facilities at a single point in time. Suppose that you found 85% of facilities in your sample were in compliance with a given requirement. In order to make an inference about the percentage of facilities in the entire population that is complying with this requirement, compute a *confidence interval* (i.e., 85% of facilities are in compliance, with a +/- X% margin of error, at a Y% confidence level). Appendix 5 contains more information on computing confidence intervals.

While analyzing baseline data need not involve comparisons, much of the analysis in ERP centers on making comparisons between different sets of data. There are three primary types of comparisons you may wish to make:

- **Comparing self-certification data to inspection data** for a sample of facilities in order to draw inferences about the reliability of self-certification forms for the entire population;
- **Comparing results between two different samples of facilities** (e.g., comparing samples before and after self-certification or comparing samples from different regions) in order to draw inferences about differences between two populations of facilities; and
- **Comparing results from one sample to an expected value** in order draw inferences about whether there is a difference between the population proportion and the expected value.

The following paragraphs discuss each of these comparisons in turn.

If your interest is in comparing self-certification data to inspection data, one approach is to compare data on a facility-by-facility basis. In other words, for each facility in the randomly selected sample of inspections that occur after self-certification, you would compare the inspection results to the information provided by the facility on the self-certification form. For example, suppose you are considering compliance with a requirement to have adequate fire equipment installed in the building. Suppose you had 150 facilities in your sample, and for 15 of those facilities, self-certification forms reported the facilities were in compliance with the requirement but inspectors found there was not adequate fire equipment installed. Thus 10% of the facilities in your sample inaccurately reported that they were in compliance with this requirement. In order to make an inference about the percentage of facilities in the entire population that inaccurately reported compliance with this requirement, you would compute a confidence interval (i.e., 10% of facilities inaccurately reported their compliance status, +/- X% margin of error, with a Y% confidence level). It is recommended that you consult with a statistician for help in setting up this type of comparison. If you are comparing the percentage of *all* facilities in compliance with a particular requirement as reported on self-certification forms to the percentage of a *sample* of facilities in compliance with that requirement as determined by inspectors, the way in which you set up the analysis may depend on whether self-certification is mandatory or voluntary.

If your interest is in comparing the results of two samples, you can use a *two-sample hypothesis test* to conduct the analysis. For example, if you were comparing a sample of facilities inspected before self-certification to a sample of facilities inspected after self-certification, you would use a

two-sample hypothesis test to infer whether compliance for a given requirement has increased among all facilities in the population after self-certification. Another case where you would be comparing two samples is if you were considering samples that were conducted in two different regions, and you wanted to compare the percentage of facilities in compliance in one region with the percentage of facilities in compliance in a second region. Again, you could use a two sample hypothesis test to infer whether compliance for all facilities in the first region differed from compliance for all facilities in the second region, based on the comparison of two samples.

Finally, if your interest is in comparing results from one sample to an expected value, you can use the formula for a *one-sample hypothesis test*. The one-sample formula allows you to take a snapshot of performance for a single group of facilities at a single point in time and compare this to a hypothesized value. Suppose, for example, that you expect that at least 90% of facilities are in compliance with a particular requirement based on past research. You could conduct a one-sample hypothesis test to compare the proportion of facilities in a single sample, at a single point in time, that are in compliance with this requirement vs. the hypothesized 90% of facilities that you expect to be in compliance.

Further information on setting up and performing hypothesis tests is included in Appendix 5.

Question 5: If you are making comparisons with the data, are you interested in determining if one population proportion is greater than (or less than) a specific value, or are you only concerned with determining whether a population proportion is equal to or different than a specified value?

This question is best illustrated by an example. Suppose you would like to know whether compliance rates for your target group of facilities for a particular requirement have improved after the self-certification process, as compared to before self-certification. In this case, you will compare a sample of inspections before self-certification to a sample of inspections after self-certification, and based on the comparison of samples, you can make inferences about changes in compliance among the total group of facilities. In this case you probably want to know whether the percentage of facilities in compliance with each requirement has *increased* between the two rounds of inspections. This type of analysis where you are interested in the direction of the difference is called a *one-sided hypothesis test*. For example, suppose you are considering compliance with a requirement to provide hazardous waste training for employees. If inspectors found that 60% of facilities in the “before” sample were in compliance with this requirement, and they found that 65% of facilities in the “after” sample of inspections were in compliance with this requirement, can you say that compliance has improved for all facilities in your target group? In order to answer this question, you should conduct a one-sided hypothesis test. In this case, the possible results are that either: 1) the proportion of all facilities in compliance with the requirement has increased or 2) the proportion of all facilities in compliance has stayed the same or decreased. With the one-sided test, in this case, you cannot distinguish between the proportion of facilities in compliance staying the same and decreasing.

In other cases, you may only be interested in determining whether there is *any difference in proportions* for the total target group of facilities, not whether the “before” proportion is greater than or less than the “after” proportion. In the example above, if you were not concerned with

testing whether compliance rates for all facilities had *improved or declined* after self-certification, but instead you just wanted to know if compliance rates had *changed*, then you would conduct a *two-sided hypothesis test*.

Another case where you could conduct a two-sided hypothesis test is where you want to see whether overall self-reported compliance rates for a given requirement are different than the rate of compliance for that requirement as determined by inspectors. In this case, you are not making a comparison on a facility-by-facility basis (as described in the previous discussion), but rather you are comparing aggregate compliance rates. For example, suppose that 80% of all facilities submitting self-certification forms reported that they were in compliance with a certain requirement, but inspectors found that 75% of facilities in the sample of inspections after self-certification were in compliance with this requirement. You could conduct a two-sided, one-sample hypothesis test to determine whether there is a statistically significant difference in self-reported vs. inspector determined compliance rates for the population as a whole. This type of aggregate comparison is less detailed and precise than making comparisons on a facility-by-facility basis, however it still may result in useful information.

Depending on whether you are constructing a one- or two-sided hypothesis test, your sample size will differ, as will the *critical values* for analyzing your statistical results. Keep in mind that both kinds of statistical tests – one-sided and two-sided – can be useful for ERP. When you are deciding on a sample size, if you think you may want to conduct any two-sided tests, you should decide on the sample size based on a two-sided test. This is because, for a given margin of error, a two-sided test requires a larger sample. In other words, you can always use a sample drawn for a two-sided test to answer one-sided questions, but you cannot necessarily do the reverse and still have statistically valid conclusions at the same level of significance. For more information on how to construct hypothesis tests, see Appendix 5.

Question 6: With what level of error are you comfortable?

As explained in the previous section, margin of error is related to sample size (i.e., for a given population level and confidence level, the smaller your desired margin of error, the larger the sample size needed). Ideally, you should select your desired margin of error and allow that figure to determine the size of your sample. The preferred margin of error typically used in statistical sampling is 5% or less (see section 2.6.1). You may have to accept a larger margin of error if resource constraints limit the size of your sample, but be aware that as your margin of error increases your estimate will become less precise. For example, suppose in order to limit the resources spent on compliance inspections a state decides to accept a margin of error of 15%. This means that if the state observes that 50% of the facilities in its sample are in compliance, the state can only estimate that somewhere between 35% and 65% of facilities in the population are in compliance. This wide margin of error may not be very useful for making informed policy decisions.

Also, as noted above, if you are interested in comparing two samples (e.g., a sample before self-certification and after self-certification), your margin of error for the comparison between the two samples will be larger than the margin of error for either sample individually. For example, if you have a margin of error of 5% for each of two samples, the minimal detectable difference

between the two samples will be a 7 percentage point difference. In this scenario, if the observed compliance rate in the second sample (after certification) was only 6% higher than the observed compliance rate in the first sample (before certification), you could not say with any statistical validity that compliance rates in the population had increased over the study period. Alternatively, for the state that chose a 15% margin of error for each sample, the minimal detectable difference between the two samples would be a 21 percentage point difference (see section 2.6.1, footnote 11). So even an observed 20% improvement in compliance between the two samples would not be statistically detectable (i.e., you would not be able to draw a statistically valid conclusion that compliance rates had increased in the population).

Based upon Massachusetts' experience with ERP, it may not be unreasonable to expect to observe some large improvements in compliance with specific requirements at the outset of the program. However, large margins of error increase the likelihood that observed changes in certain (possibly critical) parameters will not be statistically significant even in the first year. Furthermore, in later years, observed changes will likely grow smaller and be more difficult to measure with statistical confidence. In such cases, it may be necessary to compare current sample results to results from samples taken two or three rounds earlier in order to find a statistically significant change in compliance rates.

Question 7: How confident do you want to be in your estimate?

As discussed earlier, the confidence levels generally used in statistical sampling are either 90% or 95%. A 90% confidence interval means that for any given sample, there is a 10% chance that the true population proportion will not be in the range described by the margin of error. A 95% confidence interval means that for any given sample, there is a 5% chance that the true population proportion will not be in the range described by the margin of error. If you are interested in a one-sided test (see Question 4 above), a 90% confidence interval is generally acceptable. Otherwise, a 95% confidence interval is usually preferred.

Note that if you have a biased sampling population, or if you collect data in a way that biases your results (such as sampling by convenience or by targeting rather than at random), then your margin of error and confidence level are unreliable. In order to have a statistically valid sample you must ensure not only that the sample is of sufficient size, but also that your entire statistical methodology is sound.

Based on your answers to the seven questions above, you should be able to use the formulas presented in Appendix 5 to determine your sample size. However, note that regardless of the answers to the questions above and the sample size indicated from the formulas in Appendix 5, all of your samples should contain at least 30 facilities. This is because 30 facilities is a minimum number needed to carry out standard hypothesis tests without using more sophisticated, so-called "exact" methods (in which case a qualified statistician should be consulted). If your entire sampling population is less than 30 facilities you should seriously consider taking a census of the entire population rather than trying to sample it.

The following sections offer some additional practical considerations for selecting samples.

2.6.3. Available Resources and Project Timeframe

In the real world, there are often severe constraints on the resources available to carry out an inspection and evaluation program. Within these constraints, a sample should be chosen to satisfy an acceptable level of statistical accuracy. It is recommended that each individual round of inspections be carried out within a fairly tight timeframe (1 to 2 months) so as not to bias the results. Otherwise, some facilities could be accorded substantially more time and advance notice before inspections than others, unintentionally biasing measures of performance. The frequency with which the agency chooses to conduct inspections can depend on performance results. For example, if facilities seem to be performing well, based on a trend analysis, an agency may wish to extend the period between inspections in order to target inspection resources on facilities or sectors that are not performing as well.

In addition, program planners should choose timeframes between inspection periods that can remain uniform over the years. For example, one would not ideally want to have a six-month difference between the beginning of the baseline inspections and the beginning of the first set of follow-up inspections, and then have a one year elapse before the beginning of the next round of inspections. For similar reasons, all facilities should be required to provide self-certification under a common deadline, otherwise it will be difficult to interpret the impact of self-certification, because different facilities could have substantially different time periods between self-certification and the follow-up round of random inspections.

Ideally, the post-self-certification round of random inspections should occur shortly after the deadline for self-certification. If much time elapses, it will be difficult to use the results of the follow-up round of random inspections to verify the accuracy of self-certification responses, since actual performance of the facility could change between the time the facility self-reports its performance and the time the inspector visits. This final issue is discussed in further detail in Appendix 4.

2.6.4. Mandatory v. Voluntary Participation in the Program

It is recommended that the self-certification program be mandatory for two reasons: (1) so that the self-certification process can have an opportunity to impact the performance of all facilities and (2) so that self-certification data can provide a picture of the entire population of facilities targeted by the ERP. If the program is voluntary, the random samples for inspections should be drawn from the entire sampling population (not just those that volunteer). If the samples are drawn only from the facilities that volunteer to self-certify, there is a high probability of obtaining a biased sample since facilities self-select themselves into the program. A potential source of bias is that facilities with better performance may be more likely to volunteer for self-certification than facilities with worse performance. Any random sample drawn from such a volunteer population is likely to give a distorted view of the true level of compliance among facilities. A voluntary program may create other biases, as well, and it is suggested that any state considering using voluntary self-certification as part of the ERP consult with a qualified statistician to discuss potential measurement problems. It may be advisable to set up a stratified sample for inspections to enable comparisons between volunteers and non-volunteers. This will

still enable you to construct an overall portrait of performance in the sector. At the very least, if a voluntary program is used, a sufficient sample of facilities that did not volunteer must go through the inspection process so it can be determined whether or not compliance rates among volunteers and non-volunteers are significantly different.

2.7. STEP 7: Deciding on a Drop-Out Protocol

Whenever sampling methodology is applied, due care must be taken to account for the fact that facilities may drop out of the sample for various reasons (e.g., facilities may go out of business or relocate, or the database of the population of facilities may have incorrectly identified their characteristics, leaving them in the population). Such discoveries may occur before the sample is chosen (i.e., when an agency is narrowing the population based upon characteristics provided in the database), or after the sample is chosen (i.e., during pre-inspection calls, at the beginning of an inspection visit, or when facilities file non-applicability forms). A protocol to handle these situations in a consistent way should be established ahead of time to avoid a risk of bias to the sample, and program staff should be trained on the protocol. One approach is to develop a random list of “alternate” facilities that will be substitutes in case some facilities drop out from the original sample. This simple procedure, discussed in section 3.1 below, will ensure that you have a list of randomly selected substitutes for any facilities that drop out of the sample.

In addition, certain questions answered by valid facilities may be dropped out if the question is not applicable to the facility. For example, a question relevant to facilities with automotive parts washers would not be relevant to facilities without parts washers. Alternately, a question might be invalidated if the facility or inspector clearly misunderstood the question or provided inappropriate information, and follow-up could not resolve the issue. Since the exact statistical precision regarding a particular question depends on the actual number of responses, a large number of facilities in the sample for which the question is irrelevant may produce analytic results with lower than expected statistical precision. One approach to this problem is to over-sample (i.e., draw a random sample that is slightly larger than necessary, based on what the expected drop-out rate is). This approach is not ideal, since you will probably not be able to accurately predict what percentage of responses to that question will drop out. A better approach, if relevant information is available, may be to stratify the sample to reduce the number of questions with N/A responses.

Regardless of the approach(es) you choose to deal with drop-out facilities, it is very important to first establish whether facilities are dropping out for some systematic reason that might bias the results. If you do determine that facilities are dropping out for a systematic reason, you should address the source of bias in your sample rather than simply replacing the drop-out facilities.

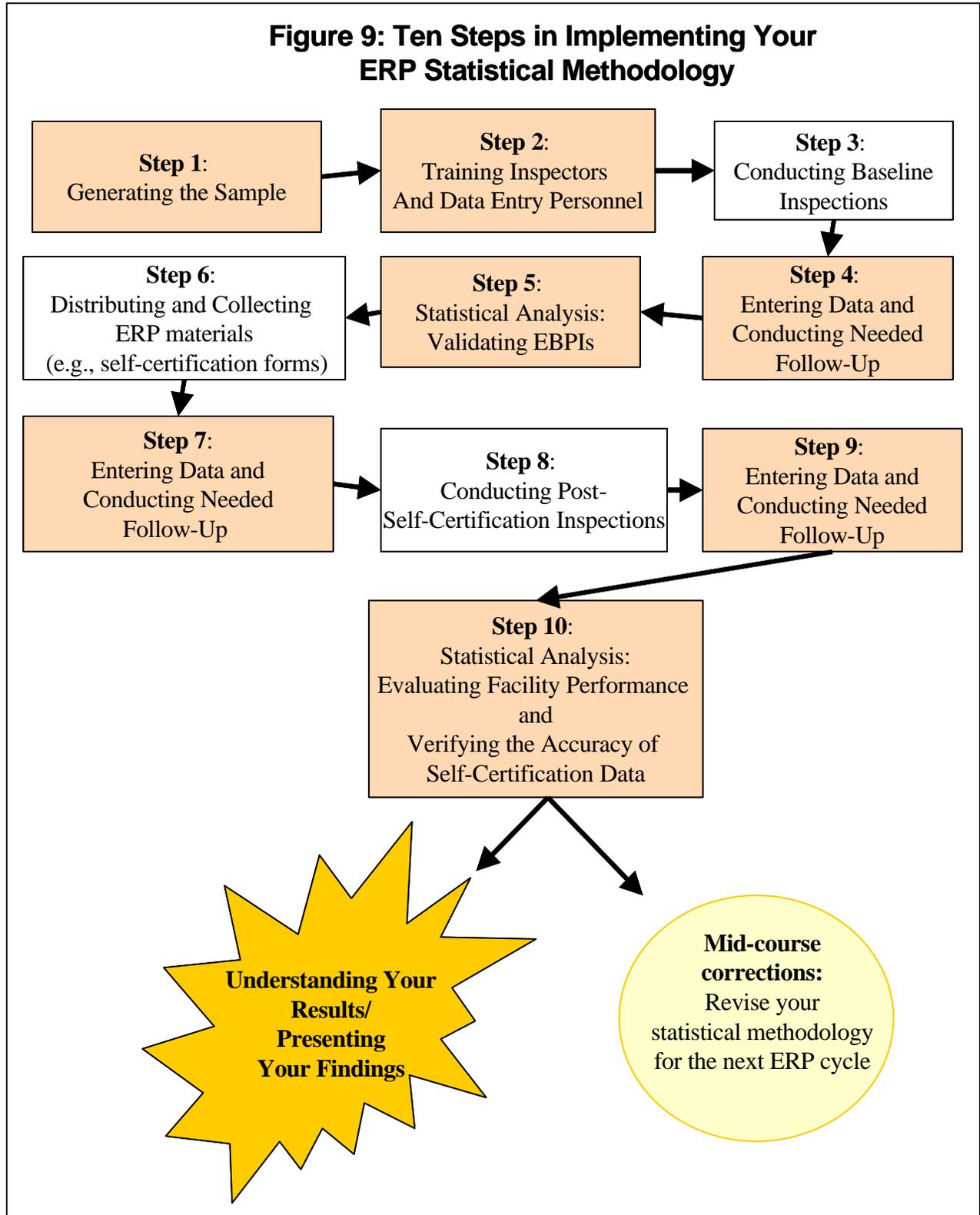
Steps in Implementing Your ERP Methodology

3. STEPS IN IMPLEMENTING YOUR ERP STATISTICAL METHODOLOGY

Once you have undertaken the seven steps to design your statistical methodology, you are ready to begin implementing your ERP methodology. There are ten steps in implementing the statistical methodology, as shown in Figure 8. (The un-shaded boxes in Figure 8 are not reviewed in this document as they are outside its scope.¹⁶) Since this is a somewhat iterative process (e.g., you will need to enter data after the baseline inspections have been completed and again after the self-certification forms have been submitted), this document describes each step only once. A key theme in many of these steps is ensuring data quality, for example, by checking for errors in data entry, ensuring that questions were correctly interpreted and truthfully answered, and validating performance measures (EPBIs). The reason that data quality is so important is that flawed data will undermine the validity of the statistical analysis.

¹⁶ The actual process of conducting inspections is not discussed here since it is beyond the scope of this document. Likewise, the process of distributing and collecting ERP materials is not described here. The Guide for Measuring Compliance Assistance Outcomes mentioned in the Introduction of this document provides very helpful guidance on distributing and collecting voluntary survey forms, much of which is relevant to the self-certification process. For more information, refer to that document's section IV. B., *Survey Implementation: The Tailored Design Method*. Most states conduct some type of compliance assistance workshop prior to or concurrent with distributing ERP materials in order to introduce compliance assistance materials and ensure that facilities understand the ERP process and how to fill out the relevant forms. Such workshops may be an important part of changing the behavior of facilities. However, a description of how to conduct compliance assistance workshops is also beyond the scope of this document.

Figure 9: Ten Steps in Implementing Your ERP Statistical Methodology



3.1. Generating the Sample

The first step of generating the sample is ensuring that reasonable efforts have been made to generate a complete, accurate and up-to-date list of all facilities in the carefully delineated population. As mentioned earlier, the list of facilities from which you draw your sample to be inspected should include *all* relevant facilities, even if they do not respond to ERP mailings. You want to be sure that facilities that do not respond are not skipped during inspections. Next, you should make decisions regarding desired confidence level, margin of error and sample stratification as described in section 2.6.

Finally, you will need to randomize your sampling population list (and subgroups if necessary) and then draw a sample of the appropriate size from it. If you are stratifying your sample (as discussed in section 2.6.2), generate separate randomized lists of facilities for each stratum. A common way to create a randomized list of facilities is to generate a random decimal number between 0 and 1 for each facility, then sort the list in increasing order of the random number. For a simple random sample of size n , just take the first n facilities in the sorted list. Random numbers can be generated in spreadsheet programs like Microsoft Excel, as well as in random number generators available on the Internet.¹⁷

If inspections are to represent the state of compliance over a long period, such as one year, then they should be spread out fairly uniformly over that period. Ideally, the order of inspections should be random and this will be the case if facilities are inspected in the order of the sorted list. Be aware that grouping facilities after a sample has been generated -- e.g., to organize a subgroup of inspections so that a particular inspector's route is efficient -- may bias the results if facilities in a subgroup have similar tendencies to be in or out of compliance at the same time. If you are concerned about the efficiency of conducting random inspections, you may wish to consider two-stage cluster sampling, which would require consultation with a qualified statistician.

3.2. Training Inspectors and Data Entry Personnel

Inspectors should be trained in order that they may properly understand the intent of the questions and how to properly and consistently complete inspector checklists, self-certification forms, and non-applicability forms. When possible, checklist and forms should be reviewed for accuracy or "problem" answers before entry into the database, and immediately flagged for follow-up to resolve the issue while fresh. Data entry personnel should also be trained so that they accurately enter data and conduct needed quality control/quality assurance procedures.

3.3. Entering Data and Conducting Needed Follow-Up

When possible, inspector checklists, self-certification forms, and non-applicability forms should be reviewed for completion and accuracy before they are entered into the database. This way, problems can be immediately flagged for follow-up. Alternatively, data may be checked once entered into the database. As mentioned earlier, automatic queries can be particularly helpful in identifying inconsistent or incomplete answers. If you discover that a form has been incompletely or incorrectly filled out, you will need to follow up with the inspectors or facilities

¹⁷ See the Research Randomizer at <http://www.randomizer.org/> or Random Number Generator Pro at <http://www.segobit.com/rng.htm>, for example.

that submitted the data in order to gather missing information or clarify the meaning of unclear answers. It is important for compliance purposes to follow up on incomplete or otherwise flawed self-certification forms. This follow-up should occur as soon as possible after the forms are submitted. Inspector checklists that are found to be incomplete or incorrect should be fixed, rather than discarded, in order to avoid biasing the sample.

The data entry process should be a two-step process to ensure data quality. When the agency is entering data originally entered on the checklist or form, one person should enter the data, and data entries should be cross-checked by another individual. Agencies should establish data entry protocols for consistently dealing with any unusual situations that occur. If facilities are using electronic submissions, the electronic submission mechanism should force the signatory to review the entry for accuracy before submission.

3.4. Conducting Statistical Analysis

There are four primary purposes in conducting a statistical analysis of ERP data: 1) to assess baseline performance, 2) to assess changes in performance, 3) to strategically target inspection resources, and 4) to increase public accountability through reporting on ERP data. In order to meet these overall goals, it is also important to verify ERP data by gauging the accuracy of the self-certification process, and validating the chosen EBPIs as indicators of broader and deeper performance. Key verifications and statistical analyses to meet these goals are introduced in the subsections below. Further information about the individual statistical tests associated with the different analyses is presented in Appendix 5.

3.4.1. Testing the Validity of Environmental Business Practice Indicators (EBPIs)

The first step that you will need to perform after baseline inspections is to validate the EBPIs. Through a careful design process, as described earlier, ERP program planners are likely to develop quite good data collection instruments. Nonetheless, it is conceivable that the EBPIs may not be as effective as desired in indicating facility compliance/performance. Therefore, it is recommended you review EBPIs once you have collected facility data to ensure the EBPIs are adequately capturing the types of performance you want to measure. Since reviewing EBPIs can be resource-intensive, it is recommended that you consider a multi-step approach. One way to do this is to ask inspectors, once they have conducted baseline inspections, whether EBPIs appear to be accurately reflecting facility performance. In addition, you may wish to set up rules-based processing systems in your response database, to ensure that responses to EBPIs are consistent with any “sub-questions” they are intended to reflect. (Keep in mind that EBPIs may be virtual indicators, rather than explicit questions on the form. Rules based processing systems can check both virtual and explicit EBPIs.)

If inspector feedback or rules-based processing checks indicate that one or more EBPIs are not adequately capturing performance, it may be worthwhile to conduct a correlation analysis to better understand how well the EBPIs in question relates to the responses on the data collection instrument. A correlation analysis can be conducted immediately after baseline inspections have been completed. Further information on setting up a correlation analysis is described at the end of Appendix 5. If resources allow, the EBPI analysis may also be repeated for the results of post-implementation inspections (and conceivably the self-certification forms), since correlation

could change over time and depending upon the data collection instrument.

If you discover that an EBPI is inadequate, you may choose to replace it with another question already on the questionnaire that demonstrates a better indicator, or to add another question already on the questionnaire to the EBPI list. It is inadvisable, unless absolutely necessary, to change the language of the EBPI or to add an entirely new question, since data collected after the change will be incomparable to data collected before the change. If you add a new EBPI or change an existing one, you will not have baseline data for this question, and so you should adjust your goals accordingly.

3.4.2. Evaluation of Facility Performance

To determine if there is any statistical difference in compliance due to the ERP approach (including inspections, self-certification, and compliance assistance), inspection data must be collected before and after the program is in place. The goal of this analysis is to detect trends in the level of overall facility environmental performance before and after ERP is introduced. An important indicator of overall performance is the proportion of facilities that are in compliance or the proportion that are out of compliance with EBPIs. However, it may also be important to analyze trends for the extent to which facilities are using “best practices” (i.e., beyond-compliance activities that reduce environmental impacts). Performance trends can be measured for different media (e.g., air, water, and hazardous wastes), individual compliance concerns, differences between districts, or even different types of facilities. (Some of these comparisons would require that fields for different facility characteristics are included in the database and that a sufficient number of facilities of each type are sampled.) Specific statistical tests for evaluating facility performance are discussed in Appendix 5.

3.4.3. Verification of the Accuracy of Self-Certification Data¹⁸

In addition to evaluating facility performance, it is also important to check the accuracy of the data provided by facilities on the self-certification forms and non-applicability forms. With regard to non-applicability forms, it is advisable to conduct random-spot check inspections for facilities that certify that they are not in ERP, in order to make sure that facilities do not use the non-applicability form as a convenient loophole to inappropriately excuse themselves from the program.

Verifying the accuracy of data submitted on the self-certification form will take place in two stages. The first stage occurs in the field, when inspectors visit facilities after ERP implementation. Inspectors may take the self-certification form for the facility they are visiting along with them on the inspection in order to immediately highlight any inconsistencies. If there are inconsistencies between self-reported certification data and conditions observed by the inspector, the inconsistencies may be due either to a change in actual compliance or performance status, or to incorrect self-certification data. It may be immediately clear to inspectors which of these scenarios is the case, but keep in mind that the longer the time lag between self-certification and subsequent inspections, the greater the likelihood that actual compliance or performance status may have changed. (Considerations for interpreting these inconsistencies are discussed in Appendix 4.) If it appears that a facility inaccurately reported its compliance status

¹⁸ This section assumes that the self-certification process is mandatory. If the self-certification process is voluntary, parts of this section can be expected to be invalid.

on the self-certification form, the inaccuracy may be the result of an accidental misunderstanding or an intentional misrepresentation of the facts. If it seems clear to the inspector that a facility intentionally submitted false information on the self-certification form, the inspector may take immediate enforcement action.

The second stage in verifying self-certification data is to review the overall consistency between self-certification forms and inspection data across all facilities in the inspection sample. This analysis is essentially an aggregation of the consistency checks performed by inspectors in the field. In other words, for each facility in the sample, and for each question, a comparison is made between the answers provided on the self-certification form and the answers provided during the inspection. This accuracy verification exercise allows the regulatory authority to understand the level of faith it may place in the reliability of self-certification data, both in terms of facilities' ability to understand self-certification materials and their willingness to report accurate compliance and performance information.

If facilities are misunderstanding questions, adjustments may be needed in the certification form, instructions, and/or compliance assistance materials. As mentioned earlier, however, agencies should be careful before revising the questionnaire, as it may create the inability to compare later data with earlier data. If many facilities misunderstood the same question, it may be appropriate to discard that question for the purpose of the statistical analysis. If however, facilities are not truthfully self-certifying, stricter enforcement or penalties may be called for.

Over the longer term, faith in self-certification validity may also allow an agency to decrease traditional permitting or enforcement resources applied toward the sector, to shift those resources to other sectors. Verification of self-certification data should be conducted regularly, however, to ensure that facilities do not increase non-compliance and data falsification in response to a decrease in traditional deterrence effect. At present, the long-term effects of the self-certification process alone are unknown.

3.4.4. Impact of Questionnaire Design on Analyses

As discussed in section 2.4 and Appendix 3, questions should be phrased in a simple, clear format that allows for only yes or no answers. This makes the tabulation of answers and any analysis much easier. If any question is answered with "N/A" or "maybe", then inspectors should ask appropriate follow-up questions that lay out the precise reasons for the answer, and provide that data on the inspector checklist. This will provide the information required by the analyst to later make a judgment on how to treat the facility or the answer. For example, an answer of "N/A" on a crucial variable of interest might signal to the analyst that the facility may not be a member of the population of interest. This can only be established if an appropriate follow-up question has been asked. Throwing out a facility or question from the analysis simply because of lack of information can seriously threaten the credibility of the analysis. As noted above, for any question being analyzed that received one or more legitimate "N/A" responses, the effective sample size is reduced by that number, in turn reducing statistical precision in inferences regarding that question.

3.5. Mid-Course Corrections

Ideally, the individuals who design the program-specific statistical methodology remain involved throughout the program implementation in order that the approach can be adapted for unexpected circumstances, which are likely to come up. In the event that staff working on the program change over time, it is important to have good documentation to keep track of program decisions and assumptions. New staff who are working on the program should be able to read the program documentation and understand how to implement the program. They should also understand when to check in with a statistician for further help. It is important to recognize that data collection and analysis under ERP will be a dynamic process. The program understandably will be refined as the agency learns more about the target population and as the target population's performance changes over time. As discussed earlier, changes to the program should be made with great care and attention to the implications for statistical validity, but some changes are likely to be necessary. It is also important to make a record of methodological decisions made, in case there is a change in statistical personnel during the program.

APPENDIX 1: CONSIDERATIONS FOR PILOT PROJECTS

You may wish to test ERP in a sector or in a smaller geographic region before rolling it out statewide or sector-wide. Such testing can make the program more manageable in the beginning and help develop staff experience with ERP. The following paragraphs describe issues agencies should consider in moving ahead with ERP pilot projects.

Use Early, Rapid Assessment to Better Understand the Problem

Since designing and implementing an ERP is resource intensive, it is important to understand early on what problems you are seeking to address through ERP. If you are not sure of compliance rates in your chosen sector, or you want to better understand why facilities are failing to improve their environmental performance, it may well be worth conducting a rapid baseline assessment for a sample of facilities before you design self-certification and compliance assistance materials. This approach could help you ensure that the sector you have targeted is worth the resources that will be required to implement ERP. Assuming you decide to go forward with ERP, an early assessment could help you better understand the characteristics of the population in order to design better compliance assistance and self-certification materials.

Targeting Larger, Statewide Populations May Be More Resource Efficient

For a given confidence interval and margin of error, the necessary sample size increases slowly as population size increases, so the sampling fraction (proportion of units sampled) has to be larger in small populations than in large populations. Therefore, achieving statistical significance for sample results can be relatively more resource-intensive relative for a small population than a large population. Consequently, be aware that, while pilot projects offer the opportunity to test new concepts and new materials on a smaller scale, ERP pilot projects will not necessarily offer the resource-efficiency that you might initially suppose.

Keep an Eye Toward the Future

You should design pilot projects with an eye toward the eventual design of the full program, to ensure that future results are statistically comparable to pilot results. Considerations may include similarities and differences between the populations targeted by the pilot program and the full program; consistency in data collected over time; and consistency in schedule of data collection.

Consider Collecting Information on Non-Environmental Facility Characteristics

Information on facility's non-environmental characteristics may help determine the scope of the full-blown program, either in terms of identifying the targeted population or stratifying the sample. For example, if you are planning a full, state-wide ERP that targets part of a sector, you may wish to design your pilot program to encompass the entire sector, so that you can better understand how to identify the facilities in your target population. A properly designed pilot program can provide information on how baseline environmental performance varies according to certain facility characteristics. This may be particularly helpful where you do not have a lot of background information on the sector. If you are considering stratifying your sample, a pilot program can help you ensure that your strata represent different and meaningful business categories.

Consider Compiling Data on the Population for the Full Program

When compiling data to identify facilities within the population for the pilot project, you may also wish to consider gathering data on all facilities likely to be in the full program. This may be a more efficient and cost-effective approach, if you are confident about the characteristics of the targeted population for the full program. For example, it may be more cost-effective to buy a statewide database for a pilot program that is targeted in a more confined geographic area, if you expect to ultimately roll-out the program statewide. The same logic may apply to purchasing a multi-sector database if you intend to target other sectors with ERP.

APPENDIX 2: EXPANDING THE POPULATION: FLORIDA DEP'S EXPERIENCE WITH IDENTIFYING NEW FACILITIES

ERP represents an opportunity for states to identify facilities that had previously been unknown to regulators, thus expanding the population of regulated facilities to encompass a much greater percentage of all facilities. However, expanding the ERP population is a complex challenge. States seek to bring as many facilities into the system as possible, while minimizing the resources necessary to do so. Many states implementing ERP for the first time struggle with determining the most effective approach for expanding the population. Florida's Department of Environmental Protection (DEP) approached the problem by combining a variety of data sources, and its experience so far with this approach offers several lessons to other states beginning ERP.

Florida DEP's Starting Point. In spring 2002, Florida DEP began to implement an ERP pilot project focused on certain types of automotive mechanical repair facilities located in the northern 35 counties of the state. When Florida DEP began planning ERP, its compliance database listed only approximately 150 mechanical repair facilities in those counties. DEP was aware that a large number of relevant facilities were likely not in the database. Moreover, DEP's database lacked sufficient detail to be able to exclude certain types of mechanical repair facilities that the pilot was not targeting -- such as dealerships, gas stations, and quick lube facilities.

Other Data Sources. Since it was clear that Florida DEP's information on facilities was incomplete, the Department investigated other data sources it could use to expand the ERP population of facilities. DEP found a database of mechanical and collision repair facilities compiled by the Florida Department of Agriculture and Consumer Services (DACS), based upon state consumer protection requirements that all automotive repair facilities register with the state. The DACS data set promised to significantly expand DEP's known population. However, discussions with DACS personnel indicated limited confidence in the extent to which the database was up-to-date and provided accurate details about facilities. Therefore, DEP elected to supplement the DACS data with a database of facilities from Dun & Bradstreet. Dun & Bradstreet is a private data provider whose records are compiled from various public and private sources, including Dun & Bradstreet proprietary business credit records. Florida DEP's decision was based in part upon the success that Massachusetts DEP has reported in using both Dun & Bradstreet and InfoUSA to build its population of facilities for other sectors. Florida DEP purchased Dun & Bradstreet data using EPA funds at a cost of 14 cents per record. In addition, because both the DACS database and the Dun & Bradstreet database cover only private businesses, Florida DEP collected information from counties to identify relevant government facilities.

Targeting Relevant Facilities. Once Florida DEP combined all of the datasets, it needed to exclude those auto repair facilities that did not meet its ERP criteria. Specifically, DEP sought to include only those auto repair facilities that conducted certain types of mechanical repairs, did not have a paint spray booth, and were not part of a gas station or car dealership. In order to identify such facilities in the database, DEP's contractor ran queries to exclude facilities that did not meet the ERP requirements. The set of tasks associated with combining all of the datasets, eliminating duplicate records, and excluding facilities that did not meet DEP's ERP criteria

required significant resources (approximately 60 staff-person hours).

Success. Florida DEP's use of these three different databases expanded its population of ERP facilities to approximately 1500 facilities, at least 1000% greater than the number of such facilities previously known to DEP.¹⁹ Furthermore, 17% of these approximately 1500 facilities would not have been identified if DEP had not purchased the Dun & Bradstreet database. Interestingly, however, the acquisition of the DACS database was even more important: 29% of facilities would not have been identified had the DACS database not been included.

Challenges. Expanding the population required that Florida DEP overcome several challenges. First, using more than one data set for identifying private businesses required elimination of duplicate records. A large number of facilities had records in both the Dun & Bradstreet database and the DACS database. By using electronic queries to match records with similar names, addresses, and/or phone numbers, DEP's contractor was able to identify most of the duplicates with limited visual inspection to confirm questionable duplicates. However, because of inconsistencies between the databases, inspectors still found that approximately 3% of facilities visited in the first round of inspections had duplicate records that had not previously been identified.

In addition to the challenge of eliminating duplicate records, eliminating facilities that were closed or were not included in the ERP population proved difficult. Florida DEP found that approximately 6% of visited facilities were closed. Another 34% of visited facilities were not included in the pilot project because they did not conduct the kinds of repairs targeted by the pilot, and/or because they were involved in business activities excluded from the ERP pilot (e.g., automotive sales). Although an effort had been made to exclude these non-targeted facilities prior to inspections by querying the dataset, the databases did not provide sufficient information needed to exclude facilities in all cases. For example, DEP wanted to exclude all facilities that did not conduct repair or maintenance on light truck or automobile engines, engine cooling systems, fuel delivery systems, brakes, or transmissions. However, the databases did not contain information on the exact types of repairs facilities conducted. Therefore, it was necessary to query the database for other types of information that would suggest whether the facility conducted the targeted types of repairs, such as if the facility had a SIC code that includes general automotive engine or transmission repair. Despite the lack of data needed to comprehensively screen out non-targeted facilities, queries on available data were able to remove thousands of non-targeted facilities from the inspection rolls. Where it was uncertain as to whether a facility should be included from the information in the database, DEP chose to err on the side of visiting such facilities in order not to improperly exclude facilities.

Lessons learned. Several lessons can be taken from Florida DEP's experience in expanding their population of ERP facilities. First, in DEP's experience, the most common reason for mistakenly including auto repair facilities in its ERP dataset was not a failure of combining different datasets, but rather a failure to exclude facilities on the basis of particular activities and services they provided. For example, it was far easier to identify facilities engaged in general types of auto repairs than to identify facilities engaged only in the repair of engines, brakes, engine cooling systems, fuel delivery systems, and transmissions. A significant percentage of

¹⁹ This figure is an estimate based upon findings during baseline inspections. A precise number of relevant facilities will not be known until after the self-certification round is complete.

the facilities in the combined dataset could not be screened out prior to inspections. This experience suggests that trying to exclude facilities on the basis of criteria that are not included in the database introduces substantial uncertainty into the process of defining the ERP population of facilities.

Despite the difficulties of excluding facilities that meet detailed ERP requirements, Florida DEP's effort to expand the population of facilities was successful in that the datasets were combined with minimal duplicate records. Moreover, both commercial and state database records appeared to be fairly up-to-date. Thus, even in a sector with reportedly high turnover rates, only 6% of facilities in the database turned out to be closed when an inspector came to visit. Furthermore, only one facility that appeared in both the DACS and Dun & Bradstreet databases turned out to be closed. This fact highlights the value of combining multiple databases in increasing the accuracy of facility identification. Facilities listed in multiple databases were more effectively screened for all exclusion criteria than facilities listed in only one database.²⁰

Florida DEP's experience that many facilities were screened out at the time of inspection suggests that it may be worthwhile for ERP inspectors to conduct additional pre-screening prior to ERP implementation. For example, states could call facilities to ensure that they meet the ERP criteria before conducting inspections. However, such an approach would require states to develop a reliable phone protocol for including and excluding facilities. In some cases (as with DEP) inspectors may not be confident that they can accurately screen facilities over the phone. In addition to pre-screening facilities prior to inspection, states may want to take additional steps to exclude facilities before sending out ERP certification materials and compliance assistance workbooks. For example, states could send out postcards to ascertain whether a facility should be included in ERP prior to sending out complete ERP certification packages.

Despite the challenges in combining multiple datasets and excluding facilities that did not meet ERP criteria, Florida DEP found that it was worth the effort to expand its ERP population of facilities by adding additional facility databases. Each database DEP added provided added a substantial number of facilities, and as a result DEP expanded the number of facilities included in its ERP program by an order of magnitude.

²⁰ It is difficult to compare the accuracy of the different database sources, because facilities that were excluded were not inspected to ensure that they should be excluded. Moreover, each database included different types of information, and as a result different exclusion criteria were applied to those facilities listed in different databases.

APPENDIX 3: REVIEW LIST FOR STATISTICALLY SOUND DATA COLLECTION

This review list is designed to help you create statistically sound data collection documents for ERP that allow data to be easily and accurately collected, interpreted, and analyzed. In order to ensure that valid conclusions can be drawn from ERP data, data collection instruments should (a) minimize respondent confusion and (b) facilitate accurate interpretation of data in the analysis stage. This review list is designed for you to use as you create ERP data collection instruments, including inspector checklist and self-certification forms.

As a guiding principle, design questions so that as many respondents can easily and precisely answer them as possible. Ensure that questions are clear, concise, and unambiguous, and write questions with their intended audience and use in mind. Format your questions to encourage consistent answers so you can easily analyze and interpret the results. Field-testing the questions with members of the intended audience can alert you to potential problems in question design. Providing both the inspectors and the regulated community with training prior to data collection can also help the users of the data collection instruments better understand the questions and therefore more reliably provide appropriate responses.

In order to evaluate whether your data collection tools meet these overall data collection principles, answer the following set of questions about your ERP materials. While questions conforming with some of these principles may increase the length and complexity of the survey, they will also increase the usefulness of the data collected. The items below are separated into two sections, those regarding specific questions and the overall questionnaire. The first section, divided into general, qualitative and quantitative questions, discusses individual question language and format. The second presents principles for the overall questionnaire design. Examples are provided, in gray, where necessary to demonstrate specific points.

QUESTION FORMAT

General Question Format

- 1. Technical language:** Do questions use language that will be easily understood by the audience?

Questions on the inspector checklist may appropriately use jargon or technical terms that the inspectors themselves are familiar with. However, this language might not be appropriate for self-certification forms. If technical terms are used, provide respondents with definitions for terms with which they may be unfamiliar. Ideally, definitions should be provided within the text of the question.

- 2. Ambiguous language:** Do the questions avoid or explain ambiguous or vague terms?

Words that are vague (e.g., many, few, very, rarely, often, adequate, appropriate, sufficient, etc.) allow the respondent to interpret them differently. Analysts will then be unclear about how to interpret the results of these questions. If ambiguous terms are used, clear definitions should be included, as in the sample question below.

<input checked="" type="checkbox"/> Are the container inspection records complete?	
Specifically, do inspection records include:	
i. The date?	<input type="checkbox"/> Yes <input type="checkbox"/> No
ii. The time?	<input type="checkbox"/> Yes <input type="checkbox"/> No
iii. Legibly written name of the inspector?	<input type="checkbox"/> Yes <input type="checkbox"/> No
iv. Number of containers?	<input type="checkbox"/> Yes <input type="checkbox"/> No
v. Condition of the containers?	<input type="checkbox"/> Yes <input type="checkbox"/> No
vi. Notes of observations made?	<input type="checkbox"/> Yes <input type="checkbox"/> No
vii. Date and nature of repairs or/and corrective actions?	<input type="checkbox"/> Yes <input type="checkbox"/> No
If you checked "yes" for each box above (questions i – vii), then check the "Yes" box to the right. Otherwise, check the "No" box to the right.	
	<input type="checkbox"/> Yes <input type="checkbox"/> No

3. Complete sentences: Are questions phrased as complete sentences?

Questions phrased as complete sentences can be easier for the respondent to understand and interpret.

<input checked="" type="checkbox"/> <i>Incomplete:</i>	Amount of yearly hazardous waste discharge to a publicly owned treatment works (POTW):	_____ lbs/month
<input checked="" type="checkbox"/> <i>Complete:</i>	How much hazardous waste did the facility discharge to a publicly owned treatment works (POTW) in the last year?	_____ lbs/month

4. Double negatives: Do questions avoid using double negatives?

Double negatives make the questions more difficult for respondents to understand, increasing the chance of error.

<input checked="" type="checkbox"/> <i>Unclear:</i>	Does the facility avoid using unapproved corrosion protection systems?	<input type="checkbox"/> Yes <input type="checkbox"/> No
<input checked="" type="checkbox"/> <i>Explicit:</i>	Does the facility use only approved corrosion protection systems?	<input type="checkbox"/> Yes <input type="checkbox"/> No

5. Scope/Timeframe: Are the scope and timeframe of questions clear?

Questions should provide enough information so that respondents are clear about both their scope (e.g., whole facility or specific process) and timeframe (e.g., activities over the last year or month, or only to present activities). While this information is sometimes implicit in questions, sometimes it may need to be directly stated. For implicit questions, include discussion of the scope in the directions (e.g., "Unless otherwise stated, questions refer to current operation on the day of the certification."). In either case, the questions should be phrased so that respondents are clear about how to answer, and analysts are clear about how to interpret answers. For instance, in the "unclear" example below, if a facility answered "yes", it would be unclear how long ago the facility may have discharged hazardous waste to a POTW. This can be particularly problematic after the first instance of data collection. Similarly, it is unclear whether a "no" means that a facility *never* discharged hazardous waste to the POTW or that it is current not discharging hazardous waste to the POTW.

<input checked="" type="checkbox"/> <i>Unclear:</i>		
Does the facility discharge hazardous waste to a publicly owned treatment works (POTW)? If No, skip to Question XXX	<input type="checkbox"/> Yes	<input type="checkbox"/> No
<input checked="" type="checkbox"/> <i>Implicit: [form directions instruct respondents to answer all questions based on activities in the past twelve months]</i>		
Does the facility discharge hazardous waste to a publicly owned treatment works (POTW)? If No, skip to Question XXX	<input type="checkbox"/> Yes	<input type="checkbox"/> No
<input checked="" type="checkbox"/> <i>Explicit [the most desirable approach]:</i>		
Did the facility discharge hazardous waste to a publicly owned treatment works (POTW) in the last twelve months? If No, skip to Question XXX	<input type="checkbox"/> Yes	<input type="checkbox"/> No

6. Exclusivity: Are answer categories mutually exclusive?

When a question provides a selection of answer choices from which the respondent is directed to choose, the choices must be mutually exclusive. If choices overlap, respondents may be unsure how to answer the question, and analysts how to interpret the answers. In the following example, if a facility engaged in engine sales and brakes repair, respondents would check all four boxes and analysts would not be able to differentiate between the activities.

Does the facility engage in any of the following activities: [mark yes to all that apply]					
<input checked="" type="checkbox"/> <i>Unclear</i>		<input checked="" type="checkbox"/> <i>Clear</i>			
Sales	<input type="checkbox"/> Yes	<input type="checkbox"/> No	Engine Sales	<input type="checkbox"/> Yes	<input type="checkbox"/> No
Repair	<input type="checkbox"/> Yes	<input type="checkbox"/> No	Engine Repair	<input type="checkbox"/> Yes	<input type="checkbox"/> No
Engine	<input type="checkbox"/> Yes	<input type="checkbox"/> No	Brakes Sales	<input type="checkbox"/> Yes	<input type="checkbox"/> No
Brakes	<input type="checkbox"/> Yes	<input type="checkbox"/> No	Brakes Repair	<input type="checkbox"/> Yes	<input type="checkbox"/> No

7. Consistency: Are questions consistent between ERP documents?

Questions on the ERP inspector checklist and self-certification form should request similar information from respondents, so that results from one questionnaire can be compared to another. While the self-certification form may only cover a sub-set of questions included in the inspector checklist, those questions that are in both documents should be phrased carefully so that they are asking for the same data and will generate consistent responses. The workbook can be useful in explaining self-certification questions to respondents, so that the self-certification responses are consistent with inspector checklist responses. Question language in the different documents may need to be tailored for various audiences.

Qualitative Questions

8. Closed-ended: Are questions closed-ended, with a defined set of potential answers from which respondents should choose?

Closed-ended questions can only be answered with a response from a defined set of possible answers, such as provided answer choices or yes/no. Open-ended questions do not provide respondents with options. Open-ended questions lead to analytical challenges, as the answers provided by various respondents will not always be comparable and are unlikely to be easily sorted or assessed. In most cases, open-ended questions should therefore be avoided if possible, or phrased as specifically as possible. Yes/no questions are preferable, even if it requires several other yes/no questions to achieve the desired answer. In the following example question, a series of yes/no questions should yield more consistent responses than the open-ended question. (Note: while recognizing the analytical limitations of open-ended questions, they can be desirable if responses cannot be anticipated, or if questionnaire designers wish to give respondents more flexibility. For instance, occasional open-ended questions can make forms feel more "user-friendly.")

<input checked="" type="checkbox"/> <i>Open-ended:</i>	
How does the facility dispose of hazardous waste? (Describe all methods used)	_____
<input checked="" type="checkbox"/> <i>Closed-ended:</i>	
Does the facility dispose of hazardous waste on site in any of the following ways:	
Septic tank?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Storm drain?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Surface water?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Ground? (soil, concrete, asphalt, other)	<input type="checkbox"/> Yes <input type="checkbox"/> No
Burning?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Dumpster?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Evaporation?	<input type="checkbox"/> Yes <input type="checkbox"/> No
If you checked "yes" for each box above, then check the "Yes" box to the right. Otherwise, check the "No" box to the right.	<input type="checkbox"/> Yes <input type="checkbox"/> No

- 9. Suggested answers:** If questions provide possible answers, are all anticipated answers listed?

To facilitate analysis, options should include all anticipated answers, as in the following example question. If it is not possible to anticipate all responses, provide a space for the respondent to include other answers

(e.g., other: _____).

<input checked="" type="checkbox"/> Does the facility have an oil/water separator? If No, skip to Question XXX	<input type="checkbox"/> Yes	<input type="checkbox"/> No
If Yes, does the oil/water separator overflow discharge to the:		
POTW?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
septic system?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
surface water?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
Other?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
If "other" is marked Yes, please describe where discharges are sent: _____		

- 10. Routing questions:** Are "routing questions" used to determine relevancy, rather than beginning questions with "If X is true..." or "When you perform X...?"

Routing questions determine whether the following question(s) is/are relevant to a particular respondent, and should be used instead "if..." or "when..." questions. These latter types of questions can confuse respondents when X is *not* true or is *not* performed. Analysis is likewise more difficult. Instead, the question should be broken into two separate yes/no questions, so that the respondent can skip the second question if the answer to the first one is no. The example question below demonstrates how the routing question can be used.

Using routing questions avoids the problem of having "n/a" as a possible response. While allowing "n/a" as an answer can simplify or streamline the data collection instrument, this approach may decrease the usefulness of the data collected. Allowing respondents to answer "n/a" introduces uncertainty into the analysis since it may not be clear when a question is not applicable. In particular, it allows respondents to independently determine whether the question is applicable to them or not, and respondents may use different decision-making criteria than the agency anticipates. Instead of including "n/a," routing questions can be used to describe the applicability of the question. If it is necessary to have "n/a" as an answer choice, respondents should be directed to explain or mark why the question is not applicable to them and/or explicitly given instructions on when they are allowed to mark "n/a."

<input checked="" type="checkbox"/> <i>"If" question:</i>		
If the facility discharges hazardous waste to a publicly owned treatment works (POTW), does it discharge more than 15 kg (33 lbs. or 4 gallons) per month?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
<input checked="" type="checkbox"/> <i>Routing question:</i>		
Does the facility discharge hazardous waste to a publicly owned treatment works (POTW)? If No, skip to Question XXX	<input type="checkbox"/> Yes	<input type="checkbox"/> No
If Yes, does the facility discharge more than 15 kg, (33 lbs. or 4 gallons) per month of hazardous waste to the POTW?	<input type="checkbox"/> Yes	<input type="checkbox"/> No

<p><input checked="" type="checkbox"/> <i>Clear directions for a N/A question :</i></p> <p>Does your shop store hazardous waste containers on surfaces designed to prevent spills?</p> <p><i>[Note: This question could be improved through use of a routing question rather than providing an N/A response. However, questions with N/A responses may be used in order to shorten the length of data collection instruments.]</i></p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No</p> <p><input type="checkbox"/> N/A (Check this box only if your shop does not generate or store any hazardous waste)</p>
--	---

11. Single question: Does each question ask only one question?

Asking multiple questions can lead to both respondent and analytical confusion. For example, if the response to “Did X and Y happen?” is no, the analyst will not know when X, Y, or neither occurred. Similarly, if the answer to “Did X or Y happen?” is yes, the analyst will not know if X, Y, or both occurred. In addition, the respondent may be confused about how to answer such double questions. For example, if asked "Did X or Y happen?" the respondent may be confused on what to answer if both X and Y happened.

<p><input checked="" type="checkbox"/> <i>Double question:</i></p> <p>Are all permits up to date and stored on site?</p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No</p>
<p><input checked="" type="checkbox"/> <i>Single questions:</i></p> <p>Are all permits up to date?</p> <p>Are all permits stored on site?</p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No</p> <p><input type="checkbox"/> Yes <input type="checkbox"/> No</p>

Quantitative Question Format

12. Units: Do quantitative questions specify units of volume or mass and time (e.g. lbs/month)?

Questions should suggest units to ensure that analysts can properly interpret the data. Questions may specify units (so that respondents must convert all answers to the given units) or may provide a selection of units from which to chose. When questions provide a selection of units, respondents may be better able to report accurate data, as they collected it. However, this may make it difficult for analysts to compare and analyze results from all respondents, especially where accurate conversion factors are not available. If respondents are asked to use specified units, all relevant conversion factors should be provided.

13. Answer range: Did you consider providing a range of answers?

Providing respondents with a range of answers from which to select can make it easier for respondents to answer the question, especially where data are needed to provide an exact value is difficult to collect or estimate and where provision of data is voluntary. However, answer ranges might not be appropriate for all data needs. For example, data from answer ranges cannot be averaged or normalized, and may not provide sufficient information for certain regulatory determinations. (See Item #15 for more information on normalizing quantitative data.) If you choose to provide answer ranges, follow general principles of answer range construction. For instance, ensure that answer ranges can accommodate likely answers from all respondents, and ensure that answer ranges are mutually exclusive. For more information, see American Statistical Association, “Designing a Questionnaire.”

How much hazardous waste (in lbs/month) does the facility discharge to a publicly owned treatment works (POTW) per month, averaged over the last six months?	<input checked="" type="checkbox"/> <i>Improper</i> <i>(because of overlapping ranges)</i>	<input checked="" type="checkbox"/> <i>Proper</i> <i>(ranges are mutually exclusive)</i>
	? 1-50	? 1-50
	? 50-100	? 51-100
	? 100-200	? 101-200

QUESTIONNAIRE CONTENT AND FORMAT

14. Exclusion rules: Does the introduction to the questionnaire include rules for excluding non-relevant facilities?

If it is possible that the data instrument will not be appropriate for all entities that receive it, exclusion rules should explain which facilities are subject to data collection. Exclusion rules should be clearly presented, e.g., in a separate section, so that respondents can easily determine if it is appropriate for them to participate in the data collection process, and so that the agency can reliably spot-check facilities that exclude themselves. An example of a question from such a section, which should be placed in the beginning of the data-gathering instrument, is shown below.

<input checked="" type="checkbox"/> Does the facility engage in the repair, maintenance, or modification of the following light truck and/or automobile components/systems?	
Engines	<input type="checkbox"/> Yes <input type="checkbox"/> No
Engine Cooling Systems	<input type="checkbox"/> Yes <input type="checkbox"/> No
Fuel Delivery Systems	<input type="checkbox"/> Yes <input type="checkbox"/> No
If you checked “yes” for ANY box above, then check the “Yes” box to the right and INCLUDE the facility.	
<input type="checkbox"/> Yes <input type="checkbox"/> No	
If you checked “no” for ALL boxes above, then check the “No” box to the right and EXCLUDE the facility.	
<input type="checkbox"/> Yes <input type="checkbox"/> No	

- 15. Normalization:** Does the questionnaire ask, if possible, for data to normalize quantitative data to production?

Normalization data allows for comparative analysis. For example, if facilities provide information on production per year and hazardous waste generated per year, then the amount of hazardous waste can be normalized per unit of production. Normalization allows for trend and efficiency analysis, as well as comparisons between facilities of different sizes and levels of production. For example, analysts could compare pounds of hazardous waste generated per automobile serviced at large and small auto repair shops. Furthermore, normalization data can allow one to understand the relationship between changes in quantitative measures over time and changes in production.

- 16. Internal verification ("red flags"):** Does the questionnaire include internal verification mechanisms, sometimes referred to as "red flags"?

Verification mechanisms in the data collection instrument can highlight potential inaccuracies in the responses, whether intentional or caused by errors or misunderstandings. For example, if system components X and Y are incompatible with each other, the checklist could ask, separately, whether X and Y are present. If respondents indicate that both X and Y are present, the discrepancy can serve as a "red flag" for regulators to follow up on. It is probably not possible to have internal verification for all or even most questions in a data collection form without greatly extending the length of the form, however a few verification questions on a few key indicators would be very helpful in ensuring data reliability.

- 17. Question order:** Is the question order logical and intuitive?

The questionnaire design should make as much intuitive sense as possible. Questions that are related should be located near each other, and respondents should be able to move through the questionnaire in a linear fashion. For example, an inspector checklist should be presented in order in which an inspector would tour the facility, or in which regulatory determinations are made; self-certification forms might be best organized into facility processes. Questions whose answers depend on each other can be "nested" as demonstrated in the example question below. This technique allows respondents to see the major questions topics and relationships between sub-questions more clearly.

<input checked="" type="checkbox"/> Is the facility mixing hazardous waste with used oil? If Yes, continue; If No, skip to Question XXX	<input type="checkbox"/> Yes <input type="checkbox"/> No
Is the hazardous waste listed (is it specifically identified as a hazardous waste in regulations)? If Yes, skip to Question XXX; If No, continue.	<input type="checkbox"/> Yes <input type="checkbox"/> No
Is the waste a characteristic hazardous waste (i.e., is it regulated as a hazardous waste because of its characteristics, namely, ignitability, corrosivity, reactivity, or toxicity)? If Yes, continue; If No, skip to Question XXX.	<input type="checkbox"/> Yes <input type="checkbox"/> No
Describe the characteristic(s) of the waste that make it hazardous _____	

- 18. Facility information:** Does the questionnaire ask for all relevant facility information (e.g. address, contact information, type of facility)?

- 19. Inter-state comparison:** If inter-state data comparisons are desirable, are questions the same as other states' ERP questions?

Sources:

American Statistical Association, section on Survey Research Methods, ASA Series: What is a Survey? "More About Mail Surveys" and "Designing a Questionnaire," 1997. Available at: <http://www.amstat.org/sections/srms/whatsurvey.html>.

Don Dillman, Mail and Internet Surveys. 2000.

Florida Department of Environmental Protection, "2002 Compliance Assistance Pilot Project Inspector Checklist," 2002.

US Environmental Protection Agency, "Guide for Measuring Compliance Assistance Outcomes," Revised June 2002.

APPENDIX 4: VERIFYING ACCURACY OF SELF-CERTIFICATION DATA

As noted in the main text of this document (section 3.4.3), an important object of analytical interest in ERP is to regularly verify the degree of accuracy of the data provided by facilities on the self-certification forms. The verification exercise compares post-self-certification inspection data to the self-certification data provided by those same inspected facilities. For each facility in the sample, and for each question, a comparison will be made between the answers provided on the self-certification form and the answers provided during the inspection. The purpose of the comparison will be to see how what the inspector reports in the field matches up with what a facility self-reported, and to make inferences about the validity of self-certification data based upon these comparisons.

Four Types of Observations

For example, let us examine a comparison of the inspector and facility responses regarding a yes/no question. There are four possible observations in making this comparison, as indicated in Table 1. In Table 1, a "yes" response indicates compliance.

Table 1

Observation Type	Facility Self-Certification Response	Inspector Response
1	Y	Y
2	N	N
3	N	Y
4	Y	N

As indicated in the following text, interpreting these different observations depends in large part upon how far apart in time the self-certification and the inspection occurred.²¹

Observation #1: Facility and Inspector Declare Compliance

- *Contemporaneous:* If the inspection occurred right after the self-certification, it is likely that the facility accurately reported being in compliance. High incidence of this observation would obviously provide assurance in the usefulness of self-certification data as a predictor for facility performance.
- *Non-contemporaneous:* As the delay between the inspections and the self-certification increases, an agency's confidence in the facility's original accuracy is reduced. That is, it is possible that the true state of compliance when the facility self-certified was actually negative, but changed before inspector arrival. For example, perhaps the facility recognized the non-compliance, decided not to report the non-compliance to the state, and subsequently corrected the violation before inspector arrival (presumably as would a facility who had reported the non-compliance). While such a situation would indicate a

²¹ This section assumes that data collected by inspectors will be highly accurate. Without such an assumption, this comparison becomes very difficult. This assumption may be difficult to make in cases in which inspectors have not been well-trained, the compliance issue is very new or complex, and/or the compliance determination requires a great deal of discretion on the part of inspectors/facilities.

reduced usefulness of self-certification data as an immediate predictor of current facility performance, it does not indicate a failure of the self-certification process to ensure facilities improve their performance.

Observation #2: Facility and Inspector Declare Non-Compliance

- *Contemporaneous:* It is likely that the facility accurately reported being out of compliance. Again, this observation provides assurance in the usefulness of self-certification data as a predictor for facility performance
- *Non-contemporaneous:* In addition to the previous interpretation, it is possible that the facility was actually in compliance at the time of self-certification, but inaccurately reported its status as non-compliant, then fell out of compliance. Presuming that a facility would not willfully mis-identify itself as non-compliant, this would indicate poor design of self-certification materials. It is likely, however, that, even if an inspection were not to occur, the agency would identify this problem because of the requirement that a facility identifying itself as non-compliant complete and implement a return-to-compliance plan.

Observation #3: Facility Reports Non-Compliance and Inspector Reports Compliance

- *Contemporaneous:* It is likely that the facility inaccurately reported itself as being out of compliance. Presuming that a facility would not willfully mis-identify itself as non-compliant, this would indicate poor design of self-certification materials. Even if the inspection had not occurred, the agency would be likely to discover this problem through the return-to-compliance process. Therefore, this instance does not create significant concerns in terms of placing faith in the self-certification process.
- *Non-contemporaneous:* In addition to the previous interpretation, it is possible that the facility correctly reported itself as being out of compliance, and has corrected the non-compliance before inspector arrival. This would be corroborated by information from the return-to-compliance plan and an interview with facility staff. Again, this instance does not create significant concerns in terms of placing faith in the self-certification process.

Observation #4: Facility Reports Compliance and Inspector Reports Non-Compliance

- *Contemporaneous:* It is likely that the facility inaccurately reported itself as being in compliance. Whether such inaccuracy resulted from conscious falsification or confusion about the question, this situation creates concerns about the reliability of self-certification data as an independent measure/predictor of facility performance.
- *Non-contemporaneous:* In addition to the previous interpretation, it is also possible that the facility accurately reported that it was in compliance at the time of self-certification, and fell out of compliance before the inspection occurred. This situation would also create serious concerns about the self-certification process, because facilities are typically required under ERP to certify that they have put in place systems to ensure that they maintain compliance and good performance in the future.

As the text above suggests, Observation #4, where a facility reports compliance but an inspector observes non-compliance, indicates the greatest concern regarding the validity of self-certification data (either due to misunderstandings or misrepresentations on the part of facilities, or failure of facilities to maintain compliance over time). If you frequently encounter observations of this type, you should consider stepping up enforcement actions and/or improving compliance assistance in order to ensure that facilities accurately self-certify and stay in compliance. While Observation #3 represents an inconsistency between what a facility reported and what an inspector found, the risk of this inconsistency is not great since follow-up with facilities will occur on the basis of Return-to-Compliance Plans. Even though Observation #1 represents consistent reports from facilities and inspectors, if there is a significant lag between self-certification and inspections, there is a risk that some facilities may have only come into compliance after submitting self-certification forms. In this case, you may not have full confidence in self-certification data, but this result does suggest that ERP as a whole is having its intended effect in encouraging facilities to come into compliance.

Impact of the Timing of Follow-up Inspections

As the discussion above makes clear, the timing of follow-up inspections can have a significant impact on the ability to interpret these observations. The sooner inspections occur after self-certification, the easier it will be to interpret the true state of facility performance at the time of self-certification. However, there is a trade-off: the sooner inspections occur, the less likely the inspection results are to be able to capture the full impact of the ERP -- because facilities who were out of compliance will not have time to return to compliance. Especially in the first year, failing to capture return-to-compliance results could significantly understate any positive impact. Furthermore, if an agency conducted inspections immediately after facilities self-certified, an agency would likely want to conduct another round of inspections within a couple of months in order to determine the extent to which facilities returned to compliance. As such, a delay in inspection time is more resource-efficient as well.

APPENDIX 5: COMMON ERP STATISTICAL CALCULATIONS AND TESTS

Introduction

This technical appendix is intended to give ERP planners a better understanding of the more common statistical calculations and tests that might be applied during ERP planning and in the analysis of ERP data. Following this introduction, this appendix provides instructions for computing a confidence interval, followed by a brief discussion of the fundamentals of hypothesis testing. This is followed by discussion of methods used to determine the minimum sample-size needed for conducting several common statistical tests used in survey analysis. The appendix concludes with a discussion of correlation analysis, as this can be an extremely useful technique for evaluating the utility of EBPI questions as performance indicators. A companion Excel file has been prepared to assist readers who are implementing the statistical analyses described in this Appendix. The file is available online at <http://www.epa.gov/permits>. **It is very important that readers first review and understand this Appendix before using the Excel file to determine a sample size or conduct hypothesis tests, since this Appendix contains background information that is essential for correctly using the Excel file.**

Because of the complexity of this subject, discussion and computational details are only provided for sampling designs based on simple random sampling. Although designs employing various means of stratifying or dividing populations into discrete sub-populations or strata may be useful in some cases, readers seeking additional details on advanced designs, or more complex analytical methods, are encouraged to consult the list of resources provided in Appendix 7. Also note that it is beyond the scope of this appendix to treat the more detailed concepts involved in hypothesis testing, beyond that which can be provided in a simple introduction. Please refer to the references listed in Appendix 7 for a more complete discussion of how to conduct hypothesis tests.

Please note that the discussion and examples provided in this appendix focus on tests based on proportions, as these are probably the most common types of analyses encountered when working with survey data. Tests based on other quantities, such as average or mean responses for a particular parameter (e.g., volume of waste generated, number of underground or aboveground chemical storage tanks for a population of facilities, etc.), may be appropriate in some situations, but the options for dealing with this type of data are more varied, and cannot be adequately addressed in a document that focuses on generic methodologies. Population proportions can be used to describe dichotomous or binary data (e.g., responses to questions that can be answered with a “Yes” or “No”). There are several options for calculating estimates of proportions in specified target populations. One commonly employed group of methods for dichotomous or binary data uses normal approximations to the binomial distribution (for large or infinite populations) or hypergeometric distributions (for small populations). This appendix describes methods based on the normal approximation to these two distributions.²² Appendix 6 discusses alternative approaches based on the analysis of contingency tables.

²² Throughout this appendix it is assumed that sample sizes are at least 30, such that the central limit theorem applies and formulas based on the normal distribution can be used.

Throughout this document, recommendations are made for consulting a qualified statistician during the planning and execution of ERPs. Readers are cautioned that while a growing list of documents and tools for designing surveys are accessible to non-technical users, considerable experience with survey methods is required to assure that data of sufficient quality are gathered and correctly analyzed and interpreted. Examples of situations in which consultation with a qualified statistician is strongly recommended include:

Dealing with Small Samples. The general test results described in this appendix use the theory of the “normal probability distribution” which requires sufficiently large sample sizes for its validity. A commonly used rule of thumb for “sufficiently large” is 30. If sample size is less than 30, so-called “exact” or “resampling” methods (e.g., Monte Carlo analysis, bootstrapping, etc.) of statistical inference must be used. These methods are beyond the scope of this appendix and a qualified statistician should be consulted in such cases. A number of commercial statistical software vendors offer packages that are tailored specifically for dealing with small sample-sizes. However, because of the added complexities of analyzing small samples, you should avoid this if at all possible. For example, you may be better off taking a census of the population, rather than using a small sample.

Treatment of Bias Using Special Calculations (e.g., Weighting). It is not uncommon to encounter situations where it may be desirable to apply “weights” or “weighting factors” to samples to adjust for various types of bias. For example, weighting may be used to adjust for non-responses or disproportionate responses in surveys. If some facilities do not respond to a survey questionnaire or some data cannot be collected, the resulting sample may no longer adequately reflect the true distribution of facilities according to some important variables. For example, if 20% of facilities in the sector are in Region X but only 10% of the sample data collected is from X, bias may result. A statistician may wish to apply weighting factors to the actual data collected in order to adjust for this bias.

Computing Confidence Intervals

As mentioned in Section 2.6.2, some of the statistical analysis in ERP, e.g., analyzing baseline data, can be accomplished by computing confidence intervals. For example, suppose you are interested in estimating the proportion of facilities that give hazardous waste training to their employees with a specified level of confidence and margin of error. Once you know the sample proportion, you can state with a specified confidence level (e.g., 90 or 95%) that the “true” proportion of the population that supplies hazardous waste training to employees is within the margin of error (e.g., +/-5%). The formula to determine the sample size for a single sample is:

$$n_o = \frac{.25(Z_{\alpha/2})^2}{d^2}, \text{ where}$$

n_o = estimate of the minimum number of samples required, assuming a large or infinite population size

δ = the margin of error or difference between the true proportion and the estimated proportion based on a sample of the population

$Z_{\alpha/2}$ = the standard normal score from statistical tables, based on a specified *significance level*, α . Specifically, the probability that the standard normal distribution exceeds the value $Z_{\alpha/2}$ is equal to $\alpha/2$. The *significance level*, corresponds to the confidence level (i.e., the confidence level = 100 (1- α)%). For example, if the confidence level is 90%, α is 0.10. Likewise if the confidence level is 95%, α is 0.05. For a 90% confidence level, $Z_{\alpha/2} = Z_{.10/2} = Z_{.05}$ is equal to 1.645. For a 95% confidence level, $Z_{\alpha/2} = Z_{.05/2} = Z_{.025}$ is equal to 1.96.

If the size of the population is less than 20 times n_0 , a finite correction factor should be used to reduce the estimate for the minimum number of samples required, as shown in the following equation:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{where,}$$

n = adjusted estimate of the minimum number of samples required.

N = number of units in the population being sampled.

In order to use the accompanying Excel file to calculate a sample size for computing a one-sample confidence interval, calculate the minimum sample size based on a one-sample, two-sided test, and use 50% as the hypothesized proportion. This will result in the most conservative (i.e., largest) sample size.

Once you have selected the sample size and conducted the sample, you can compute a confidence interval. The following equation can be used to calculate an approximate two-sided confidence interval for a single sample (where the confidence level is expressed as 100(1- α)%) for large populations based on a normal approximation to the binomial distribution.

$$p_s \pm \sqrt{\frac{p_s(1-p_s)}{n-1}} \cdot Z_{\alpha/2}, \text{ where.}$$

p_s = the proportion estimated from a sample

n = the sample size.

$Z_{\alpha/2}$ = the standard normal score from statistical tables. For a 90% confidence level, set $Z_{\alpha/2}$ equal to 1.645. For a 95% confidence level, set $Z_{\alpha/2}$ equal to 1.96.

For small populations, an approximate two-sided confidence interval (with confidence level of 100 (1- α)%) can be calculated from the following equation. All of the symbols are the same as above, and N = the number of units in the population being sampled.

$$p_s \pm \sqrt{\frac{N-n}{N}} \sqrt{\frac{p_s(1-p_s)}{n-1}} \cdot Z_{\alpha/2}$$

If you want to compute a confidence interval for estimating the difference between two population proportions, use the following formula for large samples:

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$

p_1 = proportion measured in a sample from the 1st population

p_2 = proportion measured in a sample from the 2nd population

n_1 = size of sample collected from the 1st population

n_2 = size of sample collected from the 2nd population

$Z_{\alpha/2}$ = the standard normal score from statistical tables. For a 90% confidence level, set

$Z_{\alpha/2}$ equal to 1.645. For a 95% confidence level, set $Z_{\alpha/2}$ equal to 1.96.

For small to medium samples (sample size is less than 5% of population size), use

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{((N_1-n_1)/N_1)p_1(1-p_1)/n_1 + ((N_2-n_2)/N_2)p_2(1-p_2)/n_2}$$

Constructing Hypothesis Tests

Statistical tests are formulated to address very specific questions, which are written in the form of null and alternative hypothesis statements. Null and alternative hypotheses are commonly abbreviated in the statistical literature as H_0 and H_A (or sometimes as H_1 , H_2 , etc. if there are multiple alternative hypotheses), respectively.

Several alternative forms of a hypothesis test can often be constructed. These alternative forms are referred to as one or two-sided (or alternatively, one- or two-tailed) tests, depending on the specific construction of the null and alternative hypothesis statements.

For example, suppose we wish to compare two proportions, P_1 and P_2 , from two populations (for example, compliance rates in Regions 1 and 2). We might start with a simple statement of

the question we are interested in addressing, such as “are the two proportions statistically significantly different?” This is an example of a two-sided hypothesis test, which would be written as:

$$H_0: P_1 = P_2$$

$$H_A: P_1 \neq P_2$$

In this case, we are simply asking if the two proportions are equal, and aren’t necessarily interested in the relative magnitude of P_1 and P_2 . That is, there are two ways in which the two proportions could be different: P_1 could be greater than P_2 , or P_1 could be less than P_2 .

If we were interested in a more specific statement of the relationship between P_1 and P_2 , then we might want to formulate our question as a one-sided hypothesis test. In this case, there would be two possible ways of stating H_0 and H_A .

Option 1 (“greater than” test): $H_0: P_1 \leq P_2$

$$H_A: P_1 > P_2$$

Option 2 (“less than” test): $H_0: P_1 \geq P_2$

$$H_A: P_1 < P_2$$

The preferential selection of a one- or two-sided test will depend on the specific goals of a particular survey design or study, and is another example of a situation in which a qualified statistician can provide important input during the planning phase of a project.

It is important to remember that all statistical analysis involves uncertainty and the potential for errors in drawing conclusions about a population based on a sample. There are two specific types of errors that are associated with conducting hypothesis tests. A Type I error is the mistake of rejecting the null hypothesis test when it really is true. A Type II error occurs by not rejecting a false null hypothesis. For example, suppose your null hypothesis was that there is no difference between compliance rates found in two rounds of inspections. A Type I error would occur when compliance rates for the two rounds of inspections are, in fact, the same, but you incorrectly conclude that there is a difference between them. A Type II error would occur if there was, in fact, a difference between compliance in the two rounds of inspections, but you do not detect the difference and therefore conclude that the null hypothesis is correct. The maximum probability of making a Type I error is the *significance level*. If the probability of making a Type II error is represented by β , then $1-\beta$ is known as the *power* of a hypothesis test.

Calculation of the Minimum Sample-Size Required for the One- and Two-Sample Tests of Proportions

This appendix presents two common classes of statistical tests for the analysis of survey data:

one-sample tests and two-sample tests. One-sample tests can be used to compare a sample proportion for a single random sample to a hypothesized population proportion. Two-sample tests are used to compare the results of two independent samples collected from two populations (e.g., comparing inspection results over time).

The following discussion provides general approaches that can be used to estimate the minimum number of samples required to perform the one- and two-sample test of proportions when designs are based on simple random sampling. Modified versions of these equations are available for dealing with more complex designs, and readers interested in more advanced applications are urged to consult a statistician or access the references listed in Appendix 7.

Calculation of Sample Size for the One-Sample Test of Proportions

The following equation can be used to calculate the minimum number of samples required to perform the one-sample test of proportions in cases where the target population is large, unknown, or can be assumed to be infinite. To make this concrete, suppose you are interested in estimating the proportion of facilities (in some sector) that are compliant with a certain requirement.

$$n_o = \frac{Z^2 \cdot P(1 - P)}{\delta^2}, \text{ where}$$

n_o = estimate of the minimum number of samples required, assuming a large or infinite population size

δ = the margin of error or difference between the true proportion and the estimated proportion based on a sample of the population

P = the hypothesized true proportion in the population. Note: The most conservative (i.e., maximum) estimate of the required sample size is achieved when the hypothesized proportion P is set at 0.50. As the true proportion becomes more extreme (i.e., larger or smaller), the minimum number of samples required decreases.

Z = the standard normal score from statistical tables, based on a specified confidence level. The following table presents Z values that are commonly used.

	Critical Values of Z	
	Significance Level = 10%	Significance Level = 5%
One-Sided Hypothesis Test	Z = 1.28	Z = 1.645
Two-Sided Hypothesis Test	Z = 1.645	Z = 1.96

If the size of the population is less than 20 times n_0 , a finite correction factor should be used to reduce the estimate for the minimum number of samples required, as shown in the following equation:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{where,}$$

n = adjusted estimate of the minimum number of samples required.

N = number of units in the population being sampled.

Calculation of Sample Size for the Two-Sample Test of Proportions

The following equation can be used to calculate the minimum number of samples required from each population to perform the two-sample test of proportions.

$$n = \frac{(Z_\alpha + Z_\beta)^2 [P_1(1 - P_1) + P_2(1 - P_2)]}{\delta^2}, \quad \text{where}$$

n = estimate of the minimum number of sample required from each population

P_1 = hypothesized true proportion in the 1st population

P_2 = hypothesized true proportion in the 2nd population

δ = the minimum difference between P_1 and P_2 that you want to detect. In some sources this is referred to as the minimum detectable difference or “effect size.” Note: that for a two-sided test, $\delta = |P_1 - P_2|$. For one-sided tests where $H_A = P_1 > P_2$ or $H_A = P_1 < P_2$, $\delta = P_1 - P_2$ or $\delta = P_2 - P_1$, respectively.

α = type I error probability (probability of falsely rejecting a true null hypothesis).

β = type II error probability (probability of failing to reject a false null hypothesis). β is calculated as $1 - \text{Power}$.

Power = desired probability of detecting a difference in proportions of size δ . Power is calculated as $1 - \beta$.

Z_α = the standard normal score from statistical tables, based on a specified type I error rate. **Note that one should use $\alpha/2$ for a two-sided test and α for a one-sided test.**

Z_β = standard normal score from statistical tables, based on a specified type II error rate.

Common choices of the minimum detectable difference, δ , are between 0.05 (5%) and 0.10

(10%). The smaller you make δ , the larger your sample size will be. Common choices for the Power are 0.80 (80%) or 0.90 (90%). Therefore, the value of β is 0.20 in the first case and 0.10 in the latter. The corresponding values of Z_β are 0.84 and 1.28 respectively. The table below gives the value of $(Z_\alpha + Z_\beta)^2$ for several combinations of power and desired confidence.

	Values of $(Z_\alpha + Z_\beta)^2$			
	Power = 80%		Power = 90%	
	90% Confidence	95% Confidence	90% Confidence	95% Confidence
One-Sided Hypothesis Test	4.5	6.2	6.6	8.6
Two-Sided Hypothesis Test	6.2	7.9	8.6	10.5

A conservative choice for the term $P_1(1-P_1) + P_2(1-P_2)$ in the formula for n is 0.50. Therefore, one can use the table of values for $(Z_\alpha + Z_\beta)^2$ to simplify the above expression for n to:

$$n = \frac{(Z_\alpha + Z_\beta)^2 \cdot 0.50}{\delta^2}$$

Conducting a One-sample Test of Proportions

One-sample tests can be used to compare results for a single round of random inspections to a hypothesized population proportion. For example, you may want to test that the proportion of facilities in compliance with some regulation is at least as great as shown on self-certifications. Another example: suppose inspections are performed to determine how many requirements in a checklist have been met. Perhaps there is a goal that at least 90% of applicable requirements have been met for this industry. From the sample one could test the hypothesis that the proportion of requirements that have been met is at least 0.9. In both of these examples, the tests are one-sided with the alternative that the proportion is less than the requirement.

These tests are simply a way of stating what the results of a single round of data collection from a sample imply for the general population from which the sample is drawn, with due regard to statistical properties such as the confidence level and margin of error. The test statistics presented in this appendix rely on properties of the standard normal distribution (i.e., Z-statistic).

Computational Details for the One-sample Tests of Proportions

The following formula can be used to assess the statistical significance of a one-sample test of proportions, in which a proportion estimated from a sample, p_s , is compared to an hypothesized proportion, P . The null hypothesis is that the true proportion is P . Of course, it is virtually never the case that the sample proportion will equal the true proportion. The test determines whether

or not the sample proportion p_s differs sufficiently from P to conclude that P is probably not the true proportion.

$$Z_{calc} = \frac{p_s - P}{\sqrt{\frac{P(1-P)}{n}}}, \text{ where}$$

Z_{calc} = test statistic calculated for the sample

p_s = proportion measured in the sample

P = hypothesized proportion

n = sample size

The value for Z_{calc} is then compared to a tabulated critical value, Z , which can be looked up in tables provided in most statistics textbooks (note: many software programs, such as Microsoft Excel, can also provide critical values for many standard distributions). The following table of Z is sufficient for most work.

	Critical Values of Z	
	Significance Level = 10%	Significance Level = 5%
One-Sided Hypothesis Test	Z = 1.28	Z = 1.645
Two-Sided Hypothesis Test	Z = 1.645	Z = 1.96

For example, for a two-sided hypothesis test, we would be 95% confident that we have correctly rejected our null hypothesis if the absolute value of Z_{calc} is greater than 1.96. The critical value corresponding to a confidence level of 95% for a one-sided hypothesis test is 1.645 if the hypothesis test is structured as a “greater than” test but negative 1.645 if hypothesis test is a “less than” test. In other words, if the alternative hypotheses is “greater than” you only reject the null hypothesis if Z_{calc} is greater than 1.645. If the alternative hypothesis is “less than” you only reject the null hypothesis if Z_{calc} is less than -1.645 .

Conducting a Two-sample Test of Proportions

Two-sample tests can be used to compare the results of independent samples collected from two populations (such as two different regions) or from the same population at different times (such as two different rounds of inspection to measure improvements in compliance from one round to the next).

The following example illustrates a two-sample test likely to be of interest for ERP. Suppose we are interested in measuring facility performance between two sampling rounds of compliance inspections. The first round of sampling might be used as a “baseline” for measuring improvement in subsequent sampling rounds. In this example, compliance could be compared for individual EBPIs or for groups of questions rolled up into a single “Yes” or “No” response.

We could analyze the data using either a one- or two-sided statistical test. The null and alternative hypotheses for a two-sided test would be:

H₀: The proportion of compliant facilities in round 1 is equal to the proportion of compliant facilities in round 2.

H_A: The proportion of compliant facilities in round 1 is not equal to the proportion of compliant facilities in round 2.

However, because we are really interested in knowing whether facility compliance has improved in round 2 in comparison to our baseline, it would be better to state this as a one-sided hypothesis test:

H₀: The proportion of compliant facilities in round 2 is less than or equal to the proportion of compliant facilities in round 1 (the baseline).

H_A: The proportion of compliant facilities in round 2 is greater than the proportion of compliant facilities in round 1 (the baseline).

In the one-sided example, H_0 is the hypothesis statement that we wish to reject in order to demonstrate that an improvement in compliance has been realized. If we fail to reject H_0 , then we would conclude there is not sufficient evidence to claim that improvement in compliance has resulted. Note that with small or even medium size samples, it may be difficult to get a significant result that would enable one to reject the null hypothesis, especially if the improvement in compliance is not great. For example, if compliance has increased from 75% to 85% from round 1 to round 2, the calculated Z would only be 1.77 with sample sizes of 100 in each round. This is significant for the one-tailed test at the 5% significance level (critical value of Z is 1.645). But if compliance in round 2 was 83% instead of 85%, the critical value of Z is 1.39, which is not significant. So in the latter case one would not reject the null hypothesis. This should not be interpreted as the compliance rates in both rounds are the same, but rather that there is not enough evidence to conclude the rate is greater in round 2. Note that if the sample size was 150 instead of 100 and the same 75% and 83% percentages were observed, the Z statistic would be significant (value 1.70). The moral is that a difference might be present but the sample was too small to pick it up definitively. The solution would be to either increase the round 2 sample or conduct additional rounds in subsequent time periods to see if the trend continues. A qualified statistician can help with more sophisticated tests to determine whether a trend is statistically significant.

Computational Details for the Two-sample Test of Proportions

The following formula can be used to assess the statistical significance of a two-sample test of proportions, in which proportions estimated from two samples, p_1 and p_2 , are compared.

$$Z_{calc} = \frac{p_1 - p_2}{\sqrt{p_s \cdot (1 - p_s) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } p_s = \left(\frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2}\right), \text{ and}$$

Z_{cal} = test statistic calculated for sample

p_1 = proportion measured in a sample from the 1st population

p_2 = proportion measured in a sample from the 2nd population

n_1 = size of sample collected from the 1st population

n_2 = size of sample collected from the 2nd population

As described before for the one-sample test of proportion, Z_{calc} is then compared to a critical value, Z , from standard statistical tables in order to evaluate the significance of the test, as shown in the following table:

	Critical Values of Z	
	Significance Level = 10%	Significance Level = 5%
One-Sided Hypothesis Test	Z = 1.28	Z = 1.645
Two-Sided Hypothesis Test	Z = 1.645	Z = 1.96

Software Tools for Determining Sample Size and Performing Tests of Proportions

An Excel file is available to assist readers who are implementing the statistical analyses described in this Appendix. The file is available online at <http://www.epa.gov/permits>. The file contains calculation modules that perform the following functions: 1) calculation of the minimum sample size for performing the one-sample test of proportions, 2) calculation of the minimum sample size for performing the two-sample test of proportions, 3) test of significance for the one-sample test of proportions, and 4) test of significance for the two-sample test of proportions.

The algorithms used in these modules follow the equations presented in this appendix and are taken from Snedecor and Cochran (1989). The “user interface” for accessing these modules is shown in Figures 5-1 through 5-4. Within each module, color coding and embedded comments

are used to guide users and better explain program details. Also, with the exception of data-entry cells, all cells in each of these worksheets are locked to assure that users enter information into the correct cells of each module.

Correlation Analysis

In its simplest application, correlation analysis is a means of quantifying the measure of association between two variables. Agencies implementing ERP might, for example, be interested in assessing whether they have chosen appropriate EBPIs for the purposes of measuring environmental performance. Ideally, the EBPI questions would be well correlated with one or more measures of good performance. Correlation analysis involves calculation of a correlation coefficient (often abbreviated as r in statistical literature). Values for r lie between -1 (implying perfect negative correlation) and $+1$ (implying perfect positive correlation).

The following examples illustrate how correlation analysis might be used in analyzing ERP data.

- ❖ **Example 1:** We might want to test whether each EBPI question is positively correlated with a high performance score within a particular section of a questionnaire. For example, suppose there is a section of the questionnaire with 10 questions pertaining to hazardous waste best management practices. Suppose an agency chooses the EBPI for that section based upon the presumption that facilities implementing that best management practice are likely to be implementing other best management practices. An answer of 'Yes' to this question would presumably correlate with a high score on all the individual questions in this section detailing best management practices. If this is found not to be the case, it would be questionable whether the EBPI was well chosen.
- ❖ **Example 2:** We might also test to see if a high correlation exists between the aggregate EBPI score for a facility and its aggregate questionnaire score. If this is not the case, then the EBPIs might not be doing a good job of supplying an accurate short-form indication of overall facility performance. However, since EBPIs carry more weight than other questions, a difference between EBPI scores and aggregate questionnaire score does not necessarily indicate that the EBPIs are not a good gauge of facility performance.

APPENDIX 6: ALTERNATIVE APPROACHES FOR ANALYZING SURVEY DATA - ANALYSIS OF CATEGORICAL DATA USING CONTINGENCY TABLES

The contingency table approach described here is an alternate way to perform two-sample tests. It is useful because many commercial statistical packages have menu driven procedures to perform such tests. Moreover, most of these packages contain the Fisher exact test which can test hypotheses even for small sample-sizes.

Survey results are commonly expressed as counts that fall into discrete categories, such as the number of facilities in (or not in) compliance with one or more regulatory requirements BEFORE and AFTER implementation of ERP, or the number of facilities that report they are in (or not in) compliance based on self-certification versus the number of facilities found to be in (or not in) compliance based on independent inspections. These counts can be presented in tables called “contingency tables.” The most basic contingency tables are called 2 X 2 tables and consist of four “cells” or possible outcomes for survey responses.

The use of a contingency analysis approach is illustrated in the following examples.

❖ *Example 1:* Suppose we wish to evaluate facility performance via independent inspection (with respect to a group of individual compliance indicators) BEFORE and AFTER implementation of facility self-certification.

We would start by posing the general question we are interested in addressing, then would restate the question in the form of a null (H_0) and alternative (H_A) hypothesis.

Question: For each compliance indicator (or groups of indicators that are rolled up into a single response), is there a significant difference in the proportion of compliant facilities determined via independent inspection BEFORE versus AFTER facility self-certification?

Statement of Null and Alternative Hypotheses:

H_0 : the proportion of facilities scored as compliant with respect to indicator X (or a group of indicators) via independent inspection AFTER implementation of self-certification is the same as the proportion BEFORE implementation of self-certification (that is, AFTER = BEFORE)

H_A : the proportion of compliant facilities AFTER implementation of self-certification is not the same as the proportion BEFORE implementation of self-certification (that is, AFTER \neq BEFORE).

Note: The above example is stated as a two-sided hypothesis test. However, if the objective is to determine if self-certification activities have had a positive effect on the rate of facility compliance, H_0 and H_A could be stated as a one-sided test, as shown below.

H₀: the proportion of facilities scored as compliant with respect to indicator X (or a group of indicators) via independent inspection AFTER implementation of self-certification is less than or equal to the proportion BEFORE implementation of self-certification (that is, AFTER ≤ BEFORE)

H_A: the proportion of compliant facilities AFTER implementation of self-certification is greater than the proportion BEFORE implementation of self-certification (that is, AFTER > BEFORE).

Next, the data from samples collected BEFORE and AFTER implementation of self-certification are arranged in a 2 X 2 contingency table.

Facility Response	Number of Facilities	
	AFTER	BEFORE
Compliant (C) (= <u>YES</u> Response)	(enter count)	(enter count)
Non-Compliant (NC) (= <u>NO</u> response)	(enter count)	(enter count)

❖ **Example 2:** Suppose we wish to compare the proportion of facilities that score themselves as “compliant” with respect to an individual compliance indicator (or multiple indicators that are rolled up into a single performance measure), versus the proportion of facilities judged to be compliant based on independent inspection.

Again, we would start by posing the general question we are interested in addressing, then would restate the question in the form of H₀ and H_A hypothesis statements.

Question: Is the proportion of facilities that score themselves as compliant greater than the proportion of facilities judged to be compliant based on the results of independent inspection?

Statement of Null and Alternative Hypotheses:

H₀: the proportion of facilities that score themselves as compliant with respect to indicator X (or a group of indicators rolled up into a single response) is less than or equal to the proportion of facilities judged to be compliant based on independent inspection.

H_A: the proportion of facilities that score themselves as compliant with respect to indicator X (or

a group of indicators rolled up into a single response) is greater than the proportion of facilities judged to be compliant based on independent inspection.

Note: this is a one-sided hypothesis test based on the collection of paired data (i.e., a single sample is collected, and measurements are recorded for both facility and inspector responses). The objective in this example is to determine whether facilities are over estimating their degree of compliance via the self-certification process.

The data for facility and inspector responses are then placed into a 2 X 2 contingency table.

Response	Number of Responses by Respondent	
	Facility	Inspector
Compliant (C)	(enter count)	(enter count)
Non-Compliant (NC)	(enter count)	(enter count)

Options for the Statistical Analysis of Contingency Tables

Once a survey has been carried out and the collected information has been arranged in a contingency table, the next step is to select an appropriate test statistic for evaluating the data and interpreting the results. The rows in a contingency table are commonly denoted as r and the columns as c , and the contingency table is referred to as an $r \times c$ table. Three methods are commonly employed for analyzing contingency tables: 1) chi-square analysis, 2) log-likelihood analysis, and 3) the Fisher exact test. The null hypothesis in each of these tests is that the frequency of observations found in the rows is independent of the frequencies found in the columns. This is equivalent to the hypothesis that the population proportion of observations in column 1 that lie in row 1 is the same as the population proportion of observations in column 2 that lie in row 2, which in the above example means the proportion of compliant facilities is the same for facility self-certification reports and reports prepared by independent inspectors. Details of the test statistics and mechanics involved in conducting contingency analysis are included in many statistics texts, although Zar (1999) provides an especially concise and approachable treatment of the subject.

The chi-square (X^2) is probably the most commonly employed test for analyzing contingency tables, mainly because the calculations are straightforward and computer programs are readily available for dealing with these types of designs. Chi-square analysis is based on the observed and expected frequencies from a contingency table, rather than estimates of proportions or percentages. The most important limitation of chi-square analysis is the requirement that each cell in the contingency table has a minimum expected number of counts or observations. In most texts it is suggested that each cell should have a minimum of 5 observations, however, Zar (1999) reports that a less restrictive criterion is that the average expected frequency (i.e., total

count/[rc]) be at least 6.

Log-likelihood analysis (G statistic) is an alternative to the chi-square for analyzing contingency tables. Log-likelihood analysis typically results in the same conclusions as chi-square analysis, and statisticians are not always in agreement concerning which approach is preferred. Zar (1999) provides an excellent introduction to log-likelihood analysis and presents additional details on the relative advantages and disadvantages of this approach compared to chi-square analysis.

The Fisher exact test is based on the hypergeometric probability distribution and, according to Zar (1999), is the preferred approach for significance testing whenever it can be applied. The Fisher exact test is computationally more demanding than other approaches, however, since many commercial software programs include this test, this should not be a serious barrier to its use in analyzing survey data. The Fisher exact test is based on the principal of permutation, and derives the name “exact” because it calculates the exact probability of obtaining all possible contingency tables that are more extreme than the results observed for the sample data. Many statistics texts, including Zar (1999), provide computational details and additional discussion of the development and application of this test.

Reference

Zar, J. H. 1999. *Biostatistical Analysis*. Fourth Edition. Prentice Hall. Upper Saddle River, NJ.

APPENDIX 7: ADDITIONAL RESOURCES ON SURVEY METHODS

This appendix contains an annotated list of references to books and miscellaneous documents, software packages, and resources available on the internet that readers can consult for further information on statistical methods for the design and analysis of surveys.

Books and Miscellaneous Documents

Cochran, W. G. 1977. *Sampling techniques*. 3rd Edition. John Wiley & Sons, Inc. New York, NY. [*This is a standard on sampling design, although this text is more appropriate for readers wishing an in depth treatment of the mathematical details of sampling.*]

Deming, W. E. 1966. *Some theory of sampling*. Dover Publications, New York, NY. [*This is a classic text on sampling from one of the true innovators in this field.*]

Henry, G.T. 1990. *Practical sampling*. Sage Publications. Thousand Oaks, CA.

Jessen, R.J. 1978. *Statistical survey techniques*. John Wiley and Sons, New York.

Kalton, G. 1983. *Introduction to survey sampling*. Quantitative applications in the Social Sciences Series, No. 35. Sage Publications. Thousand Oaks, CA. [*This is regarded by many as the authoritative work on survey sampling.*]

Kish, L. 1965. *Survey sampling*. John Wiley & Sons. New York, NY.

Levy, P.S. and S. Lemeshow. 1999. *Sampling of populations: methods and applications*. 3rd Edition. John Wiley & Sons, Inc. New York, NY.

Lohr, S.L. (1999) *Sampling: design and analysis*. Duxbury Press, Pacific Grove, CA.

Snedecor, G. W. and Cochran, W. G. 1989. *Statistical Methods*, Eighth Edition. Iowa State University Press. Ames, Iowa.

Stuart, A. 1994. *The ideas of sampling*. MacMillan Publishing Company, New York.

Thompson, S. 1992. *Sampling*. John Wiley & Sons, New York.

Other Documents: A wealth of information on sampling design and statistical methods can be obtained free from U.S. Government internet sites, such as EPA, Department of Interior, etc. EPA, in particular, provides a number of useful documents through their sites that focus on Quality Assurance/Quality Control (QA/QC) [www.epa.gov/quality/qa_docs.html]. As mentioned in the Introduction to this document, EPA's Guide for Measuring Compliance Assistance Outcomes provides a good additional resource for states implementing ERP [<http://www.epa.gov/compliance/resources/reports/planning/results/comeasuring2.pdf>].

Software Packages

Many commercial and non-commercial statistical software packages offer platforms suitable for the analysis of survey designs. The following are some of the more mainstream, commercial packages offered by companies that have established a reputation for excellence in providing statistical analysis software. Keep in mind that many straightforward statistical calculations can be conducted with common spreadsheet packages, such as Excel.

SAS (www.sas.com, www.jmpdiscovery.com) [*SAS is an established leader in providing statistical software for a wide variety of applications. SAS offers many packages that are tailored for specific applications, including the analysis of complex survey designs. SAS also offers a program called **JMP**, which is a scaled-down package for Windows and Macintosh computers, but that is nonetheless a very useful program with a range of capabilities. SAS has also recently acquired **StatView**, another popular statistical software package for Windows and Macintosh computers.*]

SPSS (www.spss.com) [*SPSS is another provider that offers a wide range of statistical analysis software, including packages specifically tailored for the analysis of survey data. SPSS also offers a wealth of information on survey design on their internet site, which can be accessed by searching on appropriate keywords. SPSS has also recently acquired **SYSTAT**, another software package that has become very popular on PCs*]

Statistica (www.statsoft.com) [*This is another comprehensive package for statistical analysis. Statsoft also offers a very good online statistical textbook on their site*]

Minitab (<http://www.minitab.com/>) [*This is another popular statistical software package that has been on the market for over 30 years.*]

Resources Available on the Internet

Note: Because internet sites appear and disappear with great regularity, no attempt has been made to provide a comprehensive list of sites, rather a few very good general-purpose sites are listed that offer access to a wide range of resources, including links to other statistical sites. All of these sites are updated frequently and have been “stable” over a long period of time. Readers are also urged to use internet search engines (e.g., www.google.com, www.aks.com) to conduct their own searches on the web.

Arizona State University Information Technology Page

<http://www.asu.edu/it/fyi/dst/helpdocs/statistics/> [*This is a highly recommended site—it offers access to many sites that provide information on survey statistics, access to commercial and shareware software products, statistical associations, electronic discussion groups, statistical journals, and databases.*]

Electronic Bibliography on Survey and Questionnaire Design

(<http://www.lib.cmich.edu/ocls/bibs/survey.htm>) [*This is a selected bibliography of books and internet links on survey methods prepared by Central Michigan University.*]

HyperStat Online Textbook (<http://davidmlane.com/hyperstat/index.html>) [*This is an award-winning web site managed by David Lane. This site provides an online statistical textbook as well as access to many statistical resources that may be useful in survey design.*]

Statistics.com (<http://www.statistics.com/>) [*This is another good general purpose site offering a range of information on statistics, as well as access to many other links. This site is especially useful for users interested in accessing information on resampling statistics.*]

Resources Pertaining to Survey Methods (<http://gsociology.icaap.org/methods/surveys.htm>) [*This is another useful page for accessing links and specific information pertaining to survey sampling.*]

GLOSSARY

Bias: Error caused by systematically favoring some outcomes over others. Bias can result from using judgment rather than probability methods to select the sample. It also results from excessive non-response and errors in reporting and recording responses. Bias is difficult to measure and can make results unreliable.

Cluster Sampling: An alternative to simple random sampling in which sample units are a collection or “cluster” of elements.

Confidence Interval: A confidence interval is a range of values that brackets a sample estimate and quantifies uncertainty around this estimate. In other words, for a population proportion, the confidence interval can be expressed as the observed sample proportion plus or minus the margin of error, along with a percentage that indicates the confidence level for that confidence interval.

Confidence Level: A 95% confidence interval is constructed from a sample by a procedure such that the interval will contain the true population value in 95% of all possible samples. Since, in reality, only one sample is selected, the corresponding confidence interval either contains the true population value or it doesn't. We express this by saying we have a level of confidence of 95% that this particular interval does contain the true value.

Critical Value: The value that is compared to the calculated test statistic to determine whether or not the null hypothesis should be rejected.

Descriptive Statistics: Methods for organizing and summarizing information for an entire population of entities. For example, if you design the self-certification form to collect information from all facilities about their number of employees, this information could be summarized using descriptive statistics. This is possible because you have information from *all* facilities for which you are trying to draw conclusions. Descriptive statistics are often used to describe demographic information, such as facility size or number of employees.

Inferential Statistics: Methods for drawing conclusions about a population and measuring the reliability of those conclusions based on information obtained from a sample of the population.

Margin of Error: A range of uncertainty. Margin of error indicates the range of values in which the true population value is likely to lie. The margin of error is often expressed as an interval of X percentage points above/below the sample observation (e.g., +/- 5%). The margin of error is sometimes called the sampling error. The range of values described by the margin of error is sometimes called the confidence interval.

One Sample Test: The situation where you are estimating characteristics about a population based on one sample (i.e., a single population at a single point in time).

One-Sided Hypothesis Test: A hypothesis test that is designed to determine whether a population proportion is less than a specified value (in the case of a left-sided test) or more than a specified value (in the case of a right-sided test). One-sided hypothesis tests are also called **one-tailed hypothesis tests**.

Population: The collection of all entities (e.g., facilities) under consideration in a statistical

study. The population is sometimes called the **universe** – the two terms are equivalent.

Population Mean: The mean (average) value of a variable for the population. For example, if you are interested in the average gallons of effluent produced for a population of facilities, the population mean would be the sum of effluent produced for all facilities in the population divided by the number of facilities in the population.

Population Proportion: The proportion (percentage) of entities in a population that have a specified attribute. For example, if you are interested in the percentage of facilities in a sector that are in compliance with a specific requirement, the population proportion would be the number of facilities in compliance with the requirement divided by the total number of facilities in the sector.

Power: The power of a hypothesis test is the probability of correctly rejecting a false null hypothesis. For a given sample size, there is a trade-off between power and significance level (i.e., increasing the power of a hypothesis test increases the likelihood of rejecting a true null hypothesis).

Probability Sampling: Sampling where a random device is used to decide which members of the population will constitute the sample instead of leaving such decisions to human judgment, which would likely introduce bias.

Simple Random Sampling: Simple random sampling is a special type of probability sampling in that each possible sample of a given size is equally likely to be the one selected. Simple random sampling is often the best way to select a representative sample – i.e., one that as closely as possible reflects the relevant characteristics of the population under consideration.

Sample: The part of the population from which information is collected.

Sampling Population: The group of entities from which a sample is drawn. Ideally, the sampling population is very close to the target population. The sampling population is sometimes called the **sampling frame**.

Significance Level: The maximum probability of rejecting a true null hypothesis (e.g., the probability of thinking that the population proportion is not equal to the value stated by the null hypothesis, when in fact it is). This type of error is called a Type I error. The significance level is represented by the symbol α . The confidence level is equal to $100(1-\alpha)\%$.

Stratified Sampling: A method of sampling whereby the total population is divided into subpopulations (strata), and then samples are taken separately from each strata.

Target Population: The group of entities in the universe (or population) in which you are interested.

Test Statistic: A calculated value that is compared to the critical value to determine whether or not the null hypothesis should be rejected.

Two Sample Test: The situation where you are comparing proportions or means from two samples (e.g., two different subgroups or “before” and “after” samples).

Two-Sided Hypothesis Test: A hypothesis test that is designed to determine whether a

population proportion is different than a specified value (i.e., the value stated in the null hypothesis). In a two-sided hypothesis test, you are concerned about both ends of the data distribution. Two-sided hypothesis tests are also called **two-tailed hypothesis tests**.