GRI®

EPA

United States
Environmental Protection
Agency

GRI-94 / 0257.21
EPA - 600/R-96-080d
June 1996

Research and Development

METHANE EMISSIONS FROM THE NATURAL GAS INDUSTRY

Volume 4: Statistical Methodology

Prepared for

Energy Information Administration (U. S. DOE)

Prepared by

National Risk Management
Research Laboratory
Research Triangle Park, NC 27711

FOREWORD

The U.S. Environmental Protection Agency is charged by Congress with protecting the Nation's land, air, and water resources. Under a mandate of national environmental laws, the Agency strives to formulate and implement actions leading to a compatible balance between human activities and the ability of natural systems to support and nurture life. To meet this mandate, EPA's research program is providing data and technical support for solving environmental problems today and building a science knowledge base necessary to manage our ecological resources wisely, understand how pollutants affect our health, and prevent or reduce environmental risks in the future.

The National Risk Management Research Laboratory is the Agency's center for investigation of technological and management approaches for reducing risks from threats to human health and the environment. The focus of the Laboratory's research program is on methods for the prevention and control of pollution to air, land, water, and subsurface resources; protection of water quality in public water systems; remediation of contaminated sites and groundwater; and prevention and control of indoor air pollution. The goal of this research effort is to catalyze development and implementation of innovative, cost-effective environmental technologies; develop scientific and engineering information needed by EPA to support regulatory and policy decisions; and provide technical support and information transfer to ensure effective implementation of environmental regulations and strategies.

This publication has been produced as part of the Laboratory's strategic long-term research plan. It is published and made available by EPA's Office of Research and Development to assist the user community and to link researchers with their clients.

E. Timothy Oppelt, Director
National Risk Management Research Laboratory

EPA REVIEW NOTICE

This report has been peer and administratively reviewed by the U.S. Environmental Protection Agency, and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

GRI-94/0257.21

EPA-600/R-96-080d
June 1996

METHANE EMISSIONS FROM THE NATURAL GAS INDUSTRY, VOLUME 4: STATISTICAL METHODOLOGY

FINAL REPORT

Prepared by:

Hugh J. Williamson
Mary B. Hall
Matthew R. Harrison

Radian International LLC
8501 N. Mopac Blvd.
P.O. Box 201088
Austin, TX 78720-1088

DCN: 95-263-081-05

For

GRI Project Manager: Robert A. Lott
GAS RESEARCH INSTITUTE
Contract No. 5091-251-2171
8600 West Bryn Mawr Ave.
Chicago, IL 60631

and

EPA Project Manager: David A. Kirchgessner
U.S. ENVIRONMENTAL PROTECTION AGENCY
Contract No. 68-D1-0031
National Risk Management Research Laboratory
Research Triangle Park, NC 27711

DISCLAIMER

LEGAL NOTICE: This report was prepared by Radian International LLC as an account of work sponsored by Gas Research Institute (GRI) and the U.S. Environmental Protection Agency (EPA). Neither EPA, GRI, members of GRI, nor any person acting on behalf of either:

a. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

b. Assumes any liability with respect to the use of, or for damages resulting from the use of, any information, apparatus, method, or process disclosed in this report.

NOTE: EPA's Office of Research and Development quality assurance/quality control (QA/QC) requirements are applicable to some of the count data generated by this project. Emission data and additional count data are from industry or literature sources, and are not subject to EPA/ORD's QA/QC policies. In all cases, data and results were reviewed by the panel of experts listed in Appendix D of Volume 2.

RESEARCH SUMMARY

Title

Methane Emissions from the Natural Gas Industry,
Volume 4: Statistical Methodology
Final Report

Contractor

Radian International LLC

GRI Contract Number 5091-251-2171
EPA Contract Number 68-D1-0031

Principal Investigators

Hugh J. Williamson
Mary B. Hall
Matthew R. Harrison

Report Period

March 1991 - June 1996
Final Report

Objective

This report describes the statistical methods used to quantify the annual methane emissions from the natural gas industry. The objective was to determine this quantity with an accuracy of 0.5% of production on the basis of a 90% confidence interval.

Technical Perspective

The increased use of natural gas has been suggested as a strategy for reducing the potential for global warming. During combustion, natural gas generates less carbon dioxide (CO_2) per unit of energy produced than either coal or oil. On the basis of the amount of CO_2 emitted, the potential for global warming could be reduced by substituting natural gas for coal or oil. However, since natural gas is primarily methane, a potent greenhouse gas, losses of natural gas during production, processing, transmission, and distribution could reduce the inherent advantage of its lower CO_2 emissions.

To investigate this, Gas Research Institute (GRI) and the U.S. Environmental Protection Agency's Office of Research and Development (EPA/ORD) cofunded a major study to quantify methane emissions from U.S. natural gas operations for the 1992 base year. The results of this study can be used to construct global methane budgets and to determine the relative impact on global warming of natural gas versus coal and oil.

Results

The national emissions for the base year are 314 ± 105 Bscf (± 33%), which is equivalent to 1.4% ± 0.5% of gross natural gas production. The program reached its accuracy goal and provides an accurate estimate of methane emissions that can be used to construct U.S. methane inventories and analyze fuel switching strategies.

Technical Approach

The technical approach involved several aspects, including statistical sampling, estimation of annual emission values, and quantification of uncertainty.

To facilitate the sampling process, the industry was divided into emission source categories. A target accuracy value was computed for each category as an aid in allocating sampling resources. The target accuracies were updated as the sampling process proceeded and more was known about the characteristics of the categories.

Tests were performed to identify categories for which the sampling process may have produced a bias. While these tests were designed to identify bias, they were also sensitive to sampling anomalies. If undetected, such anomalies could have led to larger than expected random errors. While no test exists that would absolutely guarantee that zero bias existed, the bias screening that was performed was of significant benefit in the study. When evidence of bias was discovered, steps were taken to remove it. Collecting more data is one possible step; other remedies are discussed in the report.

For each source category, an activity factor and an emission factor were computed. Typically, the activity factor is the number of sources (population) of a source category, and the emission factor is the average annual emissions of a source. The uncertainties of both the activity factor and the emission factor were quantified on the basis of the data.

The national emissions for a source category equals the activity factor times the emission factor. The annual emissions for the industry is the sum of the annual emissions for all the categories. Analysis of error propagation was performed to compute the uncertainty of the annual emissions by category and the uncertainty of the national emissions.

An analysis was performed to determine the sensitivity of the uncertainty in the national emissions to the presence of non-normally distributed errors and correlated errors among source categories. The uncertainty of the national annual emissions was computed under worst-case assumptions.

Project Implications

For the 1992 base year the annual methane emissions estimate for the U.S. natural gas industry is 314 Bscf ± 105 Bscf (± 33%). This is equivalent to 1.4% ± 0.5% of gross natural gas production. Results from this program were used to compare greenhouse gas emissions from the fuel cycle for natural gas, oil, and coal using the global warming potentials (GWPs) recently published by the Intergovernmental Panel on Climate Change (IPCC). The analysis showed that natural gas contributes less to potential global warming than coal or oil, which supports the fuel switching strategy suggested by IPCC and others.

In addition, results from this study are being used by the natural gas industry to reduce operating costs while reducing emissions. Some companies are also participating in the Natural Gas-Star program, a voluntary program sponsored by EPA's Office of Air and Radiation in cooperation with the American Gas Association to implement cost-effective emission reductions and to report reductions to the EPA. Since this program was begun after the 1992 baseline year, any reductions in methane emissions from this program are not reflected in this study's total emissions.

Robert A. Lott
Senior Project Manager, Environment and Safety

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Page

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

This report is one of several volumes that provide background information supporting the Gas Research Institute and U.S. Environmental Protection Agency Office of Research and Development (GRI-EPA/ORD) methane emissions project. The objective of this comprehensive program is to quantify the methane emissions from the gas industry for the 1992 base year to within ±0.5% of natural gas production starting at the wellhead and ending immediately downstream of the customer's meter.

This report presents a detailed discussion of the statistical methods used in this study. The major topics discussed include statistical sampling issues, calculation of the 1992 national emissions, and determination of the uncertainty of this value.

To facilitate the sampling process, the industry was divided into source categories. A target accuracy value was computed for each category as an aid in allocating sampling resources. The target accuracies were updated as the sampling process proceeded and more was known about the categories.

Tests were performed to identify categories for which the sampling process may have introduced a bias. While these tests were designed to identify bias, they were also sensitive to sampling anomalies. If undetected, such anomalies could have led to larger than expected random errors. While no tests exists that would absolutely guarantee that zero bias existed, the bias screening that was performed was of significant benefit in the study. When evidence of bias was discovered, steps were taken to remove it. Collecting more data is one possible step; other remedies are discussed later in the report.

For each source category, an activity factor and an emission factor were computed. Typically, the activity factor is the number of sources (population) of a source category, and the emission factor is the average annual emissions of a source. The uncertainties of both the activity factor and the emission factor were quantified on the basis of the data.

The national emissions for a source category equals the activity factor times the emission factor. The annual emissions for the industry is the sum of the annual emissions for all the categories. Analysis of error propagation was performed to compute the uncertainty of the annual emissions by category and the uncertainty of the national emissions.

An analysis was performed to determine the sensitivity of the national annual emissions to the presence of non-normally distributed errors and correlated errors among source categories. The uncertainty of the national annual emissions was computed under worst-case assumptions.

National emissions were quantified to be 314 ± 105 Bscf (± 33%), which is equivalent to 1.4% ± 0.5% of natural gas production. The accuracy goals of the project were met.

2.0 INTRODUCTION

In general, the first step for estimating methane emissions from the U.S. natural gas industry was to identify, delineate, and characterize each emission source within the industry, so that all significant sources were included. The industry characterization affected the sampling strategy and, therefore, is relevant to the statistical methodology discussed later in the report. While the industry characterization is covered in detail in other Tier 3 reports, a brief summary follows.

The industry was divided into its principal market segments: production, processing, transmission, and distribution. Within each market segment, the process facilities were identified, and within each facility, the individual pieces of equipment and components contributing to emissions (the source categories) were identified. The disaggregation ensured that no sources were overlooked or double counted and produced a manageable framework within which the study would be conducted. The industry market segments, major facilities within those segments, and the major equipment within the facilities are shown in Table 2-1.

After identifying the major equipment (source categories) in each industry market segment, all possible emissions from each source were identified by examining the operating modes of the equipment that may lead to emissions and by associating one of three possible types of emissions with the source: fugitive emissions, vented emissions, or combustion emissions.

In Section 3, sampling and methods for avoiding bias are discussed. In Section 4, the methods used to extrapolate emissions estimated for individual sources to obtain a nationwide average are discussed. Methods for quantifying uncertainty are also discussed in Section 4. In Section 5, the major statistical assumptions are summarized. Results pertaining to the attainment of the target accuracy are presented in Section 6.

TABLE 2-1. INDUSTRY CHARACTERIZATION

Segment	Facilities	Equipment at the Facility
Production	Well Sites, Central Gathering Facilities	Wellheads, Separators, Pneumatic Devices, Chemical Injection Pumps, Dehydrators, Compressors, Heaters, Meters, Pipelines
Processing	Gas Plants	Vessels, Dehydrators, Compressors, Acid Gas Removal (AGR) units, Heaters, Pneumatic Devices
Transmission	Transmission Pipeline Networks, Compressor Stations, Meter and Pressure Reg. Stations	Vessels, Compressors, Pipelines, Meters/Pressure Regulators, Pneumatic Devices
Storage	Underground Injection/ Withdrawal Facilities, and Liquefied Natural Gas (LNG) facilities	Wellheads, Vessels, Compressors, Dehydrators, Heaters, Pneumatic Devices
Distribution	Main and Service Pipeline Networks, Meter and Pressure Reg. Stations	Pipelines, Meters and Pressure Regulators, Pneumatic Devices, Customer Meters

3.0 SAMPLING AND AVOIDING BIAS

Data obtained in this project were necessarily collected from a limited number of sources. These data were extrapolated to obtain nationwide estimates for similar sources throughout the industry. The extrapolation techniques for creating nationwide emission estimates were developed so that the emissions for each source category could be estimated with a relatively high level of precision (given the nature of this study) and negligible bias.

The extrapolation approach is a method to scale up the average emissions from a source, determined by a limited sampling effort, to represent the entire population of similar sources in the gas industry. The extrapolation approach uses the concept of emission and activity factors to estimate emissions on the basis of a limited number of samples. These factors are defined in such a way that their product equals the total annual nationwide emissions from a source category in the natural gas industry.

$$EF \times AF = \text{National Emissions for a Category} \qquad (1)$$

where:

EF = emission factor for a category, and
AF = activity factor for the same category

Typically, the emission factor for a source category represents the average emissions per source, and the activity factor represents the total industry population of the source. The emission factor is the summation of all measured or calculated emissions from each of the sources sampled divided by the number of sources sampled. The activity factor is the total number of sources in the entire target population or source category. However, in applying this simplified approach to developing emission and activity factors, it is important to ensure that there is no bias in the data.

The extrapolation methodology involves more than just the scaling up of emissions data; it also includes the sampling approach, which is fundamental to the accuracy of the emissions data.

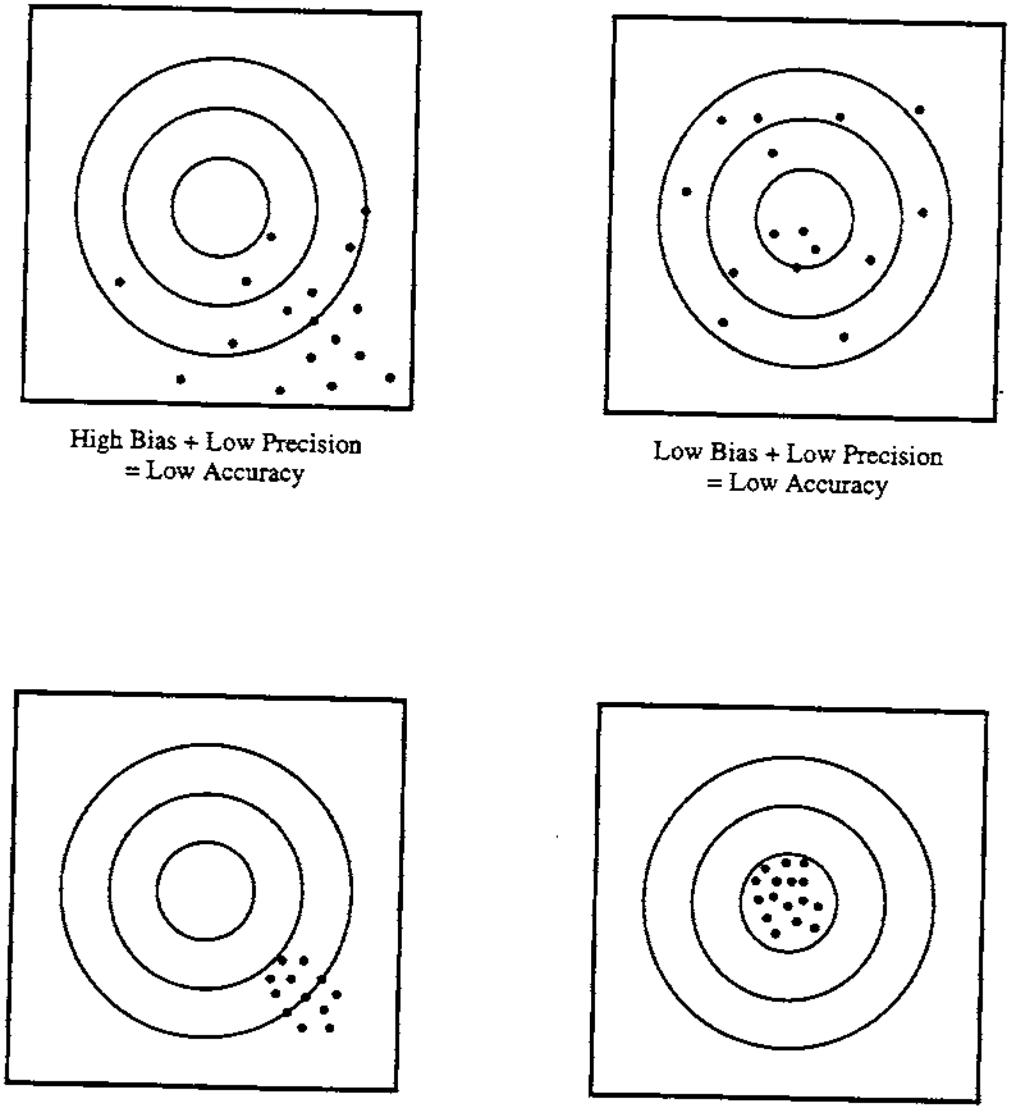
Basic issues pertaining to accuracy, precision, and bias are discussed in Section 3.1. The sampling approach designed specifically for this project is presented in Section 3.2. Calculation of target precision by source category is discussed in Section 3.3. The methods for estimating the emission factors and activity factors are discussed in Sections 3.4 and 3.5, respectively. Summary comments regarding techniques used to screen for bias are given in Section 3.6.

3.1 Accuracy, Precision, and Bias

Figure 3-1 illustrates the role of random and bias errors in the estimation process. In each of the four illustrations in this figure, the center of the concentric circles represents the correct answer. In the illustration in the upper left, there is a significant amount of random scatter in the points. The term "precision" refers to random variability alone; in this case, the precision is poor. Additionally, the points are predominantly below and to the right of the target. The systematic difference between the points and the correct answer is a bias. The term "accuracy" refers to the total error, including random and bias errors. Because of the large bias and the poor precision, the accuracy is also poor.

In the illustration in the upper right of Figure 3-1, the points are randomly scattered about the correct answer; there is little or no bias in this case, but the precision and accuracy are both poor. In the lower left, there is good precision, but there is again a large bias; thus, the accuracy is poor. In the lower right, the bias is small, and the precision is good. Thus, the accuracy is good in this case.

Sampling bias occurs if the methodology is flawed in a manner that leads to a systematic under-representation of parts of the population and a systematic over-



High Bias + Low Precision
= Low Accuracy

Low Bias + Low Precision
= Low Accuracy

High Bias + High Precision
= Low Accuracy

Low Bias + High Precision
= High Accuracy

Figure 3-1. Illustration of Random and Bias Errors

representation of other parts. Bias, in a statistical sense, can be explained as follows. Suppose it was possible to repeat the sampling and measurement process infinitely many times, and that each time the process was repeated, an independent estimate of a given emission factor was obtained. If the average of the entire infinite set of emission factor estimates equalled the true value, then a bias would not exist. If the average of these estimates differed from the true value, then the process would be wrong in a systematic sense, and a bias would be said to exist. The point here is that averaging an infinite set of independent estimates of the emission factor would remove random error altogether, leaving only bias error, if any. While it is clearly impossible to obtain an infinite set of estimates of an emission factor, the example given serves to illustrate the meaning of bias.

3.2 Sampling Approach

Even if the overall precision of an estimate is acceptable because the variability in the data is relatively low, the overall accuracy may still be poor if the data are biased. Several sampling approaches can be applied in order to avoid bias.

Random Sampling—Random sampling produces a sample set obtained in such a way that each source in the population has an equal probability of being selected. A random sample is expected to "match" the industry population because no biases are introduced when selecting the sites. The number of data points required in a random sample depends on the target precision of the final emissions estimate, the confidence with which this precision is to be met, and the underlying variability among the annual emissions of the complete set of sources.

Random sampling is not a guarantee of accurate results. It is possible, for example, that, by pure chance, random sampling would produce a disproportionately large number of sources from the Gulf Coast and an under-representation of sources from the West Coast. While such an outcome is unlikely if the sample size is sufficiently large, this particular problem can be avoided altogether by selecting an acceptable number of sources from each of a set of regions (see the discussion of stratified random sampling below).

There are two major reasons why truly random sampling was not possible in the GRI/EPA program. First, a complete list of sources did not and still does not exist. It was possible, for example, to list all compressor stations whose owners were GRI members. While this might account for 90% of the compressor stations, the list was not complete. Another example is the production segment, where it was not possible to produce a list of all the individual well owners for random selection. The second reason random sampling was not possible is that the owners of the randomly selected sources would not have been required to participate in the study. For this reason, there is no guarantee that a truly random sample of the available list could be tested.

Stratified Random Sampling—In stratified random sampling, the population of interest is divided into subsets, or strata. Then random samples are drawn from each stratum. For example, the sources of interest in this program could be stratified by geographical region, and random sampling could be applied within each region.

Strata are typically chosen so that the variable of interest (emission factor or activity factor) has a smaller variance within the strata than in the population as a whole. If this objective is achieved, stratified sampling can usually allow a given precision requirement to be achieved with a smaller sample size.

Sampling to include the different regions in the country was important. Each producing region selected for the United States had unique production characteristics. Failure to account for these regional differences in the extrapolation could have led to significant bias in the estimate.

Stratified random sampling can be performed proportionately or disproportionately. In proportionate stratified random sampling, the number of sources sampled in a stratum is in proportion to the total number of sources in that stratum. For example, if Region A had twice as many sources as Region B, then the sample would include twice as many sources from Region A as from Region B. From an intuitive point of view, a proportionate stratified random sample "matches" the population, at least with respect to the criteria used to specify the stratification.

Proportionate stratified random sampling can be used to address the issue of regional differences, but only if applied properly. In the paragraph above, it is suggested that sources could be sampled in proportion to the total number of sources by region. Alternatively, proportionality could be achieved on the basis of gas production, rather than on the basis of the number of sources. The variable or variables used to achieve proportionality must be closely related to emissions or proportionate random sampling would serve no purpose.

It is common in practice, however, to sample in such a way that the sample size for a stratum is not in proportion to the total number of sources in the stratum (and the throughput of the sampled sources is not proportional to total throughput in the stratum). This type of sample is called a "disproportionate stratified random sample." This type of sample does not "match" the population in the sense described above. As long as the disproportionality is accounted for in computing the final statistics (e.g., mean emission rate by source category), disproportionate sampling will not cause a bias in the final results.

There are various reasons for disproportionate stratified sampling. Convenience and opportunity may be factors. On a given field trip, for example, there may be the opportunity to sample more sources in a given category than are needed to achieve a proportionate random sample. Given the opportunity, it is better to obtain the available data than to restrict the sampling just to maintain a proportionate sample. Statistical issues may also lead to disproportionate sampling. For example, it may be advantageous to obtain more data points for a stratum within which the emission rate has a larger variance than within another stratum with a small variance; i.e., a more accurate estimate of the total emission rate may be achieved on the basis of a disproportionate sample.

Neither type of stratified random sampling was feasible in this study. The obstacles to random sampling, discussed earlier in this section, were also obstacles to random sampling within strata.

Further, at the outset of the program, it was not known which variables were related to emissions; thus, it was not known which variables should be used as a basis for stratification. If stratification had been performed on the basis of all variables that could possibly influence emissions, the number of strata (determined by the number of variables and the number of categories for each variable) could have become unreasonably large. For example, for leakage from underground distribution mains and services, a number of parameters were identified that potentially influence emissions: pipe material, age, operating pressure, diameter, soil type, and parameters characterizing the leak detection and repair practices of the company. The required sample size can become large because of the total number of strata, especially if proportional stratified random sampling is used. One company has embarked upon an independent program to quantify leakage from underground mains and services using a proportional sampling approach. Even within this single company, hundreds of samples were required to produce a proportionate stratified random sample for underground pipelines.

Additionally, stratified sampling is of no use unless there are activity factors that can be used to estimate the emission rate for the population. Complete information for all variables of potential interest does not exist. For example, the age of a dehydrator may not be known even by the owner of the equipment in some cases. It would be pointless to stratify dehydrator emission factors with respect to age if the necessary activity factors cannot be obtained.

Approach Selected For This Program—Thus, because of various practical limitations, neither random sampling nor stratified random sampling was feasible in this study. For this reason, an alternate approach was used. While this approach is not a textbook sampling method, it is believed to be very effective for the specific needs of this project. The selected approach is similar to disproportionate stratified random sampling, with certain differences.

Initially, some data were collected to determine if a given source was a major contributor to the methane emissions. For each source category, an initial estimate of the number of sources to be sampled was calculated based on an estimate of the target precision and the estimated standard deviation for the source category. The target precisions are based on the need ultimately to estimate the annual national emissions to within 0.5% of the annual national production (±111 Bscf) on the basis of a 90% confidence limit. The approach for determining the target precisions for the different source categories is discussed in the next subsection. Sites were selected in a random fashion from known lists of facilities, such as GRI or A.G.A. member companies. However, the companies contacted were not required to participate, and a complete list of all sources in the United States was generally not available; therefore, the final set of companies selected for sampling was not truly random. Each company that agreed to participate in the program was asked to select representative sites for sampling, rather than one-of-a-kind facilities.

After a limited set of data was collected, the data were screened for bias by evaluating the relationship between emission rate and parameters that may affect emissions. The topic of screening for bias is discussed further in Section 3.4, which pertains to the emission factor approach. If a relationship between emissions and a parameter was found, then the population, or the number of sources in the industry, was stratified by that parameter. For example, station type was determined to influence the emission rates from metering and pressure regulating stations, so the number of stations under each station type in the nation was determined. To stratify the population of sources by a parameter, data were collected from companies on the distribution of sources in each stratum, and an average covering all companies sampled was determined.

It is important to realize that just because a parameter or set of strata is identified that has a large effect on the emissions from a given source category, it does not mean that there is bias in the data. A second condition is necessary: The sampling procedure would have to produce disproportionate numbers of samples in the strata. To determine whether this has occurred, information is needed on the ratio of the total number of sources in a given stratum to the total number of sources throughout the country. If this ratio is different from the corresponding ratio for the sample data set, then there may be bias. But this bias can be eliminated by applying the correct emission factors and activity factors for the different strata.

Once the strata were identified, the precision of the emission rate extrapolated to a national basis was evaluated and compared to the target precision for the source category; the calculation of target precisions for all source categories is discussed in the next subsection. Where necessary, additional data were collected in various strata to improve the precision of the national estimate of emissions from the source. The number of additional data points needed to meet the newly calculated target precision is computed on the basis of measures of uncertainty (confidence intervals) introduced later in this report.

Tables showing the general data for the sites visited (sampled) in the production, processing, transmission, and storage segments of the gas industry are presented in Appendix A of Volume 5 on activity factors.[1] These tables are not central to the discussion of statistical methods, but they are mentioned here because they provide an indication of the magnitude of the sampling effort.

3.3 Target Precisions

A further issue regarding sampling is the number of sources to sample in each category. The ultimate objective is to estimate the industry annual emissions within 0.5% of the annual natural gas production (within 111 Bscf) on the basis of a 90% confidence interval. The sampling must be adequate to achieve this objective.

It is necessary to sample in a way that will lead to the required accuracy in the estimate of the industry total emission rate. Given this objective and the finite resources available for the project, it was not feasible to characterize the emissions from all source categories extremely accurately.

A large percentage error can be tolerated in the estimate of the emissions from categories that have small emissions without jeopardizing the accuracy of the national emissions. The percentage errors for the categories with the largest emissions would have the largest contribution to the error in the national annual emissions.

Table 3-1 illustrates these ideas. A hypothetical case has been chosen with two categories. The two source categories have emissions of 0.1 and 50 Bscf. (The range of emissions by category in the gas industry is even greater, as shown by the summary table in Appendix C.)

In Table 3-1a, the relative uncertainty of the emissions from each source is 20%. As a result, the relative uncertainty of the emissions for the two categories is 10.0 Bscf, or 20.0% of the emissions.

In Table 3-1b, the uncertainties are unequal percentages of the category emissions. For the category with smaller emissions, the uncertainty is 100% of the emissions. For the category with larger emissions, the uncertainty is only 10% of the emissions. In this case, the uncertainty of the emissions of the two categories is half the previous value, despite the 100% uncertainty for the category with the smaller emissions.

TABLE 3-1. HYPOTHETICAL ILLUSTRATION OF ISSUES RELATIVE TO TARGET ACCURACY

(a) Case with Equal Percentage Uncertainties

Source Category	Annual Emissions (Bscf)	Tolerance (Bscf)	Uncertainties (%)
1	0.1	0.02	20.0
2	50	10	20.0
Total	50.1	10.0	20.0

(b) Case with Unequal Percentage Uncertainties

Source Category	Annual Emissions (Bscf)	Tolerance (Bscf)	Uncertainties (%)
1	0.1	0.1	100.0
2	50	5	10.0
Total	50.1	5.0	10.0

In this hypothetical illustration, the percentage error for the category with larger emissions dominates the percentage error for the sum of the emissions for the two categories. Similarly, in the actual case with 86 categories, the percentage errors in the categories with larger emissions will have the greatest effect on the error in the industry annual emissions.

Note that, in Table 3.1, the uncertainty of a total is not the sum of the uncertainties for the corresponding categories. Analysis of error propagation in a sum is discussed in Section 4.4.

In view of the discussion above, it is advantageous to devote more of the project resources to sources with larger emission values. It is necessary to devote some resources, however, to all categories.

The approach taken in this project was to establish a target precision for each category, such that the required precision for the industry annual emissions is exactly met if the individual target precisions are met for all categories. The target precision was updated periodically during the program to indicate categories that require significantly more sampling to meet their target precisions.

Inherent in this discussion is the fact that the uncertainty of the estimate of the emissions for a category decreases as the number of data points increases. The relationship between the uncertainty of an estimated quantity and the number of data points (sample size) on which the estimate is based is discussed in Section 4.3.

The term "target precision" was used in the context of this section rather than "target accuracy." This is because the precision of the estimate of the emission rate for a category can be quantified on the basis of the variability in the data (again, see Section 4.3). Bias cannot be quantified. As is discussed elsewhere in the report, however, considerable efforts have been made to avoid bias. An assessment of the attainability of the desired accuracy for the industry annual emissions in view of both random errors and the possibility of undetected bias errors is presented in Section 6.2.

The equation for target precision adopted is as follows:

$$TP = 100 \left[\frac{a}{\sqrt{ER}} \right] \tag{2}$$

where TP = target precision (%),

ER = annual emissions in Bscf, and

a = coefficient determined from the data (see below).

This function is clearly unbounded as ER approaches zero. Therefore, a maximum target precision of 1500% was imposed. While this sounds like an enormous uncertainty, only the smallest categories that contribute a fraction of a percent of the industry annual emissions could be affected by this maximum.

The function is bounded below only by zero. To avoid calculating an unreasonably small target precision for categories with large emissions, a lower limit of 75% was imposed for each category. In other words, a target precision would not be set lower than 75% for any one category. However, all calculated target precision values were greater than 75%, so no values were changed as a result of this constraint.

The equation was iteratively solved for "a" so that the overall goal of ±111 Bscf precision for the national estimate was met. The constant "a" in the equation above was computed to be 6.24. That is, for this value of "a," if the target precisions were just met for all categories, the required precision would also be met for the industry annual emissions.

The target precisions were not used as absolute constraints. Suppose, for example, that on a given field trip there was an opportunity to sample more sources than were required to bring a given category into compliance with the target precision. Moreover, given that the basic travel expenses were incurred in any case, the incremental cost of obtaining the additional measurements was small. In a case such as this, a common-sense approach was used, and the additional measurements were obtained.

Further, it is not absolutely required that the target precision be met for all categories for the accuracy requirement to be met for the industry. It is possible for the industry requirement to be met given that the uncertainties for some categories are less than the target precisions and the uncertainties for other categories are greater than the target precisions.

The summary table in Appendix C includes both the uncertainties of the emission rates and the target precisions by category.

3.4 Emission Factor Approach

If it had been possible to use a random or proportionate stratified random sampling approach to collect data, then the emission factor could have been calculated by simply summing the emissions data from all sources and dividing by the number of sources sampled. In this case, the emission factor would be defined as the annual emissions per source. By the nature of true random or proportionate stratified random sampling, the resulting average emission factor would have had no inherent bias.

As discussed earlier, however, neither conventional random sampling nor conventional stratified random sampling could be used. Regardless, the emission factor is generally still defined as the annual emissions per source. In some cases, the variability of the emissions data from source to source is very large. For source types of this nature, it is normally possible to reduce variability by redefinition of the emission factor or by stratification; reducing variability reduces the number of data points needed to achieve the target precision.

Redefinition of the Emission Factor. For a few types of sources, the emissions can be more accurately estimated with fewer data points when the emission factor is defined not as a simple average for the source but in relation to key parameters that influence the emissions from the source. Since the variability is significantly reduced, fewer data points are required to achieve an acceptable level of accuracy.

For example, the internal combustion engines that drive compressors in the gas industry vary in size (i.e., horsepower rating). If data were collected on individual engines in the industry, and an average emission rate per engine was established, the variability from engine to engine would be very large because of the size differences. However, if the emission factor for the engines is defined by horsepower of the engine (i.e., annual emissions per horsepower), then the variability from engine to engine and therefore the number of samples required to reach an acceptable accuracy are both significantly reduced.

The number of data points required may also be reduced by stratifying on the basis of parameters that affect emissions. An example is quantification of the methane emissions from underground distribution mains and services. On the basis of the limited data, the variability in emission measurements for underground distribution lines was determined to be very large. By defining parameters that influence the emission rate from distribution lines and stratifying the emission factor and activity factor for this source by these parameters, the variability of emissions from source to source may be reduced, and data collection resources can be allocated to the strata that contribute the most to the overall uncertainty of the estimate. Therefore, source stratification can lead to optimization of the number of samples required to meet the target precision.

Even if there were no bias, the actual estimate of the emission factor would be expected to differ from the true value. First, the estimate is based on less than the total number of sources. Random differences between the set of sampled sources and the population of sources introduce a sampling error. Second, physical measurements have uncertainties. As is indicated previously, the term "accuracy" refers to the closeness of an estimate of a quantity to the true value. Accuracy is a measure of random error plus bias error. The term "precision" refers to random error alone. Even if a process does not produce a bias in the statistical sense described earlier, it is possible for a given segment of the population to be seriously under-represented and another segment to be over-represented by random chance (i.e., by an anomaly in the random selection of sources). The error that results is a larger-than-expected random error; an error from a correct sampling and measurement process is not a bias.

Screening for Bias in the Emission Data Set

An estimate is precise if it has a small random error, regardless of the bias. Suppose, for example, that sources had been selected only from the Gulf Coast, but that a very large number of sources had been sampled. Averaging a large number of emission measurements would lead to an emission factor estimate that had a small random error. Unless Gulf Coast sources were representative of the source type for the entire nation, the estimate could have a large bias because the sample of sources was unrepresentative of the population of sources of interest. The bias in this example, which serves for illustrative purposes, was avoided by sampling in a variety of regions of the country; more subtle potential sources of sampling bias and methods for avoiding them are discussed in this subsection.

Design, operational, and regional parameters that may cause differences in emissions across a source type were identified, and the data were analyzed to determine whether there was an established relationship between those parameters and the emission rate. Usually, these parameters were chosen on the basis of industry expertise and/or engineering judgement. If these parameters were determined to exhibit statistically different emission characteristics, then the population of sources was stratified into distinct categories by these design, operational, or regional parameters. Emission factors and activity factors were determined for each category within the source type to uniquely characterize emissions.

Metering and pressure-regulating stations provide an example where the process of screening for bias was beneficial. Table 3-2 shows the average measured emission factor for metering and pressure-regulating stations, in units of scf/station-hour, on the basis of 86 measurements performed in 19 cities in the United States. Counts of metering and pressure regulating stations were derived from data provided by distribution companies and scaled up to a national count (the activity factor) using the methods described later in Section 3.5. Assuming that the sample selection was random or representative, the extrapolated annual emissions are 104.1 Bscf, based on the average of all measurements

TABLE 3-2. ESTIMATED METHANE EMISSIONS FROM DISTRIBUTION METERING AND PRESSURE REGULATING STATIONS

Category	Location (vault or above-ground)	Emission Factor (scf/station-hr)	Activity Factor (number of stations)	Emissions (Bscf)
All Stations	--	90.2	131,970	104.1
M&R Stations	--	154.1	23,922	32.3
Reg. Stations	--	43.7	108,048	41.4
Total	--		131,970	73.7
M&R Stations				
>300 psig	A-G	179.8	3,460	5.45
100-300 psig	A-G	95.6	13,335	11.2
40-100 psig	A-G	4.31	7,127	0.269
<40 psig	A-G	--	0	0
Reg. Stations				
>300 psig	A-G	161.9	3,995	5.67
>300 psig	Vault	1.30	2,346	0.0266
100-300 psig	A-G	40.5	12,273	4.35
100-300 psig	Vault	0.180	5,514	0.0087
40-100 psig	A-G	1.04	36,328	0.332
40-100 psig	Vault	0.0865	32,215	0.0244
<40 psig	Vault	0.133	15,377	0.0179
Total			131,970	27.3

made to estimate the emission factor. However, if the data set is subdivided, or stratified, by station type (i.e., metering and pressure-regulating versus pressure-regulating), then the annual emissions from this source type decrease to 73.7 Bscf. If this source type is further subdivided by discrete operating pressure ranges and by enclosure status, the emissions decrease to 27.3 Bscf. As illustrated, the bias, which was caused by testing a disproportionate number of high-pressure stations, can be minimized by using stratification to estimate the emission and activity factors.

The screening process served to identify variables that are related to emission characteristics. Then it was possible to determine whether sources were disproportionately sampled in the different strata of these variables. Such a disproportionality need not lead to a bias in the final estimate of the emissions, if this condition is identified and accounted for properly. Moreover, the screening process was carried out during the course of the study. Thus, additional sampling to correct a disproportionality, if present, was possible.

Note that the screening process would identify unrepresentativeness in the sample, whether the problem resulted from an inadvertent bias in the sampling process or a purely random effect. The protection against both bias and anomalies in the random selection of sources is considered to be a significant benefit of the method used in this study.

3.5 Activity Factor Approach

Activity factors are an essential element in the estimation of emission rate by source category. There are many issues pertaining to the estimation of activity factors, however, that are primarily engineering, rather than statistical, in nature. For this reason, and since the subject of this report pertains to the statistical methods used in this study, only a very brief overview of activity factor estimation will be given here. A detailed discussion of activity factors is presented in Volume 5 on activity factors.[1] A much briefer summary of activity factor issues is given in Section 5.3 of Volume 3 on general methodology.[2]

In general, the activity factor is the total population of the source when the emission factor is defined as the annual emissions per source. Exceptions to this general definition of an activity factor would include only sources which have an emission factor that can be more accurately represented by one or more parameters that influence emissions (e.g., the emission factor for IC engines is in terms of annual emissions per horsepower). For these exceptions, the activity factor would apply to the parameter that influences emissions.

In some cases, existing programs track the total nationwide population of a source type, such as gas wells, miles of transmission and distribution pipelines, and total national production within the natural gas industry. However, in many cases, the total population of a source type within the gas industry is unknown. Some of the activity factors that are not tracked nationally were generated by this project.

For sources that have an unknown population, a limited number of site visits were conducted to determine the number of sources at each site and to scale up the site data to represent the total population. These site visits to collect activity factor data were typically conducted in conjunction with the data collection efforts for the emission factor. The site count data were scaled by using population data that were known and were related to the source. For example, no data were available on the nationwide population of production separators. The ratio was computed by dividing (1) the number of production separators at a site, gathered as part of the site visits, by (2) the number of wells at each site. Then the average ratio of separators to wells from all site visit data was used to extrapolate nationally by multiplying by the national well count. However, when scaling the site visit data to represent the entire population, a check for bias was made (refer to the screening-for-bias section below).

For some sources that are not tracked nationally, individual company data or regional surveys (surveys by state agencies or trade organizations) were sometimes available. Metering and pressure-regulating stations, glycol dehydrators, and compressor engines and gas turbines are tracked on a company-wide basis or through regional surveys. For regional or company-tracked activity factors, sufficient company and regional data had to be gathered to comprise a representative sample to extrapolate to a national population. In most cases, entire companies or regions could be represented by the data collected from one sample; therefore, few samples were required, in general, to represent the national population accurately.

The extrapolation of equipment activity factors from individual site data within a stratum was usually handled by selecting an "extrapolation activity factor" that was known for the site as well as regionally or nationally. Examples of extrapolation activity factors are the total production, the number of wells for production, the number of plants for processing, and the number of compressor stations for transmission. Ratios were computed by dividing (1) populations of other equipment, such as the count of separators at the site, by (2) the relevant extrapolation activity factors, allowing the resulting ratios to be easily extrapolated to a regional or national total for separators.

Where individual site data were used to determine a national activity factor, the ratio method was used to compute the activity factors. The general statistical ratio method is discussed by Cochran.[3]

To illustrate the ratio method, consider the example of estimating the number of separators in a region by using well count as the extrapolation parameter. This calculation can be accomplished by (1) summing the numbers of separators at the sites visited, (2) summing the numbers of wells at these sites, (3) dividing the total number of separators by the total number of wells, and (4) multiplying this ratio by the number of wells in the region. Extrapolation by production rather than wells can be performed in a similar manner.

The following hypothetical numerical example illustrates the calculation of the number of separators in a region.

TABLE 3-3. HYPOTHETICAL EXAMPLE DATA OF COMPILATION OF SITES IN REGION X

Site	Site Count of Separators	Site Count of Gas Wells	Site Ratio (separators/well)
1	140	138	1.01
2	324	321	1.01
3	100	100	1.00
4	5	15	0.33
5	500	1000	0.50
TOTAL	1069	1574	

In this hypothetical example, the total number of separators at the sites visited is 1,069, and the total number of gas wells is 1,574. Thus, the number of separators per well, estimated from the data, is 1,069/1,574 = 0.68. Now, suppose there are 50,000 gas wells in the region. Then the estimate of the number of separators in the region is 0.68 × 50,000 = 34,000.

In the ratio method, just described, a site with a large number of wells (e.g., site 5 in the table above) will have a larger effect on the results than will a site with a small number of wells. This method is based on the assumption that the size of the field represented by a sampled site is proportional to the number of wells at the site. The ratio method is described in much further detail in Section A.5 of Appendix A.

An alternate method was also considered. In this method, the site ratios are averaged, so the data from all sites count equally. If the ratios of separators to wells in the hypothetical example above are averaged, a value 0.77 of separators per well is obtained, compared to 0.68 separators per well produced by the ratio method. The main difference is that the ratio method places a much greater weight on site 5, which has a large number of wells and a relatively small ratio of separators per well.

The argument for the alternate method is that there is some uncertainty regarding the size of the field of which a given sampled site is representative. Changes in ownership and leasing agreements can affect how an original field can be subsequently subdivided. Thus, according to this argument, a site with a well count of 15 might be representative of a field of 10,000 wells, while a site with 200 wells might constitute an entire field.

After discussions with the industry advisors, it was decided that the number of wells at a site does provide a measure of the representativeness of a site. For this reason, the ratio method, described above, was employed.

Methods for computing a confidence interval for the number of devices estimated by the ratio method are given by Cochran.[3] The ratio method, including methods for calculation of a confidence interval, are discussed in further detail in Appendix A.

In addition, some equipment activity factors sources could be scaled up by several possible "extrapolation activity factors," called $AF_{(extrap)}$. If a known physical/technical relationship existed between the source population and one $AF_{(extrap)}$, then that factor was selected. However, where the relationship between the source population and the other parameters was not obvious from a technical perspective, many approaches having technical merit were used, and either the average of the methods was used or the resulting data from individual companies were statistically analyzed to determine the appropriate extrapolation approach. Further discussion is given in the Tier 3 methods report.[2]

In the production segment, the two $AF_{(extrap)}$ values are well count and production rate. As is discussed in Section 5.3 of the methods report[2], a tendency was observed for results from the well method to be high-biased and results from the production rate method to be low-biased. Nevertheless, averaging the two values to obtain the final estimate of the activity factor tended to allow the biases with opposite signs to counterbalance.

For example, it was not clear from a technical perspective whether to scale up the number of metering and pressure-regulating stations by miles of main pipeline or system throughput, which were the only known population statistics. The station counts from individual companies were examined both on a per-mile-main and per-system-throughput basis. A linear regression analysis showed that the data would preferentially be extrapolated using a per-mile-main basis, with lower variability in the resulting national extrapolation. In production, the number of separators appears to be technically related to both well count and throughput. Therefore, separator count was extrapolated by both methods, and the average of the two national estimates was used.

3.6 Summary Comments Regarding Screening for Bias

It is impossible to prove technically that a given dataset has no bias. Tests can be designed that are capable of revealing some bias, but there are no tests nor group of tests that would reveal all possible biases. Assuming that a given dataset has no bias, even after extensive testing, is only a theory. The following examples in this section show some of the many bias tests used in this project.

The sample sets were tested for bias by continuous technical and industry review. Numerous individual reviews and project advisors' meetings were used to review the project data with knowledgeable industry experts, so that systematic errors could be discovered and eliminated. When possible biases in the activity factor sampling plan or extrapolation method were theorized, the project was altered to test for that bias and eliminate it if it existed. All provable biases were corrected.

One example of the success of this bias review process includes the identification of regional differences in production practices. These differences were brought up by the advisors' meeting review process. The differences were then accounted for by stratifying the production data into two offshore and four onshore regions, sampling within each region, and extrapolating by region.

Some emission factor biases are eliminated by stratifying by an emission-affecting parameter. Specific examples are discussed earlier in Section 3.4. Some specific examples of eliminating activity factor bias are listed in Section 3.5.

4.0 ESTIMATION OF NATIONAL ANNUAL EMISSIONS AND UNCERTAINTY

Sampling to obtain the data necessary to estimate the activity factors and emission factors for each source category is discussed in the preceding section. Subsequent calculations leading to the estimation of the industry national emission rate and the uncertainty of this emission rate are discussed in this section.

These calculations involve several steps. First, the available data must be used to estimate the activity factors and emission factors for individual categories. Most of these estimations are made by averaging a set of values (measurements of emission rates, counts of emitters, etc.) to obtain the necessary activity or emission factor. It is necessary to avoid corrupting any such calculation by the presence of an invalid data point. Issues pertaining to outliers are discussed in Section 4.1. The calculation of the average value and uncertainty thereof to obtain a given activity or emission factor is discussed in Section 4.2. The use of these values to obtain an estimate of the annual emissions and uncertainty thereof for each source category is discussed in Section 4.3. Finally, the use of the annual emissions by category to obtain the national annual industry emissions is discussed in Section 4.4. The calculation of the uncertainty of this industry total is also discussed in Section 4.4. Further issues pertaining to error propagation are discussed in Section 4.5.

4.1 Outlier Tests

Radian did not reject any data points as outliers. However, outlier tests were performed in the distribution area. This section discusses those tests.

In the following section, the use of data to estimate the emission factor or activity factor for a source category is discussed. Suppose, for example, there are n measurements, y_i, $i=1$ to n, which are to be averaged to obtain the emission factor for a particular category. It is necessary to confirm that this data set does not contain an erroneous value that is so extreme that it invalidates the calculations.

It is possible to perform statistical tests to determine whether there is strong reason to believe that a suspected outlier could not reasonably belong to the same distribution as the other points. Even if the point was judged to be an outlier from a strictly statistical point of view, it would be very desirable to examine the point from an engineering perspective to ensure that this point did not, in fact, contain valid information.

Further, in this study, there are 86 source categories. Suppose it was decided to routinely perform outlier tests for 86 sets of emission factor measurements, for example. Suppose that the confidence level of the test was 99%, i.e., that there was a 1% chance of erroneously concluding there was an outlier. This erroneous conclusion would occur if a valid point was rejected, even though it resulted from the same statistical distribution as the other points. Valid points that appear to be significantly larger or smaller than the other points can occur by chance alone. As is discussed above, it is desirable to avoid rejecting such points if they are valid and contain important information.

It may appear that the 99% confidence level is sufficiently conservative; i.e., this confidence level appears to provide a small probability of discarding points erroneously. Consider, however, the effect of performing 86 independent tests in the case in which there were no invalid data points in any category. The probability of correctly concluding that there were no outliers in a single test would be 0.99. The probability of correctly concluding that there were no outliers in all of n independent tests would be 0.99^n. The probability of erroneously concluding that there was an outlier in at least one of the n tests, then, would be $1 - 0.99^n$. In 86 independent tests, the probability of erroneously concluding that there was at least one outlier would be 0.58; this is a high probability of error. Even if only half this number of tests were performed, the probability of erroneously concluding that there was an outlier in at least one of the categories would be 0.35, which is still high. These calculations illustrate the reason for caution regarding the blind use of outlier tests for all categories.

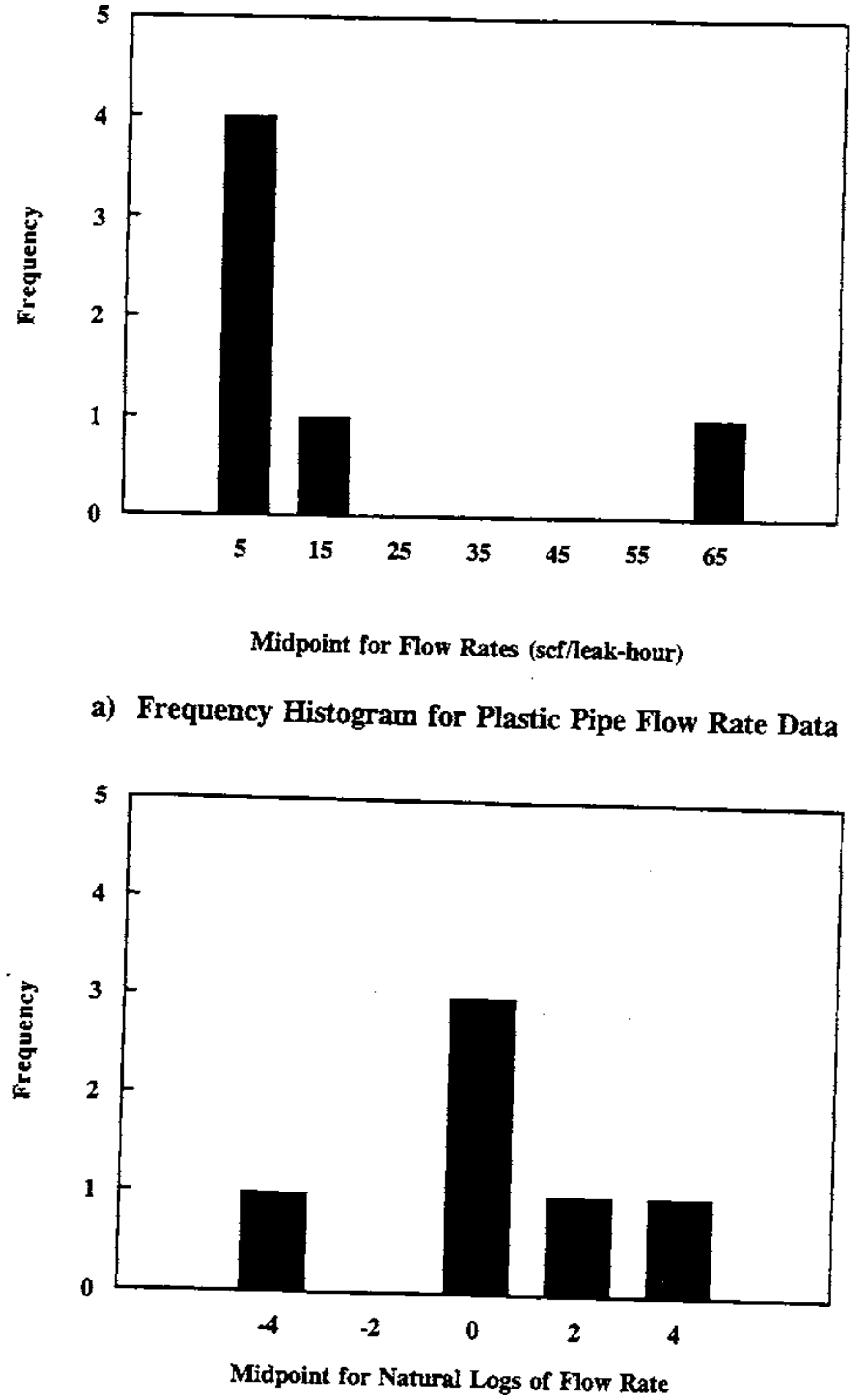
The 99% confidence level was selected for illustration in this example because it is conservative. If the 90 or 95% level were used, the probability of erroneously concluding that there was at least one outlier in the 86 categories would be much higher.

Moreover, the "outliers" may be the most important points in the data set. It would be unfortunate if valid data points corresponding to high emitters were rejected from the data base, since this could lead to a low bias in the final estimate of the industry annual emissions.

Figure 4-1a presents the histogram of the emission-rate measurements for large plastic mains. Notice that one data point is noticeably larger than the others. There has been some discussion about this data point regarding its validity and whether it should be excluded from the data set. Figure 4-1b presents the histogram of the natural logarithms of the emission rate measurements.

This data set illustrates several limitations regarding the performance of outlier tests for many of the categories. First, the data set is small, containing only six points. Any statistical test involving only six points is likely not to be very sensitive.

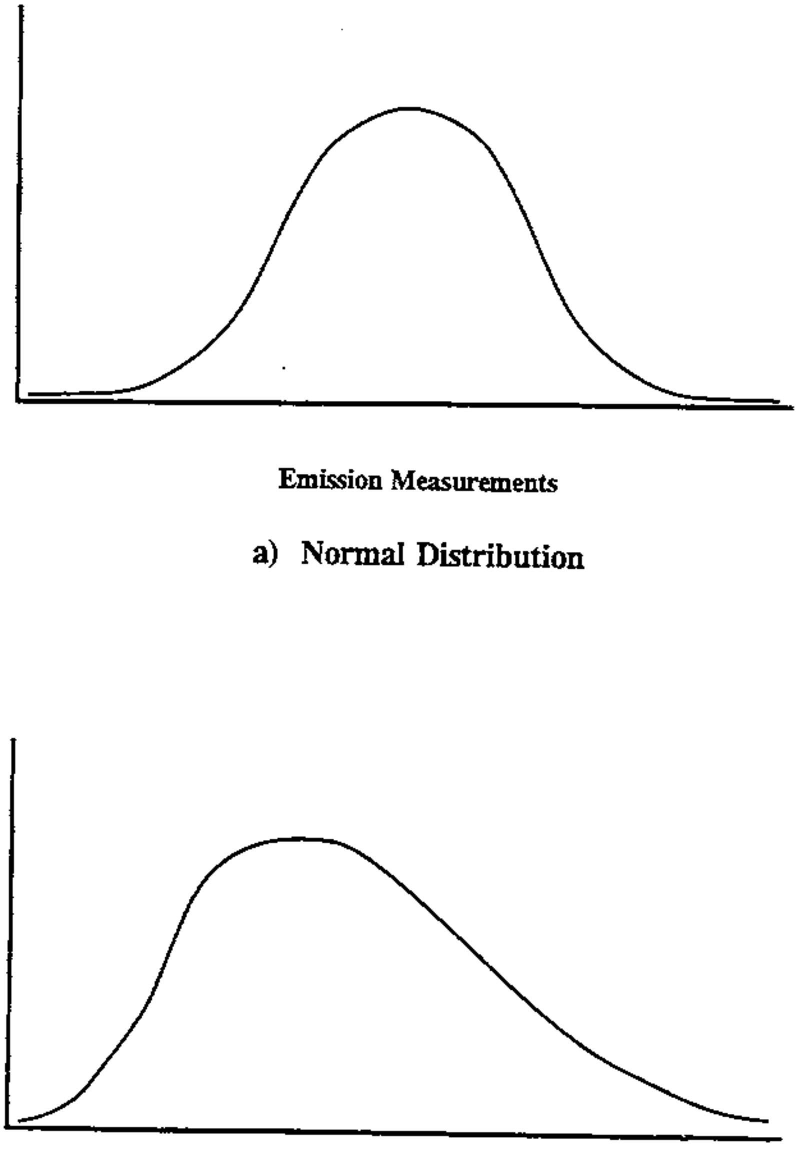
Second, most outlier tests depend on the type of statistical distribution. That is, one must assume a specific type of distribution in order to perform most tests. Figure 4-2 presents a conceptual comparison of the normal and the lognormal distributions. The normal distribution is symmetric; i.e., the likelihood that a value will occur at a given distance above the mean is the same as the likelihood that a value will occur at the same distance below the mean. The lognormal distribution is asymmetric. There is a predominance of points roughly in the vicinity of the mean, with a small number of much larger points. The distribution is bounded below by zero. While a few points may be much larger than most, there is not a corresponding chance for points much smaller than most.



a) Frequency Histogram for Plastic Pipe Flow Rate Data

b) Frequency Histogram for the Natural Logarithms of the Plastic Pipe Flow Rate Data

Figure 4-1. Frequency Histograms for the Emission Rate Data for Plastic Pipes



a) Normal Distribution

b) Lognormal Distribution

NOTE: The ordinate of these curves is a mathematical quantity called "probability density." The probability density can be used to obtain the probability that the variable falls within any given limits.

Figure 4-2. Conceptual Comparison of Normal and Lognormal Distributions

A statistical test indicated that the six emission rates for large plastic mains could not reasonably have been drawn from a normal distribution but could reasonably have been drawn from a lognormal distribution. One could debate the outcome of this test, however, since 1) the suspected outlier was included in the test, and 2) this large value would favor the lognormal over the normal distribution. With a larger sample size, the largest value would have a smaller weight in the statistical test. In any case, it is difficult to determine the type of distribution with high confidence with only six data points.

Tests were performed, however, for underground pipes with different types of materials. These tests were based on 24 to 40 data points and, therefore, provided a better opportunity to determine the type of statistical distribution. In these cases, the tests indicated that the data could not reasonably have come from a normal distribution but could reasonably have come from a lognormal distribution. Given the similar source type (underground pipes), these results support the conclusion above, to use the lognormal distribution in the outlier tests for the data for large plastic mains.

Several statistical tests were performed to determine whether the largest point in the data set should be considered an outlier from a statistical point of view. These include the Grubbs test,[4] the Dixon test,[4] the fourth-spread test,[5] and a conservative approach.[6] The Grubbs and Dixon tests require that the data be normally distributed. Given the outcome of the distributional tests discussed above, these tests were performed using the natural logarithms of the emission rate data; if the data are lognormally distributed, then the natural logarithms are normally distributed. The conservative approach does not require an a priori assumption regarding the distribution but incorporates a distributional test as a first step. Thus, the conservative approach was ultimately based on the lognormal distribution also.

The fourth-spread method relies on information from the center half of the distribution, from which limits beyond which data could be considered outliers are derived. The point here is that the center half of the distribution is relatively insensitive to outliers and, therefore, provides an effective basis for determining upper and lower limits beyond which data might be considered suspect. The fourth-spread method does not require an explicit assumption regarding the type of distribution. Since this method works best if the distribution is symmetric, the natural logarithms were again used.

None of the tests indicated that the largest point was an outlier. Further, Pacific Gas and Electric Company's (PG&E) statistician, who worked on the UAF study,[7] agrees that there is no technical or statistical justification for omitting the largest data point in this data set. Thus, the point has been retained in the data set. The details of the statistical tests are presented in Appendix B.

4.2 Emission Factor and Activity Factor Calculations

The following basic statistical calculations were performed for emission factors. A different and more complex approach, described briefly in Section 3.5 and in more detail in Appendix A, was used for activity factors. Suppose there are n individual estimates of a given emission factor. If y_i, $i=1$ to n, are the individual data points, then the factor is estimated as the average, $\bar{y}$, of the n values:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3}$$

The next step is to compute the uncertainty of this value. First, the standard deviation of, s_y, the y values is needed:

$$s_y = \frac{\sqrt{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}}{n-1} \tag{4}$$

We then calculate a 90% confidence interval for the mean value, $\bar{y}$. The confidence interval establishes lower and upper tolerances for the estimate. There is only a 5% chance that the true value falls below the lower limit of this confidence interval. There is also a 5% chance that the true value falls above the upper limit of the interval. Thus, there is a combined 10% chance that the true value falls outside the confidence interval. Since there is a 90% probability that the true value falls within the interval, it is called a 90% confidence interval. The 90% confidence interval is computed as follows:

$$\bar{y} \pm ts_y/\sqrt{n} \tag{5}$$

The t value in this equation is obtained from a standard table for the t-distribution; such tables are found in most basic statistics books. The t-value is a function of the confidence level (90% in this case) and the sample size, n.

The confidence interval computed above is strictly valid for normal populations. Even if the distribution of y values on which $\bar{y}$ is based is not normal, the average of a large enough sample of values of y_i would be approximately normally distributed; the theorem on which this is based is called the central limit theorem. If the mean value is approximately normally distributed, then the above method for computing a confidence interval is justified.

While a sample of size n produces a single mean value, it is proper to discuss the distribution of this mean value. The mean value, being based on a sample of values with random errors, is itself a random variable. The error of this mean may have a distribution that is approximately normal.

The methodology described above produced uncertainties larger than 100% for some parameters. This seems anomalous, since neither an activity factor nor an emission factor can be negative. The explanation for this effect and the reasons why the methods used are believed to be correct for estimating the uncertainty of the industry annual emissions are presented in Section 4.5, after further discussion of the issues.

The method described above for calculating uncertainty is strictly applicable for an infinite population. In fact, the number of sources in a given category is finite. The equation for the standard deviation of a finite population produces a smaller value than does the equation above for an infinite population. Thus, the method used produced a somewhat conservative (large) estimate of the uncertainties involved.

Both sampling and measurement errors contribute to the error in the estimate of the emission factor. The sampling error pertains to a finite set of sources. The measurement error pertains to an infinite population. (There is no limit on the number of replicate measurements that could be made for a given source.) The source-to-source variability is generally larger than the measurement variability, however. Thus, the statement above stands; the equation used for the standard deviation produced a somewhat conservative estimate of the uncertainties of the parameters involved.

4.3 Category Annual Emission Calculations

For most source categories, the emission value (ER) is expressed as the product of the activity factor (AF) and the emission factor (EF):

$$ER = AF \times EF \tag{6}$$

For certain source categories in the distribution segment, the emissions were estimated directly, and no separate activity and emission factors are shown in the data summary table (Appendix C). In each of these cases, several subcategories were combined to form a category. The emissions for these subcategories were summed to obtain the emissions for the category shown in the summary table.

It is necessary to obtain the uncertainty of the emission value as a function of the uncertainty of the activity and emission factors. The error propagation methods used here are based on theorems given by Mood, Graybill, and Boes[8] and quality-control practices described by Juran, et al.[9] The details of the error propagation methods are discussed further in Appendix A.

The tolerance (i.e., uncertainty) for ER as a function of the tolerance for AF and EF is as follows:

$$\mathrm{Tol(ER)} = [\mathrm{AF}^2 \times \mathrm{Tol(EF)}^2 + \mathrm{EF}^2 \times \mathrm{Tol(AF)}^2 + \mathrm{Tol(AF)}^2 \times \mathrm{Tol(EF)}^2]^{1/2} \quad (7)$$

where Tol() is the tolerance of the indicated quantity. The tolerance is the half-width of the 90% confidence interval. That is, if the confidence interval is given by

$$\bar{y} \pm ts_y/\sqrt{n} \quad (8)$$

then the tolerance is

$$\mathrm{tol}(\bar{y}) = ts_y/\sqrt{n} \quad (9)$$

Recall that confidence intervals are discussed in the preceding subsection.

4.4 Industry Annual Emission Calculations

In this subsection, three topics are covered. First, the equation for calculating the industry annual emissions is presented. Second, methods for computing the uncertainty of the industry total are discussed. Third, the effect of correlated errors for different source categories is discussed.

Industry Annual Emissions

The next step is to compute the industry annual emissions, ER_T, and its uncertainty. The industry annual emissions are simply the sum of the emissions for the different categories:

$$ER_T = \Sigma ER \quad (10)$$

Methods for Computing the Uncertainty of the Industry Total

The tolerance of ER_T is required as a function of the tolerances of the individual ER values. First, the calculation is considered on the basis of the assumption that the errors in the different ER values are independent. The errors in two quantities would be considered independent if they were estimated by entirely separate processes and there was no common source of error. The errors in two quantities would be dependent if they had a common source of error. The issues related to correlated errors are discussed later in this subsection.

The tolerance of ER_T is the square root of the sum of squares of the tolerances of the ER values:

$$\mathrm{Tol}(ER_T) = [\Sigma\{\mathrm{Tol}(ER)\}^2]^{1/2} \tag{11}$$

Method 1

This is the method for calculating the tolerance of a sum that is recommended by Juran, et al., in the *Quality Control Handbook*.[9] On the basis of the discussion by Juran and more rigorous statistical information presented by Mood, Graybill, and Boes,[8] the use of this method does not require the assumption that the separate terms in the sum have the same means, the same uncertainties, or even the same types of statistical distributions; again, see Appendix A for further details. Method 1 was used in this study.

An alternative to the method above is to express the tolerance of a sum as the sum of the tolerances:

$$\mathrm{Tol}(ER_T) = \Sigma\mathrm{Tol}(ER) \tag{12}$$

Method 2

However, this is overly conservative (overestimates the uncertainty), and was therefore not used. An analysis of this alternate method appears in Appendix A.

It can be shown that the tolerance of the industry total would equal the sum of the category tolerances if the errors for all categories were perfectly correlated. While there may be some cross-category correlations, there are many pairs of categories whose errors could not reasonably be correlated at all. For example, it is reasonable to assume that the errors in the transmission/storage categories are uncorrelated with the errors in the distribution categories. No pair of categories has perfectly correlated errors. The issues pertaining to the possibility of correlated errors among categories is addressed below.

Effect of Correlated Errors

It is mentioned above that it is not believed that the emissions for all pairs of categories are strictly independent. An analysis has been performed to assess the possible impact of correlated errors. The results of the analysis are outlined in Section 5.1 and show that the target precision was still met with correlated errors. This section outlines technical issues associated with errors.

First, certain categories have common activity factors. For such categories, the activity factors have the same errors, although the emission factors have independent or imperfectly correlated errors.

For any of a variety of other reasons, there may be correlations between the errors in the emissions for different categories. Data for different categories were collected from the same fields in some instances. It is possible, because of some characteristic of the field, that nonindependent data resulted for two or more categories. For example, the inspection and maintenance practices used for a particular field may have been significantly better than the industry average. Consequently, emissions may have been significantly lower than the industry average for all source categories for which data were collected from that field. Deviation from the industry average is a sampling error in this context, since the objective is to estimate the industry average for each category.

Nevertheless, there is typically a large source-to-source variability within a given field. For this reason, one could not say that all sources tested at a given field had a common error or even similar errors. This source-to-source variability tends to limit the correlation introduced by characteristics of a field that have a common effect on two or more categories sampled at the field. Thus, it is not believed that the correlation between errors in two categories could reasonably be very high, but the exact correlations are not known.

It is not believed that all possible source categories could reasonably have correlated errors. Categories in different segments (production, transmission/storage, and distribution segments) were assumed in most cases to have uncorrelated errors. Additionally, not all categories within a segment could reasonably have correlated errors.

The uncertainty of the industry annual emissions would be smaller if (1) the errors in the emissions for all source categories were independent than if (2) some positive correlations existed among these errors. In the first case, the maximum possibility for errors to "average out" when the emissions for 86 source categories were summed would exist. In the second case, the nonindependence of certain pairs of errors would diminish this effect.

It is stated earlier that, by the preferred method of calculation, the uncertainty of the national emissions is the square root of the sum of squares of the uncertainties of the emissions for all categories:

$$\mathrm{Tol}(\mathrm{ER}_T) = [\Sigma\{\mathrm{Tol}(\mathrm{ER})\}^2]^{1/2} \qquad \textbf{(13)}$$

If the errors are not independent, the uncertainty of the national annual emissions is increased by the addition of a term to account for each pair of categories with correlated errors:

$$\mathrm{Tol}(\mathrm{ER}_T) = [\Sigma\{\mathrm{Tol}(\mathrm{ER})\}^2 + \text{other terms}]^{1/2} \qquad \textbf{(14)}$$

If categories i and j have emissions with correlated errors, then the term that is added to account for this correlation is:

$$2r_{i,j}Tol(ER_i)Tol(ER_j) \qquad (15)$$

where

$r_{i,j}$ = correlation between the two errors,

$Tol(ER_i)$ = uncertainty of the emissions for category i, and

$Tol(ER_j)$ = uncertainty of the emissions for category j.

The correlation coefficient is a measure of the closeness of the linear relationship between two random variables (here, the errors in two emission values). If there was no association at all between the two variables, the correlation coefficient would be zero, and the added term would also be zero. If the two variables were perfectly linearly related, the correlation would be one. If the relationship between the variables were such that half the variance of one could be explained, or predicted, in terms of the other, the correlation would be approximately 0.7. Plots illustrating correlation levels considered in this analysis are presented in Section 6.1.

Negative correlations exist if one variable tends to increase as the other decreases. There is no apparent reason in this application why emissions from two categories would have negatively correlated errors, however. The reasons discussed earlier for correlated errors pertain to positive correlations. Suppose, for example, that the same activity factor was used for two source categories. If this activity factor had a positive error, the effect would be to make the emissions for both categories too large. If the activity factor had a negative error, the effect would be to make the emissions from both categories too small.

One further comment will be made regarding the interpretation of the term used to account for the correlation between the i^{th} and j^{th} errors. This comment can be skipped, but it is included here for completeness and to clarify an issue regarding the analysis discussed above. Readers familiar with analysis of error propagation may have expected to see the covariance between the errors in place of the following term:

$$r_{i,j}\mathrm{Tol}(\mathrm{ER}_i)\mathrm{Tol}(\mathrm{ER}_j) \qquad \textbf{(16)}$$

The covariance, like the correlation coefficient, is a measure of the strength of the linear association between two variables. The term used is analogous to the covariance, except that uncertainties (half-widths of 90% confidence intervals) appear above in place of standard deviations. That is, if the tolerances of ER_i and ER_j were replaced by the standard deviations of the errors in these quantities, the expression would become the covariance. The use of uncertainties (in place of standard deviations of errors) in the analysis of error propagation throughout accounts for certain effects of the finite sample sizes used to estimate the different parameters and is conservative, i.e., tends to produce larger estimates of uncertainty than alternative approaches. The mathematical reasons for this are discussed in some detail in Appendix A.

The groups of source categories with correlated activity factors are given in Appendix C, as are the groups of source categories with correlated emission factors. The correlation coefficients are also presented. In terms of these quantities, the expression used to account for correlated errors is as follows:

$$2r_{ij}Tol(ER_i)Tol(ER_j) =$$

$$2\{AF_iAF_jr_{Eij}Tol(EF_i)Tol(EF_j)+EF_iEF_jr_{Aij}Tol(AF_i)Tol(AF_j)+$$

$$r_{Eij}Tol(EF_i)Tol(EF_j)r_{Aij}Tol(AF_i)Tol(AF_j)\} \qquad \textbf{(17)}$$

where

r_{Eij} = correlation between the errors in the emission factors for the i^{th} and j^{th} categories, and

r_{Aij} = correlation between the errors in the activity factors for the same two categories.

The quantities AF and EF are used earlier, and the subscripts were required here to indicate the two separate source categories involved. The covariance term on which the expression on the right of the equals sign above is based can be derived rigorously from relationships given by Mood, Graybill, and Boes[8].

4.5 Issues Related to Statistical Distributions

In this subsection, certain issues that affect the calculation of error bounds are covered. In Section 4.2, it is mentioned that the methodology used produces an uncertainty larger than 100% for the activity factors or emission factors for some source categories. The reasons why this occurred and the justification of the methods used as a basis for estimating the uncertainty of the industry annual emissions are discussed in this subsection.

One possible reason for the uncertainties greater than 100% is as follows. The population of y_i values on which the mean value of the activity or emission factor for a given category was based may not have been normal in these cases, and the sample size may not have been sufficient to produce an estimated value whose error was approximately normally distributed. The activity factors and emission factors calculated produced emission values with uncertainties greater than 100% for several categories. Even if the data were normally distributed, but highly variable, an uncertainty of greater than 100% could have resulted from the small sample sizes that exist for some source categories. However, the sum of the emissions whose individual uncertainties were over 100% totaled less than half (about 40%) of the industry annual emissions (see the summary table in Appendix C).

The sample sizes for some of these categories were small. This is because it was not consistent with the overall goal of the program to spend large amounts of money refining the emissions of a source category that contributed a very small amount to the industry annual emissions. It was advantageous to devote more resources to categories that contributed a greater amount to the industry emissions; these issues are discussed earlier with regard to target precisions by category.

It is generally true, however, that a sum is more nearly normally distributed than are the individual terms in the sum. This statement is loosely based on the central limit theorem, mentioned earlier, which strictly applies to sums of identically distributed random variables. Thus, the sum of the emission rates with uncertainties greater than 100% will tend to be more nearly normally distributed than are the individual terms in that sum. The sum of the emission rates for all 86 source categories will tend to be more nearly normally distributed still. Moreover, the sample sizes for the source categories with larger emissions tended to be larger, and, thus, parameter estimates for these categories tended to be more nearly normally distributed for this reason.

In Section 4.2, it is indicated that, although a sample of size n produces a single mean value, it is proper to speak of the distribution of this mean. This is because the mean is based on data that have random errors, and so the mean is affected by random variability. Thus, the mean may be approximately normally distributed or may have some other distribution. Similarly, a single value serves as the estimate of the national methane emissions. By analogous reasoning, however, it is proper to talk about the statistical distribution of this value, since it is affected by random variability in the estimates of the activity factors, emission factors, and annual emissions for the categories.

Thus, even though the methodology described above may not produce a valid confidence interval for all activity and emission factors for the smaller source categories, these observations do not invalidate the methodology for the purpose of estimating the uncertainty of the industry annual emissions, which is the objective of this study. There are reasons for believing the industry annual emission value has an error that is approximately normally distributed. Rigorously proving that this is the case is not possible without knowing the distributions of the errors for the individual categories, and definitively establishing these distributions is not possible on the basis of the small sample sizes for some categories.

5.0 SUMMARY OF STATISTICAL ASSUMPTIONS

In the preceding sections, statistical issues and methods used to address sampling requirements for different types of sources, sampling requirements specific to the estimation of activity factors and to the estimation of emission factors, the analysis of error propagation, etc. are discussed. In this section, the major statistical assumptions discussed in the preceding sections are summarized.

Two major statistical assumptions are discussed that affect the calculation of the uncertainty of the industry emission rate for the baseline case. One assumption is that the error in the industry emission rate is normally distributed. Another assumption is that the errors in parameter estimates for different source categories are independent (or that the effects of any correlations present are negligible).

Calculation of the uncertainty of the industry emission rate based on these assumptions can be performed in a clearly defined manner. This is not the case if the assumptions are not satisfied. If the distribution is not normal, then the distribution is not known. (However, the lognormal distribution provides a very conservative possibility, and a result midway between the results produced by the normal and lognormal assumptions provides a more reasonable conservative outcome.) If the intercategory correlations are not zero, accurate estimates of the correlations do not exist, but approximate correlations can be assigned to specific cases on the basis of engineering judgement.

As is discussed in Section 6, calculations for conditions contrary to the baseline assumptions have been performed. These alternate calculations provide (1) an assessment of the sensitivity of the results to deviations from the baseline assumptions and (2) conservative (larger) estimates of the uncertainty of the industry annual emissions.

Assumptions regarding normality are discussed in Section 5.1. Assumptions regarding independence of errors among categories are discussed in Section 5.2.

5.1 Normality of Errors

In Section 4, the analysis of error propagation is discussed. The estimation of an activity or emission factor for a source category in most cases involved averaging a set of emission measurements from individual sources, counts of emitters from different sources of information, etc. In some instances, the uncertainty was assigned on the basis of engineering judgement.

Confidence intervals were computed for both activity factors and emission factors on the basis of the assumption that the data averaged were normally distributed. Even if the data are not normally distributed, their mean will tend to normality as the sample size increases (by the central limit theorem). Thus, for a sufficiently large sample size, methods based on the normal distribution can be used, even if the individual data averaged are not normally distributed.

In some cases, uncertainties for specific activity or emission factors have greater than 100% uncertainties, based on the 90% confidence intervals. This seems anomalous, since neither the activity factor nor the emission factor can be negative. One possible reason why this occurred is believed to be because the data on which these estimates were based were not normally distributed, and the sample sizes were not sufficient to produce estimated values that were approximately normally distributed. In some instances, the emissions computed from the activity and emission factors for an individual source category had greater than 100% uncertainty.

However, a non-normal distribution of the data is not the only possible reason why uncertainties greater than 100% could have occurred. Wide confidence intervals can occur because of small sample sizes together with a high degree of variability, even if the data are approximately normally distributed; further, there are small samples sizes in some source categories. Thus, these considerations contributed to the large uncertainties for some parameters, including the uncertainties greater than 100% in some cases.

Another consideration is that the normal distribution is unbounded below, while activity factors, emission factors, and annual emissions cannot be less than zero. Thus, while the distribution of the estimate of one of these quantities may be approximately normal, it cannot be exactly normal. For some sufficiently high confidence level, the confidence interval will extend below zero. Suppose, for example, that an estimate based on a sample of size 10 is 10,000, and the 80% confidence interval is (3,000, 17,000). The 90% confidence interval then would be (755, 19,245), which is still above zero. However, the 95% confidence interval, (-1,367, 21,367) extends below zero. The units of the numbers in the illustration presented here have been omitted, since the principle applies to the estimation of activity factors, emission factors, annual emissions for a category, and the national annual emissions.

The final result of this study, however, is the annual emissions for the entire natural gas industry. The industry annual emissions are the sum of the emissions for 86 individual source categories. First, the categories with individual emissions greater than 100% produce approximately 40% of the industry's emissions (see the summary table in Appendix C). The distribution of a sum tends to be more nearly normally distributed than are the terms in the sum. Thus, the sum of the emissions for the source categories with uncertainties larger than 100% is more nearly normally distributed as a result of the summation. Further, the relative error in this sum is reduced by an "error averaging effect." The sum of the emissions for all 86 source categories is more nearly normally distributed still. As a result of these and other considerations discussed in Section 4, it is believed that the methods used are reasonable for characterizing the uncertainty of the estimate of the industry annual emissions.

5.2 Independence of Errors Among Categories

Given the uncertainties of the emissions for individual source categories, it is necessary to compute the uncertainty of the industry annual emissions. The industry annual emissions (ER_T) are the sum of the emissions (ER) for the categories:

$$ER_T = \Sigma ER \qquad (18)$$

Thus, the uncertainty of the industry annual emissions requires analysis of the error propagation in a sum. As is discussed in Section 4, the uncertainty of a sum is the square root of the sum of the uncertainties of the terms in the sum, if the terms are independent.

The uncertainty of the sum was computed on the basis of the assumption that the errors for the different source categories are independent. An analysis was also performed to assess the impact of correlations for source categories that could reasonably have non-independent errors. Additionally, the possibility that the error in the industry annual emissions is not normally distributed was addressed. The results of this analysis are discussed in Section 6.

6.0 RESULTS PERTAINING TO THE ATTAINMENT OF THE TARGET ACCURACY

Section 6.1 presents an assessment of the uncertainty in the industry annual emissions under various assumptions. Section 6.2 presents hypothetical calculations designed to illustrate how the target accuracy can be satisfied in the presence of large random and bias errors.

6.1 Uncertainty in National Annual Emissions Under Various Assumptions

In the preceding sections, issues pertaining to the distribution of the errors and to independence or nonindependence of the errors in the emissions for different categories are discussed. An assessment has been performed of the sensitivity of the uncertainty in the estimate of the national annual emissions under different assumptions regarding these issues. It is shown that the target precision, 0.5% of national production, is achieved under any reasonable set of assumptions. While the primary purpose of this report is to present the statistical methods, rather than results as such, these results are relevant to the statistical methods and are presented here.

Calculations Under Baseline Assumptions

For the remainder of this subsection, "uncertainty" refers to the uncertainty of the national annual emissions unless otherwise indicated. In the baseline case, the uncertainty was calculated on the assumption that the error in the national annual emission value was normally distributed, and errors in the emissions for different categories were independent.

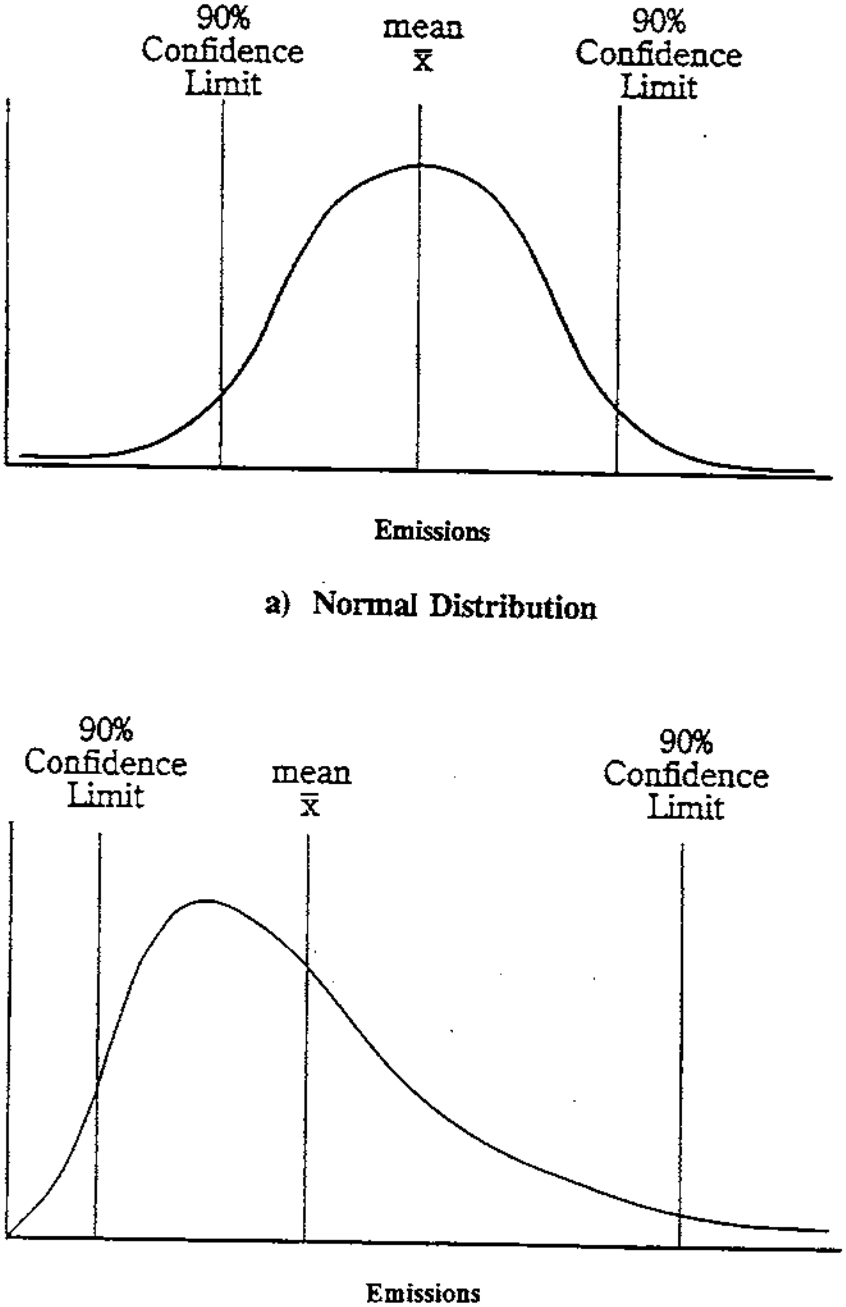
Calculations Under Alternate Assumptions

Additionally, the uncertainty was computed assuming that the error in the industry annual emission value was lognormally distributed. In this case, the standard error (i.e., standard deviation of the error) in the industry annual emissions was held constant, but the confidence interval was recalculated on the basis of a lognormal assumption. This is illustrated conceptually in Figure 6-1. Under the normal assumption, the confidence interval is symmetric about the estimated value. Under the lognormal assumption, the confidence interval is asymmetric. The lower limit is nearer the estimate than is the upper limit.

Under the normal assumption, the uncertainty could be expressed as either (1) the estimate minus the lower confidence limit or (2) the upper limit minus the estimate; the result is the same. Under the lognormal assumption, the latter uncertainty estimate is larger and quantifies the uncertainty of the estimate on the high side; this larger uncertainty estimate was used in the lognormal case. Further discussion of the relationship between the normal and lognormal distributions is given in Appendix A.

An assessment of the effect of correlated errors was also made. Source categories were identified that had either activity factors or emission factors believed to have correlated errors. These two types of errors were handled formally to derive the correlation between the errors in the emission rates for each pair of source categories. The groups of source categories with correlated activity or emission factors are shown in Appendix C.

Several levels of correlation were considered, including the following: weak (correlation coefficient of 0.2), medium (correlation coefficient of 0.5), strong (correlation coefficient of 0.8), and perfect (correlation coefficient of 1.0). A perfect correlation would exist between the errors in the activity factors for two categories if the same activity factor applied in both cases.



a) Normal Distribution

b) Lognormal Distribution

Figure 6-1. Conceptual Comparison of Normal and Lognormal Distributions with Confidence Intervals for the Mean

The correlation level was considered more uncertain in a limited number of cases. If a correlation was considered weak to medium, it was assigned a value of 0.3. If a correlation was considered medium to strong, it was assigned a value of 0.6.

To illustrate the meaning of the different levels of correlation, random samples of size 100 were generated for two variables which were both normally distributed. Figures 6-2, 6-3, and 6-4 illustrate the cases in which the true correlation is 0.2 (weak correlation), 0.5 (medium correlation), and 0.8 (strong correlation), respectively. When the correlation is 0.2, the plot of y versus x reveals a suggestion of a trend, but there is so much scatter about this trend that it is hard to discern visually. When the correlation is 0.5, there is still a lot of scatter, but the trend is apparent visually. When the correlation is increased to 0.8, the trend becomes much more clearly defined, but there is still some scatter about the trend line. In each case, the trend line (i.e., the line of "best fit," or regression line) is displayed. When the correlation is 1.0 (not shown), the two variables are perfectly linearly related; i.e., all points fall on a straight line.

Results

The results of applying the assumptions discussed above to the national emissions are shown in Table 6-1. Under the baseline case, the uncertainty is 90.4 Bscf, or 0.4% of production; under these assumptions, the target production of 0.5% of production is satisfied. The uncertainty increases somewhat when lognormal errors, correlated errors, or both are introduced.

The uncertainty is under 0.5% of production in all cases except when both the lognormal distribution and correlated errors are both introduced. In this case, the uncertainty exceeds 0.5% very slightly (by 0.007%). The lognormal assumption is considered to be excessively conservative, however, in view of the amount of averaging (averaging of data to obtain activity and emission factors for individual categories) and summing (summing of emissions for 86 categories) to obtain the industry annual emissions. Further, as discussed,

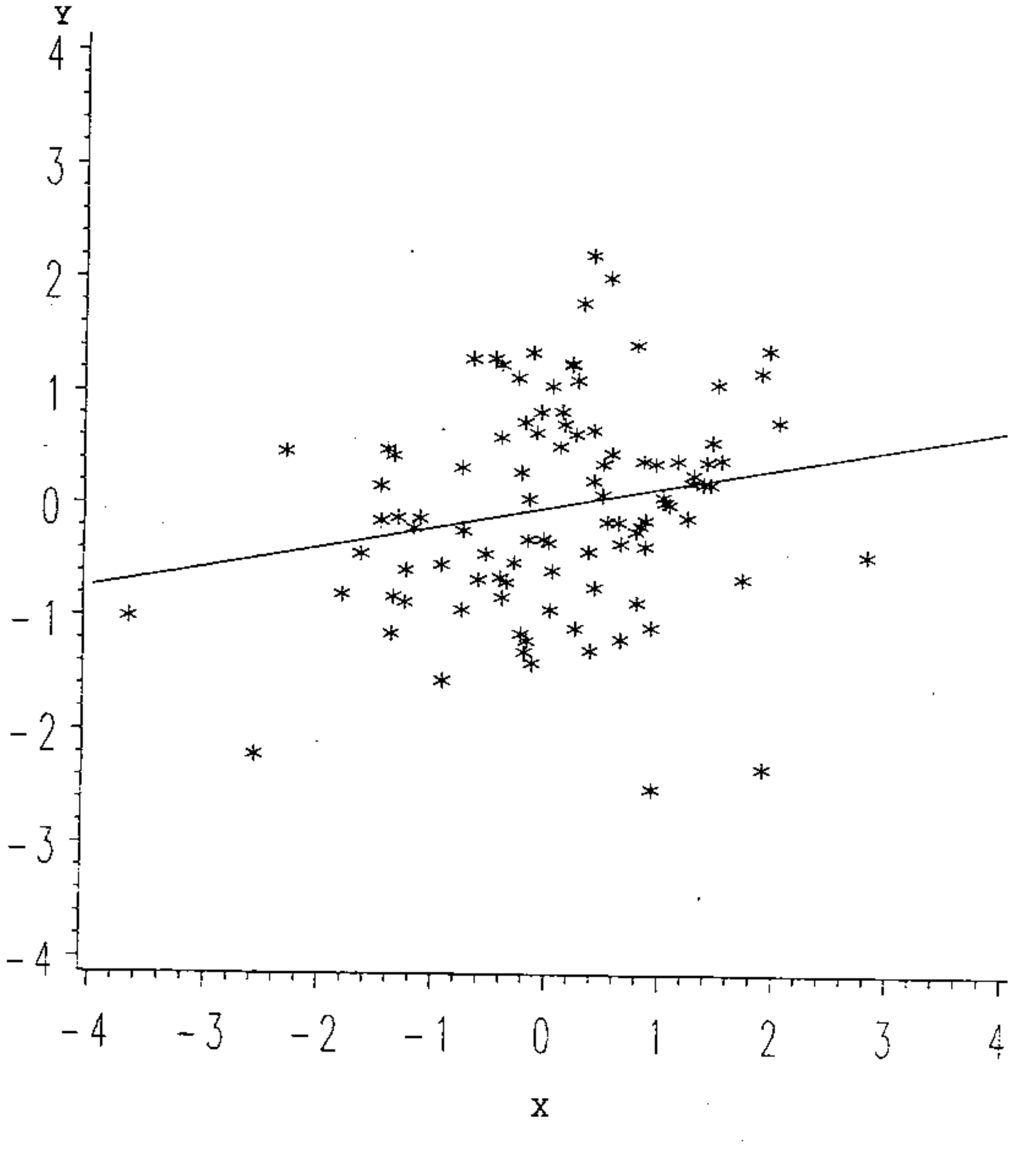


Figure 6-2. Sample from Bivariate Normal Population with Correlation Coefficient = 0.2, Regression Line Shown, Sample Size = 100

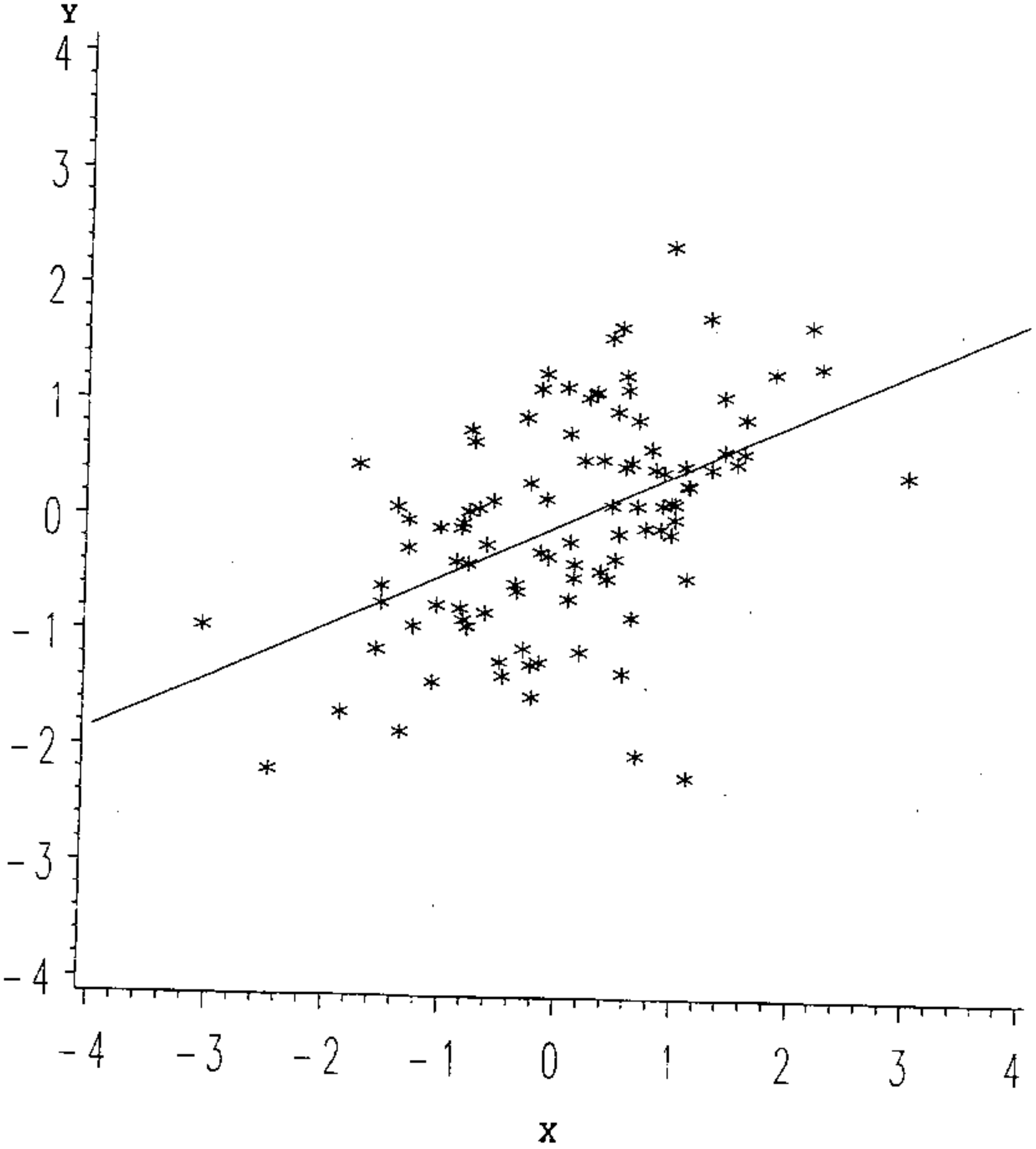


Figure 6-3. Sample from Bivariate Normal Population with Correlation Coefficient = 0.5, Regression Line Shown, Sample Size = 100

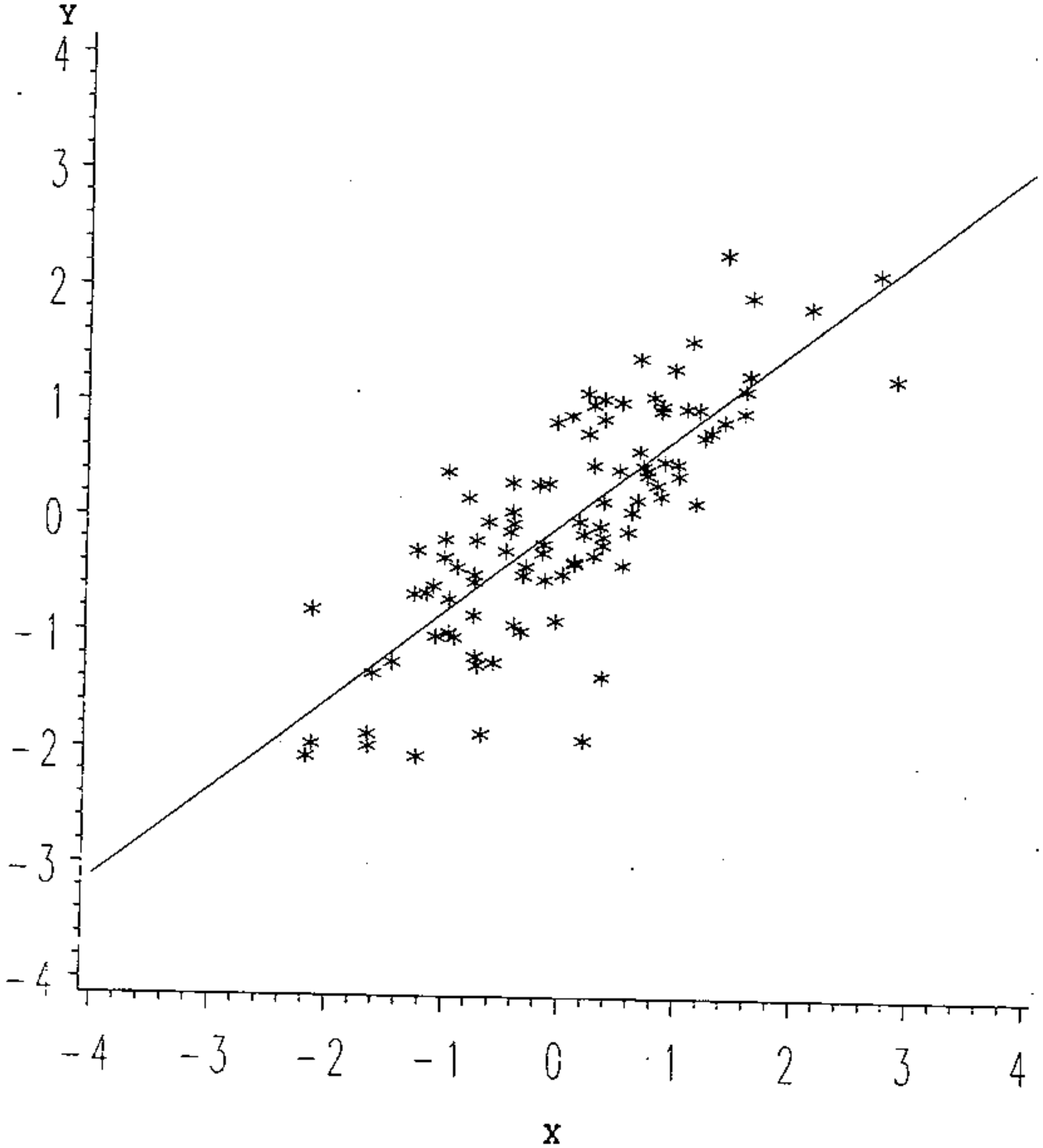


Figure 6-4. Sample from Bivariate Normal Population with Correlation Coefficient = 0.8, Regression Line Shown, Sample Size = 100

TABLE 6-1. UNCERTAINTY IN ESTIMATE OF NATIONAL ANNUAL EMISSIONS UNDER VARIOUS ASSUMPTIONS

		Uncertainty in Annual National Emissions	
Correlations Among Errors for Different Categories	Distribution of Error in Annual National Emissions	Bscf	% of National Production
Absent	Normal	89.6	0.4
Absent	Lognormal	102.8	0.5
Present	Normal	96.8	0.4
Present	Lognormal	112.3	0.5

Industry emissions for 1992: 314 Bscf
Industry production for 1992: 22,132 Bscf

the uncertainty measure based on the lognormal assumption is the most conservative one (the upper confidence limit minus the estimate, which exceeds the estimate minus the lower confidence limit and the half-width of the confidence interval).

Further, the exceedance of 0.007% is well within the uncertainty of the estimation of intercategory correlations. If all nonzero correlations are reduced by 0.1 (by 0.2 for the set considered to be more uncertain), the uncertainty becomes slightly less than 0.5% of production. If the correlations are all increased by 0.2 (0.4 for categories considered to be more uncertain), and the distribution is considered lognormal, the uncertainty of the national production rate remains within 0.54% of production.

The postulation that there are correlated errors among categories is considered reasonable. Given this assumption, it is believed that a point midway between the result for normal and lognormal errors is a more reasonable conservative case than is the result based on the lognormal assumption. The midway point represents the possibility that there is asymmetry in the distribution of the error in the industry emission rate (see Figure 4-4). While the selection of the midway point is arbitrary, it is considered a reasonable postulated conservative case, given the various issues discussed (and especially the averaging and summing of data performed to produce the industry emission rate). The midway point produces an uncertainty of 105 Bscf, which is slightly under 0.5% of national production.

Thus, the conclusion is that, under assumptions that are not unrealistically conservative, the target precision was achieved.

6.2 Attainability of the Target Accuracy

Practical considerations allow sampling only a small percentage of the large number (tens of thousands) of sources that exist nationwide. Moreover, there is typically a large amount of variability among the sources in a given category. In view of these considerations, meeting the desired accuracy may seem insurmountable. The allowed uncertainty in the emissions is 0.5% of the national methane production, on the basis of a 90% confidence limit for the emissions.

Despite these facts, the target precision for the industry emissions was achieved. The purpose of this section is to illustrate, through hypothetical calculations, how large errors in emission estimates for individual source strata can combine to allow this to occur.

As is discussed in the preceding sections, bias is minimized by randomly selecting sites (although from a limited list), analyzing the data, and creating strata in a systematic way. The estimate of total emissions is the sum of the emissions for all the strata. An essential point is that the uncertainties are not additive; the uncertainty of a sum is related to the sum of squares of the individual uncertainties. If the errors in a sum vary independently, the errors tend to "average out" to an extent; the relative error in the sum is reduced by this averaging process.

Fugitive emission sources have been split into five major segments; each segment has two to seven major source categories, and each source category is divided into 10 to 40 strata. In total, these sources have been divided into nearly 100 strata. Vented sources have been divided into approximately 40 strata. Thus, in all there are approximately 140 strata. Some of these strata (such as distribution pipe type) have been aggregated in the summary table shown in Appendix C, which shows 86 categories.

In this subsection, hypothetical calculations are presented that illustrate the effect of summing the errors in the different strata. For the purposes of the hypothetical calculations, it has been assumed that there are "n" strata with equal emissions and equal uncertainties based on random errors. While it is recognized that both the emission rate and the variability change from stratum to stratum in actuality, the simplifying assumptions facilitate a calculation that illustrates the effect of summing the emission estimates from a large number of strata.

Also, it has been assumed that undiscovered bias, if any, varies "independently" from stratum to stratum. This type of error would exist if the sources within a stratum were sampled in an unrepresentative manner, resulting in a bias error. Clearly, a systematic bias that was common to a large number of strata would have a more serious effect on the final result. The processes described earlier for screening for bias provide a protection against this (or any type of) bias error. Additionally, given the large number and diversity of strata, it is reasonable to believe that any undetected bias will exhibit a high degree of "independence" among the strata.

The bias error is represented as the stratum-to-stratum standard deviation of the biases in the emission estimates; this quantity is presented as a percent of the emissions for a stratum. In the calculations, three values have been considered for the bias: 0%, 15%, and 30%. In view of the methods used for screening for bias, 30% is considered to be a very high estimate. As indicated above, the total number of strata is approximately 140. Under one scenario modeled, it was assumed that there are approximately 100 strata with nearly equal emissions that represent the major part of the industry emissions.

Further calculations were performed assuming 40 and 20 strata, in addition to the case with 100 strata. Given that the parameters discussed above of the random and bias errors are fixed, the relative uncertainty in the final result decreases as the number of strata increases. This is because the "error averaging effect" is greater if a larger number of independent estimated quantities are summed. This does not mean that artificially increasing the number of strata would improve the accuracy. There would be fewer data points per stratum, and the uncertainty of the emission estimate for each stratum would increase.

Table 6-2 presents the results of the calculations. The random error was chosen to be as large as plus or minus 130% of the emissions for each stratum, based on a 90% confidence interval. This random uncertainty was selected so that the simulated uncertainty of the industry emission rate here for zero bias errors and 20 strata would equal the actual uncertainty in Table 6-1 for baseline assumptions. This precision value in Table 6-1 also applies in the case of zero bias errors, since it is strictly a measure of precision.

Note that, for the purpose of matching actual and simulated uncertainties, the simulated case with 20 strata was selected. This number of simulated strata is considerably less than the actual number of strata, about 140, or the actual number of source categories, 86. However, selecting the case for 20 strata as a basis of matching the actual uncertainty accounts for the fact that the 86 source categories do not have equal emissions or equal uncertainties. Thus, the reduction of the relative uncertainty achieved by an "averaging effect" when 86 category emissions are summed is less than that which would be achieved if 86 emission rates with identical statistical parameters were summed.

Thus, on the basis of points made in the last two paragraphs, the approach selected provides a reasonable (but not exact) basis of comparability between the actual and simulated results.

Table 6-2 presents the uncertainty in the simulated national emissions as a percentage of the national annual production. The uncertainty is expressed in terms of a 90% confidence interval. Since bias errors were considered as well as random errors, the numbers in Table 6-2 represent accuracy, not just precision.

TABLE 6-2. PERCENTAGE OF ERROR IN SIMULATED NATIONAL ANNUAL EMISSIONS

Bias (Percent of Emissions)	Number of Strata		
	20	40	100
0	0.40	0.29	0.18
15	0.41	0.29	0.18
30	0.43	0.31	0.19

(Percent Random Error in a Given Stratum Based Upon a 90% Confidence Interval = 130%)

The target precision is met if the percentage error is no greater than 0.5%. For all the scenarios modeled, the uncertainty is less than 0.5%. This is true even in the case in which there are only 20 strata with approximately equal emissions, and the bias is 30%. These calculations, while hypothetical, illustrate the way in which errors combine in a sum and show that meeting the target precision is feasible, even in the presence of large-percentage random errors in the individual strata and an assumed large undetectable bias error.

It must be remembered that the random and bias errors were expressed as percentages of the emissions in the strata. For these calculations, the national annual emissions were assumed to be approximately 314 Bscf. The target precision is expressed as a percentage (0.5%) of the national gas production, which was 22,132 Bscf as of 1992.

Note how small the differences are between the corresponding results for 0% bias and 30% bias. For a given number of strata, these differences are no larger than 0.03% of the national production. This is a consequence of the way independent errors combine when one error with a large uncertainty (random error assumed to be 130%) and a much smaller error (bias error) are added. Further, the 30% bias error is assumed to be very conservative (large), given the various steps taken to screen for and eliminate bias. These points imply that any remaining bias in the data probably impacted the actual final uncertainty in the national emission rate by a very small amount.

7.0 REFERENCES

1. Stapper, B.E. *Methane Emissions from the Natural Gas Industry, Volume 5: Activity Factors*, Final Report, GRI-94/0257.22 and EPA-600/R-96-080e, Gas Reasearch Institute and U.S. Environmental Protection Agency, June 1996.

2. Harrison, M.R., H.J. Williamson, and L.M. Campbell. *Methane Emissions from the Natural Gas Industry, Volume 3: General Methodology*, Final Report, GRI-94/0257.20 and EPA-600/R-96-080c, Gas Reasearch Institute and U.S. Environmental Protection Agency, June 1996.

3. Cochran, William G. *Sampling Techniques*, Third Edition. John Wiley & Sons, New York, 1977.

4. Grubbs, Frank E. "Procedures for Detecting Outlying Observations in Samples." *Technometrics*, 11(1), pp. 1-21, 1969.

5. Hoaglin, D.C., F. Mosteller, and J.W. Tukey. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York, 1983.

6. NSI Technology Services Corporation. "WHO and WMO Program Documentation." Environ. Monitoring Services Laboratory, U.S. EPA, Research Triangle Park, NC, under contract No. 68-02-4444, NSI Publication No. SP-4420-89-28, 1989.

7. Pacific Gas & Electric Company. *Unaccounted-for Gas Project*. Report No. GRI-90/0067.1, Gas Research Institute, June 7, 1990.

8. Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*, Third Edition. McGraw-Hill Book Company, New York, 1974.

9. Juran, J.M., Frank M. Gryna, Jr., and R. S. Bingham, Jr. *Quality Control Handbook*, Third Edition. McGraw-Hill Book Company, New York, 1974.

APPENDIX A

Further Details Regarding Certain Statistical Issues

APPENDIX A

FURTHER DETAILS REGARDING CERTAIN STATISTICAL ISSUES

This appendix contains certain mathematical details pertaining to statistical issues discussed in the text. The discussion in this appendix is not required in order to understand the statistical methods used or the issues involved from a conceptual point of view. The discussion here is included for completeness, as a further documentation of the basis for the methods used. In Sections A.1 and A.2, methods for analysis of error propagation in a product and in a sum, respectively, are discussed. In Section A.3, issues pertaining to the calculation of confidence intervals are discussed. In Section A.4, a method for the calculation of precision values based on consecutive assumptions is presented. In Section A.5, the use of the ratio method for the estimation of an activity factor is described. In Section A.6, the approach for combining two estimates of an activity factor obtained by the ratio method is to obtain the final estimate is discussed. In Section A.7, methods for computing the uncertainties of the industry annual emissions, given the uncertainties of the emissions for the categories, are compared. In Section A.8, a complete set of numerical examples is presented to illustrate the calculation of emission factors, activity factors, annual emissions for a source category, annual emissions for the industry, and associated uncertainties.

A.1 ERROR PROPAGATION IN A PRODUCT (EMISSION FACTOR TIMES ACTIVITY FACTOR)

In general, the product of two sample means (such as EF × AF) does not have a standard type of statistical distribution, such as the z-distribution or t-distribution. The type of distribution is not standard even if the two variables that are multiplied are both normal; the product of two lognormally distributed variables is lognormal, however. In this section, two possible ways to approximate the tolerance of a product are discussed. It is shown that the selected method more accurately accounts for the possibly different sample sizes on which the estimates of the activity and emission factors are based. Moreover, it is shown that the selected method produces a more conservative (somewhat larger) estimate of the uncertainty than does the alternate method. The rationale for the selected method is discussed.

In computing the uncertainty of the product EF × AF, it would be possible first to use standard error propagation methods[1] to obtain the variance of this product. Then, under normal theory, it would be possible to multiply this value by the appropriate z-value to obtain the half-width of a 90% confidence interval. According to this method, the tolerance of EF × AF would be obtained as follows:

$$\text{Tol (AF}\cdot\text{EF)} \ = z\sqrt{\text{var (AF}\cdot\text{EF)}}$$

$$= z[E(AF)^2 \text{ var } (EF) + E(EF)^2 \text{ var}(AF) + \text{var } (EF) \text{ var } (AF)]^{1/2}$$

$$\cong z[AF^2 \text{ var } (EF) + EF^2 \text{ var } (AF) + \text{var } (EF) \text{ var } (AF)]^{1/2}$$

where Tol ( ) signifies the half-width of a 90% confidence interval, E ( ) denotes the expected value, i.e., the true value of the parameter, and var ( ) signifies the variance of the error in the parameter estimate.

The equality is approximate in the final line because sample means have been used in place of the unknown population means. For a 90% confidence interval, the z-value required is 1.645.

The argument here is not that EF × AF has an approximately normally distributed error for most source categories. One could argue, however, that the sum of the emission rates for 86 source categories will tend to be normally distributed, because of the large number of terms added. Issues related to the error propagation of a sum are discussed in Section 4.0 and are discussed in somewhat further mathematical detail in the following section in this appendix.

Under the normality assumption, the t-statistic is the proper statistic to use for the purposes of computing a confidence interval of a mean value when the population standard deviation is not known. Tables are readily available that give the t-statistic as a function of the number of degrees of freedom and the confidence level. The number of degrees of freedom is one less than the sample size in the case of quantifying the uncertainty of a mean value (more complicated situations exist involving the comparison of two means).

One could argue that a t-statistic should be used in the equations above rather than a z-statistic, since the AF and EF values are based in most cases on averages, and the standard deviations are not known, but are estimated from the data. The sample sizes used to obtain AF and EF may be different, however; thus, the number of degrees of freedom is not clearly defined, as in the case of computing the confidence interval for the mean of a single sample. Moreover, the product of two means does not have a t-distribution.

Thus, we have used the tolerances of the individual terms (EF and AF) in the error-propagation equation:

$$\text{Tol } (AF \cdot EF) = [AF^2\{\text{Tol}(EF)\}^2 + EF^2\{\text{Tol}(AF)\}^2 + \{\text{Tol}(AF)\}^2 \{\text{Tol}(EF)\}^2]^{1/2}$$

where

$$\text{Tol}(EF) = t_{EF} \; s_{EF} \; / \sqrt{n_{EF}}$$

t_{EF} = appropriate t-value for a sample size of n_{EF},

s_{EF} = sample standard deviation of the individual EF values averaged,

n_{EF} = sample size of the EF values, and

Tol(AF), t_{AF}, s_{AF}, and n_{AF} are defined analogously.

The tolerances of AF and EF are both half-widths of 90% confidence intervals. The t-value for the appropriate sample size is used in determining the confidence interval for each factor. Thus, the effect of each of the two finite sample sizes has been explicitly taken into account in the error-propagation method.

The following derivation reveals that the method used is more conservative (produces a somewhat larger value of the uncertainty) than does the alternate method.

$$\begin{aligned}
\mathrm{Tol(AF \cdot EF)} &\cong z[\mathrm{AF}^2 \text{ var (EF)} + \mathrm{EF}^2 \text{ var (AF)} + \\
&\quad \text{var (EF) var (AF)}]^{1/2} \qquad \text{(by the alternate method)} \\
&= [\mathrm{AF}^2\, z^2 \text{ var (EF)} + \mathrm{EF}^2\, z^2 \text{ var (AF)} + \\
&\quad z^2 \text{ var (EF) var (AF)}]^{1/2} \\
&< [\mathrm{AF}^2\, z^2 \text{ var (EF)} + \mathrm{EF}^2\, z^2 \text{ var (AF)} + \\
&\quad z^4 \text{ var (EF) var (AF)}]^{1/2} \\
&< [\mathrm{AF}^2\, t^2_{EF} \text{ var (EF)} + \mathrm{EF}^2\, t^2_{AF} \text{ var (AF)} + \\
&\quad t^2_{EF} \text{ var (EF) } t^2_{AF} \text{ var (AF)}]^{1/2} \\
&= [\mathrm{AF}^2 \{\mathrm{Tol(EF)}\}^2 + \mathrm{EF}^2 \{\mathrm{Tol\ (AF)}\}^2 + \\
&\quad \{\mathrm{Tol(EF)}\}^2 \{\mathrm{Tol(AF)}\}^2]^{1/2}
\end{aligned}$$

In the derivation above, we have used the fact that $z^2 < z^4$; this follows because the z-value of interest for a 90% confidence interval, 1.645, is greater than one. Other inequalities follow from the fact that $z < t$ for any finite sample size.

A.2 ERROR PROPAGATION IN A SUM

In this section, two possible ways to approximate the tolerance of a sum are discussed. It is shown that the selected method more accurately accounts for the different sample sizes on which the estimates of the various activity and emission factors are based. Moreover, it is shown that the selected method produces a more conservative (somewhat larger) estimate of the uncertainty than does the alternate method. Again, the rationale for the selected method is given.

The alternate method of expressing the tolerance of a sum is as follows:

$$\mathrm{Tol}(\mathrm{ER}_T) = z[\sum_i \mathrm{var}\ (\mathrm{ER}_i)]^{1/2}$$

It is rigorously correct that the variance of a sum of independently distributed random variables is the sum of the variances. This is proven as a theorem by Mood, Graybill, and Boes.[1] This theorem does not depend on the distributions of the variables summed. The variables are not required to have the same distributions, the same means, or the same variances.

This expression for the half-width of a 90% confidence interval is based on the assumption that the sum of 86 separate terms will be approximately normally distributed. This expression, however, does not account for the fact that the activity and emission factors are based on different, finite sample sizes. Thus, we choose to use the following expression instead:

$$\mathrm{Tol}(\mathrm{ER}_T) = [\sum_i \{\mathrm{Tol}(\mathrm{ER}_i)\}^2]^{1/2}$$

It is easily shown that the preferred expression produces a somewhat larger uncertainty than does the alternate method:

$$\mathrm{Tol}(\mathrm{ER}_T) = z\sqrt{\sum_i \mathrm{var}\ (\mathrm{ER}_i)} \qquad \text{(by the alternate method)}$$

$$= \sqrt{\sum_i z^2\ \mathrm{var}\ (\mathrm{ER}_i)}$$

$$< \sqrt{\sum \mathrm{Tol}(\mathrm{ER}_i)^2}$$

The final inequality follows from the derivation given in the preceding section.

A.3 METHODS FOR NON-NORMAL DISTRIBUTIONS

Methodology exists for computing confidence intervals for certain types of non-normally distributed random variables. The methods discussed by Finney[2] and Patterson[3] can be used to calculate the confidence interval of the mean of a sample based on the lognormal distribution. This distribution is discussed briefly in Section 4.1. The nature of this distribution may more nearly approximate that of the emission factor estimates for an individual source category. Thus, the lognormal method may be appropriate for computing the confidence interval for the emission or activity factor for a given source category.

Because of the properties of a sum discussed above, however, it is not believed that the lognormal method is ultimately relevant for computing the confidence interval for the industry total. Moreover, the sample sizes for some of the source categories are so small that it would be difficult to confirm that the distribution was, in fact, approximately lognormal. Thus, it could not be confirmed that the lognormal method was rigorously correct, even for calculation of confidence intervals for parameters for individual source categories.

Finally, the lognormal method, while mathematically correct, is not a panacea when applied to chemical measurements. The logarithmic transformation required can produce instabilities if there are values near the detection limit of the instrument, since these values have large relative errors. Large relative errors in small values may be unimportant when calculations are based on the original data values; these large relative errors can play a major role, however, in calculations based on the logarithms of the data. The discussion here is not intended to be a complete description of the lognormal method or of the various issues regarding its use with chemical measurements. The objective is only to acknowledge that there are alternative ways of computing confidence intervals and to indicate the reasons for the method selected. The bibliographic information for the papers by Finney[2] and Patterson[3] is given in the References for readers who want to study these papers in detail.

A.4 CONSERVATIVE PRECISION ESTIMATES

In this section, an approach for computing a conservative uncertainty for an annual emission value, either for a source category or for the industry, is developed. It was felt that a derivation of the equations alone would not sufficiently convey the issues for all readers. This is not only because of the mathematical nature of the material presented here, but also because of the various nonstandard statistical issues.

Thus, Section A.4.1 presents a qualitative discussion, with graphical illustrations. This section may suffice for readers who want to know the basic qualitative issues and the objective to be achieved by computing a conservative confidence measure. Section A.4.2 presents a numerical example. Finally, the derivation of the equations is presented in Section A.4.3.

A.4.1 Qualitative Discussion

The error propagation methods discussed in Sections A.1 and A.2 lead to a confidence interval based on a normal assumption. Arguments that support the position that the error in the industry annual emissions is approximately normally distributed are given in Section 4.5. For reasons that have been discussed, it is not feasible to prove rigorously that this error is approximately normal. Moreover, for some of the categories with smaller sample sizes, the error in the emission rate may not be approximately normal.

In this section, an approach for approximating a conservative precision value is presented. This precision value is larger than that based on the normal assumption. The conservative precision value discussed in this section characterizes the uncertainty of the emission rate on the high side, which is expected to be greater than the uncertainty on the low side in this application if the error is not normally distributed.

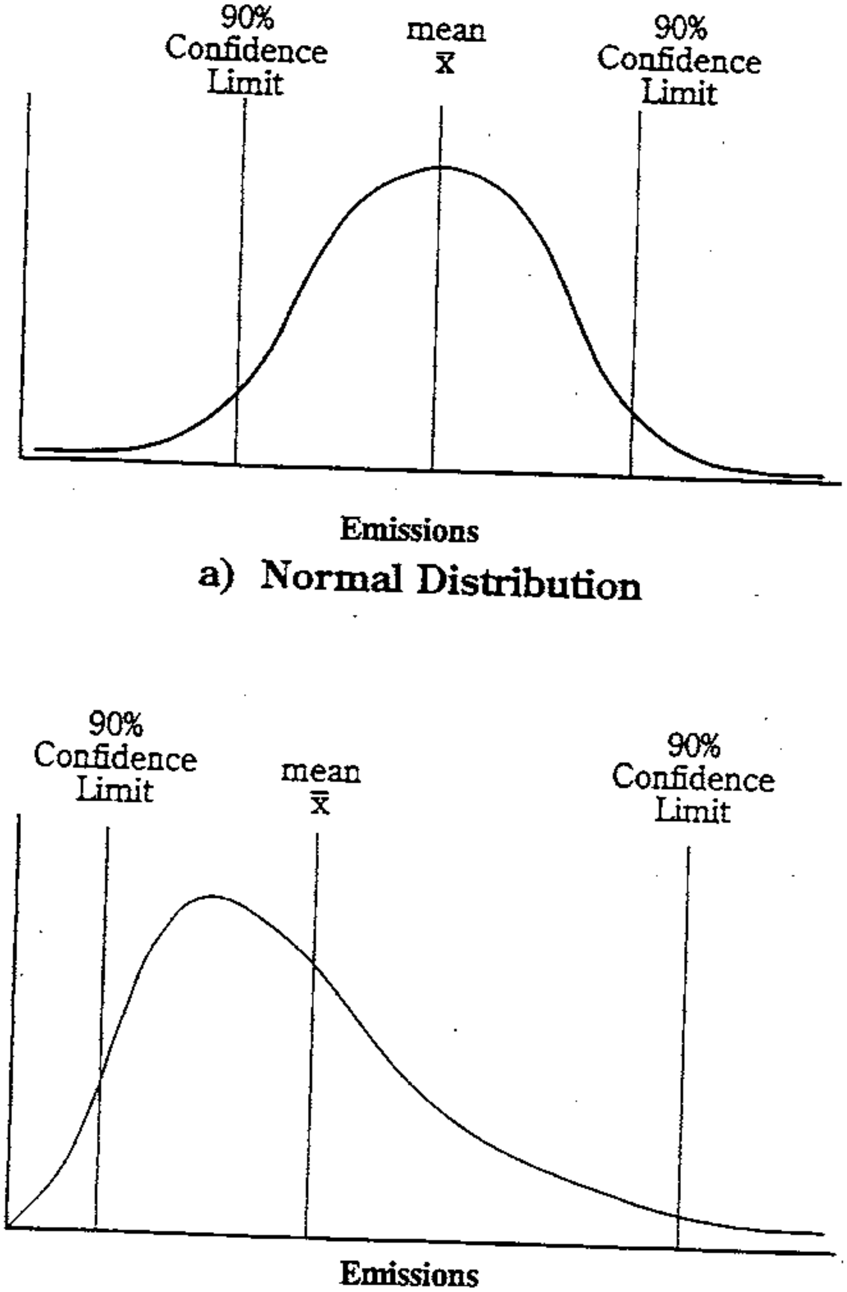
A mean value is normally distributed if the data are normally distributed. By the central limit theorem, if a sufficiently large number of non-normal data points (with the same statistical distribution) are averaged, the uncertainty in the mean value will be approximately normally distributed. The sample size required to produce an approximately normally distributed mean value is strongly dependent on the underlying distribution (especially the degree of asymmetry). Sums of large numbers of terms with non-identical distributions very often tend to be normally distributed, even though the central limit theorem does not strictly apply.

Figure A-1a illustrates the case in which the uncertainty in the sample mean is approximately normally distributed. In this case, the confidence limits are symmetrically placed about the sample mean; the distance between the lower confidence limit and the mean is the same as the distance between the upper confidence limit and the mean.

In the lognormal distribution, the majority of the points fall roughly in the vicinity of the mean, with a small percentage of much larger points. There is not a corresponding percentage of points far below the mean; thus the distribution is asymmetric (see Figure 4-2). This situation corresponds to the case in which there are a large number of sources with moderate emission rates and a small percentage of high emitters.

If the data are non-normal and the sample size is small, the uncertainty in the mean may not be approximately normally distributed. The lognormal distribution is a common type of distribution in general in emission data, and this distribution was observed in this study in the emission data for underground pipes, for example (see Section 4.1).

Figure A-1b illustrates the case in which the uncertainty in the mean is approximately lognormally distributed. Because of the asymmetry of the distribution, the 90% confidence limits are asymmetrically placed about the mean. The lower confidence limit is closer to the mean than is the upper confidence limit.



a) Normal Distribution

b) Lognormal Distribution

Figure A-1. Conceptual Comparison of Normal and Lognormal Distributions

Figure A-2 presents a plot that further illustrates the relationship between the normal and lognormal confidence limits. For the sake of illustration, the sample mean was assumed to be 20 Bscf. The standard error of the sample mean was assumed to be the same under both distributional assumptions; the standard error of the mean is the standard deviation of the uncertainty.

The confidence limits have been plotted as a function of the relative uncertainty (half-width of the 90% confidence interval) for the normal distribution. Consider, for example, the case at the far right of Figure A-2, in which the uncertainty based on the normal distribution is 100%. The confidence interval based on the normal distribution has a lower limit of 0 Bscf and an upper limit of 40 Bscf. These limits are symmetrically placed about the hypothetical sample mean of 20 Bscf.

The confidence interval based on the lognormal distribution is asymmetric. The lower limit for the lognormal distribution is closer to the mean than is the lower limit for the normal distribution. The upper limit for the lognormal distribution is larger than the upper limit for the normal distribution.

If the original data were lognormally distributed, then the sample mean would be more nearly normally distributed than were the original data. Thus, in the example discussed above, one might expect the true upper confidence limit to be between the normal and lognormal upper limits shown in Figure A-2. For this reason, by using the lognormal distribution for the uncertainty in the mean, we have computed a conservative (large) upper confidence limit.

The offset between the two confidence intervals becomes larger as the relative uncertainty increases. In the vicinity of 20% to 30% uncertainty, the difference is slight. In the vicinity of 100% uncertainty, the difference is much larger. Notice, however, that the widths of the normal and lognormal confidence intervals are approximately the same for any given uncertainty value shown in Figure A-2.

Earlier in this report, various issues are discussed that militate against a rigorously correct characterization of the error properties of the emissions for the categories or for the industry as a whole; reasons have been given, however, supporting the hypothesis that the error in the industry annual emissions is approximately normally distributed. For example, the product EF × AF does not in general have a standard type of distribution; this product is not normally distributed even if EF and AF are both normal. For another example, small sample sizes in some instances prevent rigorously establishing the type of distribution of the data averaged to obtain an activity factor or emission factor.

In view of these issues, half-widths of confidence intervals based on the normal assumption for the activity factors and emission factors have been used in error propagation analyses to approximate the uncertainty of a product (ER = EF × AF) and of a sum (the sum of the emission rates by category). In Sections A.1 and A.2, it is shown that this

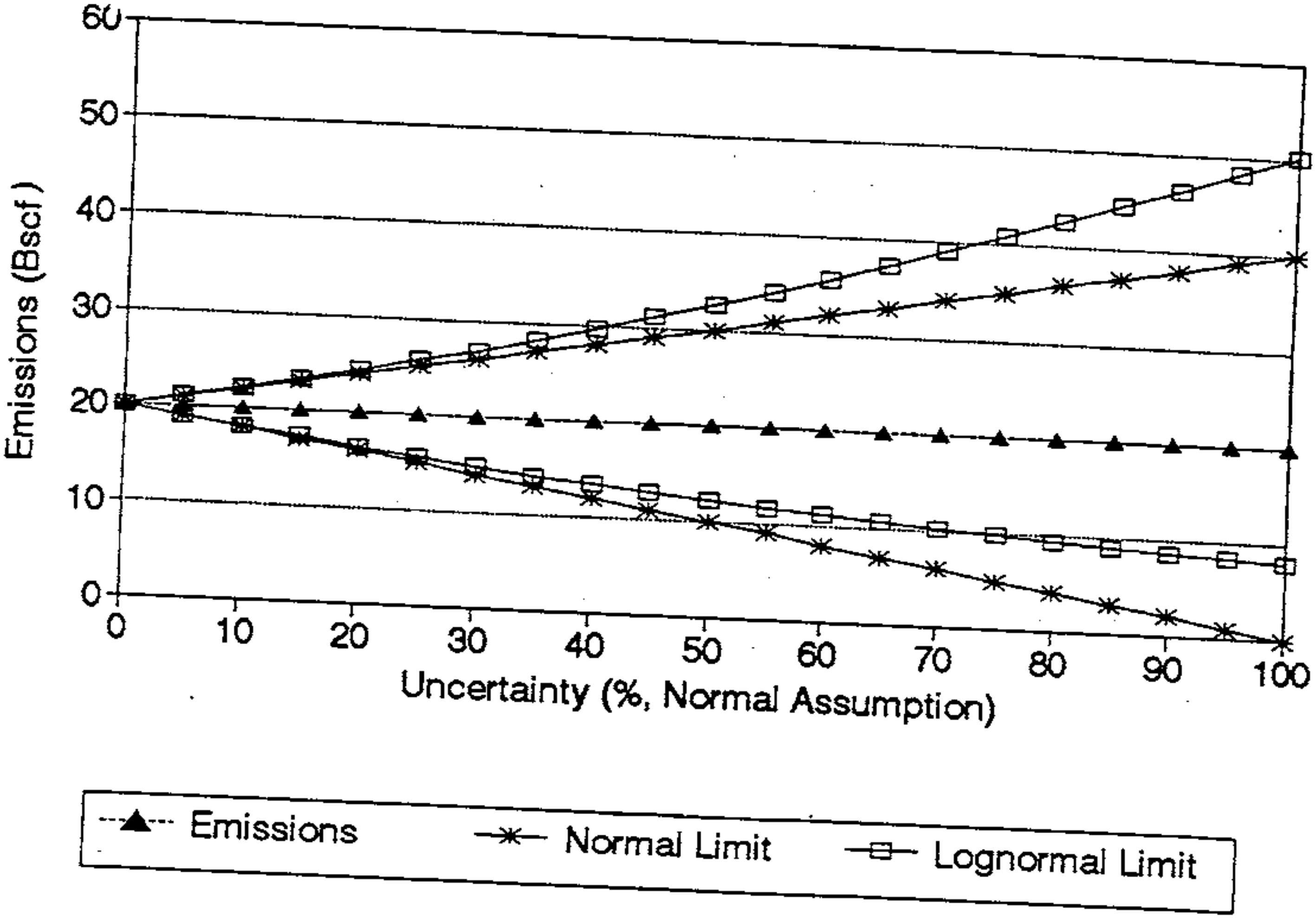


Figure A-2. Comparison of 90% Confidence Limits for Normal and Lognormal Assumptions

methodology accounts for the additional uncertainty attributable to the finite and unequal sample sizes used to estimate the different emission factors and activity factors. Moreover, it is shown that the methods used produce more conservative (larger) measures of uncertainty than would error propagation methods based on only error variances of the activity and emission factors.

In the analysis presented here, the uncertainty of the parameter in question (emissions for a category or for the total industry) was converted to a standard error, and this standard error was used as a basis for computing a confidence interval on the basis of the lognormal assumption (the mean and standard deviation completely determine either a normal or lognormal distribution). The conservative precision is the upper confidence limit minus the estimated emissions, converted to a percentage of the emissions; because of the asymmetry of the lognormal confidence limits, this precision reflects the uncertainty of the emissions on the high side. The conservative precision as calculated is larger than either of the following alternative precision values: (1) the difference between the emission estimate and the lower confidence limit, converted to a percentage of the emissions, or (2) the half-width of the lognormal confidence limit, converted to a percentage of the emissions. That is, the largest of the three precision measures mentioned here was used.

The potentially varying types of distributions of the errors in the various emission factors, activity factors, and emission rates by category have not been rigorously modeled in computing the conservative precision values discussed here. Nevertheless, the lognormal approach provides a measure of uncertainty that is conservative in several respects discussed in Sections A.1 and A.2 and earlier in this section.

A.4.2 Numerical Illustration

In the numerical illustration in this subsection, three significant figures have been reported at the intermediate points and in the final results. However, several additional digits were carried through all calculations. It was felt that three significant digits were sufficient for illustrative purposes. However, this comment is provided for the benefit of readers who may want to reproduce the numerical results. Slight differences between their results and reported values may be observed because of rounding. The practice of carrying several significant figures through all calculations and rounding only for reporting purposes was used in the analyses of the actual data in this project.

Consider a hypothetical case in which the estimated annual emission value y is 20.0 Bscf, as illustrated in Figure A-2. Suppose the precision given in the summary table based on the normal assumption is 80.0 percent. The half-width of the confidence interval based on the normal assumption is a simple conversion from a relative uncertainty in percent to an absolute uncertainty

$$(20.0 \text{ Bscf})(80.0\% \text{ uncertainty})/(100\%) = 16.0 \text{ Bscf}$$

From this uncertainty, a standard error is estimated:

$$s_y' = (16.0 \text{ Bscf})/1.645 = 9.73 \text{ Bscf}$$

The divisor, 1.645, corresponds to the half-width of a 90 percent confidence limit when the standard deviation is known; the quantity 1.645 is called a z-value. As is discussed in the earlier development, t-values were used in quantifying the uncertainties of emission factors and activity factors; the use of t-values accounts for the unknown standard deviations and produces larger uncertainties than if z-values had been used. These uncertainties were used in error propagation to obtain the uncertainties first of category annual emissions and finally of the industry annual emissions. However, a z-value was used as a divisor above because, after the error propagation, the number of degrees of freedom to use in selecting a t-value is generally not known; this is especially true in the case of the industry annual emissions, the quantity of ultimate interest in this study.

Use of the smaller divisor ($z < t$) produces a somewhat inflated estimate of the standard error, s_y'. The prime is included to signify that s_y' is not a conventional estimate of a standard error. The calculation of a somewhat inflated estimate here facilitates another calculation discussed below.

Given the estimated emissions y and the value s_y', it is possible to compute the mean and standard deviation in log space. The mean is as follows:

$$Y = -\frac{1}{2}\ln\frac{(s_y')^2+y^2}{y^4} = = -\frac{1}{2}\ln\frac{9.73^2+20.0^2}{20.0^4} = 2.89$$

The estimated standard deviation in log space is as follows:

$$(s_Y') = \sqrt{2[\ln(y)-Y]} = = \sqrt{2[\ln(20.0)-2.89]} = 0.461$$

Again, the prime was used since this standard deviation, which is dependent on s_y', is somewhat inflated.

The confidence interval for Y, the estimated logarithm of the annual emissions, is symmetric:

$$(Y-1.645s_Y', \; Y+1.645s_Y')$$

or

$$(2.13, \; 3.65)$$

It may appear that a t value, rather than the z value of 1.645, should be used, since the standard deviation is not known but is estimated from the data. Recall, however, that s_y' is inflated because it results from dividing by 1.645, and s_Y' is also inflated, since it was computed from s_y'. The use of a z-value as a multiplier here counterbalances the use of a z-value as a divisor earlier. In view of the nonlinear transformations, the two effects do not cancel exactly, but it is believed that this issue is unimportant relative to others, such as the difficulty in determining the type of statistical distribution of the errors.

Finally, the asymmetric confidence interval based on the lognormal assumption for the original annual emissions is as follows:

$$(F\ e^{Y-1.645s_Y'},\ F\ e^{Y+1.645s_Y'})$$

where F is a bias correction factor that is necessitated by the nonlinear transformation. The factor F is as follows:

$$F = e^{\frac{1}{2}(s_Y')^2}$$

The final confidence interval based on the lognormal assumption is:

(9.37 Bscf, 42.7 Bscf)

The confidence interval based on the normal assumption is:

(4.00 Bscf, 36.0 Bscf)

A.4.3 Derivation of Equations

Methods discussed briefly in Section A.3 exist for computing the confidence interval for the mean of a lognormal population. These methods are applicable for computing the mean of a single sample and do not apply to analysis of error propagation of the type involved in computing annual emissions for the categories and for the industry total. Thus, an approach specific for this application has been developed.

Suppose y is an estimate of the annual emissions for a given source category or for the industry. Further, suppose that h_y is the half-width of the absolute confidence interval for this estimate; the half-width of the confidence interval as a relative error in percent is given in the data summary table.

Now, suppose we compute the following quantity:

$$s_y' = h_y/z$$

where z is 1.645. Issues pertaining to the use of a z-value here and later in calculating the confidence interval in log space are discussed in the preceding subsection.

We are interested in a population of lognormally distributed estimates of a particular parameter. Our estimate of the mean of this population is y. The (somewhat inflated) standard deviation of the population is s_y'.

Let Y denote the estimate of the mean of the logarithm of the parameter of interest. Let s_Y' denote the standard error of this estimate. The following equations express the relationship between the parameters y and s_y' in linear space and the parameters Y and s_Y' in log space[1]:

$$y = e^{Y+\frac{1}{2}(s_Y')^2}$$

$$(s_y')^2 = e^{2Y+2(s_Y')^2} - e^{2Y+(s_Y')^2}$$

The estimate y exists in the data summary table, and s_y' can be obtained from this table as described earlier. From these values, as an initial step to computing the desired asymmetric confidence interval, we need to solve for Y and s_Y'. Taking the logarithm of both sides of the first of the two equations above yields the following:

$$\ln(y) = Y+\frac{1}{2}(s_Y')^2$$

$$(s_Y')^2 = 2[\ln(y)-Y]$$

We substitute this result into the equation for s_y' to obtain one equation in the one unknown, s_Y'.

$$(s_y')^2 = e^{2Y+4\ln(y)-4Y} - e^{2Y+2\ln(y)-2Y}$$

$$= e^{-2Y+4\ln(y)} - e^{2\ln(y)}$$

$$e^{-2Y+4\ln(y)} = (s_y')^2 + y^2$$

$$e^{-2Y} = \frac{(s_y')^2+y^2}{e^{4\ln(y)}}$$

$$= \frac{(s_y')^2 + y^2}{y^4}$$

Taking the logarithms of both sides allows us to solve for Y:

$$-2Y = \ln\frac{(s_y')^2 + y^2}{y^4}$$

$$Y = -\frac{1}{2}\ln\frac{(s_y')^2 + y^2}{y^4}.$$

The equation above allows us to calculate Y in terms of known quantities. Now we can substitute this Y value into an earlier expression to obtain a solution for $(s_Y')^2$ in terms of known quantities.

$$(s_Y')^2 = 2[\ln(y) - Y]$$

We are now in a position to compute a symmetric confidence interval for Y, from which we can obtain the desired asymmetric confidence interval for y. The confidence interval for Y is as follows:

$$Y \pm 1.645 s_Y'$$

It remains to perform the logarithmic transformation to obtain the confidence interval in the original space. Following Patterson's analysis[3], we apply the appropriate bias correction factor to both limits of the confidence interval. The resulting confidence interval is as follows:

$$(F\,e^{Y-1.645 s_Y'},\ F\,e^{Y+1.645 s_Y'})$$

where the multiplicative bias correction factor F is as follows:

$$F = e^{\frac{1}{2}(s_Y')^2}$$

A.5 RATIO METHOD FOR ESTIMATION OF AN ACTIVITY FACTOR

As is discussed in Section 3.5, the ratio method has been used to estimate activity factors on the basis of well counts or production. In that section, a numerical example is given which illustrates the ratio method for that purpose. This section provides a further description of the ratio method, including calculation of a confidence interval for an estimate obtained by this method. In Section A.5.1 estimation using the ratio method is described. In Section A.5.2, methods for computing a confidence interval for the estimate produced by the ratio method are discussed. Further discussion of the ratio method and methods for computing confidence intervals is provided by Cochran.[4]

A.5.1 Estimation Using the Ratio Method

Suppose

y_i = device count (e.g., number of separators) at the ith sampled site,

x_i = value of the extrapolation parameter (number of wells or gas production) at the ith site,

n = number of sites sampled,

X = the regional value of the extrapolation parameter, e.g. the total number of wells in the region, and

N = the total number of sites in the region.

For the purposes of illustration, we will discuss the estimation of the regional number of separators by using the well method. Then, by the ratio method, the following is the estimate of the number of separators per well:

$$\hat{R} = \frac{\bar{y}}{\bar{x}}$$

or

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

This estimated number of separators per well and the regional number of wells is then used to estimate the number of separators in the region:

$$\hat{Y}_R = \hat{R}\ X$$

A.5.2 Confidence Interval for Estimate Produced by the Ratio Method

Cochran presents two approaches for estimation on the basis of confidence intervals. First, the issues pertaining to the two methods will be described, and reasons for selecting one of these two methods will be discussed. Subsequently, the details of the selected method will be discussed. We will continue to use estimation of the number of separators in a region by the well method as an example.

One method for calculation of the confidence interval is based on the assumption that the ratio estimate, $\hat{R}$, is approximately normally distributed. In many applications, the normality assumption is satisfied only if the sample size (the number of sites visited in our application) is sufficiently large (at least 30) and the relative uncertainties (coefficients of variation) in both the average number of separators per site and the average number of wells per site are both sufficiently small (less than 10%). The suggested rules of thumb are given by Cochran. If the ratio is normally distributed, its confidence interval will be symmetric.

If the ratio itself is not approximately normally distributed, but the numerator and denominator are both normally distributed, the ratio will tend to have an asymmetric confidence interval in which the upper confidence limit is more separated from the mean than is the lower confidence limit (see Figures A-1b and A-2). A second method handles this case. As is discussed below, the cause of the asymmetry in some applications is a fundamental consideration in the selection of a method. Thus, a brief discussion of the cause will be given here.

Suppose we are concerned with a ratio a/b, such that "a" and "b" are both subject to random variability but both are non-negative. Given that "b" is subject to random variability and bounded below only by zero, a value very close to zero could occur. The ratio has no upper bound as "b" approaches zero; thus the error in the ratio is unbounded above. But the ratio has an absolute lower bound of zero. The possibility of values extremely larger than the true value, without a corresponding possibility of values extremely lower than the true value, tends to cause the uncertainty in the ratio to be asymmetric.

The method based on the assumption that the ratio is approximately normally distributed will be called Method 1. The method that produces asymmetric confidence intervals will be called Method 2. Radian has performed calculations to compare these two methods. Tests revealed that Method 2 is capable of producing an upper confidence level that is unreasonably large from an engineering point of view (see the discussion below pertaining to separators for the Central Plains Region). The confidence limits produced by Method 1 under these circumstances are much more reasonable from this perspective.

Both engineering judgement and further statistical calculations have indicated that Method 1 is preferable for this application. First, the asymmetric confidence interval is based on the general mathematical situation described above, in which the denominator can become arbitrarily close to zero. But in our application, the denominator is the sum of the production levels or of the numbers of wells for the sites visited in a region. From a practical perspective, it is not reasonable to expect that either of these sums can become arbitrarily close to zero, causing an extremely large ratio of separators per well or separators per unit of production.

The number of wells at a site of interest must be at least one. Thus, the sum of the numbers of wells has a lower bound equal to the number of sites visited. The production does not have a definable lower bound of this nature. The argument above still applies, however; it is not reasonable to expect an arbitrarily small production rate at all visited sites in a region, allowing an unbounded ratio of devices per unit of production on the basis of the data for all sites.

The argument above pertains to the possibility of an arbitrarily small denominator, which could cause extreme skewness; Figure A-1b depicts a hypothetical distribution that is skewed, or asymmetric. The relationship between the number of devices and the number of wells or amount of production is also relevant. Theoretically, positive skewness in a ratio could result from positive skewness in the numerator; this would be a special concern if the numerator could increase without bound, independently of the value of the denominator. In this application, however, it is not reasonable to expect that the number of separators attached to a given well is unbounded; similar comments apply for other device types. Further, it is not reasonable to expect that the number of separators at the visited sites in a region is independent of the total production at those sites and can become arbitrarily large, independently of the production level.

The intuitive arguments above indicate that certain mathematical causes of marked asymmetry do not exist in this application. However, these arguments do not prove that asymmetry cannot exist at all. A further investigation was performed on the basis of statistical calculations. For each of a selected set of regions and device types, the number of devices was divided by the extrapolation parameter (wells or production) for each site. This produced a ratio for each site visited for a given region and device type. In most cases, the number of sites is too small to allow a detailed characterization of the distribution. For separators for the Atlantic/Great Lakes region, however, there were 19 sites. The distribution of separators per well is displayed for this case in Figure A-3. The histogram is somewhat ragged, because of the sample size; even 19 is a small sample size to characterize a distribution. Nevertheless, there is no evidence of positive skewness. Despite the raggedness of this empirical distribution, a hypothesis test indicated that this distribution does not differ to a significant extent from a normal distribution.

Figure A-4 presents the histogram for the ratio of separators per unit of production for the same region. In this case, there is evidence of asymmetry in the distribution of the site-by-site ratios, and the hypothesis test indicated that this distribution differed significantly from a normal distribution. The primary reason for the visual impression of asymmetry is a single site with a ratio of 99.9 separators per MMcfd of production. Asymmetry in the site ratios, however, does not necessarily imply that the error in the ratio for the region is asymmetrically distributed. For the site with the large ratio, there are 1,582 separators and 16 MMcfd of production. Another site with a more moderate ratio has a much larger impact on the ratio for the region. This site has 3,227 separators and 81 MMcfd of production, so the ratio is 39.8 separators per MMcfd.

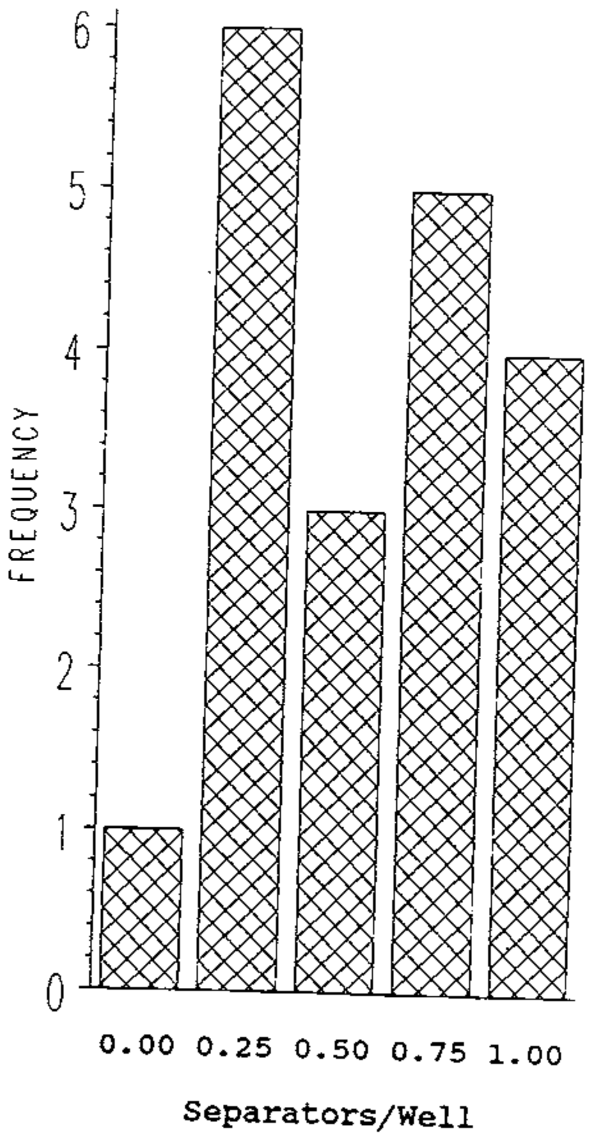


Figure A-3. Distribution of the Number of Separators per Well for 19 Sites in the Atlantic/Great Lakes Region

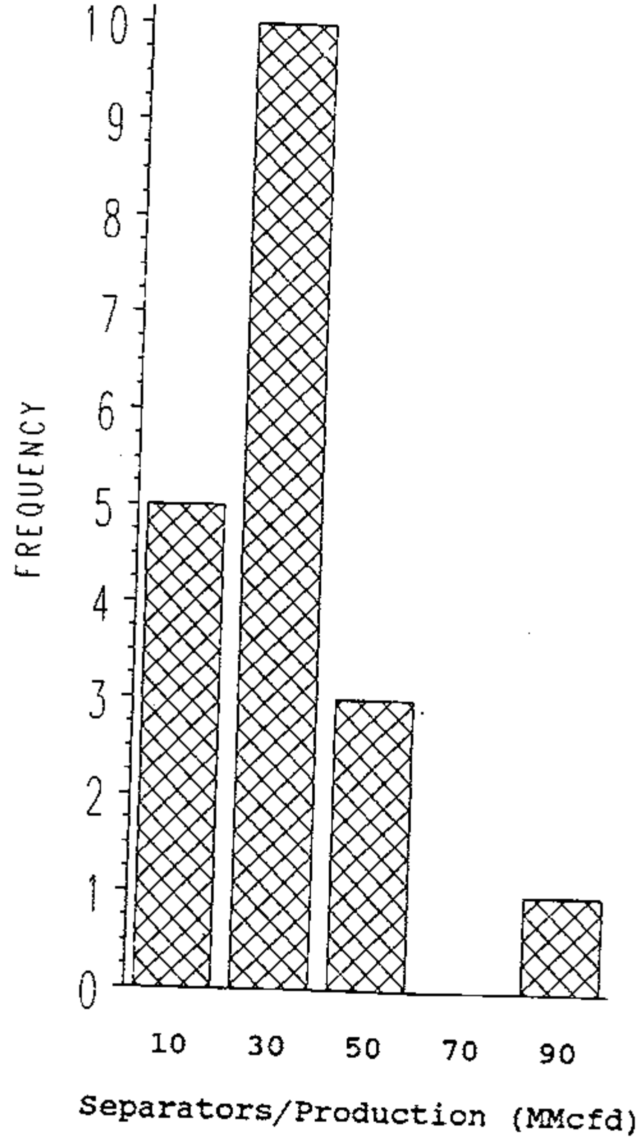


Figure A-4. Distribution of the Number of Separators per Unit of Production for 19 Sites in the Atlantic/Great Lakes Region

The site values of the number of separators per well for the Central Plains region revealed evidence of negative skewness. That is, instead of a long tail to the right, as in Figure A-4, there was some evidence of a long tail to the left. Since there were only seven sites, a histogram of this data set would not be meaningful and is not shown. In this case, Method 2 produced an upper confidence limit for the ratio of separators per well that was unreasonably large in this case from an engineering or a statistical point of view. This upper limit was several times the largest site ratio of separators per well for the Central Plains region and exceeded almost all the separator-per-well site values for several regions. In this example, Method 1 produced results that were considered to be much more reasonable. While negative skewness was the exception, this example provides another illustration of why Method 1 was preferred over Method 2 for this application.

Moreover, asymmetric uncertainties of individual parameters exist for other reasons. It is discussed elsewhere in this report (Sections 4.5 and 5.1) that confidence intervals with greater than 100% uncertainty exist for activity factors, emission factors, or emission rates for some source categories. One possible explanation is that the error in the estimated parameter is not normally distributed. The ultimate objective of the study, however, is to quantify the national annual emissions. The sum of the emissions for 86 source categories will tend toward normality, even if some of the individual values summed are nonnormal. Thus, even if some category parameters were not normal, this would not necessarily invalidate the confidence interval for the national annual emissions. Moreover, an assessment has been made of the effect of a lognormal error in the industry annual emissions. The upper confidence limits based on the normal and lognormal assumptions differ by a small amount, and the target precision is met on the basis of either assumption (Section A.4).

Based on a finite sample of size n (i.e., n sampled sites), the following is an approximation of the variance of the error in $\hat{Y}_R$:

$$v(\hat{Y}_R) = \frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n}(y_i-\hat{R}x_i)^2$$

The quantity N, the total number of sites in the region is not known and, therefore, must be estimated. The total number of separators, X, in the region is known. The quantity X divided by the average number of separators per site is an approximation of the number of sites in the region. This method of estimating N was suggested to Radian by Jonathan Cohen of ICF Kaiser in a private communication.

Thus, N is an estimate rather than a known constant. The value N is used only in quantifying the uncertainty of $\hat{Y}_R$, however, and not in estimating $\hat{Y}_R$. The quantity f is the sampling fraction, n/N.

The equation given by Cochran for a symmetric confidence interval for $\hat{Y}_R$ is as follows:

$$\hat{Y}_R \pm z \sqrt{v(\hat{Y}_R)}$$

where z is a tabulated value of the standard normal distribution selected according to the confidence level; for a 90% confidence interval, the z value is 1.645. The z value is appropriate when the quantity estimated ($\hat{Y}_R$) has an uncertainty, but the uncertainty of the variance [$v(\hat{Y}_R)$] can be neglected. The use of the z statistic is generally accepted if the sample size is greater than 30.

According to Cochran's rules of thumb, the sample size would be at least 30 when this expression for the confidence interval was used. In our case, however, the decision that the symmetric confidence interval was preferable to the asymmetric confidence interval even if the sample size was less than 30 was based on engineering considerations and data analysis, as is discussed above. To account for the uncertainty in the variance as well as in the estimate, therefore, we have replaced z in the expression above by the appropriate t value. Even though the t-distribution does not apply exactly in this context, replacing z by t provides a degree of conservatism; that is, somewhat wider confidence intervals are produced, which tends to account for the uncertainty in $v(\hat{Y}_R)$. The resulting confidence interval is as follows:

$$\hat{Y}_R \pm t \sqrt{v(\hat{Y}_R)}$$

A.6 COMBINATION OF ESTIMATES OF AN ACTIVITY FACTOR

The methods discussed in the preceding section were used to estimate the activity factor and its uncertainty on the basis of both well counts and production for some source categories. The arithmetic average of the two estimates was computed to obtain the final estimate.

The two estimates are based on different extrapolation factors (values of the x_i) but common device counts (values of the y_i). The device counts vary by site and are subject to sampling error. Thus, this source of sampling error was common to the two estimates of the activity factor. It has been discussed elsewhere that separate measured quantities (e.g., emission rates from different types of devices) may have correlated sampling errors. The evidence here for correlation is much stronger, however, since common data are used in the two estimates. Thus, steps were taken to account explicitly for the correlation. To address this issue, we introduce the following notation:

$x_{w,i}$ = number of wells at the i^th site,

$x_{p,i}$ = production at this site,

R_w = estimate of R on the basis of wells,

$\hat{R}_p$ = estimate of R on the basis of production,

$\hat{Y}_{R,w}$ = estimate of $\hat{Y}_R$ on the basis of wells, and

$\hat{Y}_{R,p}$ = estimate of $\hat{Y}_R$ on the basis of production.

By substituting $x_{w,i}$ for x_i in the appropriate equations in the preceding section, for example, one obtains the estimate of the number of devices in the region on the basis of wells and the confidence interval for this estimate. The following is a sample estimate of the covariance between the errors in the two estimates:

$$cov(\hat{Y}_{R,w}, \hat{Y}_{R,p}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^{n} (y_i - \hat{R}_w x_{w,i})(y_i - \hat{R}_p x_{p,i})$$

This expression satisfies important required properties of the covariance, such as the symmetry property:

$$\text{cov}(\hat{Y}_{R,w}, \hat{Y}_{R,p}) = \text{cov}(\hat{Y}_{R,p}, \hat{Y}_{R,w})$$

Additionally, the covariance between a quantity and itself equals the variance of that quantity. This can be confirmed simply by replacing all "w" subscripts with a "p" subscript, to obtain the variance of $\hat{Y}_p$.

In a textbook application of the ratio method, the quantity N would be known. As is discussed earlier, N must be estimated in this application. In estimating the covariance above, the average of the two estimates of N was used, both where N appears explicitly in the covariance equation and in calculating f.

Now, the expression for the final estimate of the activity based on the arithmetic average approach is as follows:

$$\hat{Y}_{R,avg} = \frac{\hat{Y}_{R,w} + \hat{Y}_{R,p}}{2}$$

The variance of this expression is:

$$var(\hat{Y}_{R,avg}) = \frac{var(\hat{Y}_{R,p}) + 2cov(\hat{Y}_{R,p},\ \hat{Y}_{R,w}) + var(\hat{Y}_{R,w})}{4}$$

The confidence interval for the final estimate is as follows:

$$\hat{Y}_{R,avg} \pm t \sqrt{var(\hat{Y}_{R,avg})}$$

In some instances, the number of sites for which data existed for both wells and production did not coincide exactly. In these cases, the covariance was computed on the basis of the sites for which common data did exist. This provided a somewhat conservative (large) estimate of the covariance. This calculation of the covariance represents the case in which the sites in common for the two extrapolation parameters are the only sites. But the fact that sites exist with data for wells but not production (or vice versa) introduces an element of independence between the estimates of $\hat{Y}_R$ based on the two extrapolation parameters. The somewhat conservative covariance estimate produces a somewhat conservative confidence interval for the final estimate of $\hat{Y}_R$.

To account for this case, the correlation between the errors in the two estimates was computed as follows:

$$r = \frac{cov(\hat{Y}_{R,w},\ \hat{Y}_{R,p})}{\sqrt{var(\hat{Y}_{R,w})var(\hat{Y}_{R,p})}}$$

Then the half-width of the confidence interval for $\hat{Y}_R$ was computed as follows:

$$\frac{1}{2}\sqrt{t_p^2 Var(\hat{Y}_{R,p}) + 2r[t_p\sqrt{var(\hat{Y}_{R,p})}][t_w\sqrt{var(\hat{Y}_{R,w})}] + t_w^2 Var(\hat{Y}_{R,w})}$$

where t_p and t_w are the t-values appropriate for the sample sizes for the two extrapolation parameters. The expression involving the correlation coefficient was written in the manner shown to emphasize that this is approximately an error propagation using half-widths of confidence intervals, as has been used elsewhere (see Sections A.1 and A.2). Each t-value is grouped with its respective standard error (the square root of an error variance is a standard error). The expression above can be simplified algebraically to the following:

$$\frac{1}{2}\sqrt{t_p^2 var(\hat{Y}_{R,p}) + 2t_p t_w cov(\hat{Y}_{R,w},\ \hat{Y}_{R,p}) + t_w^2 Var(\hat{Y}_{R,w})}$$

This expression involving different sample sizes for the two estimates reduces to the simpler expression for the half-width of the confidence interval given earlier if the sites for which data exist for wells and production are the same.

A.7 UNCERTAINTY OF INDUSTRY ANNUAL EMISSIONS

This section provides the details of the reasons for the selection of an approach for computing the uncertainty of the industry annual emissions, given the uncertainties of the emissions by source category. Recall from Section 4.4 that Method 1 involves computing the sum of squares of the uncertainties of the terms summed to obtain the industry emission rate. In Method 2, the uncertainty of the sum equals the sum of the uncertainties, or tolerances, of the terms. The terms summed are the emission rates for the 86 source categories.

The sum of the tolerances is apparently used in some applications and provides a conservative (large) estimate of the tolerance of a sum. Since the possibility of using Method 2 has been raised as an issue, a brief comparison of the two methods and the reasons for selecting Method 1 will be given. The discussion will show that Method 2 is inappropriate for this application.

Juran, et al.,[5] give the following simple example to illustrate why the Method 2 is "often too conservative." Given the mechanical assembly shown in Figure A-5, suppose it is necessary to compute the uncertainty of the sum of the three lengths. Suppose that there is one chance in 100 that a given one of the three parts will be less than its lower tolerance, and the errors in the three lengths are uncorrelated.

Now, suppose a lower tolerance for the sum of the three lengths is computed by summing the three lower tolerances. The probability that all three will be less than their lower tolerances simultaneously is:

$$1/100 \times 1/100 \times 1/100 = 1/1{,}000{,}000$$

That is, there is only one chance in a million that all three components will fall below their respective lower tolerances simultaneously. Thus, the sum of the tolerances produces a very conservative estimate, in that there is no recognition of the fact that the probability that all errors will be extreme in magnitude and have the same sign is very low. In Method 1, the fact that the errors in the different terms in a sum may have different magnitudes and even different signs is recognized.

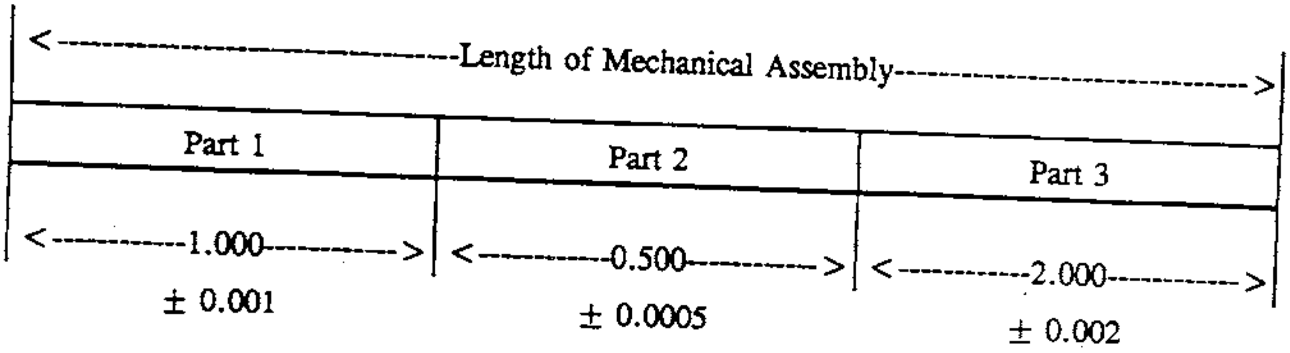
If Method 2 produces very conservative results in the case of the sum of three terms, this method produces unreasonably conservative results in the case involving 86 source categories. The tolerances used in this project are based on 90% confidence limits. The probability that all 86 true emission values will fall below the lower confidence limits simultaneously is 1.3×10^{-112}. The probability that all true emission values will fall above the upper confidence limits is the same.

A.8 EXAMPLE CALCULATIONS

In this section, a set of numerical examples illustrating the calculation of emission factors, activity factors, annual emissions for a source category, annual emissions for the industry, and associated measures of uncertainty is presented. This is an appropriate place for this set of examples, given the development of equations presented earlier in this appendix.

Numerical examples illustrating specific points are given at various points earlier in the report. The purpose of this section, however, is to combine an extensive set of examples in one place.

A hypothetical example will be presented to illustrate the calculations for an individual category. It was desired to provide an example that was representative of the basic case.



$\sqrt{\Sigma\ T_i^2}$ = 0.0023 (Preferred Method)

$\Sigma\ T_i$ = 0.0035 (Conservative Method)

T_i = Tolerance of Part i

Figure A-5. Illustration of Methods for Computing the Tolerance of a Sum

However, many actual categories involve individual characteristics and exceptions. For example, categories for which the ratio method was used to estimate the activity factor involve application of this method for each geographical region. It was considered undesirable to show the same basic calculations performed with different sets of numbers for the different regions. The repetitious aspect of this type of example would have added length to this section, but the repetition would have contributed nothing to the illustration and might even have tended to obscure the message. Further, the use of relatively small data sets facilitated presenting all data used in the calculations, without requiring large tables that would have contributed nothing extra to the illustration.

As is discussed in Section A.4.2, three significant figures have been reported at intermediate steps and in the final results. Several additional digits were carried through all calculations, however.

In Section A.8.1, calculation of the emission factor for a hypothetical source category is illustrated. In Section A.8.2, calculation of the activity factor for this category is discussed. In Section A.8.3, calculation of the category annual emissions is presented.

In Section A.8.4, the industry annual emission calculations are discussed. Both the real data given in Appendix C and hypothetical data are used as needed to illustrate different aspects of the calculations.

A.8.1 Emission Factor Calculations

Table A-1 presents the hypothetical data for the calculation of the emission factor.

TABLE A-1. HYPOTHETICAL EMISSION FACTOR DATA

Measurement Number	Annual Emissions (Scf)
1	18,000
2	17,000
3	3,000
4	10,000
5	15,000
6	7,000
7	2,000
8	13,000
9	2,000
10	13,000
Mean	10,000
Standard Deviation	6,160

Each measurement represents the annual emissions for one device for one year in Scf. The symbol e_i will be used to denote the i^th emission measurement, and n will denote the sample size, which is 10 in this example. The mean value of the measurements equals the emission factor, EF:

$$EF = \bar{e} = \frac{\sum_{i=1}^{n} e_i}{n} = \frac{18{,}000 + \ldots\ 13{,}000}{10} = 10{,}000\ Scf/device$$

The standard deviation is as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{n-1}}$$

$$= \sqrt{\frac{(18{,}000-10{,}000)^2 + \ldots + (13{,}000-10{,}000)^2}{10-1}} = 6{,}160\ Scf/device$$

The standard deviation of the error in the mean, or the standard error of the mean, equals the standard error of the emission factor, s_{EF}:

$$s_{EF} = \frac{s}{\sqrt{n}} = \frac{6{,}160}{\sqrt{10}} = 1{,}950\ Scf/device$$

The confidence interval based on the assumption that the errors are normally distributed involves the t-statistic. The parameter of the t-distribution is called the "number of degrees of freedom," which is n-1, or 9, in this context. From standard tables, the appropriate t-value for a 90% confidence interval for 9 degrees of freedom is 1.83. The uncertainty of the estimated emission factor is calculated as follows:

$$Tol(EF) = t\ s_{EF} = (1.83)(1{,}950) = 3{,}570\ Scf/device$$

This value is readily converted to an uncertainty in percent:

$$Tol(EF)\ (\%) = \frac{(100\%)Tol(EF)}{EF} = \frac{(100)(3{,}570)}{10{,}000} = 35.7\%$$

A.8.2 Activity Factor Calculations

Table A-2 presents the data for the hypothetical calculation of the activity factor. The notation and methodology used to calculate the activity factor using the ratio method are given in Section A.5. Activity-factor issues are also discussed in Section 3.5. The activity factor is the total number of devices (for example, separators) for this source category. Gas production, a common extrapolation parameter, has been employed in this example.

TABLE A-2. HYPOTHETICAL ACTIVITY FACTOR DATA

Site	Marketed Gas (x) (MMscfd)	Number of Devices (y)
1	20.0	4
2	30.0	2
3	80.0	8
4	10.0	2
Totals	140.0	16

The quantity n in this context is the number of sites visited. Both the use of n as the sample size in the preceding section and the use of n as the number of sites visited in this section are consistent with standard notation.

The average number of devices per site is:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{16}{4} = 4.00 \ \textit{devices/site}$$

The average production per site is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{140.0}{4} = 35.0 \ \textit{MMcfd/site}$$

Then $\hat{R}$, the activity-factor ratio, is as follows:

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{4.00}{35.0} = 0.114 \ devices/MMcfd$$

Suppose the published value of the total marketed gas for this source category is X = 14,000 MMscfd. The activity factor, AF, is denoted $\hat{Y}_R$ in the context of the ratio method:

$$AF = \hat{Y}_R = X\hat{R} = (14{,}000)(0.114) = 1{,}600 \ devices$$

The quantity N, the total number of sites, is not known and must be estimated:

$$N = \frac{X}{\bar{x}} = \frac{14{,}000}{35} = 400 \ sites$$

The ratio f of sites visited, n, to total sites, N, is:

$$f = \frac{n}{N} = \frac{4}{400} = 0.0100$$

We are now ready to calculate the error variance, $V(\hat{Y}_R)$, of the activity factor:

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n}(y_i-\hat{R}x_i)^2$$

$$= \frac{400^2(1-0.0100)}{(4)(4-1)}\{[(4-(0.114)(20)]^2 + \ . \ . \ . \ +[2-(0.114)(10)]^2\}$$

$$= 92{,}700$$

The standard error $S(\hat{Y}_R)$, which is the standard deviation of the error of the activity factor, is simply the square root of the error variance:

$$S(\hat{Y}_R) = \sqrt{V(\hat{Y}_R)} = 304 \ devices$$

The uncertainty in the activity factor is obtained by multiplying the standard error by the appropriate t-value from a standard statistical table. The appropriate t-value in this context for a sample size, n, of four (i.e., for three degrees of freedom) and for a 90% confidence level is 2.35. As is discussed earlier, the application of t here is not exact but is more conservative than using a z value. The use of t provides for the extra uncertainty attributable to the fact that the standard error of $\hat{Y}_R$ is estimated from the data.

The uncertainty, Tol($\hat{Y}_R$), is estimated as follows:

$$Tol(\hat{Y}_R) = tS(\hat{Y}_R) = (2.35)(304) = 715 \ devices$$

The uncertainty of $\hat{Y}_R$ in percent is obtained simply by multiplying this result by 100% and dividing by $\hat{Y}_R$:

$$Tol(\hat{Y}_R)\ (\%) = \frac{(100\%)Tol(\hat{Y}_R)}{\hat{Y}_R} = \frac{(100)(715)}{1{,}600} = 44.7\%$$

A.8.3 Category Annual Emission Calculations

From the previous subsections, we have the emission factor and the activity factor for the hypothetical category. Thus, we can calculate the annual emissions and uncertainty measures.

First, it is necessary to convert the emission factor and its uncertainty from Scf to Bscf.

$$\text{EF} = 10{,}000 \text{ Scf} = 0.0000100 \text{ Bscf}$$

$$\text{Tol(EF)} = 3{,}570 \text{ Scf} = 0.00000357 \text{ Bscf}$$

The activity factor and its uncertainty are as follows:

$$\text{AF} = 1{,}600 \text{ devices}$$

$$\text{Tol(AF)} = 715 \text{ devices}$$

The annual emissions value is as follows:

$$\text{ER} = \text{(AF)(EF)} = (1{,}600)(0.00001) = 0.016 \text{ Bscf}$$

The uncertainty of this quantity is as follows:

$$Tol(ER) = \sqrt{AF^2 Tol(EF)^2 + EF^2 Tol(AF)^2 + Tol(AF)^2 Tol(EF)^2}$$

By direct substitution of the values above, we obtain the following:

$$Tol(ER) = 0.00950 \ Bscf$$

This is readily converted to a percent:

$$Tol(ER)(\%) = \frac{(100\%)Tol(ER)}{ER} = \frac{(100)(0.00950)}{0.016} = 59.4\%$$

It remains to calculate the conservative uncertainty. As is discussed in Section A.4, this measure is based on the upper confidence limit, assuming the error in ER is lognormally distributed.

In Section A.4, the term y was used to denote the annual emissions; the methodology developed applies to either the emissions for a category (y equals ER) or the emissions for the industry (y equals ER_T). The development required transformations between linear space and log space. The notation used facilitated the association of corresponding quantities in the two spaces; expressions involving y correspond to linear space, while expressions involving Y correspond to log space. For consistency and convenience, we retain the notation of Section A.4 for the purposes of calculating the conservative uncertainty. Here, y is the annual emissions ER for a hypothetical source category. First, we approximate the standard error of the annual emissions:

$$s_y' = \frac{Tol(y)}{1.645} = \frac{0.00950}{1.645} = 0.00578 \ Bscf$$

Regarding the use of the value 1.645 here and below, see Section A.4. Given the estimated annual emissions y and the value s_y', it is possible to compute the mean and standard deviation of the natural logarithm of emissions:

$$Y = -\frac{1}{2}\ln\frac{(s_y')^2+y^2}{y^4} = -\frac{1}{2}\ln\frac{0.00578^2+0.0160^2}{0.0160^4} = -4.20$$

The estimated standard deviation in log space is as follows:

$$s_Y' = \sqrt{2(\ln(y)-Y)} = \sqrt{2\{\ln(0.0160)-(-4.20)\}} = 0.350$$

The upper confidence limit for Y is as follows:

$$Y+1.645s_Y' = -4.20+(1.645)(0.350) = -3.62$$

The bias correction factor is:

$$F = e^{\frac{1}{2}(s_Y')^2} = e^{\frac{1}{2}(0.350)^2} = 1.06$$

The conservative upper confidence limit is as follows:

$$U_{cons} = Fe^{Y+1.645s_Y'} = 1.06e^{-4.20+(1.645)(0.350)} = 0.0285\ Bscf$$

The conservative uncertainty based on this upper confidence limit is:

$$Conservative\ Tol(y) = U_{cons}-y = 0.0285-0.0160 = 0.0125\ Bscf$$

This is readily converted to a percentage value:

$$Conservative\ Tol(y)\ (\%) = \frac{(100\%)Conservative\ Tol(y)}{y}$$

$$= \frac{(100)(0.0125)}{0.0160} = 77.8\%$$

A.8.4 <u>Industry Annual Emission Calculations</u>

The data set used for the industry annual emission calculations is included in Appendix C. While this data set is not ideal for illustrative purposes in view of its size, it was felt that illustration of the industry emissions calculations using the actual data in the summary table would be beneficial. These calculations do not involve category-to-category special cases, as do calculations of annual emissions and uncertainties for individual categories. The role of correlated errors is illustrated through both a hypothetical example and calculations with real data.

The first step is to calculate the industry annual emissions, ER_T. This value is simply the sum of the emissions for the 86 categories. Using values from the summary table in Appendix C, ER_T is computed as follows:

$$ER_T = \sum_{i=1}^{86} ER_i$$

$$= 0.3352+0.0013+\ .\ .\ .\ +2.0631 = 314\ Bscf$$

where ER_i denotes the annual emissions for the ith source category. The emissions for the first two categories and the last category listed in the table in Appendix C are shown explicitly here.

As is discussed in Section 6.1, uncertainties were computed on the basis of several assumptions to illustrate the effect of certain factors and to arrive at a final measure of uncertainty. In the baseline case, the error in the industry annual emissions is assumed to be normally distributed, and the errors for different source categories are assumed to be uncorrelated. In this case, the uncertainty of the industry annual emissions is the square root of the sum of squares of the uncertainties of the emissions for the categories:

$$Tol(ER_T) = \sqrt{\sum_{i=1}^{86} Tol(ER_i)^2}$$

$$= \sqrt{0.2749^2 0.3352^2 + 2.1764^2 0.0013^2 + \ . \ . \ . \ + 19.2441^2 2.0631^2}$$

$$= 89.6 \ Bscf$$

where the uncertainty for a given category is the precision, expressed as a fraction, times the annual emissions in Bscf. Again, the values shown are from Appendix C. It is felt that the conversion from an absolute uncertainty to an uncertainty in percent is basic and has been sufficiently illustrated. The absolute uncertainty above is equivalent to an uncertainty of 28.5% of annual emissions. This uncertainty is better than the target precision of 0.5% of production (see Section 6.1).

The uncertainty given above can be converted to a conservative uncertainty. This calculation is analogous to the conversion to a conservative uncertainty for category annual emissions, which is illustrated in Section A.8.3. As shown in the data summary table in Appendix C, the conservative uncertainty of the industry annual emissions is 32.7% of emissions. This uncertainty is also better than the target precision.

The calculation of the uncertainty based on the assumption of correlated errors is the same as the uncertainty calculation in the baseline case, except that additional terms are involved. The following is the simplest expression for the additional terms:

$$2r_{ij}\mathrm{Tol}(\mathrm{ER}_i)\mathrm{Tol}(\mathrm{ER}_j)$$

where r_{ij} is the correlation coefficient between the errors in the annual emissions in the ith and jth categories. Further intuitive discussion of correlated errors is given in Section 4.4. Plots illustrating different levels of correlation are presented in Section 6.1. The exact role of these other terms is discussed in Section 4.4. The number of terms, including the 86 squared tolerances and the terms accounting for correlated errors, is large.

The errors in the emissions for two categories may be uncorrelated or may be correlated because of a common influence on their activity factors or a common influence on their emission factors. The correlation coefficients considered are given in Appendix C, as are categories postulated to have correlated errors.

The calculations associated with correlated errors will be illustrated in two steps. First, a simple hypothetical numerical example involving two sources will be used to illustrate the manner in which the term used to account for the correlation is combined with the uncertainties of the annual emissions for the categories. Second, the actual calculation of the term used to account for the correlated errors using data from the summary table in Appendix C is illustrated.

The following hypothetical numerical example involving two sources is more manageable for illustrative purposes than is the actual case involving 86 source categories. Consider two sources with emissions $E_1 = 3.00$ Bscf and $E_2 = 4.00$ Bscf. Suppose the uncertainties are $Tol(E_1) = 1.00$ Bscf and $Tol(E_2) = 2.00$ Bscf. The total emissions are as follows:

$$E_T = E_1 + E_2 = 7.00 \ Bscf$$

If the errors were uncorrelated, the uncertainty of the total emissions would be:

$$Tol(E_T) = \sqrt{Tol(E_1)^2 + Tol(E_2)^2} = 2.24 \ \ Bscf$$

Now, suppose the errors in the emissions for the two categories have a correlation coefficient, r, of 0.5. A plot illustrating the strength of the relationship that exists when the correlation coefficient is 0.5 is given in Figure 6-3. Then the uncertainty of the total emissions would be as follows:

$$Tol(E_T) = \sqrt{Tol(E_1)^2 + Tol(E_2)^2 + 2rTol(E_1)Tol(E_2)} = 2.65 \ \ Bscf$$

Thus, the correlation term increased the uncertainty of the total emissions by 0.041 Bscf, from 2.24 Bscf to 2.65 Bscf, in this hypothetical example.

Now the actual calculation of the uncertainty term using data in the summary table in Appendix C will be illustrated. The two sources are the first and third source categories listed on the first page of the table; this page pertains to the production segment. The first category (category i) includes gas wells (Eastern on shore). The other category considered here (category j) includes separators, listed under field separation equipment (Eastern on shore).

Emission factors are given in Scfd/well in the table. These are converted to annual Bscf/well for use in the calculations. The uncertainties are given in percentages. These are converted to absolute uncertainties for use in the calculations. The values needed are as follows:

$$EF_i = 7.11 \text{ Scfd/well} = 2.60 \times 10^{-6} \text{ Bscf/well}$$

$$Tol(EF_i) = 27\% = 7.01 \times 10^{-7} \text{ Bscf/well}$$

$$AF_i = 129{,}157 \text{ Wells}$$

$$Tol(AF_i) = 5\% = 6{,}460 \text{ Wells}$$

$$EF_j = 0.900 \text{ Scfd/Separator} = 3.29 \times 10^{-7} \text{ Bscf/Separator}$$

$$Tol(EF_j) = 27\% = 8.87 \times 10^{-8} \text{ Bscf/Separator}$$

$$AF_j = 91{,}670 \text{ Separators}$$

$$Tol(AF_j) = 23\% = 21{,}100 \text{ Separators}$$

The numerical value at the right of each equation above was used in the calculations.

In this case, both activity factors are involved in intercategory correlations but are not correlated with each other. The errors in the two emission factors have a weak correlation coefficient of 0.2. Thus,

$$r_{Eij} = 0.2$$

and

$$r_{Aij} = 0$$

where

r_{Eij} = correlation between the errors in the emission factors for the i^th^ and j^th^ categories, and

r_{Aij} = correlation between the errors in the activity factors for the same two categories.

From Section 4.4, the expression actually used to quantify the contribution of the correlated errors appears after the equals sign in the following equation:

$$2r_{ij}Tol(ER_i)Tol(ER_j) =$$

$$2\{AF_iAF_jr_{Eij}Tol(EF_i)Tol(EF_j)+EF_iEF_jr_{Aij}Tol(AF_i)Tol(AF_j)+$$

$$r_{Eij}Tol(EF_i)Tol(EF_j)r_{Aij}Tol(AF_i)Tol(AF_j)\}$$

Direct substitution of the data values given above into the expression for the correlation term produces the value 0.000294. This term, and other correlation terms, are added to the sum of squares of the uncertainties of the category annual emissions. The square root of the resulting sum is the uncertainty of the industry annual emissions in Bscf (see equation 14 in Section 4.4 and the two-category example given earlier in this subsection).

The uncertainty of the industry emissions based on correlated errors is converted to a conservative uncertainty in exactly the same way that the category annual emissions are so converted; this conversion is illustrated numerically in Section A.8.3.

The results of the calculations of the uncertainty for the industry annual emissions for all sets of assumptions considered are presented in Section 6.1. The uncertainties are given in Bscf and as percentages of national production.

References

1. Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*, Third Edition. McGraw-Hill Book Company, New York, 1974.

2. Finney, D.J. "On the Distribution of a Variate Whose Logarithm is Normally Distributed." Supplement to *Journal of the Royal Statistical Society*, 7, pp. 155-161, 1941.

3. Patterson, R.L. "Difficulties Involved in Estimation of a Population Mean Using Transformed Sample Data." *Technometrics*, 8(3), pp. 535-537, August 1966.

4. Cochran, William G. *Sampling Techniques*, Third Edition. John Wiley & Sons, New York, 1977.

5. Juran, J.M., Frank M. Gryna, Jr., and R. S. Bingham, Jr. *Quality Control Handbook*, Third Edition. McGraw-Hill Book Company, New York, 1974.

APPENDIX B

Further Details Regarding Outlier Test

APPENDIX B

FURTHER DETAILS REGARDING OUTLIER TEST

MEMORANDUM

TO: Bob Lott, GRI
David Kirchgessner, EPA

FROM: David Epperson and Lisa Campbell, Radian

COPY: Mike Cowgill, Radian
Hugh Williamson, Radian
Mike Campbell, Radian
Matt Harrison, Radian

DATE: November 10, 1994

SUBJECT: **Results of Statistical Outlier Tests for Plastic Main Leakage Data**

Attached is a brief document that discusses the results of the statistical tests performed to determine whether the large plastic main data point is an outlier. As you know, the issue of omitting the very large leak test data point for plastic mains was brought up in the August industry review meeting in Austin. The industry reviewers were concerned that a large overall leak rate for plastic mains would be misinterpreted, even though the contribution from plastic mains to the overall leakage from mains and services in the U.S. is very small.

On the basis of the results of the outlier tests performed, there is no statistical justification for omitting the large data point from the plastic leak measurements. Furthermore, PG&E's statistician who worked on the UAF study confirms that there is no technical or statistical justification for omitting that data point. Consequently, we recommend that the data point remain part of the distribution leak measurement database.

Results of Outlier Tests for Plastic Pipe Leakage Data

Overview

The GRI gas data for plastic pipes were screened for potential outliers. The Grubbs test,[1] the Dixon test,[1] the Fourth-Spread test,[2] and a conservative approach,[3] were used to identify potential outliers in the plastic pipe data. The Grubbs and Dixon tests require that the data being screened are normally distributed. The Fourth-Spread test does not strictly require normality, but it could produce spurious results if the data distribution were markedly asymmetric. The conservative approach addresses cases of normality and non-normality.

The largest value and the smallest value in the plastic pipe dataset were tested separately. Table 1 lists the results of the four outlier tests for both the largest and smallest plastic pipe data values. The smallest value is identified as a potential outlier only in the Fourth-Spread test; all other tests indicate no outliers. However, the test criteria from both the Grubbs and Dixon tests suggest that the smallest value is closer to being a potential outlier than the largest value.

Data

The plastic pipe flow rate data and the natural logarithms of these data, as well as the means and standard deviations, are shown in Table 2. The data in Table 2 are arranged so that the smallest value appears in the first row and the largest value appears in the last row of the table. Only six data points comprise the plastic pipe data and these six points span five orders of magnitude, ranging from 0.008 SCF/leak-hour to 61.000 SCF/leak-hour.

The Shapiro-Wilk W statistic, generated by the SAS UNIVARIATE[4] procedure, was used to determine whether the nontransformed and natural log-transformed plastic pipe data were normally distributed. For the nontransformed data, the W-statistic was 0.6068 and the associated p-value was 0.0001, indicating that the nontransformed data were not normally distributed. However, for the natural log-transformed data, the W-statistic was 0.9396 and the associated p-value was 0.6747, indicating that the natural log-transformed data were normally distributed, within random variability. Because of the small sample size (consisting of 6 data points), however, this test is not highly sensitive. Small or moderate deviations from normality might not be detected on the basis of a hypothesis test with this sample size. Figure 1 shows the frequency histogram for the nontransformed data and Figure 2 shows the frequency histogram for the natural log-transformed data to illustrate the

TABLE 1. RESULTS OF THE OUTLIER TESTS

Outlier Test	Data Value Tested (natural logarithm)	Criteria[a]	Result
Grubbs	Minimum: −4.8283 (ID 2014)	1.71<1.82	not an outlier
	Maximum: 4.1109 (ID 2002)	1.26<1.82	not an outlier
Dixon	Minimum: −4.8283 (ID 2014)	0.50<0.56	not an outlier
	Maximum: 4.1109 (ID 2002)	0.20<0.56	not an outlier
F-Spread	Minimum: −4.8283 (ID 2014)	outside bounds: −4.3850 to 6.3571	OUTLIER
	Maximum: 4.1109 (ID 2002)	inside bounds: −4.3850 to 6.3571	not an outlier
Conservative Approach	Minimum: −4.8283 (ID 2014)	inside bounds: −8.7334 to 9.3532	not an outlier
	Maximum: 4.1109 (ID 2002)	inside bounds: −8.7334 to 9.3532	not an outlier

[a] The criteria are based on the 5% significance level for the Grubbs and Dixon tests

TABLE 2. PLASTIC PIPE FLOW RATE DATA AND NATURAL LOGARITHMS OF THE FLOW RATES

Test ID Number	Standard Flow Rate (SCF/leak-hour)	Natural Log of Standard Flow Rate
2014	0.008	−4.8283
3020	0.700	−0.3567
3019	1.130	0.1222
3039	1.620	0.4824
11002	10.266	2.3288
2002	61.000	4.1109
Mean	**12.454**	**0.309894**
Standard Deviation	**24.084**	**3.014434**

results suggested by the W-statistics. The nontransformed data are obviously skewed and not normally distributed, while the natural log-transformed data are much more symmetric and appear to be closer to the normal distribution.

On the basis of the results of the normal distribution tests, the natural logarithms of the plastic pipe flow rates were tested for outliers using the Grubbs, Dixon, and Fourth-Spread tests. Following is a discussion of outlier screening in general, followed by specific details pertaining to each of the outlier tests used in this analysis.

Outlier Screening

Outliers have been defined as observations that do "not conform to the pattern established by other observations,"[5] or as observations that appear "to deviate markedly from other members of the sample in which" they occur.[1] Outliers may be caused by transcription, keypunch, or data-coding errors, instrument breakdowns, calibration problems, and power failures, or they may be manifestations of a greater amount of inherent spatial or temporal variability than expected.[6]

Many different tests exist to screen for outliers, some of which have certain limitations that prevent them from being applied to all datasets. Some tests require that the data be distributed normally because statistical parameters are used in the outlier test, while other tests rely on other types of information from the data to perform the outlier test. Because of the variety and number of different outlier tests, it is therefore important that no datum be discarded solely on the basis of a single statistical test. There should always be some plausible explanation other than the test result that warrants the exclusion or the replacement of an outlier.[6] If possible, several different types of tests should be applied to validate the results of the outlier screening process.

The four different tests applied to the GRI plastic pipe data represent some of the different types of outlier tests. The Grubbs test relies on statistical parameters (mean and

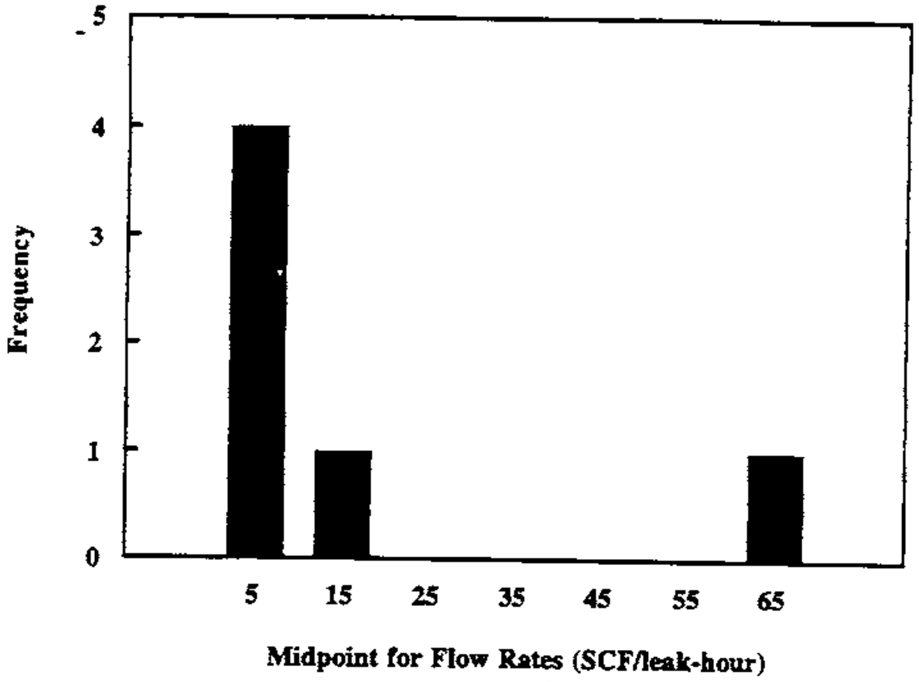


Figure 1. Frequency Histogram for Plastic Pipe Flow Rate Data

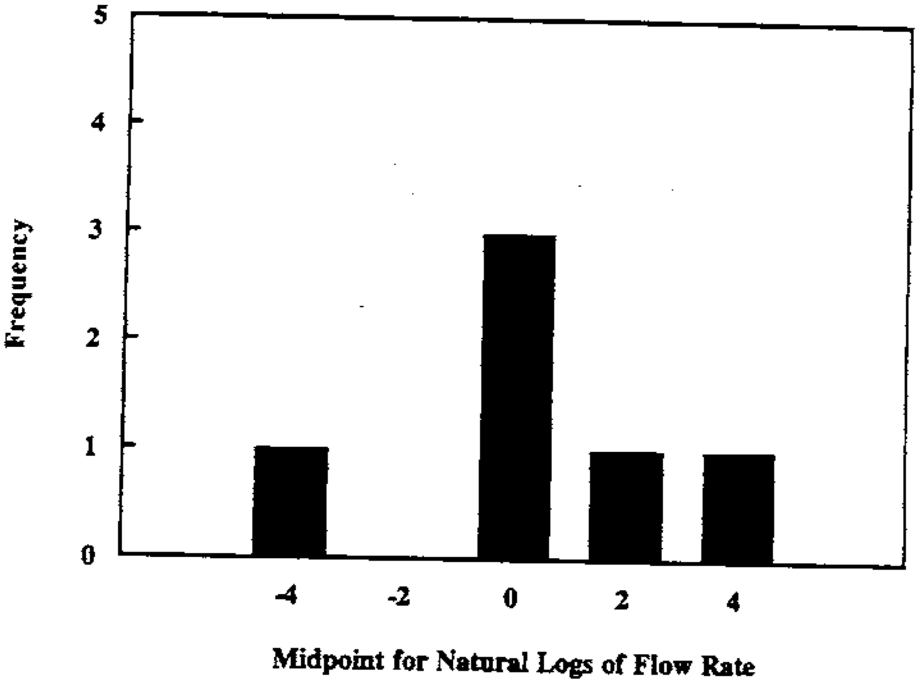


Figure 2. Frequency Histogram for the Natural Logarithms of the Plastic Pipe Flow Rate Data

standard deviation), the Dixon test relies on ratios of values in the tails, the Fourth-Spread test relies on the spread of the central half of the data, and the conservative approach is capable of handling any data distribution. Following are specific details regarding how each of these tests were applied to the plastic pipe data.

Grubbs Test[1]

The hypothesis tested in the Grubbs test is that all observations in the sample come from the same normal population. Thus, the transformation of skewed data, such as taking the natural logarithms, may be necessary. The data are ordered from smallest to largest for the Grubbs test, such that:

$$\{X_1 \leq X_2 \leq X_3 \leq ... \leq X_n\} \qquad (1)$$

The Grubbs test is then applied to a single suspect value–either the largest value (X_n) or the smallest value (X_1). For the largest value (X_n), the test statistic (T_n) is calculated as follows:

$$T_n = \frac{X_n - \overline{X}}{s} \qquad (2)$$

where:

X_n = the largest data value,
$\overline{X}$ = the arithmetic average of all n values, and
s = the sample standard deviation, with $n-1$ degrees of freedom.

For the smallest value (X_1), the test statistic (T_1) is calculated as follows:

$$T_1 = \frac{\overline{X} - X_1}{s} \qquad (3)$$

where:

$X_1 =$ the smallest data value, and

$\bar{X}$ and $s =$ the same as for Equation (2).

The test statistic (T_1 or T_n) is compared with the appropriate critical value for the statistic. When the test statistic is larger than the critical value, then the suspect data point is deemed a potential outlier.

Using the mean and standard deviation shown in Table 2 for the plastic pipe data, T_1=1.71 and T_n=1.26 for the natural logarithms of the flow rates. The critical value for a one-sided test using the 5% significance level for a sample size of six is 1.82, and the critical value using the recommended 1% significance level is 1.94.[1] Therefore, neither the largest nor the smallest of the natural logarithms of the plastic pipe flow rates were considered outliers by the Grubbs test.

Dixon Test[1]

The Dixon test is an alternative system that does not rely on the calculation of statistical parameters (e.g., the mean or standard deviation), and is based entirely on ratios of differences between some of the observations. As with the Grubbs test, the Dixon test requires a normal data distribution because the ratios of differences are calculated from both tails. One drawback to the Dixon test is that not all of the data are utilized–only data from the tails are used. Similarly to the Grubbs test, the data are ordered from smallest to largest for the Dixon test, as shown in Equation (1). The Dixon test is then applied to a single suspect value, either the largest or smallest of all of the data values. A test statistic (r_{xx}) that depends on sample size is calculated. The formula for the largest value (X_n) from a sample size of 6 (the plastic pipe data sample size) is:

$$r_{10} = \frac{X_n - X_{n-1}}{X_n - X_1} \qquad (4)$$

where:

$X_n =$ the largest data value,

$X_{n-1} =$ the second largest data value, and

X_1 = the smallest data value.

The corresponding formula for the smallest value (X_1) from a sample size of 6 data points (the plastic pipe data sample size) is:

$$r_{10} = \frac{X_2 - X_1}{X_n - X_1} \tag{5}$$

where:

X_2 = the second smallest data value,
X_1 = the smallest data value, and
X_n = the largest data value.

The test statistic (r_{10}) is compared to the appropriate critical value for the statistic. When the test statistic is larger than the critical value, the suspect data point is deemed a potential outlier.

Using the data shown in Table 2 for the plastic pipe data, r_{10}=0.20 for the largest value and r_{10}=0.50 for the smallest value of the natural logarithms of the flow rates. The critical value for a one-sided test using the 5% significance level for a sample size of six is 0.560, and the critical value using the recommended 1% significance level is 0.698.[1] Therefore, neither the largest nor the smallest of the natural logarithms of the plastic pipe flow rates were considered outliers by the Dixon test.

Fourth-Spread Test[2]

The Fourth-Spread (F-Spread) test does not rely on the calculation of the mean or standard deviation, rather it relies on information from the center half of the data mass to define the distance, beyond which, data points should be considered potential outliers. The center half of the distribution is relatively insensitive to outliers and, therefore, provides a reasonable basis for characterizing the distribution under the hypothesis that no outliers are present. As with the Grubbs and Dixon tests, the data must be arranged from smallest to largest, as shown in Equation (1). The data need not be normally distributed, but the distribution should be symmetric. First, the lower and upper fourths (also called the 25th and 75th percentiles, respectively) for the data distribution are calculated. The F-Spread (d_F) is then calculated by subtracting the lower-fourth (F_L) from the upper-fourth (F_U). Any data points larger than F_U+(1.5×d_F) and any data points smaller than F_L−(1.5×d_F) are then considered potential outliers. Figure 3 shows the relationship between the fourths and cutoffs used to define outliers with the F-Spread method.

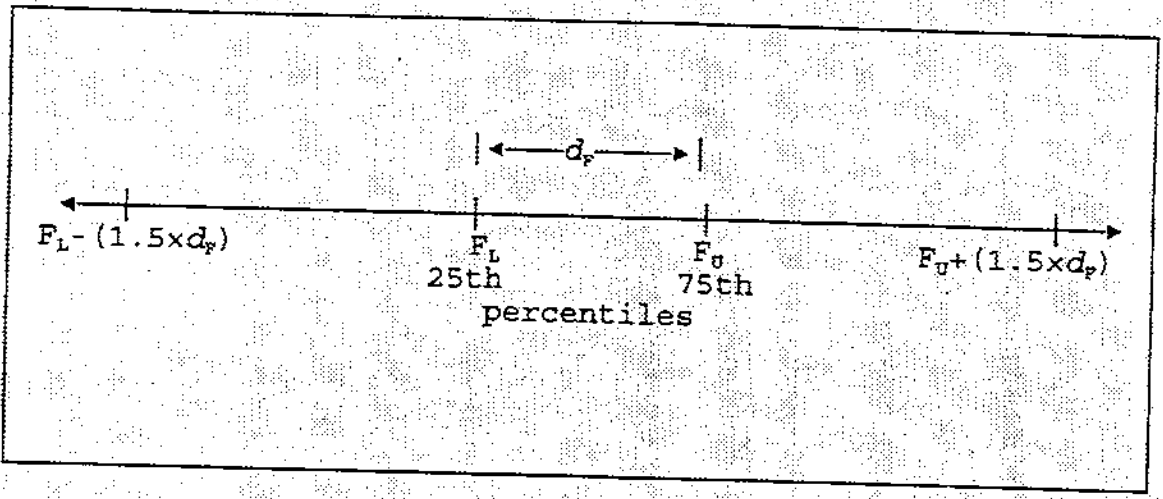


Figure 3. Depiction of the fourths (F_L and F_U), fourth-spread (d_F), and boundaries (F_L-1.5×d_F; F_U+1.5×d_F) for the F-Spread outlier detection method.

The F-Spread for the plastic pipe flow rate data was 2.6855 (F_U=2.3288 and F_L=−0.3567). Thus, data values smaller than −4.3850 or larger than 6.3571 should be considered potential outliers. One of the six plastic pipe data points, the smallest (ID=2014, value=0.008 SCF/leak-hour, ln value=−4.8283), was therefore considered a potential outlier.

Conservative Approach[3]

This approach is conservative because it screens for only the most blatant outliers. Thus, data points that may be considered outliers in other methods, may not be considered outliers by this approach, unless they are separated by a rather large distance from the main data mass. The histogram for the nontransformed data and the histogram from the natural log-transform of the data are used as visual aids in this method. Some measure of the normality (e.g., the Shapiro-Wilk W-statistic) for the data distributions shown in the two histograms is also used in this method. The following four steps are applied in sequence until the conditions are met and the criteria are defined for identifying outliers:

(1) The untransformed data distribution is normal. Values more than 3 standard deviations from the mean (mean±3×standard deviation) are considered potential outliers.

(2) The natural log-transformed data distribution is normal. Values more than 3 standard deviations from the mean of the natural logarithms (mean±3×standard deviation) are considered potential outliers.

(3) The untransformed data distribution is visually symmetric, but not normal. Values more than 3 standard deviations from the mean (mean±3×standard deviation) are considered potential outliers.

(4) The untransformed data distribution is not normal and not visually symmetric. Values more than 6 standard deviations from the mean (mean±6×standard deviation) are considered potential outliers. For the plastic pipe data, this method produced the following results for the first two steps (at which point the conditions were met and outlier criteria were established):

(1) The untransformed data distribution is not normal. Go to step 2.

(2) The natural log-transformed data distribution is normal. Therefore, values more than 3 standard deviations from the mean are considered potential outliers. Thus, using the mean and standard deviation shown in Table 2 for the natural logarithms of the flow rates, values more than 9.3532 or less than −8.7334 should be considered potential outliers. None of the data points met these criteria and therefore there were no outliers.

References

1. Grubbs, Frank E. "Procedures for Detecting Outlying Observations in Samples." *Technometrics*, 11(1), pp. 1-21, 1969.

2. Hoaglin, D.C., F. Mosteller, and J.W. Tukey. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, Inc., New York, 1983.

3. NSI Technology Services Corporation. *WHO and WMO Program Documentation*. Environ. Monitoring Systems Laboratory, U.S. EPA, Research Triangle Park, NC, under contract No. 68-02-4444. NSI Publication Number SP-4420-89-28, 1989.

4. SAS Institute, Inc. *SAS® Procedures Guide, Version 6, Third Edition*. Cary, NC: SAS Institute, Inc., 1990.

5. Hunt, W.F., Jr., G. Akland, W. Cox, T. Curran, N. Frank, S. Goranson, P. Ross, H. Sauls, and J. Suggs. *U.S. Environmental Protection Agency Intra-Agency Task Force on Air Quality Indicators*, EPA-450/4-81-015 (NTIS PB81-177982). Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC, 1981.

6. Gilbert, Richard O. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York, 1987.

APPENDIX C

Summary Data Table

APPENDIX C

SUMMARY DATA TABLE

Appendix C presents the data summary table referenced in various places in the body of this report. This table summarizes the final results of the study, including the estimate of the industry annual emissions and the uncertainty thereof.

The primary focus of this report is to discuss the statistical methods used in the study, not to present final results. For this reason, a detailed discussion of results for different source categories will not be presented here; Appendix C is presented here for completeness. The results are discussed in other project reports, such as Volume 1, executive summary,[1] and Volume 2, technical report.[2]

The column titled "Precision of Annual Emissions" is calculated as described in some detail in earlier parts of the report. This precision measure is based on the assumption that the error is normally distributed. A second uncertainty measure, in the column titled "Conservative Precision of Annual Emissions," was also calculated. This measure is based on the assumption that the error is lognormally distributed. The purpose of reporting this second precision value is to provide an approximate assessment of the uncertainties of the different emission rates if the normal assumption is not satisfied. The conservative precision value is larger than the normal precision estimate.

It is possible that the different emissions (the emissions by category and the industry annual emissions) have errors with different distributions. Thus, both types of precision measures are provided for each emission value given in the data summary table.

The conservative precision measure is briefly described as follows.

In computing the conservative precision, the same standard error was used as in the precision based on the normal assumption. The standard error is the estimated standard deviation of the error in the emission value. However, an approximate 90% asymmetric confidence interval was computed on the assumption that the error was lognormally distributed. The conservative precision is based on the upper confidence limit of this interval, i.e.,

$$p_{conservative} = 100\% \times (ER_{U,conservative} - ER)/ER$$

where

$p_{conservative}$ = conservative precision (%),

$ER_{U,conservative}$ = upper confidence limit based on the lognormal assumption, and

ER = estimated emission value.

Because of the asymmetry of the lognormal distribution (see Figure 4-2 and the accompanying discussion), the conservative precision value is larger than the precision based on the normal assumption. A lognormal precision based on the upper confidence limit was provided as a conservative (large) upper bound for the emissions. The analogous precision based on the lognormal lower confidence interval would be smaller than the precision based on the normal distribution. The issues related to the calculation of the conservative precision value are discussed in Section A.4. Figures A-1 and A-2 depict confidence intervals based on the normal and lognormal assumptions.

The data summary table also provides information concerning source categories with activity factors or emission factors that have possibly correlated errors. First, consider the activity factor groupings. All source categories with group 1, for example, are postulated to have activity factors with weakly correlated errors. However, a source category with group 1 and a category with group 2 are postulated to have activity factors with uncorrelated errors. If no group number is listed for a source category, its activity factor is assumed to be uncorrelated with that of any other category. A similar scheme was used for identifying groups of sources with emission factors whose errors may be correlated. The groups of categories shown were identified through engineering judgement and discussions between Radian and GRI staff.

These correlated groups were used to assess the impact of correlated errors among source categories on the uncertainty of the industry annual emissions. The results of this analysis are discussed in Section 6.1.

The groups for the activity factors are numbered from 1 to 16; i.e., 16 groups of categories were identified such that all members of a group had activity factors with possibly correlated errors. The groups for emission factors are numbered from 20 through 30. The group numbers have no quantitative meaning whatsoever. The fact that there are no groups numbered 17 through 19 has no importance.

The data summary table lists weak, medium, strong, and perfect correlations. The numerical values used are as follows:

weak	0.2
medium	0.5
strong	0.8
perfect	1.0
weak-medium	0.3
medium-strong	0.6

These correlations are postulated values, not accurate quantitative estimates, which are not available. Nevertheless, the postulated correlations provide a basis for assessing the sensitivity of the results to correlated errors.

METHANE EMISSION AND ACCURACY ESTIMATES

PROCESS SEGMENT Emission Type Source	1992 Emissions (Tg)	1992 Emissions (Bscf)	Percent of Total Emissions (%)	Percent of Total Production % (a)	Activity				Emission				Precision of Annual Emissions	Conservative Precision of Annual Emissions (d)	Target Precision (%) (c)
					Value	Units	Upper Bound (b)	Activity Factor Correlated Groups	Value	Units	Upper Bound (b)	Emission Factor Correlated Groups			
PRODUCTION															
Normal Fugitives															
Gas Wells (Eastern on shore)	0.0064	0 3352	0.11	0.002	129,157	wells	5%	1 weak	7.11	Scfd/well	27%	20 weak	27.49%	31.39%	1077.82
Field Separation Equipment (Eastern on shore)															
Heaters	0 0000	0.0013	0 00	0 000	260	heaters	196%	2 medium	14 21	scfd/heater	43%	20 weak	217.64%	423 13%	1500.00
Separators	0 0006	0 0301	0.01	0 000	91,670	separators	23%	2 medium	0.90	scfd/sep	27%	20 weak	38.01%	42.74%	1500.00
Gathering Compressors															
Small Recip. Compr.	0 0000	0 0006	0 00	0 000	129	compressors	33%	2 medium	12 1	scfd/comp	27%	20 weak	43.56%	53.45%	1500.00
Meters/Piping	0 0048	0 2508	0.08	0 001	78,262	meters	100%	1 weak	9.01	scfd/meter	30%	20 weak	108.63%	169 02%	1248 02
Dehydrators	0.0002	0.0083	0.00	0.000	1,047	dehydrators	20%	2 medium	21.75	scfd/dehy	35%	20 weak	40.91%	49 63%	1500.00
Gas Wells (Rest of US on shore)	0.0365	1 8969	0.60	0.009	142,771	wells	5%	1 weak	38.40	Scfd/well	24%	20 weak	24.54%	27.65%	453 08
Gulf of Mexico (offshore pltfrms)	0.0223	1.1615	0.37	0.005	1,092	platforms	10%	3 weak	2914	Scfd/plat	27%		28.92%	33 24%	579.01
Rest of US (offshore platforms)	0 0002	0 0095	0.00	0.000	22	platforms	10%	3 weak	1178	Scfd/plat	38%		37.54%	44.88%	1500.00
Field Separation Equipment (Rest of US on shore)															
Heaters	0.0206	1.0686	0.34	0.005	50,740	heaters	95%	2 medium	57 7	scfd/heater	40%	21 weak	109.86%	171.56%	603.64
Separators	0.0639	3.3252	1.06	0.015	74,674	separators	57%	2 medium	122.0	scfd/sep	33%	21 weak	68.50%	93.05%	342.20
Gathering Compressors															
Small Recip. Compr.	0.0318	1.6534	0 53	0.007	18,915	compressors	52%	2 medium	267.8	scfd/comp	68%	21 weak	92 62%	137.08%	485.29
Large Recip Compr.	0 0102	0.5328	0.17	0.002	96	compressors	100%	4 medium	15205.0	scfd/comp	65%	22 weak-med.	135.83%	227.42%	854.90
Large Recip. Stations	0.0007	0.0381	0.01	0.000	12	stations	100%	4 medium	8247.0	scfd/station	102%	21 weak	175.52%	319.82%	1500.00
Meters/Piping	0.1118	5.8153	1.85	0.028	301,180	meters	100%	1 weak	52.9	scfd/meters	30%	21 weak	108.63%	169 02%	258 78
Dehydrators	0.0235	1.2229	0.39	0.006	38,777	dehydrators	20%	2 medium	91.1	scfd/dehy	25%	21 weak	32.40%	37.84%	564.28
Pipeline Leaks	0.1269	6.6000	2.10	0.030	340,200	miles	10%	5 perfect	53.2	scfd/mile	107%	23 med -strong	108.00%	187.72%	242.89
Vented and Combusted															
Drilling and Well Completion															
Completion Flaring	0.0000	0.0006	0.00	0.000	844	compl/yr	10%		733	scf/compl	200%		201.25%	382.35%	1500.00
Normal Operations															
Pneumatic Device Vents	0.6037	31.3948	9.99	0.142	249,111	controllers	48%	2 medium	345	Scfd/device	40%	29 weak	64 99%	87.10%	111.37
Chemical Inj Pumps	0.0295	1.5365	0.49	0.007	16,971	active pumps	143%	2 medium	248 05	Scfd/pump	83%		203 53%	388.00%	503.41
Kimray Pumps	0.2108	10.9616	3.49	0.050	1.105E+07	MMscf/yr	82%	13 medium	992.00	scf/MMscf	77%		110.03%	171.90%	188.47
Dehydrator Vents	0.0657	3.4171	1.09	0.015	1.240E+07	MMscf/yr	62%	13 medium	275 57	scf/MMscf	154%		191.90%	359.36%	337.57
Compressor Exhaust Vented															
Gas Engines	0.1267	6.5904	2.10	0.030	27,460	MMHPhr	200%		0 240	scf/HPhr	5%	27 perfect	200.31%	380.04%	243.07
Routine Maintenance															
Well Workovers															
Gas Wells	0.0004	0 0230	0.01	0.000	9,392	w.o./yr	258%		2,454	scfy/w.o.	459%		1296.00%	2748.84%	1500.00
Well Clean Ups (LP Gas Wells)	0.1088	5.6579	1.80	0.026	114,139	LP gas wells	45%	2 medium	49570	scfy/LP well	344%	24 med.-strong	379.90%	834.58%	262.34
Blowdowns															
Vessel BD	0 0004	0.0200	0.01	0.000	255,996	vessels	26%	2 medium	78	Scfy/vsl	266%	24 med.-strong	276.07%	571.10%	1500 00
Pipeline BD	0.0020	0.1051	0 03	0.000	340,000	miles(gath)	10%	5 perfect	309	Scfy/mile	32%	24 med.-strong	33.68%	39.58%	1500.00
Compressor BD	0.0012	0.0646	0.02	0.000	17,112	compressors	52%	2 medium	3774	Scfy/comp	147%	24 med -strong	173.66%	315.14%	1500.00
Compressor Starts	0.0028	0.1445	0.05	0.001	17,112	compressors	52%	2 medium	8443	Scfy/comp	157%	24 med -strong	184 44%	341.16%	1500.00
Upsets															
Pressure Relief Valves	0.0003	0.0180	0.01	0.000	528,440	PRV	53%	2 medium	34	Scfy/PRV	252%	24 med.-strong	290.09%	608.86%	1500.00
ESD	0.0055	0 2864	0.09	0.001	1,115	platforms	10%	3 weak	256888	Scfy/plat	200%		201 25%	382.35%	1185.95
Mishaps (Dig-ins)	0 0044	0.2275	0.07	0.001	340,000	miles	10%	5 perfect	689	scf/mile/yr	1925%	25 medium	1934.63%	3768.88%	1308.38

(a) Based on a total gross national production of 22132 Bscf for 1992

(b) Precision based on a 90% confidence interval.

(c) Target Precision = 100*(8 24/SQRT(ER)), where ER = emissions in Bscf Overall TP is +/- 110 88 Bscf.
Maximum Relative Category TP is +/- 1500%, Minimun Relative Category TP is +/- 75%, where TP = target precision.

(d) Conservative precision based on upper limit of a 90% confidence interval. This confidence interval is based on a lognormal assumption.

METHANE EMISSION AND ACCURACY ESTIMATES

PROCESS SEGMENT Emission Type Source	1992 Emissions (Tg)	1992 Emissions (Bscf)	Percent of Total Emissions (%)	Percent of Total Production % (a)	Activity				Emission				Precision of Annual Emissions	Conservative Precision of Annual Emissions (d)	Target Precision (%) (c)
					Value	Units	Upper Bound (b)	Activity Factor Correlated Groups	Value	Units	Upper Bound (b)	Emission Factor Correlated Groups			
Gas Processing Plants															
Normal Fugitives															
Plants	0.0403	2 0950	0.67	0.009	726	plants	2%	6 perfect	7906	scfd/plant	48%	30 weak	48.05%	60.11%	431.12
Recip. Compressors	0 3216	16 7251	5 32	0.076	4,092	compressors	48%		11198	scfd/comp	74%	22 weak-med	95.09%	141.87%	152.58
Centrifugal Compressors	0.1082	5 6257	1.79	0.025	726	compressors	77%		21230	scfd/comp	39%	22 weak-med	91.39%	134.71%	263.09
Vented and Combusted															
Normal Operations															
Compressor Exhaust															
Gas Engines	0.1281	6.6624	2.12	0.030	27,760	MMHPhr	133%	9 medium	0.240	scf/HPhr	5%	27 perfect	133.26%	221.71%	241.75
Gas Turbines	0 0036	0.1876	0.06	0.001	32,910	MMHPhr	121%	9 medium	0.0057	scf/HPhr	30%	28 perfect	129.84%	214.17%	1440.74
AGR Vents	0 0158	0.8237	0.26	0.004	371	AGR units	20%		6083	scfd/AGR	105%		108.65%	169.48%	687.54
Kimray Pumps	0.0033	0.1703	0.05	0.001	957900	MMscf/yr	192%	14 medium	177 75	scf/MMscf	57%		228.00%	449 12%	1500.00
Dehydrator Vents	0.0202	1.0490	0.33	0.005	8,630,000	MMscf/yr	22%	14 medium	121.55	scf/MMscf	202%		208.20%	399.58%	609.28
Pneumatic Devices	0.0023	0.1196	0.04	0.001	726	gas plants	2%	6 perfect	164721	scfy/plant	133%	29 weak	133.04%	221.23%	1500.00
Routine Maintenance															
Blowdowns/Venting	0.0567	2.9475	0.94	0.013	726	gas plants	2%	6 perfect	4060	Mscfy/plant	282%	28 strong	262.16%	535.66%	363.46

(a) Based on a total gross national production of 22132 Bscf for 1992.
(b) Precision based on a 90% confidence interval.
(c) Target Precision = 100*(6.24/SQRT(ER)), where ER = emissions in Bscf. Overall TP is +/- 110.66 Bscf.
Maximum Relative Category TP is +/- 1500%, Minimun Relative Category TP is +/- 75%, where TP = target precision.
(d) Conservative precision based on upper limit of a 90% confidence interval. This confidence interval is based on a lognormal assumption.

METHANE EMISSION AND ACCURACY ESTIMATES

PROCESS SEGMENT Emission Type Source	1992 Emissions (Tg)	1992 Emissions (Bscf)	Percent of Total Emissions (%)	Percent of Total Production % (a)	Activity				Emission				Precision of Annual Emissions	Conservative Precision of Annual Emissions (d)	Target Precision (%) (c)
					Value	Units	Upper Bound (b)	Activity Factor Correlated Groups	Value	Units	Upper Bound (b)	Emission Factor Correlated Groups			
TRANSMISSION/STORAGE															
Fugitives															
Pipeline Leaks	0.0031	0.1600	0.05	0.001	284,500	miles	5%	7 perfect	1.541	scfd/mile	89%	23 med.-strong	89.00%	130.14%	1500.00
Compressor Stations (TRANS)															
Station	0.1047	5.4467	1.73	0.025	1,700	stations	10%		8778	scfd/station	102%	30 weak	103.00%	157.55%	267.37
Recip. Compressor	0.7256	37.7333	12.01	0.170	6,799	comp.	17%		15205	scfd/comp.	65%	22 weak-med.	68.09%	92.35%	101.58
Centrifugal Compressor	0.1449	7.5328	2.40	0.034	681	comp.	26%		30305	scfd/comp.	34%	22 weak-med.	43.71%	53.87%	227.36
Compressor Stations (STOR)															
Station	0.0717	3.7268	1.19	0.017	475	stations	5%		21507	scfd/station	100%	30 weak	100.25%	152.05%	323.15
Recip. Compressor	0.2069	10.7594	3.42	0.049	1,396	comp.	58%		21118	scfd/comp	48%	22 weak-med	80.27%	113.88%	190.24
Centrifugal Compressor	0.0292	1.5178	0.48	0.007	138	comp.	119%		30573	scfd/comp.	34%	22 weak-med.	130.21%	214.97%	506.53
Wells (STOR)	0.0145	0.7522	0.24	0.003	17,999	wells	5%		114.5	scfd/well	78%		78.26%	106.64%	719.47
M&R (Trans. Co Interconnect)	0.0708	3.6834	1.17	0.017	2,533	stations	776%		3984	scfd/station	80%		996.98%	2197.40%	325.14
M&R (Farm Taps + Direct Sales)	0.0159	0.8271	0.28	0.004	72,630	stations	780%		31.2	scfd/station	80%		1002.09%	2207.28%	686.13
Vented and Combusted															
Normal Operations															
Dehydrator Vents (TRANS)	0.0020	0.1018	0.03	0.000	1,066,000	MMscf/yr	144%		93.72	scf/MMscf	208%		391.75%	884.25%	1500.00
Dehydrator Vents (STOR)	0.0045	0.2344	0.07	0.001	2,000,000	MMscf/yr	25%		117.16	scf/MMscf	160%		166.56%	298.24%	1288.98
Compressor Exhaust															
Engines (TRANS)	0.1884	9.6912	3.08	0.044	40,380	MMHPhr	17%	10 medium	0.240	scf/HPhr	5%	27 perfect	17.74%	19.35%	200.45
Turbines (TRANS)	0.0011	0.0549	0.02	0.000	9,635	MMHPhr	33%	10 medium	0.0057	scf/HPhr	30%	28 perfect	45.68%	56.58%	1500.00
Engines (STOR)	0.0227	1.1813	0.38	0.005	4,922	MMHPhr	27%	11 medium	0.240	scf/HPhr	5%	27 perfect	27.49%	31.39%	574.13
Turbines (STOR)	0.0002	0.0099	0.00	0.000	1729	MMHPhr	826%	11 medium	0.0057	scf/HPhr	30%	28 perfect	654.25%	1485.73%	1500.00
Generators (Engines)	0.0091	0.4748	0.15	0.002	1,978	MMHPhr	45%	12 medium	0.240	scf/HPhr	5%	27 perfect	45.25%	55.94%	905.60
Generators (Turbines)	0.0000	0.0001	0.00	0.000	23.3	MMHPhr	1114%	12 medium	0.0057	scf/HPhr	30%	28 perfect	1163.33%	2510.01%	1500.00
Pneumatic Devices	0.2720	14.1446	4.50	0.064	87,206	devices	38%		162197	scfy/device	44%	29 weak	80.49%	79.85%	165.92
Routine Maintenance/Upsets															
Pipeline Venting	0.1732	9.0044	2.87	0.041	284,500	miles	5%	7 perfect	31.65	Mscfy/mile	236%		236.25%	469.92%	207.95
Station Venting	0.1823	9.4800	3.02	0.043	2,175	cmp stations	8%		4358	Mscfy/station	282%	26 strong	282.66%	538.93%	202.67

(a) Based on a total gross national production of 22132 Bscf for 1992.

(b) Precision based on a 90% confidence interval.

(c) Target Precision = 100*(8.24/SQRT(ER)), where ER = emissions in Bscf. Overall TP is +/- 110.66 Bscf.
Maximum Relative Category TP is +/- 1500%, Minimum Relative Category TP is +/- 75%, where TP = target precision.

(d) Conservative precision based on upper limit of a 90% confidence interval. This confidence interval is based on a lognormal assumption.

METHANE EMISSION AND ACCURACY ESTIMATES

PROCESS SEGMENT Emission Type Source	1992 Emissions (Tg)	1992 Emissions (Bscf)	Percent of Total Emissions (%)	Percent of Total Production % (a)	Activity				Emission				Precision of Annual Emissions	Conservative Precision of Annual Emissions (d)	Target Precision (%) (c)
					Value	Units	Upper Bound (b)	Activity Factor Correlated Groups	Value	Units	Upper Bound (b)	Emission Factor Correlated Groups			
DISTRIBUTION															
Normal Fugitives															
Pipeline Leaks															
Mains - Cast Iron	0.2538	13.1992	4 20	0.060	55,288	miles	5%		238.7	Mscf/mile-yr	84%		63 97%	85.39%	171 76
Mains - Unprotected Steel	0.1740	9.0478	2.88	0.041	174,857	equiv. leaks	58%	15 weak	51.8	Mscf/leak-yr	93%		122 42%	196 05%	207.45
Mains - Protected Steel	0.0266	1.3846	0.44	0.006	88,308	equiv. leaks	62%	15 weak	20.3	Mscf/leak-yr	85%		118.00%	188.59%	530.30
Mains - Plastic	0.0945	4.9150	1 56	0.022	49,226	equiv. leaks	116%	15 weak	99.8	Mscf/leak-yr	166%		282.18%	586.66%	281.47
Services - Unprotected Steel	0.1781	9.2630	2.95	0.042	458,476	equiv. leaks	109%	15 weak	20.2	Mscf/leak-yr	105%		189 27%	352.92%	205.03
Services - Protected Steel	0.0691	3.5922	1.14	0.016	390,628	equiv. leaks	135%	15 weak	9.20	Mscf/leak-yr	61%		188.90%	303.79%	329.24
Services - Plastic	0.0032	0.1644	0.05	0.001	68,903	equiv. leaks	87%	15 weak	2.39	Mscf/leak-yr	143%		221.59%	433.02%	1500.00
Services - Copper	0.0011	0.0593	0.02	0.000	7,720	equiv. leaks	110%	15 weak	7.68	Mscf/leak-yr	72%		154.25%	269.35%	1500.00
Meter/Regulator (City Gates)															
M & R > 300	0.1048	5.4510	1.73	0.025	3,460	stations	71%	16 weak	179.6	scfh/station	39%		85 46%	123.47%	287.27
M & R 100-300	0.2149	11.1731	3.56	0.050	13,335	stations	106%	16 weak	95.6	scfh/station	112%		194.97%	366.89%	186.68
M & R < 100	0.0052	0.2693	0.09	0.001	7,127	stations	116%	16 weak	4.31	scfh/station	227%		370.94%	812.06%	1202.35
Reg > 300	0.1090	5.6655	1.80	0.026	3,995	stations	68%	16 weak	161.9	scfh/station	58%		97.37%	148.35%	262.16
R-Vault > 300	0.0005	0.0266	0 01	0.000	2,346	stations	66%	16 weak	1 30	scfh/station	162%		230 44%	455.26%	1500.00
Reg 100-300	0.0837	4.3520	1.38	0.020	12,273	stations	61%	16 weak	40.5	scfh/station	86%		98.47%	148.52%	299.12
R-Vault 100-300	0.0002	0 0087	0.00	0.000	5,514	stations	61%	16 weak	0 180	scfh/station	94%		128.14%	206.09%	1500.00
Reg 40-100	0.0064	0.3317	0 11	0.001	36,326	stations	64%	16 weak	1.04	scfh/station	74%		109.09%	169.96%	1063.42
R-Vault 40-100	0.0005	0 0244	0.01	0.000	32,215	stations	64%	16 weak	0.0665	scfh/station	64%		98.97%	149.51%	1500.00
Reg < 40	0 0003	0 0179	0.01	0.000	15,377	stations	65%	16 week	0 133	scfh/station	135%		173.87%	315.67%	1500.00
Customer Meters															
Residential	0.1067	5.5468	1.76	0.025	40,049,306	outdr meters	10%		138 5	scfy/meter	17%		19.80%	21.80%	264.95
Commercial/Industry	0.0042	0 2207	0.07	0.001	4,606,000	meters	5%		47.9	scfy/meter	35%		35.40%	41.91%	1328.20
Vented															
Routine Maintenance															
Pressure Relief Valve Releases	0.0008	0.0418	0.01	0.000	836,760	mile main	5%		0 050	Mscf/mile	3914%		3918.89%	6199.19%	1500.00
Pipeline Blowdown	0 0025	0.1324	0.04	0.001	1,297,569	miles	5%	8 perfect	0.102	Mscfy/mile	2521%		2524.15%	4579.78%	1500.00
Upsets															
Mishaps (Dig-ins)	0.0397	2.0631	0.66	0.009	1,297,569	miles	5%	8 perfect	1 59	Mscfy/mile	1922%	25 medium	1924.41%	3751.65%	434.43
INDUSTRY TOTAL EMISSIONS	6.0437	314.2714	100.0000	1.4200											
UNCERTAINTY (+/-)	0.0172	89.6029											28.51%	32 71%	35 21

(a) Based on a total gross national production of 22132 Bscf for 1992.

(b) Precision based on a 90% confidence interval.

(c) Target Precision = 100*(6.24/SQRT(ER)), where ER = emissions in Bscf. Overall TP is +/- 110.66 Bscf.
Maximum Relative Category TP is +/- 1500%, Minimun Relative Category TP is +/- 75%, where TP = target precision.

(d) Conservative precision based on upper limit of a 90% confidence interval This confidence interval is based on a lognormal assumption.

References

1. Harrison, M.R., T.M. Shires, J.K. Wessels, and R.M. Cowgill. *Methane Emissions from the Natural Gas Industry, Volume 1: Executive Summary,* Final Report, GRI-94/0257 and EPA-600/R-96-080a, Gas Research Institute and U.S. Environmental Protection Agency, June 1996.

2. Harrison, M.R., L.M. Campbell, T.M. Shires, and R.M. Cowgill. *Methane Emissions from the Natural Gas Industry, Volume 2: Technical Report,* Final Report, GRI-94/0257.1 and EPA-600/R-96-080b, Gas Research Institute and U.S. Environmental Protection Agency, June 1996.

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/R-96-080d

2.

3. RECIPIENT'S ACCESSION NO.

4. TITLE AND SUBTITLE Methane Emissions from the Natural Gas Industry, Volumes 1-15 (Volume 4: Statistical Methodology)

5. REPORT DATE June 1996

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S) L. Campbell, M. Campbell, M. Cowgill, D. Epperson, M. Hall, M. Harrison, K. Hummel, D. Myers, T. Shires, B. Stapper, C. Stapper, J. Wessels, and *

8. PERFORMING ORGANIZATION REPORT NO. DCN 96-263-081-17

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Radian International LLC
P.O. Box 201088
Austin, Texas 78720-1088

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO. 5091-251-2171 (GRI)
68-D1-0031 (EPA)

12. SPONSORING AGENCY NAME AND ADDRESS

EPA, Office of Research and Development
Air Pollution Prevention and Control Division
Research Triangle Park, NC 27711

13. TYPE OF REPORT AND PERIOD COVERED Final; 3/91-4/96

14. SPONSORING AGENCY CODE EPA/600/13

15. SUPPLEMENTARY NOTES EPA project officer is D. A. Kirchgessner, MD-63, 919/541-4021. Cosponsor GRI project officer is R. A. Lott, Gas Research Institute, 8600 West Bryn Mawr Ave., Chicago, IL 60631. (*)H. Williamson (Block 7).

16. ABSTRACT The 15-volume report summarizes the results of a comprehensive program to quantify methane (CH4) emissions from the U.S. natural gas industry for the base year. The objective was to determine CH4 emissions from the wellhead and ending downstream at the customer's meter. The accuracy goal was to determine these emissions within +/-0.5% of natural gas production for a 90% confidence interval. For the 1992 base year, total CH4 emissions for the U.S. natural gas industry was 314 +/- 105 Bscf (6.04 +/- 2.01 Tg). This is equivalent to 1.4 +/- 0.5% of gross natural gas production, and reflects neither emissions reductions (per the voluntary American Gas Association/EPA Star Program) nor incremental increases (due to increased gas usage) since 1992. Results from this program were used to compare greenhouse gas emissions from the fuel cycle for natural gas, oil, and coal using the global warming potentials (GWPs) recently published by the Intergovernmental Panel on Climate Change (IPCC). The analysis showed that natural gas contributes less to potential global warming than coal or oil, which supports the fuel switching strategy suggested by the IPCC and others. In addition, study results are being used by the natural gas industry to reduce operating costs while reducing emissions.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Pollution	Pollution Prevention	13B
Emission	Stationary Sources	14G
Greenhouse Effect	Global Warming	04A
Natural Gas		21D
Gas Pipelines		15E
Methane		07C

18. DISTRIBUTION STATEMENT

Release to Public

19. SECURITY CLASS *(This Report)* Unclassified

20. SECURITY CLASS *(This page)* Unclassified

21. NO. OF PAGES 132

22. PRICE

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Research and Development
National Risk Management Research Laboratory
Technology Transfer and Support Division
Cincinnati, Ohio 45268

Publication No. EPA- 600/R-96-080d