Confidence Intervals and Hypothesis Tests

A

CONFIDENCE INTERVALS and formal hypothesis tests are two statistical methods that can be used for decisionmaking. A hypothesis test controls both the false positive decision error rate (α) and false negative decision error rate (β). A confidence interval controls only the probability of making a false positive decision error (α) (for example, concluding that a site is clean when it is truly dirty). However, the probability of making a false negative decision error (β) is fixed at 50% for confidence intervals (i.e., $\beta = 0.5$).

A confidence interval and a hypothesis test can be very similar. Consider the problem of determining whether the mean concentration (μ) of a site exceeds a cleanup standard (CS), where the contaminant is normally distributed. A confidence interval could be constructed for the mean, or a t-test could be used to test the statistical hypothesis:

$$H_0\text{: } \mu > CS \text{ vs. } H_a\text{: } \mu < CS$$

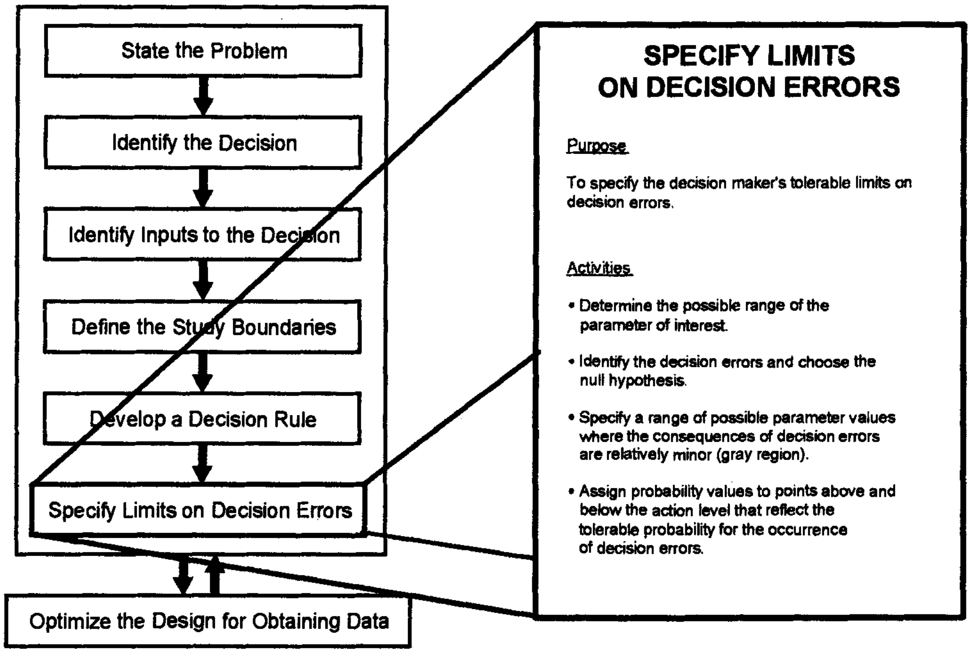
If the site manager's false negative decision error rate is 0.5 (i.e., $\beta = 0.5$), these methods are the same. Additionally, with a fixed α, the sample size of a confidence interval influences only the width of the interval (since $\beta = 0.5$). Similarly, the sample size of a t-test influences β and δ (where δ = upper value of the gray region minus the lower value of the gray region). However, by solving for the sample size using a t-test, one can substitute back into the sample size equation for a confidence interval and compute a width corresponding to this sample size. Then the results of the two methods will be identical.

Although the results of the hypothesis test and the confidence interval may be identical, the hypothesis test has the added advantage of a power curve. The power curve is defined as the probability of rejecting the null hypothesis. An ideal power curve is 1 for those values corresponding to the alternative hypothesis (all $\mu < CS$ in the example above) and 0 for those values corresponding to the null hypothesis (all $\mu > CS$ in the example above). The power curve is thus a way to tell how well a given test performs and can be used to compare two or more tests. Additionally, if the null hypothesis is not rejected, the power curve gives the decisionmaker some idea of whether or not the design could actually reject the null hypothesis for a given level (μ).

There is no corresponding idea of a power curve in terms of confidence intervals. To derive a power curve, one would need to translate the confidence interval into the corresponding test (i.e., a t-test) and then compute the power curve. Additionally, whereas a statistical test accounts directly for the false negative decision error, a confidence interval does not ($\beta = 0.5$). Finally, a confidence interval and a statistical test almost always are based on distributional assumptions, independence assumptions, etc. If these assumptions are violated, it may be easier to select an alternative test (for example, a non-parametric test) than it is to derive an alternative confidence interval. For these reasons, this document concentrates its discussion on hypothesis testing.

Step 6: Specify Tolerable Limits on Decision Errors

THE DATA QUALITY OBJECTIVES PROCESS


State the Problem
Identify the Decision
Identify Inputs to the Decision
Define the Study Boundaries
Develop a Decision Rule
Specify Limits on Decision Errors
Optimize the Design for Obtaining Data

SPECIFY LIMITS ON DECISION ERRORS

Purpose

To specify the decision maker's tolerable limits on decision errors.

Activities

- Determine the possible range of the parameter of interest.
- Identify the decision errors and choose the null hypothesis.
- Specify a range of possible parameter values where the consequences of decision errors are relatively minor (gray region).
- Assign probability values to points above and below the action level that reflect the tolerable probability for the occurrence of decision errors.

PURPOSE

The purpose of this step is to specify the decisionmaker's tolerable limits on decision errors, which are used to establish performance goals for the data collection design.

EXPECTED OUTPUTS

- The decisionmaker's tolerable decision error rates based on a consideration of the consequences of making an incorrect decision.

BACKGROUND

Decisionmakers are interested in knowing the true state of some feature of the environment. Since data can only *estimate* this state, decisions that are based on measurement data could be in error (decision error). Most of the time the correct decision will be made; however, this chapter will focus on controlling the less likely possibility of making a decision error. The goal of the planning team is to develop a data collection design that reduces the chance of making a decision error to a tolerable level. This step of the DQO process will provide a mechanism for allowing the decisionmaker to define tolerable limits on the probability of making a decision error.

There are two reasons why the decisionmaker cannot know the true value of a population parameter (i.e., the true state of some feature of the environment):

(1) The population of interest almost always varies over time and space. Limited sampling will miss some features of this natural variation because it is usually impossible or impractical to measure every point of a population. *Sampling design error* occurs when the sampling design is unable to capture the complete extent of natural variability that exists in the true state of the environment.

(2) Analytical methods and instruments are never absolutely perfect, hence a measurement can only estimate the true

[*] Pages 32–36 from EPA's QA/5-4 (Ch. 2, Ref *1*).

value of an environmental sample. *Measurement error* refers to a combination of random and systematic errors that inevitably arise during the various steps of the measurement process (for example, sample collection, sample handling, sample preparation, sample analysis, data reduction, and data handling).

The combination of sampling design error and measurement error is called *total study error,* which may lead to a decision error. Since it is impossible to eliminate error in measurement data, basing decisions on measurement data will lead to the possibility of making a decision error.

The probability of decision errors can be controlled by adopting a scientific approach. In this approach, the data are used to select between one condition of the environment (the *null hypothesis,* H_0) and an alternative condition (the *alternative hypothesis,* H_a). The null hypothesis is treated like a baseline condition that is presumed to be true in the absence of strong evidence to the contrary. This feature provides a way to guard against making the decision error that the decisionmaker considers to have the more undesirable consequences.

A decision error occurs when the decisionmaker rejects the null hypothesis when it is true, or fails to reject the null hypothesis when it is false. These two types of decision errors are classified as *false positive* and *false negative* decision errors, respectively. They are described below.

False Positive Decision Error—A false positive decision error occurs when the null hypothesis (H_0) is rejected when it is true. Consider an example where the decisionmaker presumes that a certain waste is hazardous (i.e., the null hypothesis or baseline condition is "the waste is hazardous"). If the decisionmaker concludes that there is insufficient evidence to classify the waste as hazardous when it truly is hazardous, then the decisionmaker would make a false positive decision error. A statistician usually refers to the false positive error as a "Type I" error. The measure of the size of this error is called alpha (α), the level of significance, or the size of the critical region.

False Negative Decision Error—A false negative decision error occurs when the null hypothesis is *not* rejected when it is false. In the above waste example, the false negative decision error occurs when the decisionmaker concludes that the waste is hazardous when it truly is *not* hazardous. A statistician usually refers to a false negative error as a "Type II" error. The measure of the size of this error is called beta (β), and is also known as the complement of the *power* of a hypothesis test.

The definition of false positive and false negative decision errors depends on the viewpoint of the decision maker.[1] Consider the viewpoint where a person has been presumed to be "innocent until proven guilty" (i.e., H_0 is "innocent"; H_a is "guilty"). A false positive error would be convicting an innocent person; a false negative error would be not convicting the guilty person. From the viewpoint where a person is presumed to be "guilty until proven innocent" (i.e., H_0 is "guilty"; H_a is "innocent"), the errors are reversed. Here, the false positive error would be not convicting the guilty person, and the false negative error would be convicting the innocent person.

While the possibility of a decision error can never be totally eliminated, it can be controlled. To control the possibility of making decision errors, the planning team must control total study error. There are many ways to accomplish this, including collecting a large number of samples (to control sampling design error), analyzing individual samples several times, or using more precise laboratory methods (to control measurement error). Better sampling designs can also be developed to collect data that more accurately and efficiently represent the population of interest. Every study will use a slightly different method of controlling decision errors, depending on where the largest components of total study error exist in the data set and the ease of reducing those error components. Reducing the probability of making decision errors generally increases costs. In many cases controlling decision error within very small limits is unnecessary for making a decision that satisfies the decisionmaker's needs. For instance, if the consequences of decision errors are minor, a reasonable decision could be made based on relatively crude data (data with high total study error). On the other hand, if the consequences of decision errors are severe, the decisionmaker will want to control sampling design and measurement errors within very small limits.

To minimize unnecessary effort controlling decision errors, the planning team must determine whether reducing sampling design and measurement errors is necessary to meet the decisionmaker's needs. These needs are made explicit when the decision maker specifies probabilities of decision errors that are tolerable. Once these tolerable limits on decision errors are defined, then the effort necessary to analyze and reduce sampling design and measurement errors to satisfy these limits can be determined in Step 7: Optimize the Design for Obtaining Data. It may be necessary to iterate between these two steps before finding tolerable probabilities of decision errors that are feasible given resource constraints.

ACTIVITIES

Determine the possible range of the parameter of interest. Establish the possible range of the parameter of interest by estimating its likely upper and lower bounds. This will help focus the remaining activities of this step on only the relevant values of the parameter. For example, the range of the parameter shown in Figs. 6-1 and 6-2 at the end of this chapter is between 50 and 200 ppm. Historical and documented analytical data are of great help in establishing the potential parameter range.

Identify the decision errors and choose the null hypothesis. Define where each decision error occurs relative to the action level and establish which decision error should be defined as the null hypothesis (baseline condition). This process has four steps:

(1) *Define both types of decision errors and establish the true state of nature for each decision error.* Define both types of decision errors and determine which one occurs above and which one occurs below the action level. A decision error occurs when the data mislead the decisionmaker

[1] Note that these definitions are not the same as false positive or false negative instrument readings, where similar terms are commonly used by laboratory or field personnel to describe a fault in a single result; false positive and false negative *decision* errors are defined in the context of hypothesis testing, where the terms are defined with respect to the null hypothesis.

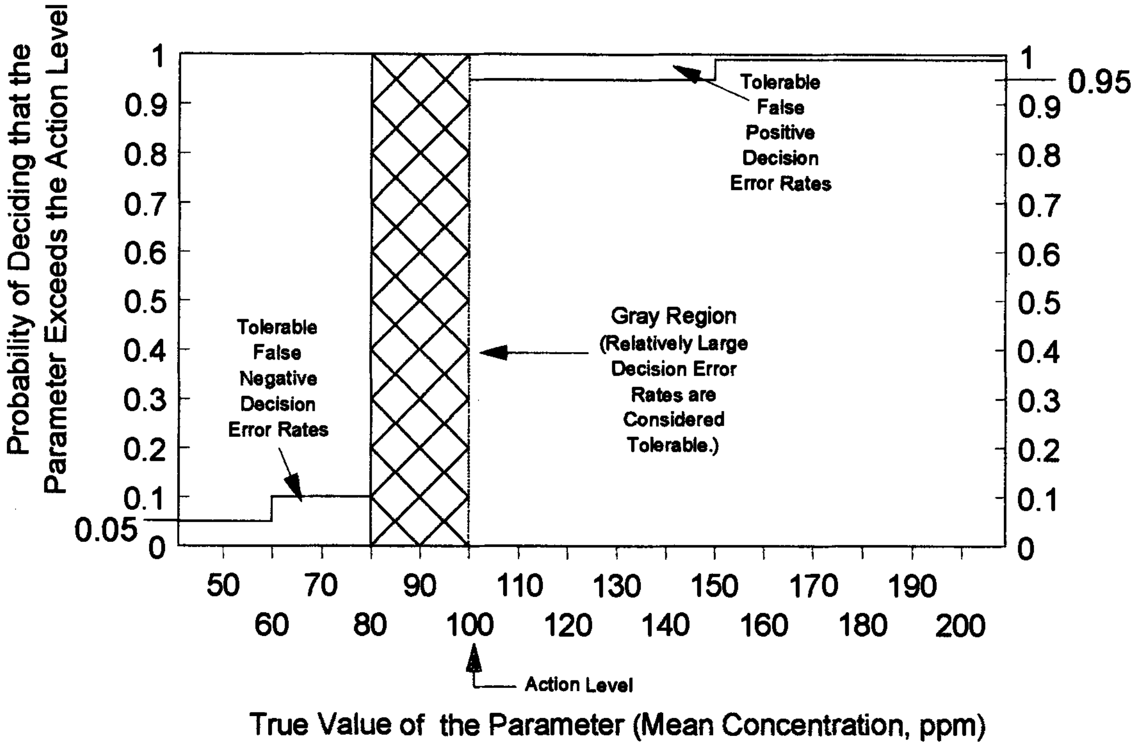


FIG. 6.1—An example of a decision performance goal diagram baseline condition: Parameter exceeds action level.

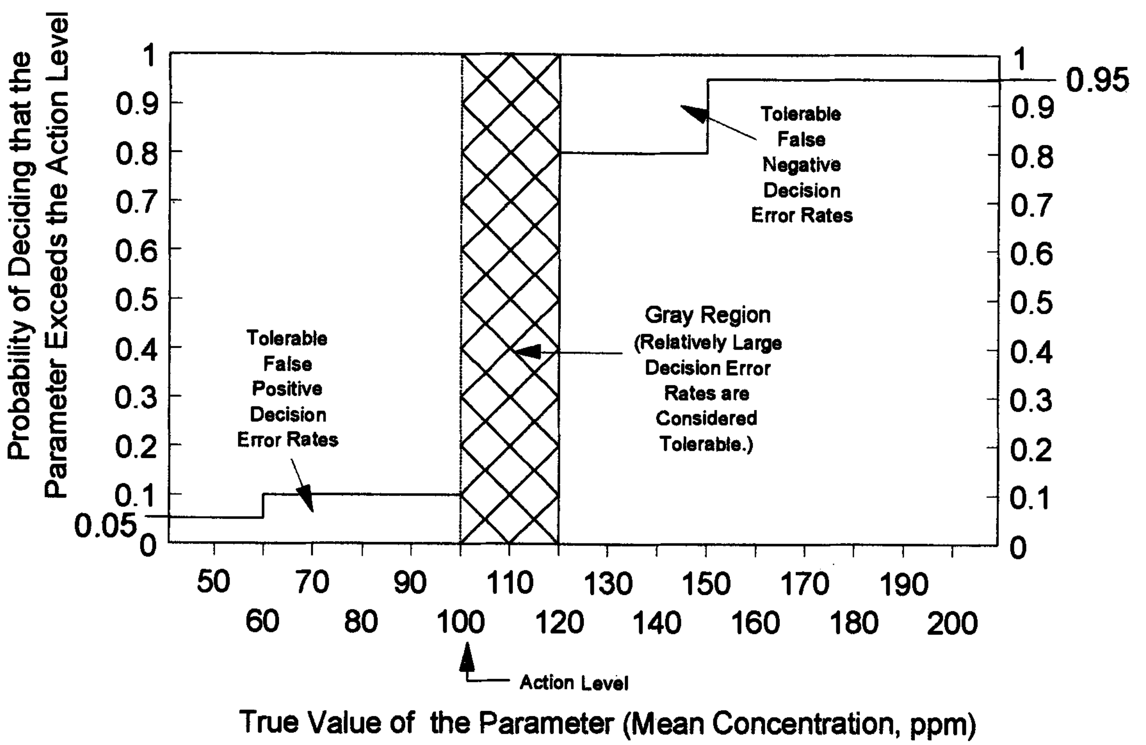


FIG. 6.2—An example of a decision performance goal diagram baseline condition: Parameter is less than action level.

into concluding that the parameter of interest is on one side of the action level when the true value of the parameter is on the other side of the action level. For example, consider a situation in which a study is being conducted to determine if mercury contamination is creating a health hazard and EPA wants to take action if more than 5% of a population of fish have mercury levels above a risk-based action level. In this case, a decision error would occur if the data lead the decisionmaker to conclude that 95% of the mercury levels found in the fish population were below the action level (i.e., the parameter is the "95th percentile" of mercury levels in the fish population) when the true 95th percentile of mercury levels in the fish population was above the action level (which means that more than 5% of the fish population contain mercury levels greater than the action level). The other decision error for this example would be that the data lead the decisionmaker to conclude that the 95th percentile of mercury levels in the fish population is greater than the action level when the true 95th percentile is less than the action level. The "true state of nature" is the actual condition or feature of the environment that exists, but is unknown to the decisionmaker. Each decision error consists of two parts, the true state of nature and the conclusion that the decisionmaker draws. Using the example above, the true state of nature for the first decision error is that the 95th percentile of mercury levels in the fish population is above the action level.

(2) *Specify and evaluate the potential consequences of each decision error.* Specify the likely consequences of making each decision error and evaluate their potential severity in terms of economic and social costs, human health and ecological effects, political and legal ramifications, and so on. Consider the alternative actions that would be taken under each decision error scenario, as well as secondary effects of those actions. For example, in determining whether or not 95% of a fish population contain mercury levels above a risk-based action level, there may be a variety of potential consequences of committing a decision error. In the first decision error described above, where the decisionmaker concludes that the 95th percentile is below when the true 95th percentile was above the action level, the decisionmaker may decide to continue to allow fishing in the waters and not undertake any cleanup activity. The resulting consequences might include human health and ecological effects from consumption of contaminated fish by humans and other animals, economic and social costs of health care and family disruption, and damaged credibility of EPA when (and if) the decision error is detected. If the other type of decision error is committed, where the decisionmaker decides that the 95th percentile exceeds the action level when the true 95th percentile is below the action level, the decisionmaker might ban all fishing in the local waters and initiate cleanup activities. The consequences might include economic and social costs of lost revenues and job displacement in the fishing industry, damaged credibility for EPA when the cleanup activities expose the nature of the decision error, and the threat of lawsuits by fishing interests.

Evaluate the severity of potential consequences of decision errors at different points within the domains of each type of decision error, since the severity of consequences may change as the parameter moves further away from the action level. Consider whether or not the consequences change abruptly at some value, such as a threshold health effect level; the decisionmaker may want to change the tolerable limit on the decision error at such a point.

(3) *Establish which decision error has more severe consequences near the action level.* Based on the evaluation of potential consequences of decision errors, the decisionmaker should determine which decision error causes greater concern when the true parameter value is near the action level. It is important to focus on the region near the action level because this is where the true parameter value is most likely to be when a decision error is made (in other words, when the true parameter is far above or far below the action level, the data are much more likely to indicate the correct decision). This determination typically involves value judgments about the relative severity of different types of consequences within the context of the problem. In the fish contamination problem above, the decisionmaker would weigh the potential health consequences from allowing people to consume contaminated fish versus the economic and social disruption from banning all fishing in the community. In this case, the decisionmaker might carefully consider how uncertain or conservative the risk-based action level is.

(4) *Define the null hypothesis (baseline condition) and the alternative hypothesis and assign the terms "false positive" and "false negative" to the appropriate decision error.* In problems that concern regulatory compliance, human health, or ecological risk, the decision error that has the most adverse potential consequences should be defined as the null hypothesis (baseline condition).[2] In statistical hypothesis testing, the data must conclusively demonstrate that the null hypothesis is false. That is, the data must provide enough information to authoritatively reject the null hypothesis (disprove the baseline condition) in favor of the alternative. Therefore, by setting the null hypothesis equal to the true state of nature that exists when the more severe decision error occurs, the decisionmaker guards against making the more severe decision error by placing the burden of proof on demonstrating that the most adverse consequences will *not* be likely to occur.

It should be noted that the null and alternative hypotheses have been predetermined in many regulations. If not, the planning team should define the null hypothesis (baseline condition) to correspond to the true state of nature for the more severe decision error and define the alternative hypothesis to correspond to the true state of nature for the less severe decision error.

Using the definitions of null and alternative hypotheses, assign the term "false positive" to the decision error in which the decisionmaker rejects the null hypothesis when it is true, which corresponds to the decision error with the more severe consequences identified in task (3). Assign the term "false negative" to the decision error in which the decisionmaker fails to reject the null hypothesis when it is false, which corresponds to the decision error with the less severe consequences identified in task (3).

[2] Note that this differs somewhat from the conventional use of hypothesis testing in the context of planned experiments. There, the alternative hypothesis usually corresponds to what the experimenter hopes to prove, and the null hypothesis usually corresponds to some baseline condition that represents an "opposite" assumption. For instance, the experimenter may wish to prove that a new water treatment method works better than an existing accepted method. The experimenter might formulate the null hypothesis to correspond to "the new method performs no better than the accepted method," and the alternative hypothesis as "the new method performs better than the accepted method." The burden of proof would then be on the experimental data to show that the new method performs better than the accepted method, and that this result is not due to chanc

Specify a range of possible parameter values where the consequences of decision errors are relatively minor (gray region). The gray region is a range of possible parameter values where the consequences of a false negative decision error are relatively minor. The gray region is bounded on one side by the action level and on the other side by that parameter value where the consequences of making a false negative decision error begin to be significant. Establish this boundary by evaluating the consequences of not rejecting the null hypothesis when it is false. The edge of the gray region should be placed where these consequences are severe enough to set a limit on the magnitude of this false negative decision error. Thus, the gray region is the area between this parameter value and the action level.

It is necessary to specify a gray region because variability in the population and unavoidable imprecision in the measurement system combine to produce variability in the data such that a decision may be "too close to call" when the true parameter value is very near the action level. Thus, the gray region (or "area of uncertainty") establishes the minimum distance from the action level where the decisionmaker would like to begin to control false negative decision errors. In statistics, the width of this interval is called the "minimum detectable difference" and is often expressed as the Greek letter delta (Δ). The width of the gray region is an essential part of the calculations for determining the number of samples needed to satisfy the DQOs, and represents one important aspect of the decision maker's concern for decision errors. A more narrow gray region implies a desire to detect conclusively the condition when the true parameter value is close to the action level ("close" relative to the variability in the data). When the true value of the parameter falls within the gray region, the decisionmaker may face a high probability of making a false negative decision error, since the data may not provide conclusive evidence for rejecting the null hypothesis, even though it is actually false (i.e., the data may be too variable to allow the decisionmaker to recognize that the presumed baseline condition is, in fact, *not* true).

From a practical standpoint, the gray region is an area where it will not be feasible or reasonable to control the false negative decision error rate to low levels because of high costs. Given the resources that would be required to reliably detect small differences between the action level and the true parameter value, the decisionmaker must balance the resources spent on data collection with the expected consequences of making that decision error. For example, when testing whether a parameter (such as the mean concentration) exceeds the action level, if the *true* parameter is near the action level (relative to the expected variability of the data), then the imperfect data will tend to be clustered around the action level, with some values above the action level and some below. In this situation, the likelihood of committing a false negative decision error will be large. To determine with confidence whether the true value of the parameter is above or below the action level, the decisionmaker would need to collect a large amount of data, increase the precision of the measurements, or both. If taken to an extreme, the cost of collecting data can exceed the cost of making a decision error, especially where the consequences of the decision error may be relatively minor. Therefore, the decisionmaker should establish the gray region, or the region where it is not critical to control the false negative decision error, by balancing the resources needed to "make a close call" versus the consequences of making that decision error.

Assign probability limits to points above and below the gray region that reflect the tolerable probability for the occurrence of decision errors. Assign probability values to points above and below the gray region that reflect the decisionmaker's tolerable limits for making an incorrect decision. Select a possible value of the parameter; then choose a probability limit based on an evaluation of the seriousness of the potential consequences of making the decision error if the true parameter value is located at that point. At a minimum, the decisionmaker should specify a false positive decision error limit at the action level, and a false negative decision error limit at the other end of the gray region. For many situations, the decision maker may wish to specify additional probability limits at other possible parameter values. For example, consider a hypothetical toxic substance that has a regulatory action level of 10 ppm, and which produces threshold effects in humans exposed to mean concentrations above 100 ppm. In this situation, the decisionmaker may wish to specify more stringent probability limits at that threshold concentration of 100 ppm than those specified at 10 ppm. The tolerable decision error limits should decrease further away from the action level as the consequences of decision error become more severe.

Given the potentially high cost of controlling sampling design error and measurement error for environmental data, Agency decision making is rarely supported by decision error limits more stringent than 0.01 (1%) for both the false positive and false negative decision errors. This guidance recommends using 0.01 as the starting point for setting decision error rates. The most frequent reasons for setting limits greater (i.e., less stringent) than 0.01 are that the consequences of the decision errors may not be severe enough to warrant setting decision error rates that are this extreme. The value of 0.01 should *not* be considered a prescriptive value for setting decision error rates, nor should it be considered as the policy of EPA to encourage the use of any particular decision error

TABLE 6.1—Decision Error Limits Table Corresponding to Figure 6-1. (Action Level = 100 ppm).

True Concentration	Correct Decision	Type of Error	Tolerable Probability of Incorrect Decision
<60 ppm	Not exceed	F(−)	5%
60 to 80	Not exceed	F(−)	10%
80 to 100	Not exceed	F(−)	gray region
100 to 150	Does exceed	F(+)	5%
>150	Does exceed	F(+)	1%

TABLE 6.2—Decision Error Limits Table Corresponding to Figure 6-2. (Action Level = 100 ppm).

True Concentration	Correct Decision	Type of Error	Tolerable Probability of Incorrect Decision
<60 ppm	Not exceed	F(+)	5%
60 to 100	Not exceed	F(+)	10%
100 to 120	Does exceed	F(−)	gray region
120 to 150	Does exceed	F(−)	20%
>150	Does exceed	F(−)	5%

rate. Rather, it should be viewed as a starting point from which to develop limits on decision errors that are applicable for each study. If the decisionmaker chooses to relax the decision error rates from 0.01 for false positive or false negative decision errors, the planning team should document the reasoning behind setting the less stringent decision error rate and the potential impacts on cost, resource expenditure, human health, and ecological conditions.

The combined information from the activities section of this chapter can be graphed onto a "Decision Performance Goal Diagram" or charted in a "Decision Error Limits Table" (see Figs. 6-1 and 6-2 and Tables 6-1 and 6-2). Both are useful tools for visualizing and evaluating all of the outputs from this step. Figure 6-1 and Table 6-1 illustrate the case where the null hypothesis (baseline condition) is that the parameter of interest exceeds the action level (e.g., the waste is hazardous). Figure 6-2 and Table 6-2 illustrate the case where the null hypothesis (baseline condition) is that the parameter is less than the action level (e.g., the waste is not hazardous).

C

Waste Pile Example

INTRODUCTION

IN THIS EXAMPLE five case studies with varying waste pile characteristics and alternate sampling designs are presented through the planning (DQO process), implementation, and assessment phases. For purposes of these case studies, the stakeholders have different prior knowledge for each case. However, for consistency and to clearly present the development of the alternate sampling designs, each waste pile has the same characteristics, as described in the following paragraph.

The waste pile in these examples consists of material that has been generated from a metals recovery process. The dimensions of the waste pile are approximately 100 by 100 ft (38.48 m) with a maximum height of 10 ft (3.048 m); however, more material was deposited in the front corner of the pile (see Fig. 1—Topographic Base Map). The material in the pile was generated from the same source and contaminated with lead. It is also known that no containerized waste has been disposed of in the waste pile. The waste pile is now a Solid Waste Management Unit (SWMU) under investigation as part of a RCRA Facility Investigation (RFI). Specific guidance is provided in ASTM's Standard Guide for Sampling Waste Piles, D 6009. Note that the sampling design for each case is denoted in the text of the example for clarification purposes; the appropriate sampling design is actually selected at Step Seven in the DQO process.

For *Case 1 (authoritative)*, the stakeholders expect the lead concentration to be extremely elevated due to process knowledge (perhaps several times the Toxicity Characteristic (TC) Rule regulatory level of 5.0 mg/L), and it is likely that the TCLP results will designate the material as hazardous. If the lead concentration in the TCLP greatly exceeds the TC Rule regulatory level, then a statistical evaluation of the data would not be necessary. Thus, a complex sampling design would probably not be warranted in this case. In this case, the stakeholders have set a limit of $2,000 for the analytical costs of the study.

For *Case 2 (simple random)*, preliminary data indicate that the mean lead concentration is near the regulatory limit. The stakeholders expect the pile to be relatively homogeneous; therefore, information on the distribution of lead is not important. (The entire waste pile will be considered the "remediation unit" in this case. (See Identifying Inputs to Decision section).) Although the degree of stratification is not known (either over space or by component), it is not expected to be significant because the recovery process that generated the waste was reportedly constant over the time period that the pile was generated and the particle sizes of the material in the pile could be considered homogeneous for the purposes of this investigation (also known as practically homogeneous). The stakeholders have decided that a limit of $8,000 for the analytical costs of the study will be set in this case.

For *Case 3 (systematic grid)*, a minimal amount of data exists on the material in the waste pile so that no assumptions concerning probable contaminant concentrations can be made initially. Information regarding contaminant distribution across the waste pile is a primary objective of the study. The stakeholders have decided that a limit of $5,000 for the analytical costs of the study will be set in this case.

For *Case 4 (systematic grid with compositing)*, a minimal amount of data exists on the material in the waste pile so that no assumptions concerning probable contaminant concentrations can be made initially. Specific information regarding distribution of contamination across the waste pile is not an objective of the study. The degree of stratification is not known, but it is not expected to be significant. The stakeholders have set a limit of $2,000 for the analytical costs of the study in this case.

For *Case 5 (stratified with systematic grid)*, it is discovered that a recent process change was incorporated in the metals recovery process which significantly increased the lead concentration in the waste. Information exists suggesting that approximately the front 20% of the pile (note slightly greater elevation) was generated by the new process, while the material generated by the previous process is located in the remainder of the pile. Although two areas of different concentrations, or strata, exist within the waste pile, the two individual strata are internally homogeneous. One decision will be made on the entire waste pile. The stakeholders have decided on an analytical cost limit of $5,000.

PLANNING PHASE

The DQO process and sampling design optimization process are outlined in the Planning Step section of this manual. The following information pertains to all five cases described in the introduction unless otherwise stated. Figures illustrating the location of the samples for each case are included at the end of the example.

Data Quality Objectives (DQO) Process

Step One: Stating the Problem

The waste pile contains material that may be considered hazardous due to elevated lead content. Therefore, in each case the

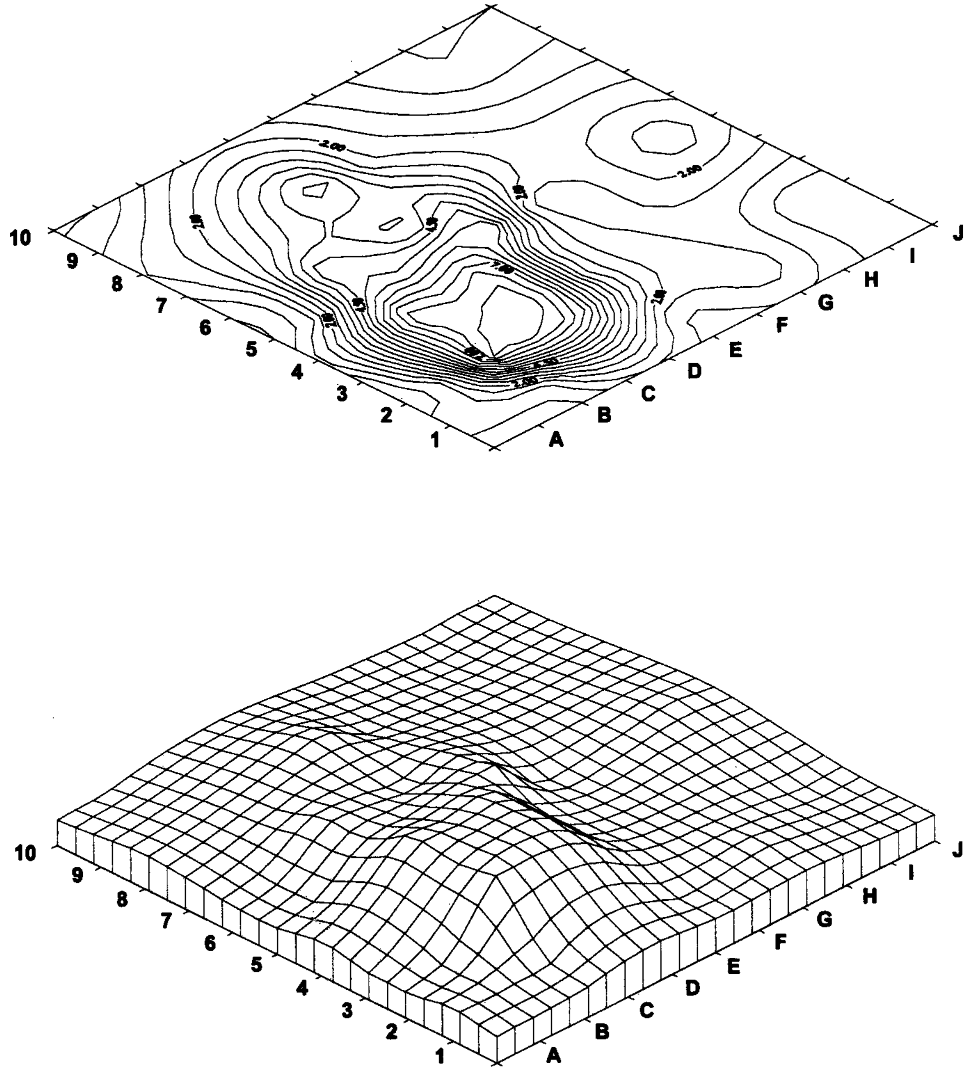


FIG. 1—Topographic base map.

company needs to determine if the material should be disposed of in a hazardous waste landfill under Subtitle C of RCRA (@ $500 per ton) versus a Subtitle D landfill (@ $50 per ton). The stakeholders in this study are the company that generated the waste (and will be conducting the sampling and analysis), the appropriate regulatory agencies, and in some cases representatives from local communities. The company will be required to develop a sampling design that meets the objectives of the study and satisfies all pertinent regulatory requirements.

Step Two: Identifying Possible Decisions

The principal study question is: Is the material in the waste pile a RCRA hazardous waste (per 40 CFR 261.24)? The potential alternate actions are: (a) the material must be managed under Subtitle C of RCRA as hazardous waste or (b) the material may be disposed of in a permitted Subtitle D Municipal Solid Waste Landfill (MSWLF).

Step Three: Identifying Inputs to the Decision

- The decision on whether the material is hazardous or not will depend on the results of the Toxicity Characteristic Leaching Procedure (TCLP) test on the samples collected. The regulatory level for lead under the TC Rule is 5.0 mg/L. If the sample results exceed this value, the material will be considered hazardous. Totals results may be used to determine if the lead concentration is elevated enough—at least 20 times the regulatory level—to warrant completion of the TCLP test. (See EPA Method 1311, Section 1.1.) Note that the totals results may also be necessary to provide information for a subsequent risk assessment to determine the need to characterize soil and/or groundwater in areas adjacent to the waste pile if it is determined to be non-hazardous, and, in the case when the material is determined to be hazardous, for characterization required for off-site disposal by a permitted Treatment, Storage, Disposal Facility (TSDF). For purposes of this example, only Cases 1 and 2 will include totals results; however, they may be included during the planning step based on the objectives of the study.
- In each case, the decision will be based on the entire waste pile; in other words, there will not be smaller "remediation units" within the pile where a Subtitle C versus D decision will be made. Either the entire pile is hazardous, or the entire pile is not. In certain situations, however, it may prove advantageous to employ different scales of decisionmak-

ing, such as with a two-part decision rule. An example of a two-part decision rule that could be used in this situation would be to (1) compare the mean of the pile to a regulatory level and (2) make a decision on smaller remediation units of the pile if they contained lead greater than three standard deviations above the regulatory level.

- For Cases 1–4 the material in the waste pile was generated by the same process, while two different processes were used in Case 5.
- Lead is the contaminant of concern, although the exact distribution across the pile is unknown.
- Access to the pile is not limited, and traditional sampling equipment is expected to be adequate.
- The analytical methods for lead (SW-846 Method 6010B for total lead and SW-846 Method 1311 for the TCLP) should be able to meet the required detection limits as the sample matrix is not expected to be difficult from a sample preparation or analysis standpoint. The totals results, if being used for a subsequent risk assessment, must meet the quantitation limits required for the assessment. Also, an acceptable approach for addressing non-detects must be decided upon prior to the investigation (see Data Quality Assessment section in the Manual).
- The particle size of the material in the waste pile (approximately 0.05 cm) could be considered homogeneous for purposes of this investigation.
- "Real-time" field analytical techniques and innovative approaches (such as XRF, field atomic adsorption or gas chromatography, immunoassay-based test kits, direct push technologies, etc.) could be used to improve decision-making in the field. These techniques would be incorporated into the DQO process to provide flexibility in the field based on the information being generated on-site. They would also assist the investigators in determining the presence and nature of contaminant heterogeneity.

Step Four: Defining Boundaries

The waste pile will be sampled using an appropriate design and analyzed for lead (totals and TCLP). The spatial boundary of the waste pile has been defined by the obvious elevation above the surrounding terrain, the discoloration associated with the material, and the practically homogeneous particle size of the material. The samples will be collected from the surface to a 1-ft (0.30 m) depth, although in every case locations should be sampled to the base of the waste pile to obtain information about potential vertical stratification (Case 1 illustrates this approach). Samples will be collected within a reasonable time frame; however, a temporal boundary for an inorganic contaminant such as lead is generally not a concern.

Step Five: Developing Decision Rules

The decision rule will differ depending on the case under consideration.

With an authoritative design (Case 1), the decision rule will be:

If the average lead concentration for the data set, based on a judgmental approach, greatly exceeds the regulatory level of 5.0 mg/L using the TCLP, then the material in the waste pile will be considered hazardous, and it will be managed under Subtitle C of RCRA. If the average concentration is near or below the regulatory level, a more complex sampling design will be developed. Since an authoritative design is being considered for this investigation, a statistical test would not be applicable and, in fact, unnecessary if the results significantly exceed the regulatory level.

With a probabilistic design (Cases 2–5), the decision rule will be:

If the 90% (one-tailed) upper confidence level (UCL) of the mean concentration is equal to or exceeds the regulatory level of 5.0 mg/L using the TCLP, then the material in the waste pile will be considered hazardous, and it will be managed under Subtitle C of RCRA. If the 90% UCL is below the regulatory level, the material will not be considered hazardous and will be managed under Subtitle D for Municipal Solid Waste Landfills. The use of the term "mean" assumes a normal distribution of the data, an assumption that must be checked. A lognormal distribution could also be evaluated, but the UCL would be computed differently. (See Data Quality Assessment section of this example.)

Step Six: Specifying Limits on Decision Errors

The sampling design error and measurement error will be minimized by using a well-prepared Project Plan (QAPP). The acceptable decision error is decidedly smaller for a Type I error (the material is actually hazardous when the study indicates it is not); therefore, the stakeholders have decided that any outcome where the lead concentration is near or below the regulatory level will result in the need for further investigation using a more complex sampling design. However, because the risk associated with a Type II error (the material is determined to be hazardous when it is not) from an environmental or human health standpoint is less, a result that is significantly above the regulatory level will result in a decision that is protective. Note that the decision error is more important when the mean of the data set is near the regulatory level of 5.0 mg/L of lead.

For a study implementing a probabilistic design, limits on decision errors will be set as follows:

In the case of making a hazardous waste determination, we are comparing the 90% UCL of the mean concentration of the TCLP results for the sample to the Toxicity Characteristic (TC) Rule regulatory level of 5 mg/L. SW-846 suggests that the decision be based on a 90% one-tailed test [*1*]. The Type I error rate is set at 0.10 (10%). That is the probability of rejecting the null hypothesis when it is actually true. See Appendix B for additional information on hypothesis testing.

Step Seven: Optimizing Data Collection Design

Initial Design Selection

The initial design selection for the Case 1 study is:

Since available information strongly suggests that the lead concentration in the waste pile is elevated, an authoritative design is chosen initially for this case. However, if the sample results reveal values close to the regulatory limits, the sample design will need to be reconsidered in light of the new data. Two types of authoritative designs are to be considered: *biased*, where the investigation targets worst case conditions, or *judgmental*, where the investigator uses professional judgment and site information/observations to collect samples that reflect average conditions on the site. The determination of average conditions would be appropriate in this case because the facility has conceded that the lead concentrations are elevated. Note that worst case conditions would be difficult to determine in a waste unit such as this but would be appropriate when process or site knowledge can be used to identify areas of highest contamination. Therefore, the specific sampling locations and the number of samples will be determined by the investigators in the field. As a general rule, at least four to six samples should be collected. This number allows for one sample to be taken in each of four quadrants and provides a minimum degree of coverage for the pile.

The initial design selection for the Case 2 study is:

The stakeholders expect the lead concentration to be near the regulatory limit; thus, a probabilistic approach will be chosen to validate data results. Simple random, stratified, and systematic (grid-based) designs provide information on the mean concentration of lead. Since the existence of strata is not expected (although could be discovered during the investigation), the stratified design is at this time eliminated from consideration. Information on spatial distribution of lead in the pile is not a primary objective of this study, although it would confirm the investigators, assumptions concerning a non-stratified contaminant distribution. A simple random design is the simplest of the probabilistic sampling methods, but it is not ideally suited for providing information on spatial distribution. The systematic design, both without compositing or with compositing, provides some spatial distribution information and is typically easy to implement. Compositing may increase precision and reduce decision errors by reducing the variability of the estimated mean. The design team will further consider all three alternatives in the Practical Evaluation step of the optimization process.

The initial design selection for the Case 3 and Case 4 study is:

The stakeholders do not have enough information to predict the lead concentration; thus, a probabilistic approach will be chosen to validate data results. Simple random, stratified random, and systematic (grid-based) designs will provide information on the mean concentration of lead. Since the existence of distinct strata is not expected, the stratified design is at this time eliminated from consideration. The design team will further consider the remaining alternatives in the Practical Evaluation step.

The initial design selection for the Case 5 study is:

Due to the existence of a process change that affected the characteristics of the waste, and the expected stratification of the waste pile, a stratified sample design is chosen.

Practical Evaluation

The practical considerations that should be reviewed for each alternative include site access and conditions, equipment selection/use, experience needed, special analytical needs, and scheduling. The remaining alternatives do not have significant practical considerations that would limit their potential use for this study. However, the systematic design may result in sampling locations that are easier to survey and locate in the field, and it would provide better spatial coverage, if needed. Problems with access to all sampling locations, difficult matrices (resistant to penetration by an auger, for example, or containing large pieces of debris or material), and sampling into native material below the pile should all be considered during the development of the Quality Assurance Sampling Plan. A standard operating procedures (SOP) manual for conducting the field sampling will influence the collection of a representative sample.

Estimating the Number of Samples Required for the Study

The designs are evaluated for the number of samples that will be required:

Step One: Determination of the Number of Samples

Based on the use of an *authoritative* approach (Case 1):

Samples will be collected within each quadrant of the waste pile and at the center of the pile. The boring at the center will be advanced to the base of the pile at two-foot intervals to provide information on the vertical concentration profile. The TCLP will be conducted on the top one-foot interval of the boring.

Based on the use of a *probabilistic* approach (Cases 2 to 5):

Simple random design (Case 2):

An acceptable margin of error (Δ) and acceptable probability of exceeding that error (α) must be set. Then the appropriate number of samples to collect may be calculated by [*1*]:

$$n = \frac{(t_{1-\alpha} + t_{1-\beta})^2 s^2}{\Delta^2}$$

where:

n = number of samples to collect,

$t_{1-\alpha}$ = percentile value for the Student t distribution for $n - 1$ degrees of, freedom where α is the probability of making a Type I error,

$t_{1-\beta}$ = percentile value for the Student t distribution for $n - 1$ degrees of, freedom where β is the probability of making a Type II error,

s^2 = estimate of the variance (for individual samples), and

$\Delta = RT - \bar{x}$ (RT is the regulatory threshold, $\bar{x}$ is the estimated mean).

Note that values of the Student t distribution may be obtained from Table 3 in Appendix D. Because the Type II error rate (the chance of deciding the waste is hazardous when it is not) is set at 50% (i.e., β = 0.50), the associated t value becomes zero and the $t_{(1-\beta)}$ term drops from the equation. The discussion in Appendix B addresses the advantages obtained by setting the Type II error rate at a value less than 0.50. The resulting equation is used to calculate the number of samples:

$$n = \frac{t_{1-\alpha}^2 \cdot s^2}{\Delta^2}$$

In a preliminary pilot study, five samples were collected at random. Results for TCLP were 5.8, 10.5, 4.9, 2.1, and 5.4 mg/L. The mean and standard deviation were estimated to be 5.74 and 3.03, respectively. Note that the regulatory level for lead is 5.0 mg/L, and α was set at 0.10. Thus, the acceptable margin of error is calculated as $\Delta = RT - \bar{x} = -0.74$. Using this sample size equation and the t value with $n - 1 = 4$ degrees of freedom,

$$n = \frac{1.533^2 \cdot 3.03^2}{(5 - 5.74)^2} = 40$$

An iteration of the equation is then performed to stabilize the result using $n = 40$ and a t value for $n - 1 = 39$ degrees of freedom. The final sample size is calculated as:

$$n = \frac{1.303^2 \cdot 3.03^2}{(5 - 5.74)^2} = 29$$

Systematic grid design (Case 3):

The minimum number of samples for a systematic grid sampling design may be estimated using the same approach described above for the Simple Random design. Such an approach should provide acceptable results if no strong cyclical patterns, periodicities, or significant spatial correlations exist between sample locations [*1*].

In Case 3, a preliminary pilot study was utilized to calculate the number of samples using the method described above for Case 2. With five samples, the estimated mean and standard deviation were 4.42 and 1.37, respectively. The "*n*" necessary to achieve a 10% probability of exceeding the absolute margin of error was calculated (after several iterations to stabilize the result) to be 11 samples.

Systematic grid design with compositing (Case 4):

Compositing samples, when appropriate, reduces decision errors and increases the precision of the estimated sample mean by reducing variability associated with that mean. With the assumption that the analytical variation is negligible compared to the spatial variation, the sample variance with compositing is equal to the variance without compositing divided by the number of aliquots (k). The necessary number of samples to achieve a desired α is inversely proportional to the number of aliquots. The number of aliquots (k) refers to the number of individual grab samples used to form each composite. For a simple random design, the number of samples may be calculated by:

$$n = \frac{t_{1-\alpha}^2 \cdot (s^2/k)}{\Delta^2}$$

Using the same pilot study data for this case as used for Case 3 and choosing k to be 5, the number of samples necessary with compositing would be reduced to 4. In summary, four composite samples will be collected and each will be comprised of five aliquots that are distributed in four quadrants around a center point, with the last aliquot for each sample coming from the center point.

Stratified systematic design (Case 5):

It is known that the waste pile consists of two different types of internally homogeneous material, so the total waste pile is divided into $L = 2$ nonoverlapping strata. The number of population units in each of the two strata is denoted by N_1 and N_2, and the number of necessary samples in h^{th} stratum may be calculated by $N_h = N \cdot W_h$, where W_h represents the weight or volume of material in the h^{th} stratum. Since it is known that approximately 20% of the waste pile was generated by a new process, W_1 will be set equal to 0.2 and W_2 will be 0.8. Preliminary data was collected from the pile. Three samples were collected from Strata 1, and five samples were collected from Strata 2. The mean and standard deviation for Strata 1 was calculated to be 9.9 and 0.7, respectively. For Strata 2, the mean and standard deviation were 3.5 and 0.7, respectively. The optimum number of samples may be determined using proportional allocation by [*1*]:

$$n = \frac{(t_{1-\alpha,df} + t_{1-\beta,df})^2}{\Delta^2} \cdot \sum_{h=1}^{L} W_h \cdot s_h^2$$

where

$t_{1-\alpha}$ = percentile value for the Student t distribution for $n - 1$ degrees of freedom where α is the probability of making a Type I error,

$t_{1-\beta}$ = percentile value for the Student t distribution for $n - 1$ degrees of freedom where β is the probability of making a Type II error,

$\Delta = RT - \bar{x}$ (RT is the regulatory threshold, $\bar{x}$ is the estimated mean),

s^2 = estimate of the variance (for individual samples),

W_h = weight or volume of material in the h^{th} stratum,

df = the degrees of freedom connected with each t-quantile.

The value of df may be calculated by:

$$df = \left(\sum_{h=1}^{L} W_h \cdot s_h^2\right)^2 \bigg/ \left(\sum_{h=1}^{L} \frac{W_h^2 \cdot s_h^4}{(n \cdot W_h) - 1}\right)$$

Using the preliminary pilot data results and the weighting values for the two strata, df is calculated to be 2, and the corresponding number of samples is 30. The equations must be solved iteratively, so the same calculations are repeated using $n = 30$. After several iterations, the total number of samples is set at 17. Using proportional allocation with $n = 17$ samples, $0.2 \cdot 17 = 3$ samples should be taken from Stratum 1, while $0.8 \cdot 17 = 14$ samples should be collected from Stratum 2. The pilot study data may be used as a portion of the final data set. Thus, no additional samples need to be collected from Stratum 1, and nine additional samples are needed from Stratum 2.

The mean of the data set will be evaluated using the approach in SW-846, Chapter Nine, where the upper bound of the 90% (one-tailed) UCL of the mean is compared to the regulatory level (in this case 5.0 mg/L for lead using the TCLP). The 90% one-tailed approach has been determined by the EPA to provide an adequate margin of safety against making a wrong decision.

Cost Evaluation

This section evaluates the cost associated with the alternate sampling designs.

For Case 1 (authoritative sampling design):

A judgmental authoritative design meets the requirements for the study; that is, it estimates the average lead concentration (via the TCLP) for the material in the waste pile. "Average" is used here rather than "mean," which is associated with a probabilistic design. Seven samples will be collected at an analytical cost of $250 per sample plus an additional 10% for various quality assurance samples. The total analytical cost for each remaining sampling design will be approximately $1,925, which is under the analytical budget target of $2,000. Because a judgmental authoritative design provides information on the average concentration of lead in the waste pile (without the establishment of a confidence interval), it is selected as the preferred sampling design. Note that if this simple design did not meet the study objectives, then a modification in either the design or the study objectives would be required.

For Case 2 (simple random sampling design):

The simple random design as well as both approaches to the systematic design (with and without compositing) meet the statistical requirements for the study in determining the estimated mean lead concentration (via the TCLP) for the material in the waste pile. If a simple random design or a systematic grid design without compositing is chosen, 30 samples will be collected. The analytical cost per sample is $250 including the totals and TCLP, and various quality assurance samples would increase the cost by approximately 10%. Both the simple random design and the systematic grid design without compositing would generate a total analytical cost of about $8,250 (30 samples at $250 for the totals and TCLP plus 10% for quality assurance). The stakeholders decide on the simple random design because they expect the waste pile to be relatively homogeneous; therefore, information on the distribution of lead is not important.

For Cases 3–4 (systematic grid sampling designs):

Again the simple random design and both approaches to the systematic design (with and without compositing) meet the statistical requirements for the study in determining the estimated mean lead concentration (via the TCLP) for the material in the waste pile. If a simple random design or a systematic grid design without compositing is chosen, 15 samples will be collected, to exceed the estimated number of necessary samples. The analytical cost per sample is $250 for the TCLP, and various quality assurance samples would increase the cost by approximately 10%. Both simple random design and the systematic grid design without compositing would generate a total analytical cost of about $4,125 (15 samples at $250 each for the TCLP plus 10% for quality assurance). A systematic grid design with compositing may improve precision over the systematic design without compositing. For Case 3, the analytical costs of each of the alternate sample designs are within the budget of $5,000. The stakeholders decide to use the systematic grid design because spatial information is desired. For Case 4, the systematic grid with compositing is chosen to improve precision and study efficiency (fewer samples collected). Four composite samples will be collected. The cost for that design is approximately $1,100).

For Case 5 (stratified random sampling design):

A stratified random approach is chosen due to the expected stratification of the waste pile. This approach should improve the efficiency of the final determination on the entire waste pile. The analytical costs are estimated at $4,675 (17 samples at $250 each for the TCLP plus 10% for quality assurance) and are within the proposed analytical budget of $5,000.

(What if the Alternate Designs Do Not Meet the DQOs?)

Note that if the sampling designs do not meet the study objectives for each case, then a modification in either the design (more samples, use of sampling tools such as compositing or double sampling) or study objectives (change in the confidence interval, study boundaries, allowable decision error, or budget constraints) will then be required.

IMPLEMENTATION PHASE

For All Cases

Implementation of the authoritative design, simple random design, systematic grid design, and the stratified random design should not present any significant problems. The samples will be collected using decontaminated hand augers, and glass pans will be used for sample mixing. The samples will be collected to a depth of 1 ft (0.61 m) at each location. Note that for Case 1 information will be collected to evaluate the potential presence of vertical stratification in the waste pile. In that Case, samples for vertical profiling will be collected at one location by a boring advanced to the base of the waste pile. Individual samples will be collected at 2-ft (0.61 m) intervals. The simple and stratified random samples may require careful surveying to determine the location of the specific sampling locations. See Figs. 5–9 at the end of this chapter for the sample locations.

ASSESSMENT PHASE

This section illustrates some of the graphical and statistical techniques available for completing the data quality assessment (DQA) step of a data collection activity. The U.S. EPA publication on Data Quality Assessment (QA/G-9) and the accompanying software (DataQUEST) may be utilized as a tool by the investigator in this step [2,3]. Other references provided in Chapter 4 of the manual should also be consulted. More detail is presented for Case 2 in order to illustrate a range of graphical and statistical assessment options.

Review of the DQOs and the Sampling Design

In each case, the data collected during the study have met the DQOs. Sampling error was minimized through the selection and use of correctly designed sampling devises, careful implementation of the field sampling and handling procedures, and use of minimally biased subsampling procedures within the laboratory (e.g., using guidance in ASTM D 6051) as specified in the QAPP and SOPs. The material that was sampled does not appear to have presented any special problems concerning access to sampling locations, equipment usage, particle size distribution, or matrix interferences. The analytical package has been validated and the data generated are acceptable for their intended purpose.

FOR CASE 1—AUTHORITATIVE SAMPLING DESIGN:

Preliminary Data Review

Results for the data collection effort are listed in Table 1-1.

Statistical Quantities:

Table 1-2 lists the totals and TCLP mean and range of values for lead. As expected, the TCLP concentration for lead greatly exceeds the TC Rule regulatory level of 5.0 mg/L. Totals and TCLP results of the vertical boring indicate that there is not a discernable difference in the lead concentration at the 1 to 3 and 3 to 5 ft intervals versus the surface interval (0 to 1 ft). This confirms the original assumptions concerning vertical stratification that was based on knowledge of the waste generated and the management practices of the facility.

Graphical Representation for Case 1 data:

Because of the limited amount of data collected and the authoritative nature of the study design, no graphical depictions were prepared.

Conclusion

Based on the established decision rule, the material in the waste pile was determined to be hazardous for lead for Case 1. The totals results could be used for profiling the waste to ensure compliance with the Subtitle C permit (see Identifying Inputs to the Decision).

TABLE 1-1—Total and TCLP Results for Case 1.

Location	C3	C7	E5	G3	G7
Totals result (mg/kg)	1400	975	1420	1800	1500
TCLP result (mg/L)	26	20	30	42	32

Vertical Boring	Total Results, mg/kg	TCLP Results, mg/L
E5 (1–3 feet)	1600	28
E5 (3–5 feet)	1350	32

NOTE: 1 ft = 0.3048 m.

TABLE 1-2—Totals and TCLP Statistical Results—Case 1.

Totals Results, mg/kg		TCLP Results, mg/L	
Average	Range	Average	Range
1419	975–1800	30	20–42

FOR CASE 2—SIMPLE RANDOM SAMPLING DESIGN

FOR CASE 2 CONSIDER TWO DIFFERENT DATA SETS, TERMED 2A (NORMAL DISTRIBUTION) AND 2B (NON-NORMAL DISTRIBUTION)

FOR CASE 2A (NORMAL DISTRIBUTION):

Preliminary Data Review

The results for the data collection effort are listed in Table 2a-1. Thirty samples were collected to exceed twenty nine (the number of samples calculated to achieve the specified margin of error). Note that the samples collected from the two vertical cores (Locations H8 and C4) indicate that no significant vertical stratification is present.

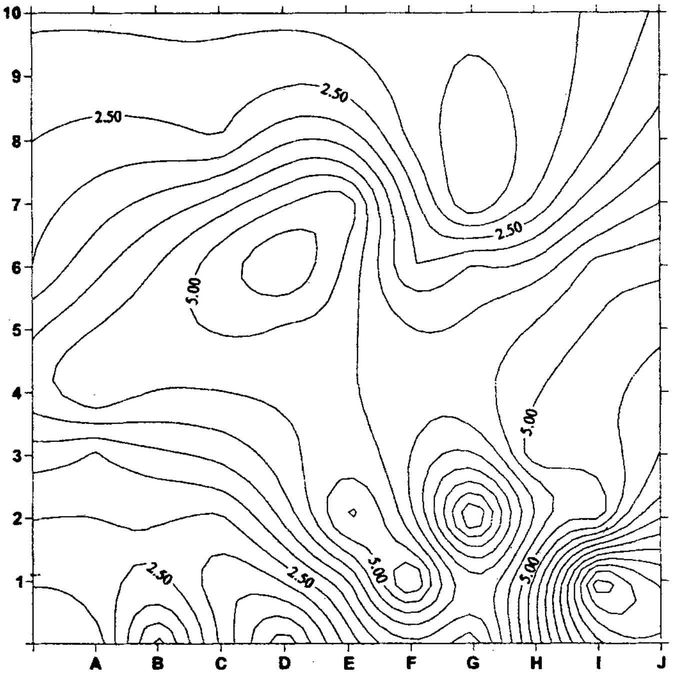


FIG. 2a-1—Lead concentration distribution—Case 2a.

TABLE 2a-1—Totals and TCLP Analytical Results for Case 2a.

Location	Totals Result, mg/kg	TCLP Result, mg/L	Location	Totals Result, mg/kg	TCLP Result, mg/L
A5	1574	4.34	F3	1478	5.73
A7	1047	2.95	F8	1678	5.36
B1	405	1.58	G2	1415	6.34
B4	328	2.86	G7	452	3.05
B5	1234	5.03	G9	24	1.92
B9	661	2.65	H1	219	2.57
C1	1359	4.31	H3	189	0.74
D2	327	1.61	H7	358	3.57
D3	129	2.40	H8	89	1.00
D7	924	5.29	I4	1592	5.36
D9	1012	2.54	I8	2015	10.50
E1	24	0.11	J2	861	6.30
E6	1310	4.89	J3	654	4.61
E7	605	6.04	J7	1014	4.70
F2	1319	3.42	J9	689	2.55

Graphical Representation:

Figure 2a-1 shows the lead concentration isopleth based on the data generated. Although the graphical depiction has inherent limitations, the distribution of lead across the waste pile can be readily observed. No spatial trends or distinct strata are apparent.

Statistical Evaluation of the Data

TCLP versus Totals Results

Figure 2a-2 is provided to evaluate the general relationship between the TCLP and Totals results. The data presented is provided for illustrative purposes, and conclusions should not be drawn about any relationship between the totals and the TCLP data for other data sets. However, the information concerning this relationship could be useful in the future to estimate in very general terms at what totals concentration is this waste likely to exceed the TCLP regulatory level (approximately ≥1,600 mg/kg). Remember, use the results of this comparison with caution, even with a similar waste stream. Note also that in most cases the investigators would not have completed the TCLP on samples collected at the following locations since the Total results were below 100 mg/kg—E1, G9, and H8.

Histogram

Figure 2a-3 is a histogram of the totals data, which provides a picture of the shape of the data and aids in identifying the symmetry and variability of the data set. Using a histogram, one may visually estimate the underlying distribution using binned data plotted against relative frequency of occurrence. If the data are symmetric, then the structure of the histogram will be symmetric around a central point, such as the mean, if the data set is sufficiently large ($n > 25$). Thus, using a histogram, a normal distribution or a skewed distribution may be visually identified. The histogram provides a tool for preliminary data assessment but is inadequate for verification of distributional assumptions. TCLP data is used to test distributional assumptions since the final decision will be made using this data set. EPA's QA/G-9 (Guidance for Data Quality Assessment) provides guidance in creating a histogram. In this case, the histogram appears to display symmetric data [*2*].

Coefficient of Variation

The coefficient of variation (CV) may be used to quickly check if the data may be modeled by the normal curve by comparing the sample CV to 1. If the CV is greater than 1, then the data should not be modeled by a normal curve. However, this method should not be used to conclude the opposite. (If CV < 1, the test is inconclusive). The CV is computed by dividing the standard deviation by the mean of the data set. In this case, the CV of the TCLP data is computed to be 0.6, so the test is inconclusive.

Box and Whiskers Plot

An additional visual method of evaluating the shape of the data is a box and whiskers plot; it is useful in determining the

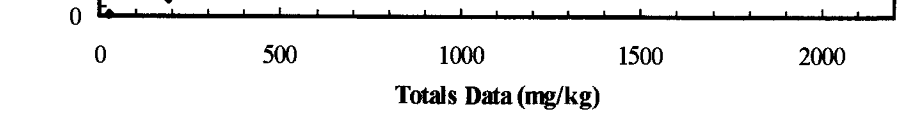


FIG. 2a-2—TCLP vs. total data—Case 2a.

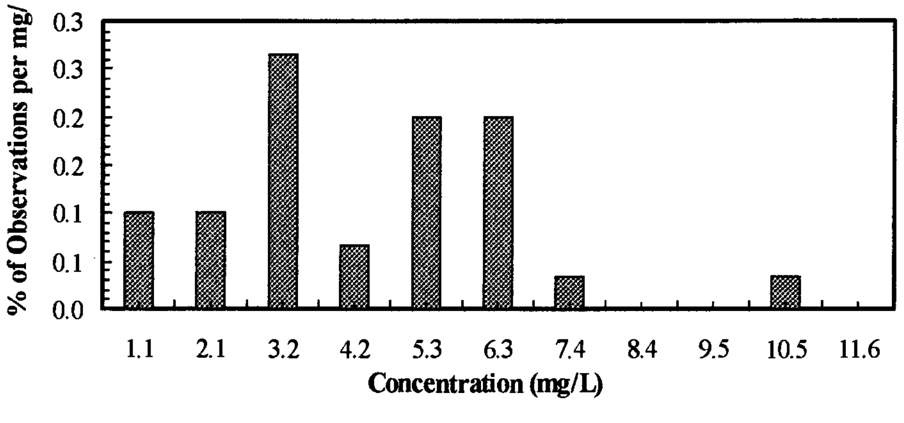


FIG. 2a-3—Histogram—Case 2a.

symmetry of the data. See QA/G-9 for guidance on constructing Box and Whiskers plots. The TCLP data was used to generate the box and whiskers plot for Case 2a seen in Fig. 2a-4.

The box and whiskers plot consists of a central box, whose length denotes the spread of the bulk of the data (the central 50%) and whiskers, whose length indicates the spreading of the distribution tails. The width of the box is arbitrary. The plus sign marks the sample mean, and the sample median is displayed as a line through the box. Any outlying data points are marked by a "*" on the plot. In Case 2 the identified "outlier" is the TCLP result at Location J2 (10.5 mg/L). Techniques and approaches for determining when to keep or discard an identified outlier are discussed in Chapter 4 of the manual. Just because this technique identifies the data point as an outlier does not mean that the data point should be discarded. It could be an actual hot-spot within the pile rather than an error introduced through cross contamination of the sample or laboratory problems. If a valid reason for the "outlier" cannot be identified, then further investigation at this location in the waste pile may be warranted.

If the distribution is symmetrical, the box is divided into two equal halves; the whiskers are about the same length, and any extreme data points are equally distributed. According to the box and whiskers plot shown here, the data set appears to be symmetrical with one identified outlier.

Normal Probability Plot (Quantile-Quantile Plot)

A normal probability plot, or Q-Q plot (Fig. 2a-5), may be used to visually check if a sample data set fits a specified probability model. The n TCLP data values, x_i, are plotted against the expected data value, y_i, from the parent model probability distribution. A normal probability plot, which may be used to test the assumption of normality, is the graph of the quantiles of a data set against the quantiles of the normal distribution. If the data follow an approximate linear trend on the plot, the validity of the normality assumption is probable. Refer to EPA QA/G-9 for guidance on generating a normal probability plot. The data set appears to be normally distributed from the Q-Q plot in Fig. 2a-5. However, the plot is a visual quantifier of the data and may not be used to finalize distributional assumptions.

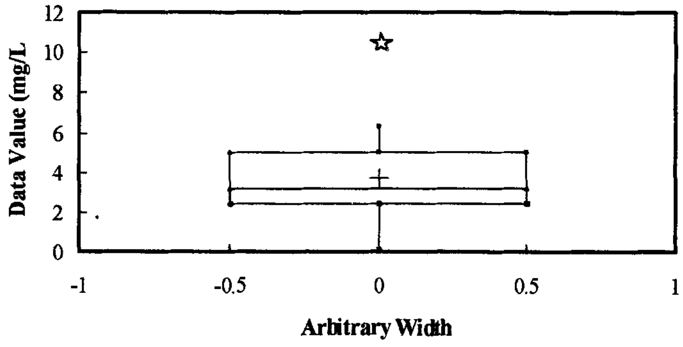


FIG. 2a-4—Box and whiskers plot—Case 2a.

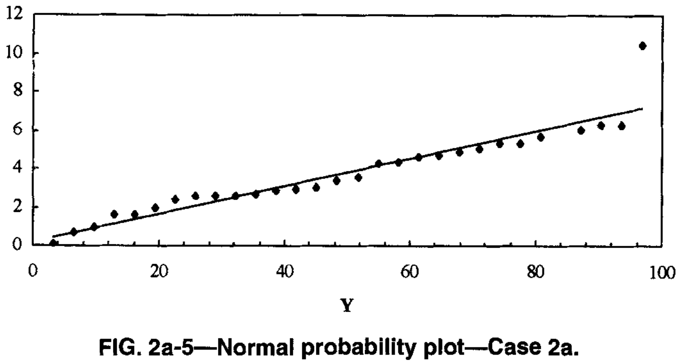


FIG. 2a-5—Normal probability plot—Case 2a.

Shapiro-Wilk Test for Normality

A more precise test for distributional assumptions is the Shapiro-Wilk test, which is conducted on the TCLP data to check for normality as follows:

Compute d, the denominator of the test statistic, using the n data.

$$d = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 132$$

Compute k, where
$k = n/2$ If n is even.
$k = (n - 1)/2$ If n is odd.

In this case, $n = 30$ and $k = 15$. From Table 1 in Appendix D (Table A-6 in Gilbert's Statistical Methods for Environmental Pollution Monitoring (1989)), the coefficients for the test may be obtained as $a_1, a_2, \ldots, a_k$. [*4*]. Then compute the W value.

$$W = \frac{1}{d}\left[\sum_{i=1}^{k} a_i\left(x_{[n-i+1]} - x_{[i]}\right)\right]^2 = 0.948$$

If the computed W value is greater the tabled quantile at the given alpha significance level, then the assumption of normality cannot be rejected. In this case, alpha is taken to be 0.01. Because the W value for this example is higher than the 0.01 quantile of 0.900, the assumption of normality cannot be rejected. W values may be obtained from Table 2 in Appendix D of this manual (also found in Gilbert, Table A-7 "Shapiro-Wilk Tables").

Characterization of the Distribution

The statistical analysis of the TCLP data upheld the distributional assumption of normality. Statistical quantities may now be calculated based on the assumption of normality. The results are displayed in Table 2a-2.

To calculate the 90% UCL when the true standard deviation is not known, use the t distribution from Table 3 in Appendix D. Calculate the 90% UCL by

$$\begin{aligned} 90\%\ \text{UCL} &= \bar{x} + t_{1-a}\left(\frac{s}{\sqrt{n}}\right) \\ &= \bar{x} + t_{0.90}\left(\frac{s}{\sqrt{n}}\right) \\ &= 3.8 + 1.311\left(\frac{2.1}{\sqrt{30}}\right) \\ &= 4.3\ \text{mg/L} \end{aligned}$$

The tabulated "t value" (1.311) is based on a 90% one-tailed confidence interval with a probability of 0.10, $t_{0.90}$ (see Table 1 in Appendix D).

TABLE 2a-2—Totals and TCLP Results—Case 2a.

	Mean	Range	Standard Deviation	Variance	Coefficient of Variation	90% UCL (one-tailed)
Totals Result, mg/kg	833	24–2015				
TCLP Result, mg/L	3.8	0.1–10.5	2.1	4.6	0.6	4.3

TABLE 2b-1—Totals and TCLP Analytical Results—Case 2b.

Location	Totals Result, mg/kg	TCLP Result, mg/L	Location	Totals Result, mg/kg	TCLP Result, mg/L
A5	308	1.7	F3	1283	3.4
A7	474	1.7	F8	320	1.7
B1	570	2.3	G2	869	3.2
B4	709	1.9	G7	331	3.0
B5	415	2.7	G9	540	1.6
B9	363	1.1	H1	502	1.7
C1	516	3.0	H3	1118	4.3
D2	72	1.2	H7	268	2.4
D3	654	2.4	H8	348	1.5
D7	643	2.0	I4	498	5.2
D9	336	1.2	I8	461	4.6
E1	777	2.2	J2	2259	7.1
E6	234	1.0	J3	453	1.4
E7	334	1.5	J7	2587	6.9
F2	474	4.5	J9	283	1.9

Conclusion

The 90% UCL for the mean of the TCLP data is calculated to be 4.3 mg/L, which is less than the regulatory level of 5.0 mg/L. Thus, in Case 2a the material in the waste pile is determined not to be hazardous for lead based on the established decision rule. Note that the TCLP result for the pilot study (5.7 mg/L) indicated that the waste pile was hazardous; however, the more comprehensive evaluation using a simple random approach shows that the waste pile is actually non-hazardous. This illustrates the potential advantage of an expanded characterization effort based on a probabilistic sampling design.

A quick check may be performed to determine if an adequate number of samples was collected to satisfy specified error limits. Refer to Chapter 2 of the Manual to review the sample size equation. The standard deviation and sample mean are entered into the sample size equation with $n - 1 = 29$ degrees of freedom and $\alpha = 0.10$.

$$n = \frac{t_{1-\alpha}^2 \cdot s^2}{\Delta^2} = \frac{1.311^2 \cdot 2.1^2}{(5 - 3.8)^2} = 6$$

Five is less than thirty; therefore, the test was sufficiently powerful and achieves the Type I error rate specified in the DQOs.

FOR CASE 2B (NON-NORMAL DATA DISTRIBUTION):

Preliminary Data Review

The results for the data collection effort are listed in Table 2b-1.

Graphical Representation:

See Fig. 2a-1 for an example of concentration isopleths based on the data generated.

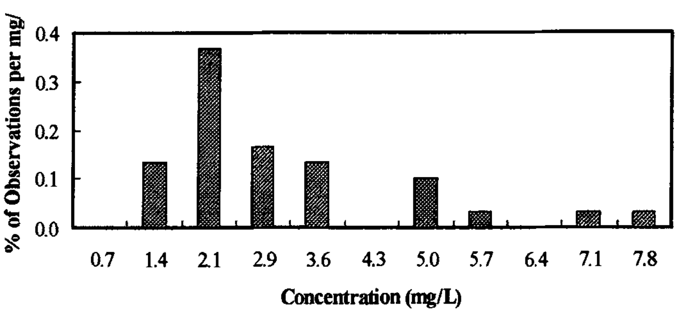


FIG. 2b-1—Histogram—Case 2b.

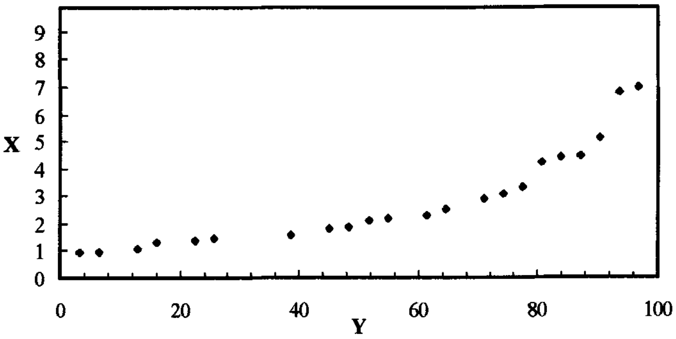


FIG. 2b-2—Normal probability plot—Case 2b.

Statistical Evaluation of the Data

The CV test yields a value of 0.6 for the TCLP data. The CV value is less than 1. Thus, this method is inconclusive, and additional statistical evaluation is needed. Figure 2b-1 is a histogram of the totals data.

The histogram does not appear to display normally distributed data. A normal probability plot is constructed to further test the distribution (Fig. 2b-2).

The data set does not follow a linear trend; thus, the distribution may not be normal. The Shapiro-Wilk test is performed to further verify the deviation from normality at a 0.01 significance level. The test estimated a *W* value of 0.827, which is less than the 0.01 quantile, 0.900 (found in Appendix D). Thus, the Shapiro-Wilk test confirms the non-normality of the data. To check for lognormality, a lognormal probability plot may be created (Fig. 2b-3) in which the natural logarithms of the data are plotted against the calculated *Y*. If the data lie linearly on the lognormal plot, the assumption of a lognormal distribution is strengthened.

The natural logarithms of the data follow an approximately linear trend on a logrithmic scale. Thus, the plot agrees with the assumption of log-normality. The Shapiro-Wilk test is a more accurate way to access lognormality by conducting the test on the natural logrithms of the data. This method produces a *W* value of 0.946. Because the *W* value for this example is higher than the 0.10 quantile of 0.939 (found in Appendix D), the assumption of log-normality may be accepted as valid.

Characterization of the Distribution

The statistical analysis of the data indicates a log-normal data distribution. Statistical quantities are calculated for the TCLP data assuming a log-normal data distribution. The resulting values are displayed in Table 2b-2. The 90% upper confidence limit for the mean is then compared to the regulatory limit of 5.0 mg/L. Several methods exist for estimating the mean of a log-normal distribution [4]. A simple method for estimating the mean and variance of lognormally distributed data is illustrated below.

Compute the log-transformed data set $y_i = \ln x_i$ where x_i is the original data set. Then compute the mean and variance of the log-transformed data.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 0.8$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 = 0.3$$

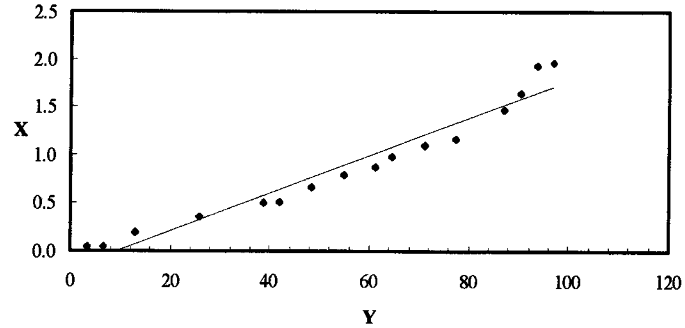


FIG. 2b-3—Lognormal probability plot—Case 2b.

The upper one-sided $100(1 - \alpha)\%$ confidence limit for the mean of log-normally distributed data is calculated by:

$$\mathrm{UCL}_{1-\alpha} = \exp\left(\bar{y} + 0.5s_y^2 + \frac{s_y H_{1-\alpha}}{\sqrt{n-1}}\right)$$

where $\bar{y}$ and s_y^2 are the mean and the variance, respectively, of the log-transformed data set, n is the number of samples, and $H_{1-\alpha}$ is an empirical constant that is provided in tables by Land and Gilbert [4]. For $\alpha = 0.1$, $H_{1-\alpha} = 1.505$, and the UCL_{90} is calculated to be 3.1 mg/L. Note that this formula for estimating the UCL on the mean of a lognormal distribution can give unreliable results if n is small even when the data are truly lognormally distributed. Refer to Singh for further information on the lognormal distribution [5].

Conclusion

The 90% UCL for the mean of a log-normal distribution was calculated to be 3.1 mg/L, which is less than the regulatory level of 5.0 mg/L. Thus, in Case 2b the material in the waste pile was determined not to be hazardous for lead based on the established decision rule.

FOR CASE 3—SYSTEMATIC GRID WITHOUT COMPOSITING SAMPLING DESIGN:

Preliminary Data Review

Fifteen samples were collected to exceed eleven (the calculated number of samples to achieve the desired margin of error). The results for the data collection effort are listed in Table 3-1.

Graphical Representation:

A graphical depiction of the data could be completed. (See Case 2a for an example.)

Statistical Evaluation of the Data

A histogram is not constructed because the number of samples is too small to accurately use this quantifier ($n < 25$). A normal probability plot is constructed to test the assumption

TABLE 3-1—Totals and TCLP Results—Case 3.

Location	TCLP Result, mg/L	Location	TCLP Result, mg/L
B2	0.7	F2	3.6
B4	4.5	F4	5.2
B6	7.9	F6	6.1
B8	6.0	F8	7.4
D2	4.1	H2	1.1
D4	2.3	H4	9.6
D6	5.2	H6	5.6
D8	9.2		

TABLE 2b-2—Totals and TCLP Statistical Result—Case 2b.

	Mean	Range	Standard Deviation	Variance	Coefficient of Variation	90% UCL (one-tailed)
Totals Results, mg/kg	633	72–2587				
TCLP Results, mg/L	2.7	1.0–7.1	1.6	2.6	0.6	3.1

of normality (Fig. 3-1). Again, the TCLP data is used to test for normality.

The data set appears to be normally distributed from the Q-Q plot. The Shapiro-Wilk test is conducted on the TCLP data to further validate the distributional assumption of normality. The *W* value is 0.939, which is higher than the 0.01 quantile of 0.855 (found in Table 2 of Appendix D), so the assumption of normality cannot be rejected.

Characterization of the Distribution

The statistical analysis of the data upheld the distributional assumption of normality. Statistical quantities may now be calculated based on the assumption of normality. The results are displayed in Table 3-2.

To calculate the 90% UCL, use the t-distribution:

$$
\begin{aligned}
90\% \text{ UCL for TCLP data} &= \bar{x} + t_{1-\alpha,n-1}\left(\frac{s}{\sqrt{n}}\right) \\
&= 6.3 + t_{0.90,14}\left(\frac{s}{\sqrt{n}}\right) \\
&= 6.3 + 1.345\left(\frac{2.6}{\sqrt{15}}\right) \\
&= 7.2 \text{ mg/L}
\end{aligned}
$$

The tabulated "*t* value" (1.345) is based on a 90% one-tailed confidence interval with a probability of 0.10 and 14 degrees of freedom, $t_{0.90,14}$ (Table 3 in Appendix C).

Conclusion

The 90% UCL for the mean of the TCLP data is 7.2 mg/L, which is greater than the regulatory level of 5.0 mg/L. Thus, in Case 3 the material in the waste pile is determined to be hazardous for lead based on the established decision rule.

A quick check is performed to determine if a sufficient number of samples were collected to satisfy specified decision error limits on the test for whether the waste pile is hazardous. The standard deviation and sample mean are entered into the sample size equation with $n - 1 = 14$ degrees of freedom and $\alpha = 0.10$. The calculated number is six samples, which is less than fifteen, therefore a sufficient number of samples was collected.

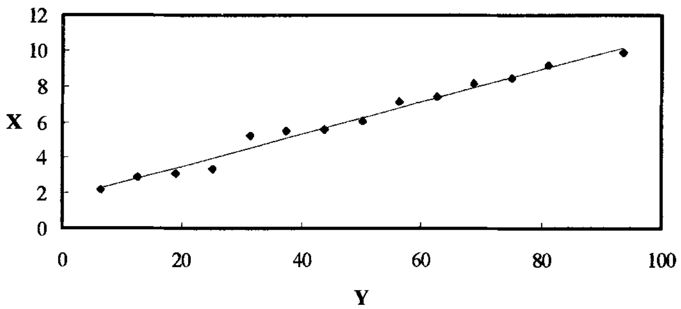


FIG. 3-1—Normal probability plot.

FOR CASE 4—SYSTEMATIC GRID WITH COMPOSITING SAMPLING DESIGN:

Preliminary Data Review

Four samples were collected as specified by the sample size equation. The results for the data collection effort are listed in Table 4-1.

Statistical Evaluation of the Data

A histogram is not constructed because the number of samples is too small to accurately use this quantifier. A normal probability plot is constructed on the TCLP data to test the assumption of normality (Fig. 4-1).

The data set appears to be normally distributed from the normal probability plot. The Shapiro-Wilk test is conducted to further validate the distributional assumption. The *W* value (Table 2 in Appendix D) is 0.903, which is higher than the 0.01 quantile for the sample size of 0.707, so the assumption of normality cannot be rejected. However, it should be noted that both the Q-Q plot and the Shapiro-Wilk test have low power to detect small deviations from normality when *n* is so small.

Characterization of the Distribution

The statistical analysis of the totals data upheld the distributional assumption of normality. Statistical quantities may

TABLE 4-1—Totals and TCLP Results for Case 4.

Location	TCLP Result, mg/L
C2	4.8
C8	3.4
H2	4.1
H8	4.9

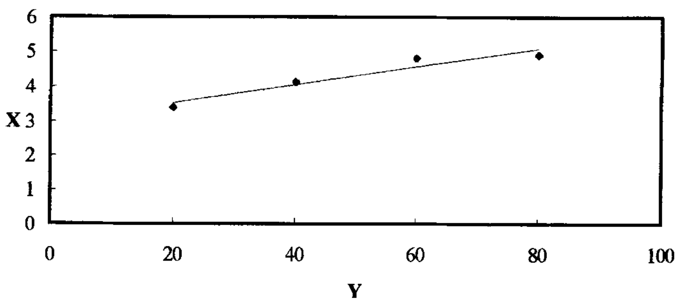


FIG. 4-1—Normal probability plot for Case 4.

TABLE 3-2—Totals and TCLP Statistical Result—Case 3.

	Mean	Range	Standard Deviation	Variance	Coefficient of Variation	90% UCL (one-tailed)
TCLP Results, mg/L	6.3	2.2–9.9	2.6	6.6	0.4	7.2

TABLE 4-2—Totals and TCLP Statistical Results—Case 4.

	Mean	Range	Standard Deviation	Variance	Coefficient of Variation	90% UCL (one-tailed)
TCLP Results, mg/L	4.3	3.4–4.9	0.3	0.1	0.1	4.6

now be calculated based on the assumption of normality. The results are displayed in Table 4-2.

Conclusion

The 90% UCL for the mean of the TCLP data is 4.6 mg/L, which is less than the regulatory level of 5.0 mg/L. Thus, in Case 4 the material in the waste pile is determined to be nonhazardous for lead based on the established decision rule.

A quick check is performed to determine if a sufficient number of samples were collected to satisfy specified decision error limits on the test for whether the waste pile is hazardous. The standard deviation and sample mean are entered into the sample size equation with $n - 1 = 3$ degrees of freedom and $\alpha = 0.10$. The calculated number is one sample, which is less than four, therefore a sufficient number of samples was collected.

FOR CASE 5—STRATIFIED RANDOM SAMPLING DESIGN:

Preliminary Data Review

Three samples are collected for stratum one, and fourteen samples are collected from Stratum 2 as calculated in the sample size equation for proportional allocation. The results for the data collection effort are listed in Table 5-1.

Characterization of the Distribution

Statistical quantities may now be calculated. The results are displayed in Table 5-2.

For a stratified design which considers multiple strata, the overall mean concentration for the waste pile, $\bar{x}_{\text{total}}$, may be calculated using the following formula [6]:

$$\bar{x}_{\text{total}} = \sum_{h=1}^{L} W_h \cdot \bar{x}_h = 0.8 \cdot 3.7 + 0.2 \cdot 9.9 = 4.9$$

where $\bar{x}_h$ is equal to the mean of the individual stratum (computed as shown above for Case 2a—Simple Random), W_h is equal to the weight of the individual stratum, h is the individual stratum, and L is the total number of strata.

The standard deviation of the overall waste pile may be calculated by:

$$s_{\text{total}} = \sqrt{\sum_{h=1}^{L} W_h^2 \cdot \frac{s_h^2}{n_h}} = 0.2$$

where $_{N_h}$ is the number of samples collected in the h^{th} stratum. To calculate the upper confidence limit (UCL) on the mean, the degrees of freedom (df) must first be calculated using the formula

$$df = \frac{s_{\text{total}}^2}{\sum_{h=1}^{L} \frac{(W_h \cdot s_h)^4}{n_h^2\,(n_h - 1)}} = 469$$

The upper confidence limit on the mean can then be calculated using the specified alpha error rate and the degrees of freedom calculated using the above equation.

$$\text{UCL}_\alpha = \bar{x}_{\text{total}} + t_{1-\alpha,df} \cdot s_{\text{total}} = 4.9 + 1.284 \cdot 0.2 = 5.1 \text{ mg/L}$$

Conclusion

The 90% UCL for the mean of the TCLP data is 5.1 mg/L, which is greater than the regulatory level of 5.0 mg/L. Thus, material in the waste pile is determined to be hazardous for lead based on the established decision rule.

TABLE 5-1—Totals and TCLP Results—Case 5.

Location	TCLP Result, mg/L	Location	TCLP Result, mg/L
Stratum 1 (A1):	9.2	Stratum 2 (F4):	4.8
Stratum 1 (B3):	10.5	Stratum 2 (F7):	3.0
Stratum 1 (C2):	9.9	Stratum 2 (G8):	4.4
Stratum 2 (A8):	3.5	Stratum 2 (H1):	3.7
Stratum 2 (B7):	4.2	Stratum 2 (H6):	3.1
Stratum 2 (C5):	3.8	Stratum 2 (I9):	5.0
Stratum 2 (D7):	3.6	Stratum 2 (J3):	2.8
Stratum 2 (E9):	2.3	Stratum 2 (J6):	3.4
Stratum 2 (F2):	4.0		

REFERENCES

[1] U.S. EPA, "RCRA Waste Sampling Draft Technical Guidance SW-846 Chapter Nine—Planning, Implementation, and Assessment," EPA/530/R-99/015, Solid Waste and Emergency Response, Washington, DC, 1999.

[2] U.S. EPA, "Guidance for Data Quality Assessment—Practical Methods for Data Analysis," QA/G-9, EPA/600/R-96/084, Office of Research and Development, Washington, DC, 1998.

[3] U.S. EPA, Data Quality Assessment Statistical Toolbox (DataQUEST), User's Guide and Software, <http://es.epa.gov/

TABLE 5-2—Totals and TCLP Statistical Results for Case 5.

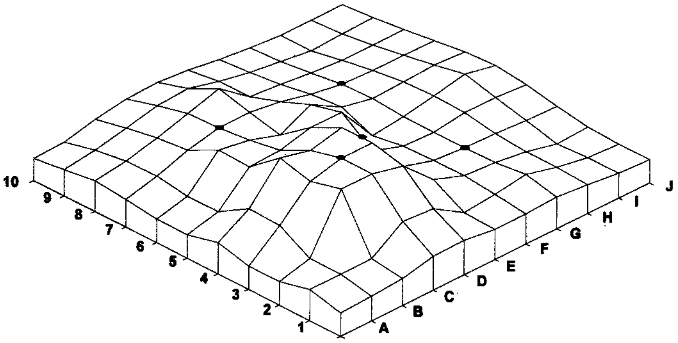
Standard	Coefficient of	90% UCL

ncerqa/qa/qa_docs.html#G-9d> EPA QA/G-9D, EPA/600/R-96/085, December 1996.

[4] Gilbert, R. O., *Statistical Methods for Environmental Pollution Monitoring,* John Wiley and Sons, New York, NY, 1987.

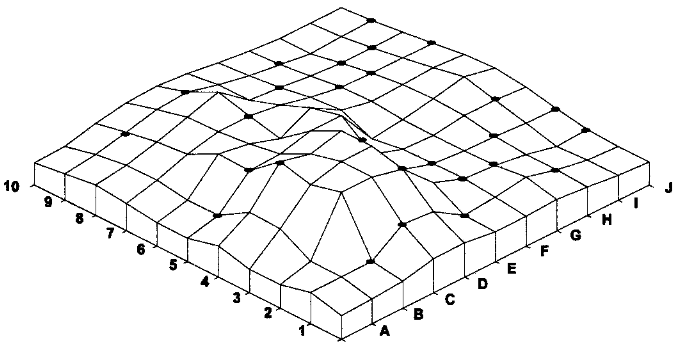
[5] Singh, A. K., Singh, A., and Engelhardt, M., "The Lognormal Distribution in Environmental Applications," (EPA/600/R-97/006), Technology Support Center Issue, USEPA Office of Research and Development, 1997.

[6] U.S. EPA, "Methods For Evaluating the Attainment of Cleanup Standards—Volume 1: Soils and Solid Media," EPA 230/02-89-042, Office of Policy Planning and Evaluation, Washington, DC, 1989.



● Sample Location

FIG. 5—Sample location map, Case 1: Authoritative Sampling Design.



● Sample Location

FIG. 6—Sample location map, Case 2a and 2b: Simple Random Design.

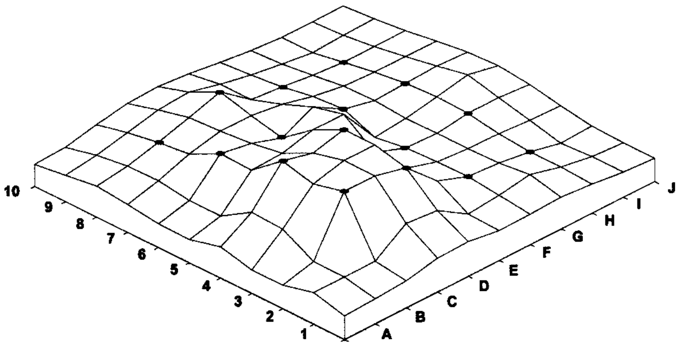
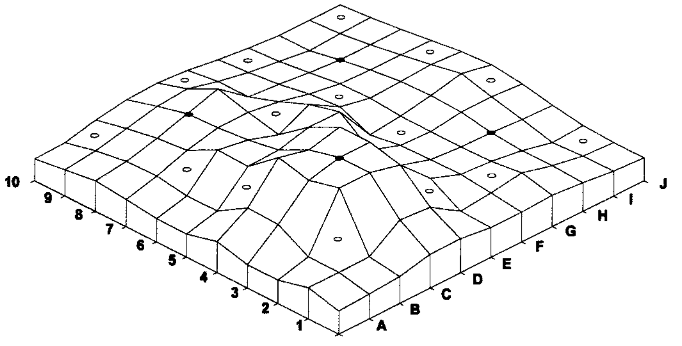


FIG. 7—Sample location map, Case 3: Systematic Grid Sampling Design (without compositing).



● Sample Location (center point)
○ Alliquot Locations

FIG. 8—Sample location map, Case 4: Systematic Grid Sampling Design (with compositing).

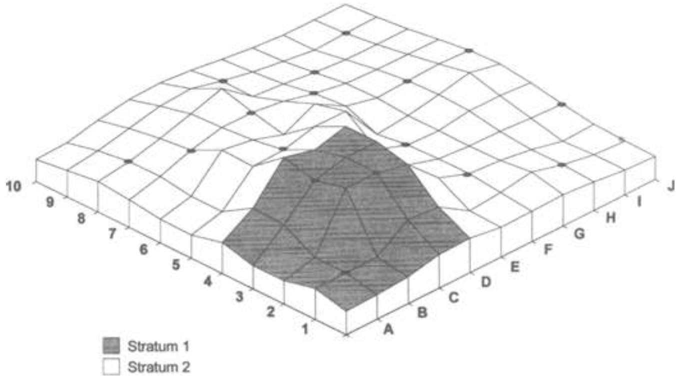


FIG. 9—Sample location map, Case 5: Stratified Random Sampling Design.

TABLE 1—Coefficients of a_i for the Shapiro-Wilk Test for Normality.

i\n	2	3	4	5	6	7	8	9	10	
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	
2	—	0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	
3	—	—	—	0.0000	0.0875	0.1401	0.1743	0.1976	0.2141	
4	—	—	—	—	—	0.0000	0.0561	0.0947	0.1224	
5	—	—	—	—	—	—	—	0.0000	0.0399	
i\n	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7	—	—	0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8	—	—	—	—	0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9	—	—	—	—	—	—	0.0000	0.0163	0.0303	0.0422
10	—	—	—	—	—	—	—	—	0.0000	0.0140
i\n	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944
3	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.2510	0.2499	0.2487
4	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148
5	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836	0.1848	0.1857	0.1864	0.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316	0.1346	0.1372	0.1395	0.1415
8	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219
9	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036
10	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672	0.0728	0.0778	0.0822	0.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12	—	—	0.0000	0.0107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537
13	—	—	—	—	0.0000	0.0094	0.0178	0.0253	0.0320	0.0381
14	—	—	—	—	—	—	0.0000	0.0084	0.0159	0.0227
15	—	—	—	—	—	—	—	—	0.0000	0.0076
i\n	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964
2	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774	0.2755	0.2737
3	0.2475	0.2462	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391	0.2380	0.2368
4	0.2145	0.2141	0.2137	0.2132	0.2127	0.2121	0.2116	0.2110	0.2104	0.2098
5	0.1874	0.1878	0.1880	0.1882	0.1883	0.1883	0.1883	0.1881	0.1880	0.1878
6	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691
7	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513	0.1520	0.1526
8	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356	0.1366	0.1376
9	0.1066	0.1093	0.1118	0.1140	0.1160	0.1179	0.1196	0.1211	0.1225	0.1237
10	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075	0.1092	0.1108
11	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986
12	0.0585	0.0629	0.0669	0.0706	0.0739	0.0770	0.0798	0.0824	0.0848	0.0870
13	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706	0.0733	0.0759
14	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592	0.0622	0.0651
15	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481	0.0515	0.0546
16	0.0000	0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444
17	—	—	0.0000	0.0062	0.0119	0.0172	0.0220	0.0264	0.0305	0.0343
18	—	—	—	—	0.0000	0.0057	0.0110	0.0158	0.0203	0.0244
19	—	—	—	—	—	—	0.0000	0.0053	0.0101	0.0146
20	—	—	—	—	—	—	—	—	0.0000	0.0049

TABLE 1—(continued).

$i \backslash n$	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
3	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
4	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
8	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
9	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
10	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
11	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
13	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
21	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22	—	—	0.0000	0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23	—	—	—	—	0.0000	0.0039	0.0076	0.0111	0.0143	0.0174
24	—	—	—	—	—	—	0.0000	0.0037	0.0071	0.0104
25	—	—	—	—	—	—	—	—	0.0000	0.0035

Source: From Shapiro and Wilk, 1965. Used by permission.
This table is used in Section 12.3.1

TABLE 2—Quantiles of the Shapiro-Wilk *W* Test for Normality (values of *W* such that 100*p*% of the distribution of *W* is less than W_p).

n	$w_{0.01}$	$w_{0.02}$	$w_{0.05}$	$w_{0.10}$	$w_{0.50}$
3	0.753	0.756	0.767	0.789	0.959
4	0.687	0.707	0.748	0.792	0.935
5	0.686	0.715	0.762	0.806	0.927
6	0.713	0.743	0.788	0.826	0.927
7	0.730	0.760	0.803	0.838	0.928
8	0.749	0.778	0.818	0.851	0.932
9	0.764	0.791	0.829	0.859	0.935
10	0.781	0.806	0.842	0.869	0.938
11	0.792	0.817	0.850	0.876	0.940
12	0.805	0.828	0.859	0.883	0.943
13	0.814	0.837	0.866	0.889	0.945
14	0.825	0.846	0.874	0.895	0.947
15	0.835	0.855	0.881	0.901	0.950
16	0.844	0.863	0.887	0.906	0.952
17	0.851	0.869	0.892	0.910	0.954
18	0.858	0.874	0.897	0.914	0.956
19	0.863	0.879	0.901	0.917	0.957
20	0.868	0.884	0.905	0.920	0.959
21	0.873	0.888	0.908	0.923	0.960
22	0.878	0.892	0.911	0.926	0.961
23	0.881	0.895	0.914	0.928	0.962
24	0.884	0.898	0.916	0.930	0.963
25	0.886	0.901	0.918	0.931	0.964
26	0.891	0.904	0.920	0.933	0.965
27	0.894	0.906	0.923	0.935	0.965
28	0.896	0.908	0.924	0.936	0.966
29	0.898	0.910	0.926	0.937	0.966
30	0.900	0.912	0.927	0.939	0.967
31	0.902	0.914	0.929	0.940	0.967
32	0.904	0.915	0.930	0.941	0.968
33	0.906	0.917	0.931	0.942	0.968
34	0.908	0.919	0.933	0.943	0.969
35	0.910	0.920	0.934	0.944	0.969
36	0.912	0.922	0.935	0.945	0.970
37	0.914	0.924	0.936	0.946	0.970
38	0.916	0.925	0.938	0.947	0.971
39	0.917	0.927	0.939	0.948	0.971
40	0.919	0.928	0.940	0.949	0.972
41	0.920	0.929	0.941	0.950	0.972
42	0.922	0.930	0.942	0.951	0.972
43	0.923	0.932	0.943	0.951	0.973
44	0.924	0.933	0.944	0.952	0.973
45	0.926	0.934	0.945	0.953	0.973
46	0.927	0.935	0.945	0.953	0.974
47	0.928	0.936	0.946	0.954	0.974
48	0.929	0.937	0.947	0.954	0.974
49	0.929	0.937	0.947	0.955	0.974
50	0.930	0.938	0.947	0.955	0.974

Source: After Shapiro and Wilk, 1965.

The null hypothesis of a normal distribution is rejected at the α significance level if the calculated *W* is less than W_α.

This table is used in Section 12.3.1

TABLE 3—Quantiles of the *t* Distribution (values of *t* such that 100*p*% of the distribution is less than t_p).

Degrees of Freedom	$t_{0.60}$	$t_{0.70}$	$t_{0.80}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.990}$	$t_{0.995}$
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763
29	.256	.530	.854	1.311	1.699	2.045	2.462	2.756
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576

Source: From Fisher and Yates, 1974. Used by permission.
This table is first used in Section 4.4.2