63

PNL-8989
UC-630

73773

Comparing Statistical Tests for Detecting Soil Contamination Greater Than Background

J. W. Hardin
R. O. Gilbert

December 1993

Prepared for the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Operated for the U.S. Department of Energy
by Battelle Memorial Institute

Battelle

PNL-8989

3617

RECEIVED
JAN 31 2003
By RPY-

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes any **warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.** Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST LABORATORY
operated by
BATTELLE MEMORIAL INSTITUTE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC06-76RLO 1830

Printed in the United States of America

Available to DOE and DOE contractors from the Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831; prices available from (615) 576-8401. FTS 626-8401.

Available to the public from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161.

The contents of this report were printed on recycled paper

PNL-8989
UC-630

COMPARING STATISTICAL TESTS FOR DETECTING SOIL CONTAMINATION GREATER THAN BACKGROUND

J. W. Hardin
R. O. Gilbert

December 1993

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

SUMMARY

The Washington State Department of Ecology (WSDE) recently issued a report (WSDE 1992) that provides guidance on statistical issues regarding investigation and cleanup of soil and groundwater contamination under the Model Toxics Control Act Cleanup Regulation. Included in the report are procedures for determining a background-based cleanup standard and for conducting a 3-step statistical test procedure (denoted here as the State test) to decide if a site is contaminated greater than the background standard (WSDE 1992, Figure 12). The guidance specifies that the State test should only be used if the background and site data are lognormally distributed. The guidance in WSDE (1992) allows for using alternative tests on a site-specific basis if prior approval is obtained from WSDE.

This report presents the results of a Monte Carlo computer simulation study conducted to evaluate the performance of the State test and several alternative tests for various contamination scenarios (background and site data distributions). The primary test performance criteria are (1) the probability the test will indicate that a contaminated site is indeed contaminated, and (2) the probability that the test will indicate an uncontaminated site is contaminated.

The simulation study was conducted assuming the background concentrations were from lognormal or Weibull distributions. The site data were drawn from distributions selected to represent various contamination scenarios. The statistical tests studied are the State test, t test, Satterthwaite's t test, five distribution-free tests, and several tandem tests (wherein two or more tests are conducted using the same data set).

The main conclusions of the Monte Carlo study are (see Section 5.0 for related discussion):

- No single test performs well for all contamination scenarios and numbers of samples.

- When the background and site data distributions are highly asymmetric with a long tail of high concentrations, the State test has an extremely high probability (equal to 1 in some cases) of indicating an uncontaminated site is contaminated. For this reason, the State test is not recommended for general testing purposes.

- This report recommends that the Wilcoxon Rank Sum test, in combination with the Quantile test, be routinely used in place of the State test, and that graphical procedures also be used to visually display and compare the background and site samples.

- The Wilcoxon Rank Sum test is the best over-all performing test. It performance advantage is particularly high when the site is uniformly contaminated greater than the background region.

- The Quantile test performs better than the Wilcoxon Rank Sum test when a small portion of the site is highly contaminated.

- In general, all tests have little ability to identify sites that have only a small fraction of their area contaminated only slightly greater than background. The situation improves as the concentration levels become large and a large number of samples are collected.

ACKNOWLEDGEMENTS

This study was conducted by the Pacific Northwest Laboratory for the Westinghouse Hanford Corporation (WHC), Richland, Washington. We wish to acknowledge the able direction and technical assistance provided by Tim LeGore, the WHC technical manager for the project. Peer review of this paper was provided by A. M. Liebetrau, D.J. (Rick) Bates, and Robert F. O'Brien of the Pacific Northwest Laboratory (PNL). We also thank D. Bruce Rex, Computer Sciences Department, Applied Physics Center, PNL, for reviewing the simulation computer code.

CONTENTS

SUMMARY iii

ACKNOWLEDGEMENTS v

1. INTRODUCTION 1.1

1.1 Scope 1.1
1.2 Hypotheses 1.3
1.3 Decision Errors 1.3

2. STATISTICAL TESTS 2.1

2.1 State Test 2.3

2.1.1 Background Cleanup Standard 2.3
2.1.2 Upper Confidence Limit of Lognormal Mean 2.4
2.1.3 Percentage Exceedance Test 2.4
2.1.4 Magnitude of Exceedance Test 2.5

2.2 Wilcoxon Rank Sum Test 2.5
2.3 Quantile Test 2.5
2.4 Slippage Test 2.6
2.5 Median Test 2.6
2.6 Savage Test 2.6
2.7 Student's t Test 2.7
2.8 Satterthwaite's t Test 2.7
2.9 Tandem Testing 2.7

3. DATA MODELS AND THE SIMULATION PROCEDURE 3.1

3.1 Background Distributions 3.1
3.2 Site Distribution (F_s) 3.4
3.3 Shift Parameter (P_τ) 3.5
3.4 Monte Carlo Simulation Procedure 3.6

4. RESULTS 4.1

4.1 Pervasive Contamination ($\epsilon = 1$) 4.2

4.1.1 Slightly Skewed Distributions (CV = 0.1) 4.2
4.1.2 Highly Skewed Distributions (CV = 1 and 2) 4.8
4.1.3 Tandem Testing 4.19

4.2 Ten Percent Contamination at the Site (ϵ = 0.1) 4.33

4.2.1 Slightly Skewed Distribution (CV = 0.1) 4.33
4.2.2 Highly Skewed Distributions (CV = 1 and 2) 4.38
4.2.3 Tandem Testing 4.39

4.3 Summary of Type I Error Rates 4.49

4.3.1 State Test 4.49
4.3.2 Other Tests 4.51

4.4 Cases where the State Test Performs Well 4.57

5. DISCUSSION 5.1

6. REFERENCES 6.1

APPENDIX: SIMULATION METHODS A.1

A.1 Determining Distribution Parameters A.1

A.1.1 Lognormal Distribution A.1
A.1.2 Weibull Distribution A.2

A.2 Simulating Random Normal Values A.2
A.3 Simulating Random Lognormal Values A.3
A.4 Simulating Random Weibull Values A.4

LIST OF FIGURES

3.1 The five lognormal background distributions used in the simulation 3.2

3.2 The five Weibull background distributions used in the simulation 3.3

3.3 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted lognormal distribution with 100(1-ϵ)% of a lognormal background distribution with CV = 0.1 . . . . . 3.9

3.4 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted lognormal distribution with 100(1-ϵ)% of a lognormal background distribution with CV = 0.5 . . . . . 3.10

3.5 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted lognormal background with 100(1-ϵ)% of a lognormal background distribution with CV = 1.0 . . . . . 3.11

3.6 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted lognormal background with 100(1-ϵ)% of a lognormal background distribution with CV = 1.5 . . . . . 3.12

3.7 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted lognormal background with 100(1-ϵ)% of a lognormal background distribution with CV = 2.0 . . . . . 3.13

3.8 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted Weibull distribution with 100(1-ϵ)% of a Weibull background distribution with CV = 0.1 . . . . . 3.14

3.9 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted Weibull distribution with 100(1-ϵ)% of a Weibull background distribution with CV = 0.5 . . . . . 3.15

3.10 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted Weibull distribution with 100(1-ϵ)% of a Weibull background distribution with CV = 1.0 . . . . . 3.16

3.11 Illustration of five site distributions obtained by mixing 100ϵ% of a shifted Weibull distribution with 100(1-ϵ)% of a Weibull background distribution with CV = 1.5 . . . . . 3.17

3.12 Illustration of five site distributions obtained by mixed 100ϵ% of a shifted Weibull distribution with 100(1-ϵ%) of a Weibull background distribution with CV = 2.0 . . . . . 3.18

4.1 Power curves for 5 test procedures for a lognormal background distribution, CV = 0.10, ϵ = 1, 10 background samples and 10 site samples 4.3

4.2 Power curves for 5 test procedures for a Weibull background distribution, CV = 0.10, ϵ = 1, 10 background samples and 10 site samples 4.4

4.3 Power curves for 5 test procedures for a lognormal background distribution, CV = 0.10, ϵ = 1, 60 background samples and 50 site samples 4.6

4.4 Power curves for 5 test procedures for a Weibull background distribution, CV = 0.10, ϵ = 1, 60 background samples and 50 site samples 4.7

4.5 Power curves for 5 test procedures for a lognormal background distribution, CV = 1.0, ϵ = 1, 10 background samples and 10 site samples 4.9

4.6 Power curves for 5 test procedures for a Weibull background distribution, CV = 1.0, ϵ = 1, 10 background samples and 10 site samples 4.10

4.7 Power curves for 5 test procedures for a lognormal background distribution, CV = 1.0, ϵ = 1, 60 background samples and 50 site samples 4.11

4.8 Power curves for 5 test procedures for a Weibull background distribution, CV = 1.0, ϵ = 1, 60 background samples and 50 site samples 4.12

4.9 Power curves for 5 test procedures for a lognormal background distribution, CV = 2.0, ϵ = 1, 10 background samples and 10 site samples 4.13

4.10 Power curves for 5 test procedures for a Weibull background distribution, CV = 2.0, ϵ = 1, 10 background samples and 10 site samples 4.14

4.11 Power curves for 5 test procedures for a lognormal background distribution, CV = 2.0, ϵ = 1, 60 background samples and 50 site samples 4.15

4.12 Power curves for 5 test procedures for a Weibull background distribution, CV = 2.0, ϵ = 1, 60 background samples and 50 site samples 4.16

4.13 Power curves for 3 tandem tests and the state procedure for a lognormal background distribution, CV = 0.10, ϵ = 1, 10 background samples and 10 site samples 4.20

4.14 Power curves for 3 tandem tests and the state procedure for a Weibull background distribution, CV = 0.10, ϵ = 1, 10 background samples and 10 site samples 4.21

4.15 Power curves for 3 tandem tests and the state procedure for a lognormal background distribution, CV = 0.10, ϵ = 1, 60 background samples and 50 site samples 4.22

4.16 Power curves for 3 tandem tests and the state procedure for a Weibull background distribution, CV = 0.10, ϵ = 1, 60 background samples and 50 site samples 4.23

4.17 Power curves for 3 tandem tests and the state procedure for a lognormal background distribution, CV = 1.0, ϵ = 1, 10 background samples and 10 site samples 4.24

4.18 Power curves for 3 tandem tests and the state procedure for a Weibull background distribution, CV = 1.0, ϵ = 1, 10 background samples and 10 site samples 4.25

4.19 Power curves for 3 tandem tests and the state procedure for a lognormal background distribution, CV = 1.0, ϵ = 1, 60 background samples and 50 site samples 4.26

4.20 Power curves for 3 tandem tests and the state procedure for a Weibull background distribution, CV = 1.0, ϵ = 1, 60 background samples and 50 site samples 4.27

4.21 Power curves for 3 tandem tests and the state procedure for a lognormal background distribution, CV = 2.0, ϵ = 1, 10 background samples and 10 site samples 4.28

4.22 Power curves for 3 tandem tests and the state procedure for a Weibull background distribution, CV = 2.0, ϵ = 1, 10 background samples and 10 site samples 4.29

4.23 Power curves for 3 tandem tests and the state procedure for a lognormal background distribution, CV = 2.0, ϵ = 1, 60 background samples and 50 site samples 4.30

4.24 Power curves for 3 tandem tests and the state procedure for a Weibull background distribution, CV = 2.0, ϵ = 1, 60 background samples and 50 site samples 4.31

4.25 Power curves for 5 test procedures for a lognormal background distribution, CV = 0.10, ϵ = 0.1, 10 background samples and 10 site samples 4.34

4.26 Power curves for 5 test procedures for a Weibull background distribution, CV = 0.10, ϵ = 0.1, 10 background samples and 10 site samples 4.35

4.27 Power curves for 5 test procedures for a lognormal background distribution, CV = 0.10, ϵ = 0.1, 60 background samples and 50 site samples 4.36

4.28 Power curves for 5 test procedures for a Weibull background distribution, CV = 0.10, ϵ = 0.1, 60 background samples and 50 site samples 4.37

4.29 Power curves for 5 test procedures for a lognormal background distribution, CV = 1.0, ϵ = 0.1, 10 background samples and 10 site samples 4.40

4.30 Power curves for 5 test procedures for a Weibull background distribution, CV = 1.0, ϵ = 0.1, 10 background samples, and 10 site samples 4.41

4.31 Power curves for 5 test procedures for a lognormal background distribution, CV = 1.0, ϵ = 0.1, 60 background samples, and 50 site samples 4.42

4.32 Power curves for 5 test procedures for a Weibull background distribution, CV = 1.0, ϵ = 0.1, 60 background samples, and 50 site samples 4.43

4.33 Power curves for 5 test procedures for a lognormal background distribution, CV = 2.0, ϵ = 0.1, 10 background samples, and 10 site samples 4.44

4.34 Power curves for 5 test procedures for a Weibull background distribution, CV = 2.0, ϵ = 0.1, 10 background samples, and 10 site samples 4.45

4.35 Power curves for 5 test procedures for a lognormal background distribution, CV = 2.0, ϵ = 0.1, 60 background samples, and 50 site samples 4.46

4.36 Power curves for 5 test procedures for a Weibull background distribution, CV = 2.0, ϵ = 0.1, 60 background samples, and 50 site samples 4.47

4.37 Power curves for State test and its component parts for the lognormal distribution, CV = 0.10, ϵ = 1, 10 background samples and 10 site samples 4.52

4.38 Power curves for State test and its component parts for the lognormal distribution, CV = 0.10, ϵ = 1, 60 background samples and 50 site samples 4.53

4.39 Power curves for State test and its component parts for the lognormal distribution, CV = 0.50, ϵ = 1, 10 background samples and 10 site samples 4.54

4.40 Power curves for State test and its component parts for the lognormal distribution, CV = 2.0, ϵ = 1, 10 background samples and 10 site samples 4.55

LIST OF TABLES

3.1 Parameter Values Used in the Monte Carlo Simulation Study 3.6

4.1 Estimated Type I Error Rates for the State Test 4.50

4.2 Range of Estimated Type I Error Rates for All Tests 4.56

4.3 Contamination Scenarios for which the State Test has only Moderately Elevated Type I Error Levels and Generally Better Power than Other Tests 4.57

1. INTRODUCTION

In August 1992, the Washington State Department of Ecology (WSDE) issued a report (WSDE 1992) that provides guidance on statistical issues regarding investigation and cleanup of soil and groundwater contamination under the Model Toxics Control Act (MTCA) Cleanup Regulation. The WSDE (1992) guidance document describes and illustrates statistical procedures that are recommended by WSDE for routine use at most sites. The guidance also allows for alternative testing procedures on a site-specific basis if prior approval is obtained from WSDE.

The purpose of this document is to report the results of a Monte Carlo computer simulation study conducted to evaluate the performance of several statistical tests, including a state test described in WSDE (1992) for deciding whether a site is contaminated above background levels. The goal of this report is to provide information about which tests have the best performance for general application, and which tests work best for selected contamination scenarios. This report will illustrate the limitations as well as the strengths of formal statistical tests. This study was conducted by staff of the Statistical Modeling and Display Group and the Statistical Design and Analysis Group in the Analytic Sciences Department, Applied Physics Center, Pacific Northwest Laboratory (PNL)[a], Richland, Washington.

1.1 Scope

This document evaluates the performance of several statistical tests regarding their ability to correctly indicate whether a site is contaminated, i.e., has concentrations significantly greater than those in the background area. These tests are discussed in Section 2. This performance assessment is conducted assuming that measurements from the defined background area have either a lognormal or a Weibull distribution, which are defined mathematically in the Appendix. A mixture model is used to generate site measurements that

---

[a]Operated for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76RLO 1830.

reflect the extent and magnitude of contamination at the site. This model and the Monte Carlo simulation procedure used are discussed in Section 3.

The ability (power) of the tests to detect contamination on the basis of 10 or more background and site measurements is determined. The probability that each test will incorrectly indicate that an uncontaminated site is contaminated is also determined. The simulation results apply to the particular assumptions and site contamination scenarios described in Section 3, including the assumptions that samples are collected using simple random sampling and measurements are independent and hence uncorrelated.

The tests evaluated here include several distribution-free procedures. These tests do not require that measurements be normally distributed. Distribution-free tests may perform better or worse than tests based on a distribution assumption, depending in part on the number of samples, the extent and magnitude of contamination at the site relative to the background area, and the underlying distribution. The performance of all the tests is discussed and summarized graphically in Section 4.0.

Section 5.0 provides a summary discussion of the results of this paper, including caveats and limitations. That section also outlines a recommended strategy for comparing a site with background. This process includes using visual aids, formal statistical tests, and professional judgment.

No attempt is made here to consider the performance of tests when data sets are censored, i.e., when background or site data sets include measurements that are less than the detection limit. The WSDE has recently issued a report (WSDE 1993) as a supplement to WSDE (1992) that provides guidance on the analysis of censored background or site data sets. The performance of statistical testing methods applicable to censored data sets is not considered in this report.

1.2 Hypotheses

All statistical tests evaluated in this report test the null hypothesis

H_o: site is not contaminated more than the background area

versus the alternative hypothesis

H_a: site is contaminated more than the background area

These hypotheses can be restated in terms of the distribution of measurements for the background and site areas:

H_o: The background and site distributions are identical

(1)

H_a: At least part of the site distribution is located to the right of the background distribution

This latter formulation emphasizes that what we really want to know is how the site and background distributions compare in location and shape. Each test in this study uses the available site and background data in a different way to evaluate whether H_o should be rejected. If a test indicates H_o should be rejected, then H_a is accepted.

1.3 Decision Errors

There are two types of decision errors that can be made when a statistical test is performed:

- Type I Error (False Positive Decision):

- when a test rejects H_o (see Equation 1) when H_o is actually true, i.e., when a test incorrectly indicates an uncontaminated site is contaminated

- Type II Error (False Negative Decision):

  - when a test fails to reject H_o when H_a is really true, i.e., when a test incorrectly indicates the site is not contaminated

These definitions indicate that a Type I error may lead to performing unneeded remedial action, whereas a Type II error may lead to not performing needed remedial action. Now, the probability of making a Type II error is related to the "power" of a test as follows:

Power = 1 - probability of a Type II error

= probability of rejecting H_o when H_o is false and H_a is really true

Stated in the context of this report, the power of a test is the probability that the test correctly identifies when a site distribution lays to the right of the background distribution. The purpose of the Monte Carlo study was to estimate for each test the power and the probability of making a Type I error for a variety of background and site contamination scenarios.

Before a statistical test can be conducted, it is necessary to specify an acceptable probability of making a Type I error. For the Monte Carlo study of this report, this probability was specified to be 0.05 (5 percent chance). One purpose of the Monte Carlo study was to determine which tests, if any, had Type I error probabilities that differed significantly from the specified (target) 0.05 rate. The actual Type I error rate of a test can deviate from the specified rate because assumptions underlying the test are not fulfilled or because the test does not have a sound theoretical basis.

For example, the t test and Satterthwaite's t test (Section 2) require that measurements be normally distributed. As the Monte Carlo study considered lognormal and Weibull background distributions, this assumption is not true for the contamination scenarios considered here. Hence, for these two test procedures, the deviation of the actual Type I error rate from the nominal (specified) 0.05 rate is due in part to the fact that the normality assumption is false. Similarly, the Type I error rate of the State test (which assumes a lognormal distribution) may vary from the nominal 0.05 rate whenever the Weibull distribution is used. Actually, as will be seen in Section 4, the Type I error rate of the State test seriously deviates from the nominal 0.05 rate even for some lognormal distributions. As a final example, it is not possible to conduct the Quantile test at a 0.05 Type I error rate for some sample sizes. Due to the discrete nature of the Quantile test, its specified Type I error rate is closer to 0.01 than to 0.05 for some sample sizes.

The power of a test depends on the number of samples collected in the background and site areas, on the position and shape of the distribution of site measurements as compared with the distribution of background measurements, and the particular way that the test uses the site and background data to make comparisons. This Monte Carlo study estimates the power of various tests for various combinations of site and background sample sizes when the background and site distributions take on different shapes and relative locations.

2. STATISTICAL TESTS

The statistical tests evaluated in this report are:

- State test
- Wilcoxon Rank Sum test
- Quantile test
- Slippage test
- Median test
- Savage test
- Student's t test
- Satterthwaites's t test

When background and site data can be assumed to be lognormally distributed, WSDE (1992) indicates that the default 3-step State test described in Section 2.1 is used. This test, as illustrated in Figure 12 of WSDE (1992), consists of computing an upper confidence limit on the site mean and comparing it to a fixed background value (standard), followed by a frequency of exceedance test and a magnitude of exceedance test. However, regardless of the distribution of the data, the WSDE background testing approach (WSDE 1992, Figure 12) allows for the use of alternative tests if adequate supporting materials are first provided to WSDE for their review and approval. The other tests evaluated in this report are potential alternative tests.

The Wilcoxon test, Savage test, Median test, Quantile test and Slippage test are distribution-free tests. These tests can be conducted regardless of the distribution of the background or site data. The default 3-step State test, the t test, and Satterthwaite's t test are not distribution-free. As mentioned above, the State test assumes the background and site data are lognormally distributed. The two t tests assume the data to be normally distributed. If distribution assumptions are not met in practice, the Type I error rates of the tests may deviate from specified target values.

The performance of the State test was evaluated by the WSDE prior to its inclusion in WSDE (1992). Those evaluations form the basis for the test as

discussed in WSDE (1992) and Section 2.1 in this report. However, WSDE (1992, page 94) indicates that when background data sets are highly skewed to the right (to high concentrations), it may not be possible to simultaneously achieve desirable false positive error rates and statistical power to detect contamination when the State test is used and the number of samples is small. For that case, WSDE (1992) suggests that the Wilcoxon test and the Quantile tests may be more appropriate. The study reported here provides additional information on the performance of the State test relative to that of the other tests.

The power of the Wilcoxon test is known to be only slightly less than that of the t test when data are normally distributed (Lehmann 1975). Gilbert and Simpson (1992) compared the performance of the Wilcoxon test and the Quantile test when background measurements are normally distributed and the site distribution is a mixture of the background distribution and a contamination distribution. This mixture model is the same model that is used in this report (described by Equation (2) in Section 3). They found that the Wilcoxon test had more power than the Quantile test when the site distribution has the same shape as the background distribution but is shifted to higher concentrations. The Quantile test had more power than the Wilcoxon test when only a relatively small part of the site was highly contaminated. Gilbert and Simpson (1992) suggest using both the Wilcoxon test and the Quantile test when the contamination pattern is not known in advance. In that case, the null hypothesis is rejected if either test indicates the null hypothesis should be rejected. This procedure is one of the tandem tests studied in this report.

Gilbert and Simpson (1990) compared the power of the Quantile test and the Slippage test for normal background distributions and the mixture model. They found that the Quantile test had more power than the Slippage test except when the site distribution is bimodal and one mode is at a higher concentration than most of the background distribution. The Savage test is known to have good power properties when the background and site measurements have Weibull distributions (Hsieh 1988).

2.1 State Test

WSDE (1992) proposes that the default 3-step State test outlined in Sections 2.1.2 through 2.1.4 be used when the background and site data are lognormally distributed. The performance of this State test and all the other tests was evaluated for lognormal and Weibull distributions and for various mixtures of distributions, as described in Section 3.

2.1.1 Background Cleanup Standard

The default WSDE background cleanup standard (WSDE 1992) is the estimated 90th percentile of the distribution of background measurements, unless that estimate is greater than four times the estimated median of the background distribution. In the latter case, the cleanup standard is four times the median of the background distribution. For this simulation study, this rule implies that the cleanup standard is the estimated 90th percentile whenever the coefficient of variation (standard deviation divided by the mean), CV, of the background distribution is less than 1.5. Hence, for this study, if the background CV is greater than or equal to 1.5, then the cleanup standard is 4 times the estimated median. It will be seen in Section 3.1 that distributions in this report that have a CV greater than 1.5 are highly skewed to the right.

WSDE (1992) specifies that once the background standard is determined on the basis of background measurements, it is considered to be a fixed constant. Hence, the uncertainty in estimating the standard on the basis of a finite number of background samples is not considered in the test. For this Monte Carlo study, the background standard was determined from the specified background distribution and did not change from iteration to iteration, even though the background data set changed with each iteration. That is, the background standard was not estimated for each of the 10,000 iterations. Hence, the performance results for the State test do not include variability in estimating the background standard from background data.

It should be noted that the alternative tests evaluated in this report do not use the 90th percentile or any other fixed (constant) parameter of the background distribution as the standard. Each alternative test evaluates the background and site data in a somewhat different way, taking into account the variability of the background and site data, to determine if the site distribution is positioned to the right of the background distribution.

2.1.2 Upper Confidence Limit of Lognormal Mean

The first stage of the State test procedure is to compute the upper one-sided 95% confidence limit of the mean of the lognormal site distribution. This is done by first computing the mean and variance (s^2) of the logarithms (base e) of the n site data. The upper confidence limit is then calculated as

$$UCL = \exp[\text{mean} + 0.5s^2 + sH_{0.95}/(n-1)]$$

where $H_{0.95}$ is a value, based on n and s^2, that is taken from the tables in Land (1975).

If UCL is greater than the background cleanup standard as determined using the procedure in Section 2.1.1, then the site is considered to be contaminated. If the calculated value is not greater than the cleanup standard, then the second stage of the State test is conducted, as described in Section 2.1.3.

2.1.3 Percentage Exceedance Test

If the number of samples from the site is less than or equal to 30 and the 90th percentile of the background distribution is the cleanup standard, then no more than 20% of the site sample measurements may exceed the cleanup standard. If more than 20% exceed the standard, then the site is considered to be contaminated. If the number of site measurements is greater than 30 or the cleanup standard is 4 times the median of the background distribution, then the analyst is instructed to contact WSDE for guidance in determining the allowable percentage of sample site observations that may exceed the cleanup standard. If this second step does not indicate the site is contaminated, then the third step

of the procedure, the "magnitude of exceedance test," described in Section 2.1.4 is used.

2.1.4 Magnitude of Exceedance Test

If the CV of the site measurements is less than 0.5, the number of site measurements is ≤ 30, and the 90th percentile of the background distribution is the cleanup standard, then no measurement from the site may exceed twice the cleanup standard. If any site measurement exceeds twice the cleanup standard, then the site is considered to be contaminated. The procedure described in Attachment 2 (page 43) of the state guidance (WSDE 1992) is used when the cleanup standard is defined to be 4 times the median and the sample size exceeds 30. This procedure alters the allowable magnitude of exceedance value. When applicable, this alternative procedure was used in the Monte Carlo study reported here.

2.2 Wilcoxon Rank Sum Test

The distribution-free Wilcoxon Rank Sum test is performed by first listing the combined site and background measurements from smallest to largest. Then the measurements are replaced by their ranks, where the ranks are the integer values from 1 through N, where N is the total number of combined measurements. The rank 1 is assigned to the smallest value, 2 to the second smallest observation, etc. Next, the sum of the ranks of the site measurements is computed. This sum is used to compute the Wilcoxon test statistic which is compared to a critical value to determine whether to reject H_0 and accept H_a (Equation 1.). The Wilcoxon test procedure is explained in many statistical books and papers including Conover (1980), Lehmann (1975), Gilbert and Simpson (1992), and Gilbert (1987).

2.3 Quantile Test

The Quantile test (Johnson et al. 1987; Gilbert and Simpson 1990, 1992) is a distribution-free test that is performed by first listing the combined background and site measurements from smallest to largest. Then one examines

only the largest r measurements (r is typically less than 10). If k or more of the r measurements are from the site, then the site is considered to be contaminated. The values of k and r may be determined from Tables A.2 through A.8 in Gilbert and Simpson (1992).

2.4 Slippage Test

The distribution-free slippage test can be used to rapidly identify when the site distribution is shifted to the right of the background distribution. The test consists of counting the number of site measurements that exceed the maximum background measurement. If this number of site measurements is too large, as determined by reference to the tables in Rosenbaum (1954), it is concluded that the site is contaminated.

2.5 Median Test

The distribution-free median test (Conover 1980, pp. 171-176) is designed to test if the site and background distributions have the same median. The test is conducted by first computing the grand median, which is the median of the combined set of background and site measurements. Then a 2x2 contingency table is constructed for which the first column shows the number of background measurements that are above and below the grand median. The second column shows the same information for the site measurements. A chi-square test is then applied to the 4 frequencies in the contingency table to decide whether the median of the site distribution is larger than the median of the background distribution. If so, the site is considered to be contaminated.

2.6 Savage Test

The Savage Test [Savage 1956, Hsieh 1988 (pp. 267-270)] is a distribution-free procedure that uses the ranks of the combined data sets, which are obtained in the same way as for the Wilcoxon test (Section 2.2). The Savage test statistic combines the ranks in a manner such that theoretically the test has good power when the underlying distribution is

Weibull. Hsieh (1988, p. 268) provides an example of how to compute the test statistic and conduct the test. If the test statistic is too large, the site is considered to be contaminated.

2.7 Student's t test

The t-test is a well known test for comparing the means of two populations. This test assumes the background and site measurements are normally distributed and that both distributions have the same variance. All measurements are assumed to be independent and hence uncorrelated, an assumption shared with all tests discussed in this report. Critical values for conducting the test are taken from a table of the t-distribution, which is found in most statistics books. Iman and Conover (1983, pp. 272-276) provide an excellent description of this test.

2.8 Satterthwaite's t test

Satterthwaite's t test is a modified form of the standard t test that is appropriate when the background and site distributions have unequal variances. The performance of this test was evaluated and compared with that of the standard t test because unequal variances are expected in practice. This test assumes the site and background data are normally distributed. Iman and Conover (1983, pp. 276-278) discuss and illustrate the test.

2.9 Tandem Testing

Tandem testing is a term that refers to conducting two or more tests on the same set of data and rejecting the null hypothesis, H_o, if one or more of the tests reject H_o. Note that the State test (Section 2.1) is a tandem test because H_o is rejected if any one of the 3 tests in the 3-step State test rejects H_o.

The Monte Carlo study reported here determined the Type I error rate and power of tandem-testing for the case where two tests were conducted on the same background and site data sets. This was done for all combinations of the tests in Sections 2.2 through 2.8. The results of these assessments are discussed in Section 4.

We note that the performance of a "modified Slippage test" (denoted by Slippage*) was also evaluated. The standard Slippage test (Section 2.4) consists of determining if K or more site measurements exceed the maximum background measurement, where K is determined from Rosenbaum (1954) on the basis of the number of site and background measurements and the target Type I error rate (0.05). The modified test consists of using either K or K-1 depending on which value provides a Type I error rate closer to the target value. This modified test was evaluated because for some sample sizes the Type I error rate for the Slippage test when K is used is much smaller than 0.05. When K-1 instead of K is used the Type I error rate will be closer (although somewhat larger) to 0.05 and the power of the test will increase slightly. The performance results for the modified test are on microfiche (available from the first author), which show results for all tests.

3. DATA MODELS AND THE SIMULATION PROCEDURE

This section begins with a description of the models that were used to generate multiple background and site data sets to assess the performance of the statistical tests. This is followed by a discussion of the Monte Carlo simulation procedure and the various conditions for which the performance assessment was conducted.

3.1 Background Distributions

The performance of statistical tests was determined when measurements taken on samples collected in the background area were assumed to have either a lognormal or Weibull distribution. These distributions were selected primarily because measurements of natural analytes in background Hanford soil have been shown to be well fit by either distribution (DOE, 1993). Also, these distributions cover a range of shapes from nearly symmetrical to highly skewed to the right. Hence, by selecting lognormal and Weibull distributions with a range of parameters for this study, the performance of statistical tests was evaluated for a wide range of distribution shapes.

The Weibull and lognormal distributions that were used as models for background concentrations are displayed in Figures 3.1 and 3.2. As seen in these figures, five shapes of both distribution types were used. Each shape was determined by specifying a mean and a coefficient of variation (CV) (standard deviation divided by the mean). For a constant mean value, the shape of the distribution changes with the CV. Five values of the CV were used (CV = 0.1, 0.5, 1.0, 1.5, and 2.0) and the mean was set at 100. Note from Figures 3.1 and 3.2 that the lognormal and Weibull distributions are almost symmetrical when CV = 0.1 and become more skewed to the right as the CV increases. Appendix A shows how the specified mean and CV were used to determine the shape of the background distributions.

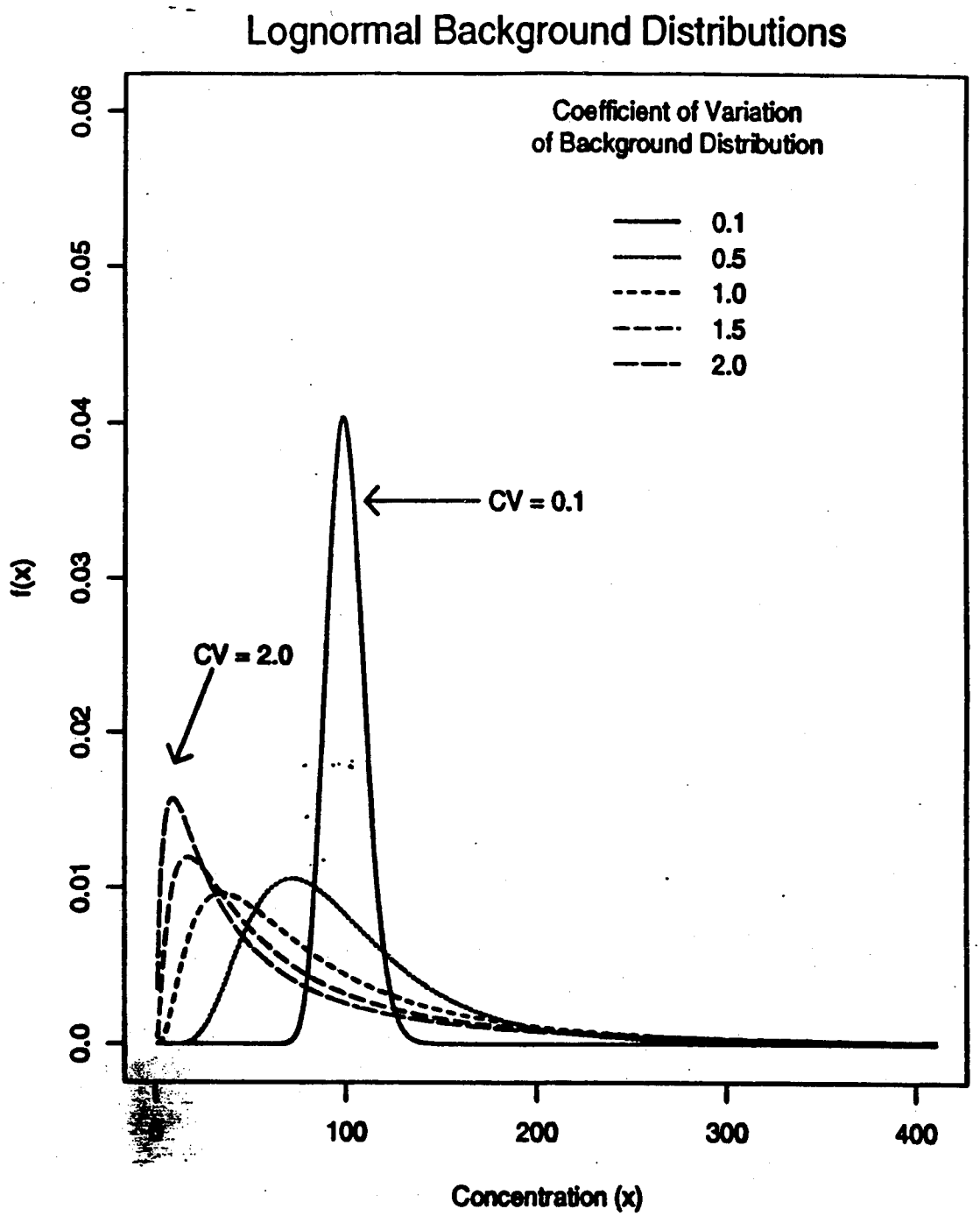


Figure 3.1: The five lognormal background distributions used in the simulation.

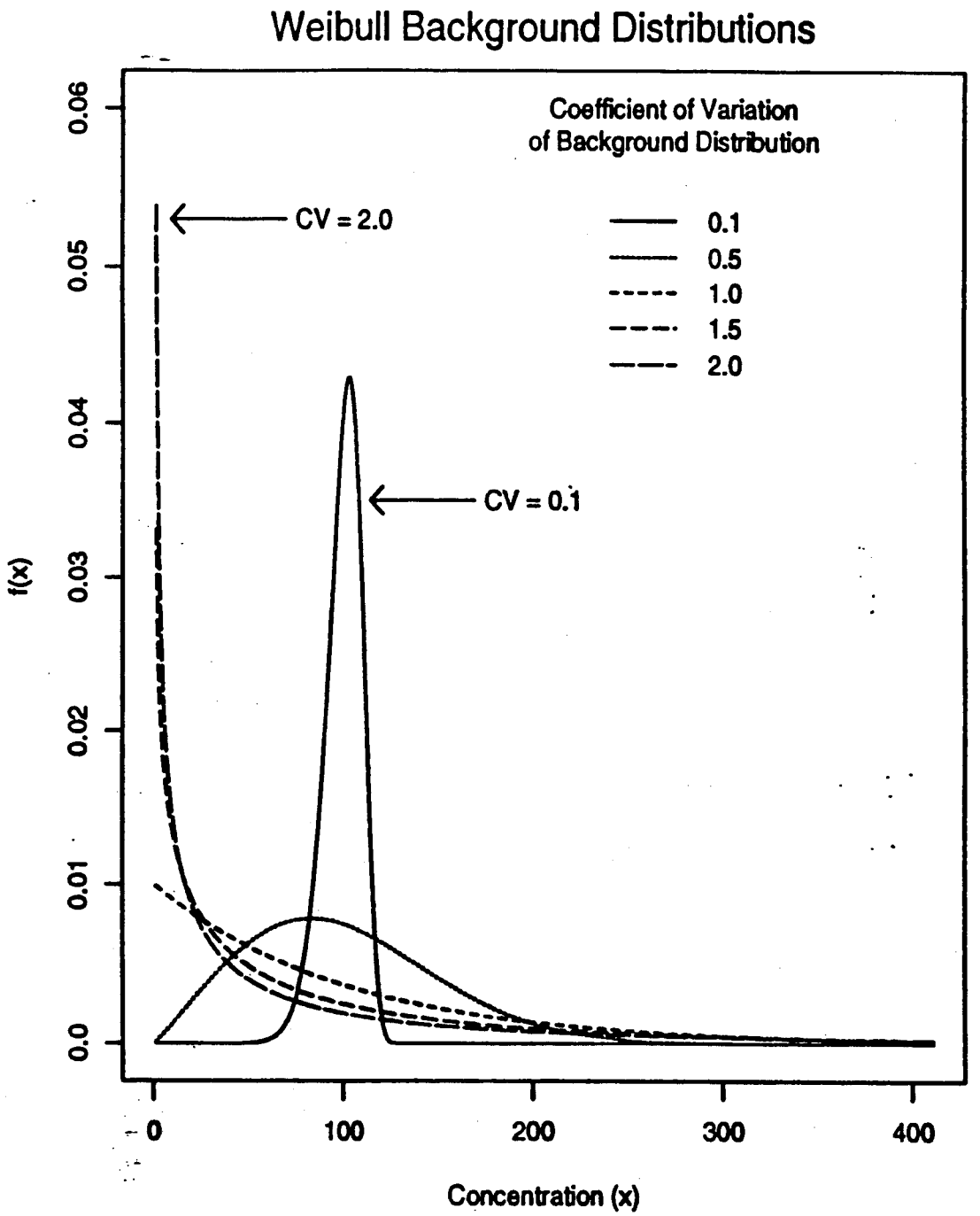


Figure 3.2: The five Weibull background distributions used in the simulation.

3.2 Site Distribution (F_s)

The distribution (F_s) of site measurements was obtained using a mixture model wherein the site distribution is a mixture of the background distribution and the "contamination" distribution, F_c (also referred to as the "shifted" distribution) that represents contamination above background. This approach permitted a wide variety of site distributions to be generated and used for the simulation study. The mixture model is (Johnson et al., 1987; Gilbert and Simpson, 1990):

$$F_s = \epsilon F_c + (1 - \epsilon) F_{bk} \qquad (2)$$

where

F_s = site distribution

F_{bk} = background distribution

F_c = contamination (shifted) distribution

ϵ = proportion of the site that is contaminated with concentrations from F_c.

The site distribution (F_s) was obtained by specifying F_{bk}, F_c, and ϵ in Equation (2). As indicated in Section 3.1, F_{bk} was specified in this study to be lognormal or Weibull. The contamination distribution (F_c) is a shifted background distribution. That is, F_c is identical in shape to the background distribution (F_{bk}), but it is shifted an amount Δ to the right of F_{bk}. F_c can be thought of as the distribution of measurements drawn from portions of the study site that had a constant contamination added to background.

The parameter ϵ can be thought of as the fraction of the total number of site samples that are collected from the contaminated portion of the site. Consequently, $1-\epsilon$ is the fraction collected from F_{bk}, i.e., from the portion of the site that is not contaminated greater than the background area. When $\epsilon = 0$, the site and background distributions are identical. When $\epsilon = 1$, then Equation (2) indicates that $F_c = F_s$. Hence, when $\epsilon = 1$, the site and background distributions have the same shape, but the site distribution is shifted to the right of the background distribution by the amount Δ.

If $\epsilon = 0.5$, then 50 percent of the n site observations are from the contamination (shifted) distribution, F_c, and 50 percent are from the background distribution, F_{bk}. The Monte Carlo study of this report was repeated for ϵ equal to 0, 0.1, 0.3, 0.5, 0.7, and 1.0 for various values of Δ. The Appendix discusses the mathematical methods used to determine the parameters of the background and shifted distributions.

3.3 Shift Parameter (P_r)

The contamination distribution, F_c, was defined in Section 3.2 to be the background distribution shifted to the right an amount Δ. The amount of shift can be characterized either according to the distance Δ or by the probability P_r, where

P_r = probability that a random value drawn from the contamination (shifted) distribution, F_c, is greater than a random value drawn from the background distribution, F_{bk}

All of the performance results presented in this report (see the figures in Section 4.0) are in terms of P_r rather than Δ. This approach was taken because different values of Δ are obtained for the same specified value of P_r as the shape of the background distribution changes. Hence, comparing the performance of tests on the basis of Δ can be difficult.

The value of Δ that corresponds to a specified value of P_r can be obtained analytically for some distributions, e.g., the normal distribution (see Equation 6.10 in Gilbert and Simpson 1992). However, analytical solutions are not available for lognormal and Weibull distributions. Therefore, the value of Δ that corresponds to a specified value of P_r was obtained via a recursive search method. The parameter P_r is discussed in some detail by Lehmann (1975) (he uses the notation P_1 instead of P_r).

When $\epsilon = 0$ so that the site and background distributions are identical, then $P_r = 0.5$. When $\epsilon > 0$, as P_r is specified to be closer and closer

to 1, the contamination distribution is shifted more and more to the right of the background distribution.

3.4 Monte Carlo Simulation Procedure

The Monte Carlo simulation study was conducted for all combinations of the parameter values in Table 3.1.

Table 3.1 Parameter Values Used in the Monte Carlo Simulation Study

PARAMETER	VALUES
Background Distributions	Lognormal, Weibull
Mean	100
CV of Background Distribution	0.1, 0.5, 1.0, 1.5, 2.0
P_r	0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.99, 0.999
ϵ	0, 0.1, 0.3, 0.5, 0.9, 1.0
Number of Background Samples (n_b)	10, 30, 50, 70
Number of Site Samples (n_s)	10, 30, 60, 90, 120
Number of iterations	10,000

The simulation proceeded as follows using parameter values from Table 3.1:

1. Specify the shape and location of the background distribution by specifying a lognormal (or Weibull) distribution that has a mean of 100 and a specified value of the CV.

2. Determine the shape and location of the site distribution by specifying a value for P_r and for ϵ and using Equation (2).

3. Specify the number of background samples and site samples.

4. Generate on the computer the specified number of background and site samples.

5. Conduct all statistical tests using the generated background and site data sets.

6. Repeat Steps 4 and 5 a total of 10,000 times, each time using a new set of generated samples.

7. Store in a computer file for each statistical test the proportion of the 10,000 iterations for which the test rejected the null hypothesis.

The above procedure was carried out for all combinations of sample sizes, CV, P_r, and ϵ for the lognormal distribution. Then the entire process was repeated for the Weibull distribution. The performance of each test procedure for each combination of the parameters in Table 3.1 was measured by the results stored in Step 7, i.e., by the fraction of the 10,000 data sets for which the test procedure indicated the null hypothesis [Equation (1)] should be rejected. When $P_r = 0.5$, which means that the background and site distribution are identical, this fraction is the estimated probability of a Type I decision error (probability that the test will incorrectly indicate the site is contaminated). When $P_r > 0.5$, the fraction from Step 7 is an estimate of the power of the test to correctly indicate when the site is contaminated greater than background. As P_r takes on larger and larger values (approaching 1.0) the power of any valid test procedure should increase to one, the maximum possible value. The performance of the test procedures was compared graphically by plotting the estimated power versus the value of P_r for a specified background distribution, ϵ and number of samples. These results are the subject of Section 4.

Figures 3.3 through 3.7 and 3.8 through 3.12 show for lognormal and Weibull distributions, respectively, the site distributions when $\epsilon = 0.1$, 0.3, 0.5, 0.7, 0.9 and 1 for CV = 0.1, 0.5, 1, 1.5, and 2 when $P_r = 0.85$. These figures also show the lognormal or Weibull background distribution for each CV value. These figures illustrate the wide variety of unimodal and bimodal (two-humped) site distributions for which the performance of tests was assessed when $P_r = 0.85$. Of course, as discussed above, test performance was

also assessed for a much wider suite of site distributions corresponding to values of P_r listed in Table 3.1.

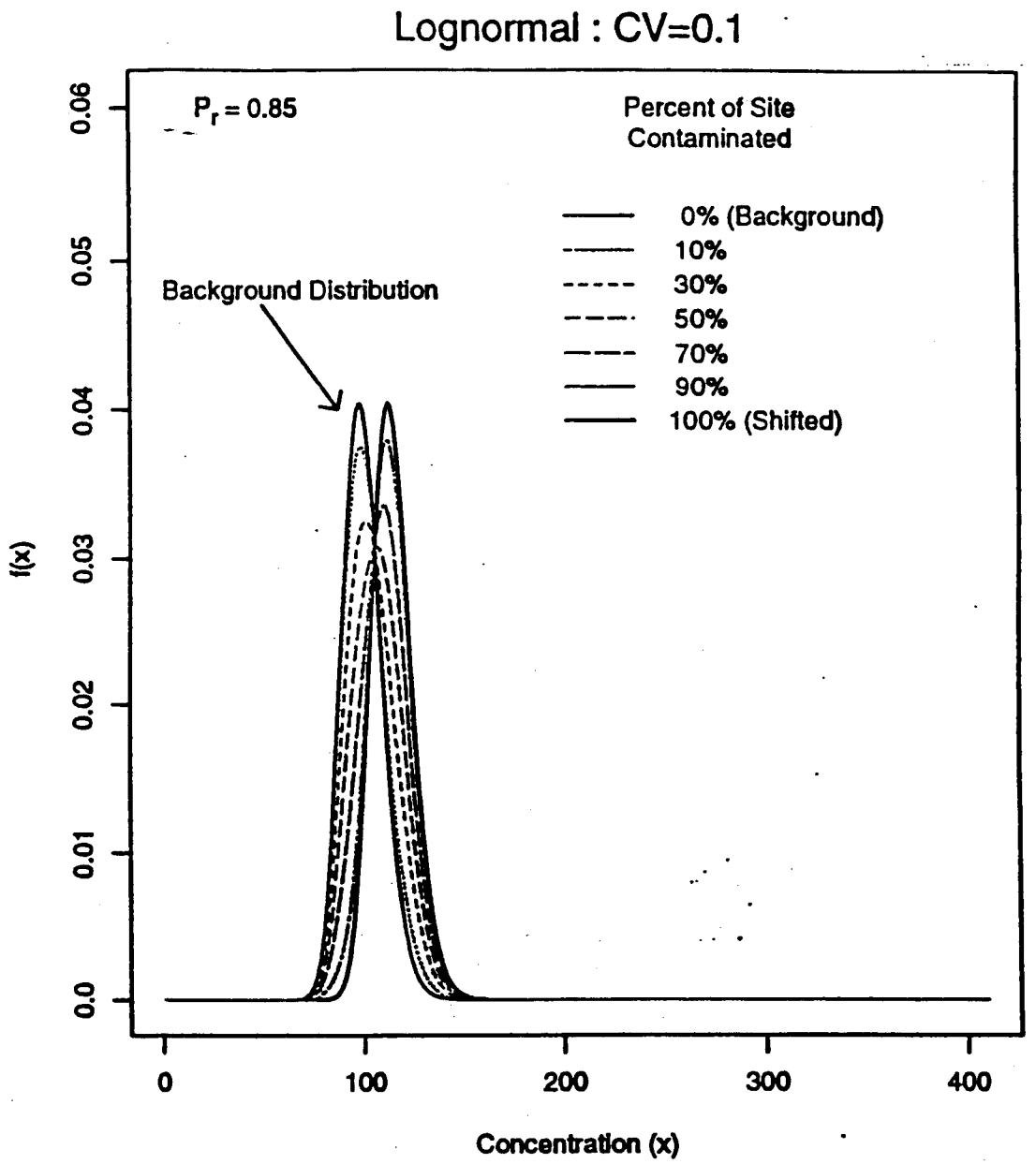


Figure 3.3: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for $\epsilon = 0.1, 0.3, 0.5, 0.7$, and 0.9. The background distribution is lognormal with CV=0.1 and the shifted distribution is characterized by $P_r = 0.85$, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

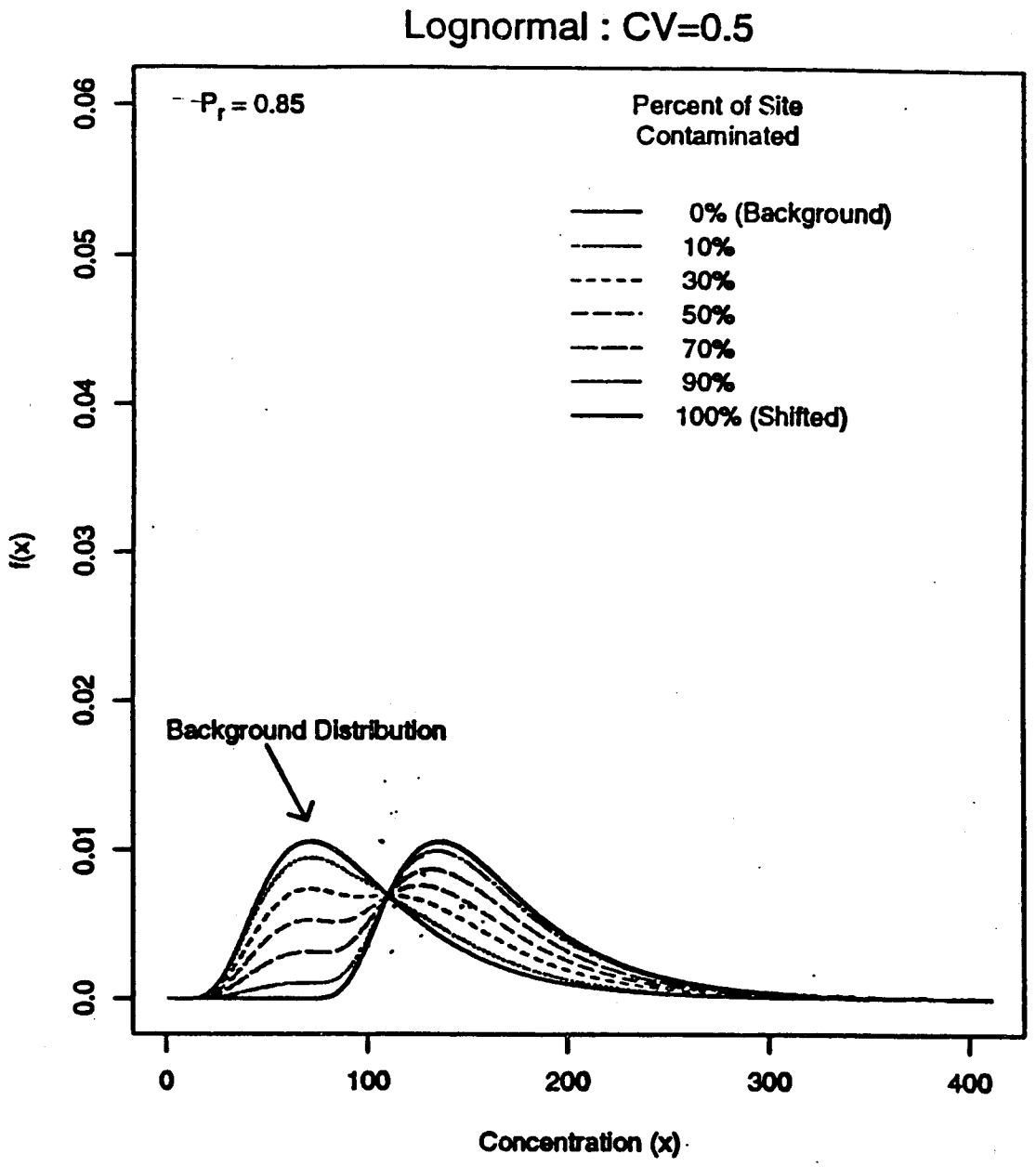


Figure 3.4: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for ϵ = 0.1, 0.3, 0.5, 0.7, and 0.9. The background distribution is lognormal with CV=0.5 and the shifted distribution is characterized by P_r = 0.85, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

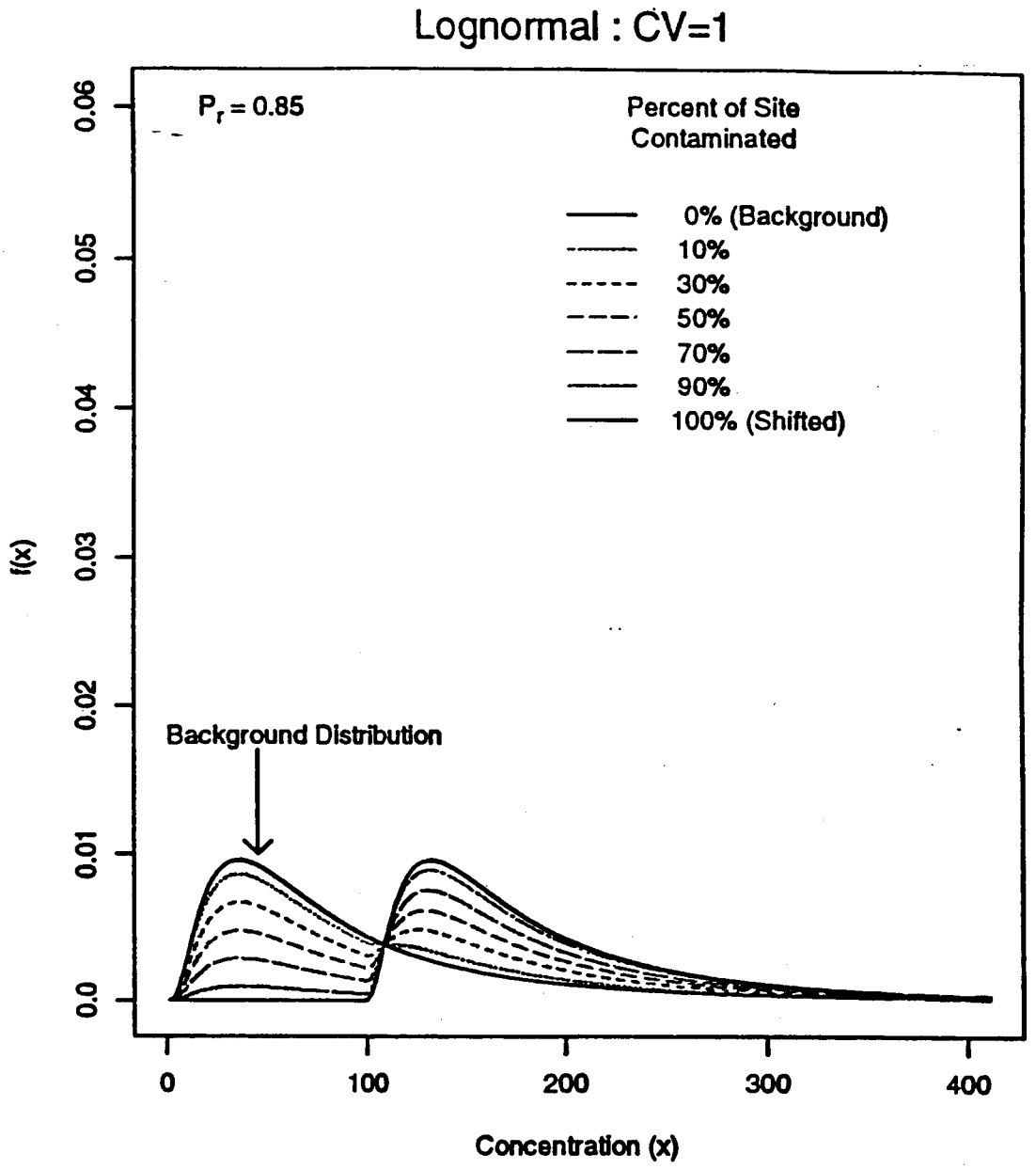


Figure 3.5: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for $\epsilon = 0.1, 0.3, 0.5, 0.7$, and 0.9. The background distribution is lognormal with CV=1.0 and the shifted distribution is characterized by $P_r = 0.85$, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

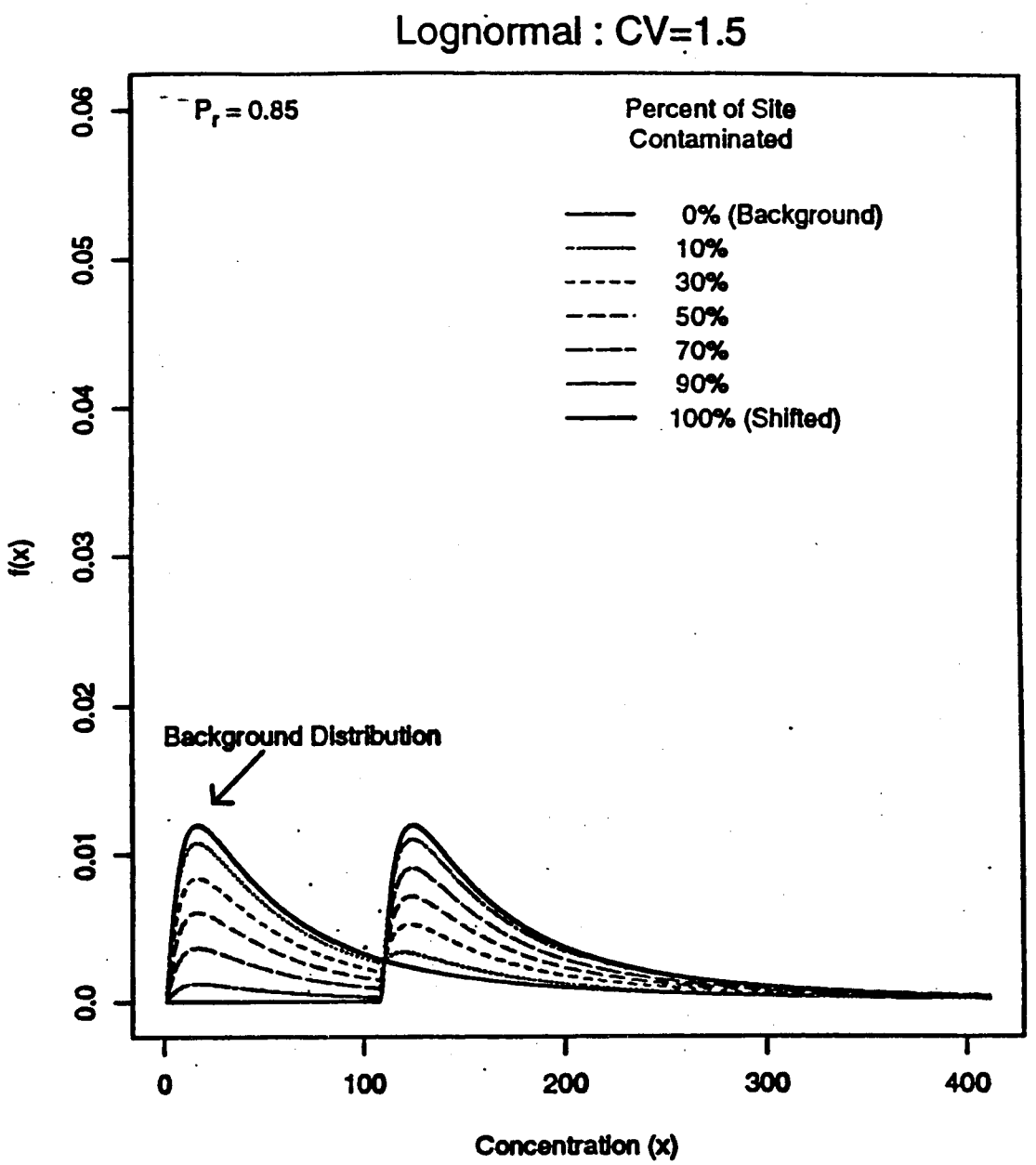


Figure 3.6: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for $\epsilon = 0.1, 0.3, 0.5, 0.7$, and 0.9. The background distribution is lognormal with CV=1.5 and the shifted distribution is characterized by $P_r = 0.85$, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

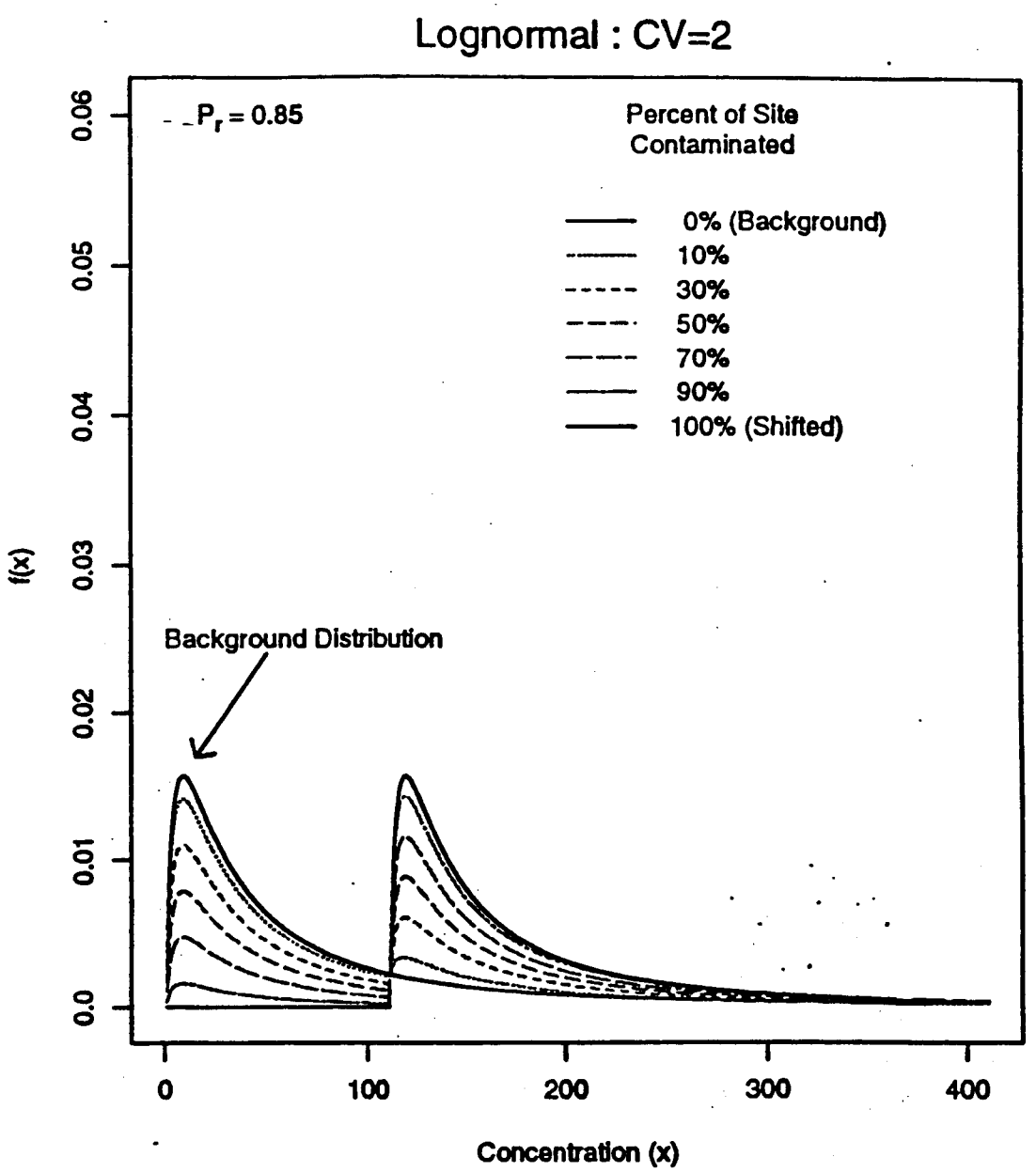


Figure 3.7: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for $\epsilon = 0.1, 0.3, 0.5, 0.7$, and 0.9. The background distribution is lognormal with CV=2.0 and the shifted distribution is characterized by $P_r = 0.85$, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

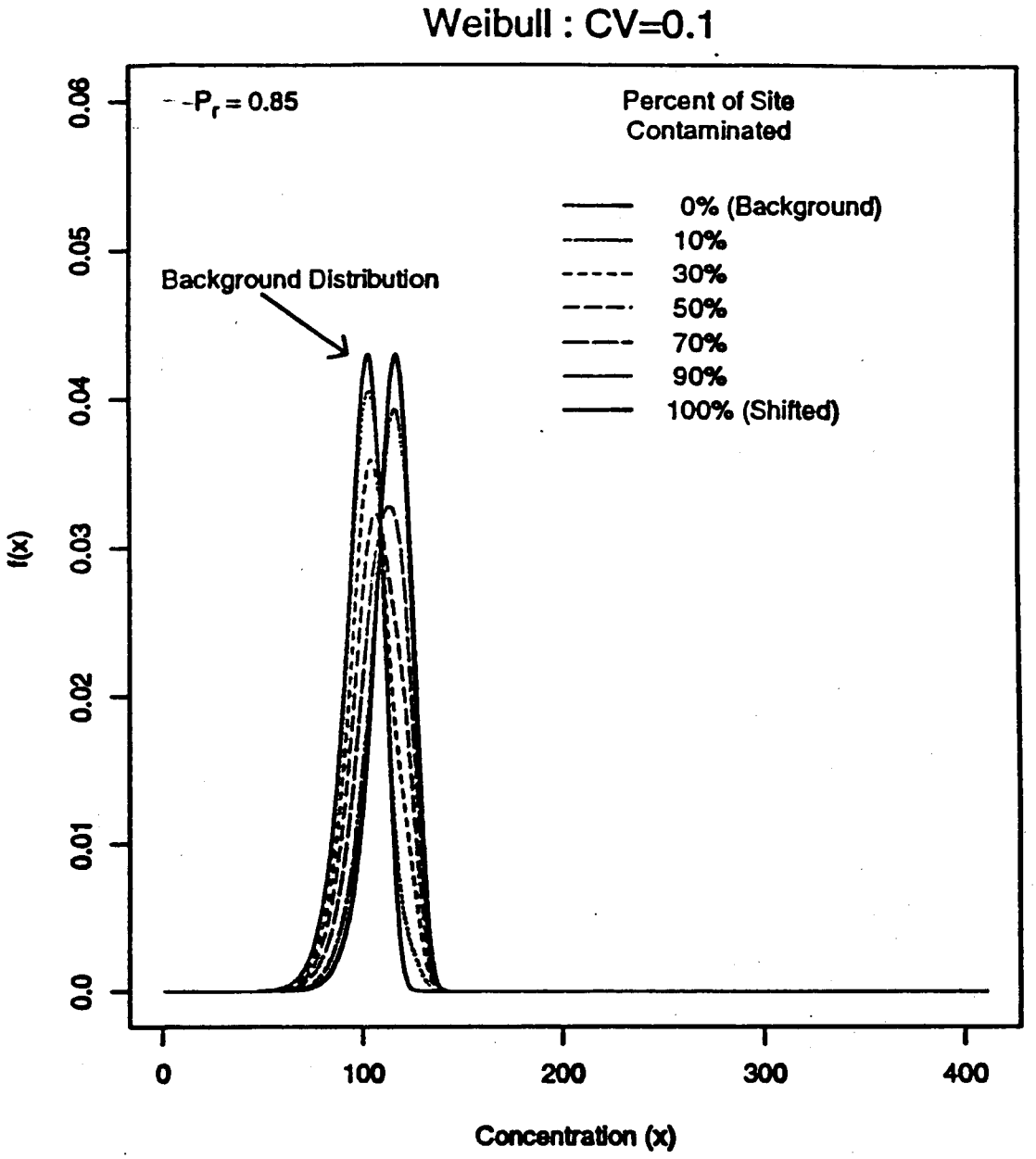


Figure 3.8: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for $\epsilon = 0.1, 0.3, 0.5, 0.7$, and 0.9. The background distribution is Weibull with CV=0.1 and the shifted distribution is characterized by $P_r = 0.85$, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

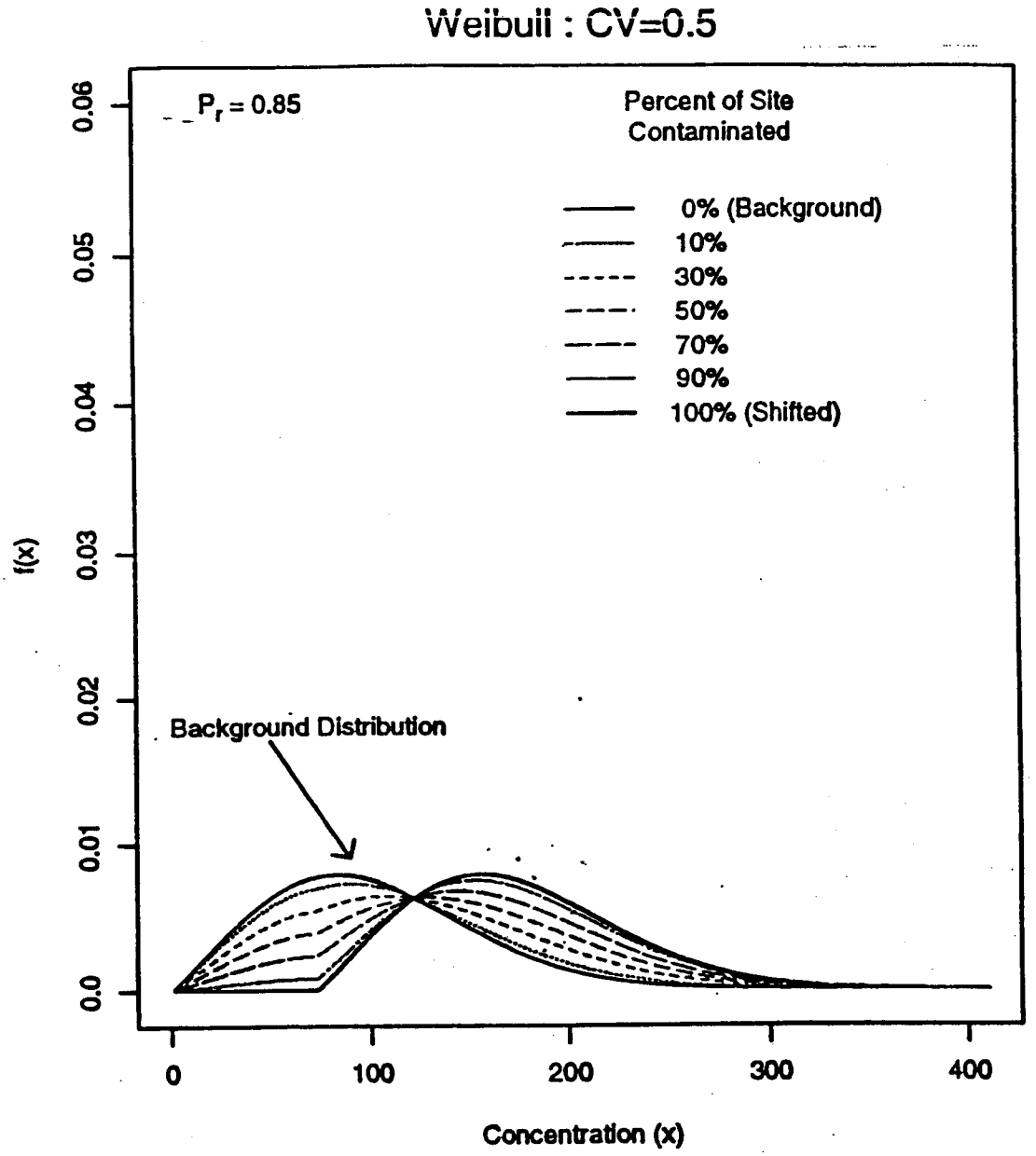


Figure 3.9: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for $\epsilon = 0.1, 0.3, 0.5, 0.7$, and 0.9. The background distribution is Weibull with CV=0.5 and the shifted distribution is characterized by $P_r = 0.85$, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

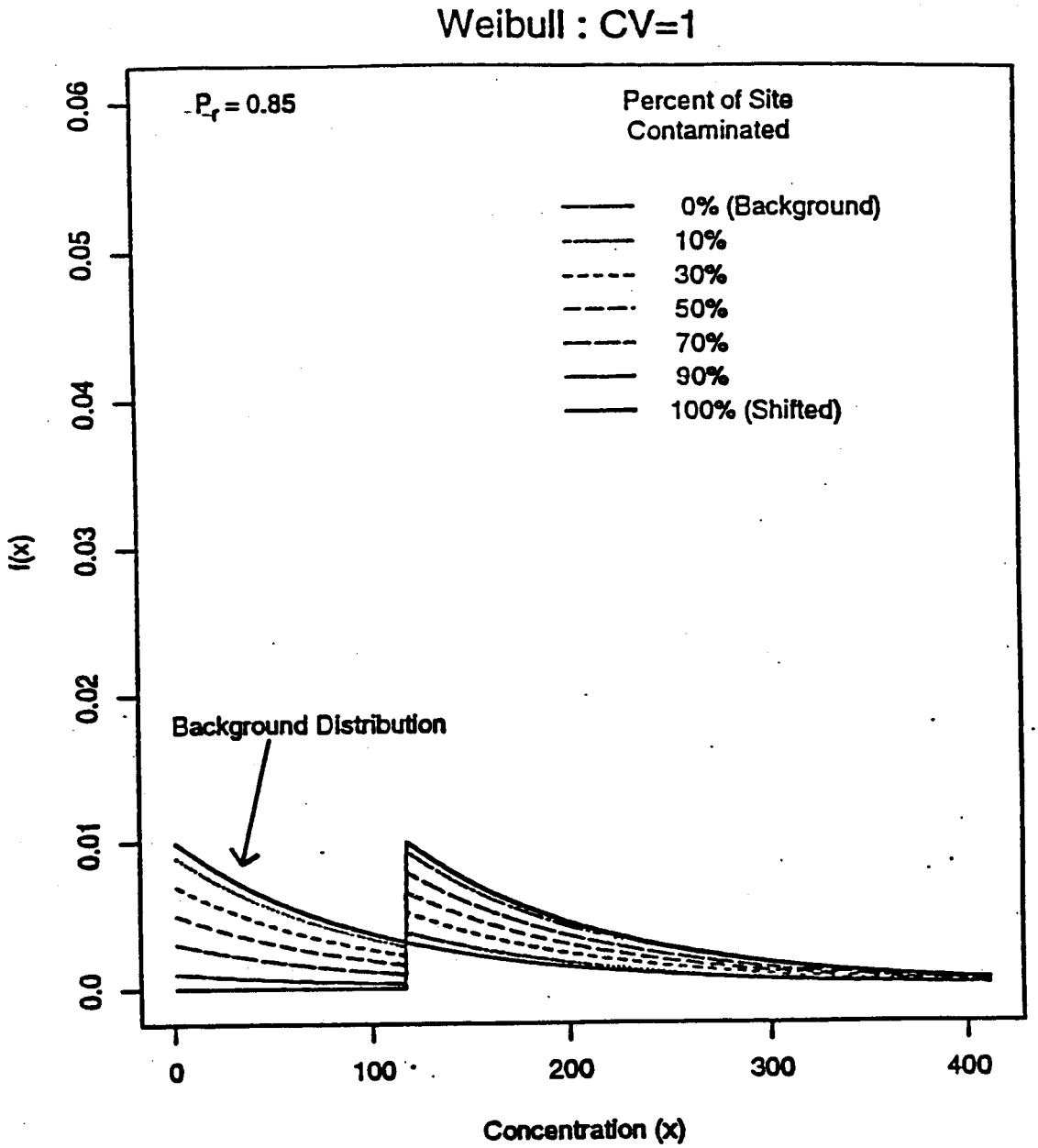


Figure 3.10: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for ϵ = 0.1, 0.3, 0.5, 0.7, and 0.9. The background distribution is Weibull with CV=1.0 and the shifted distribution is characterized by P_r = 0.85, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

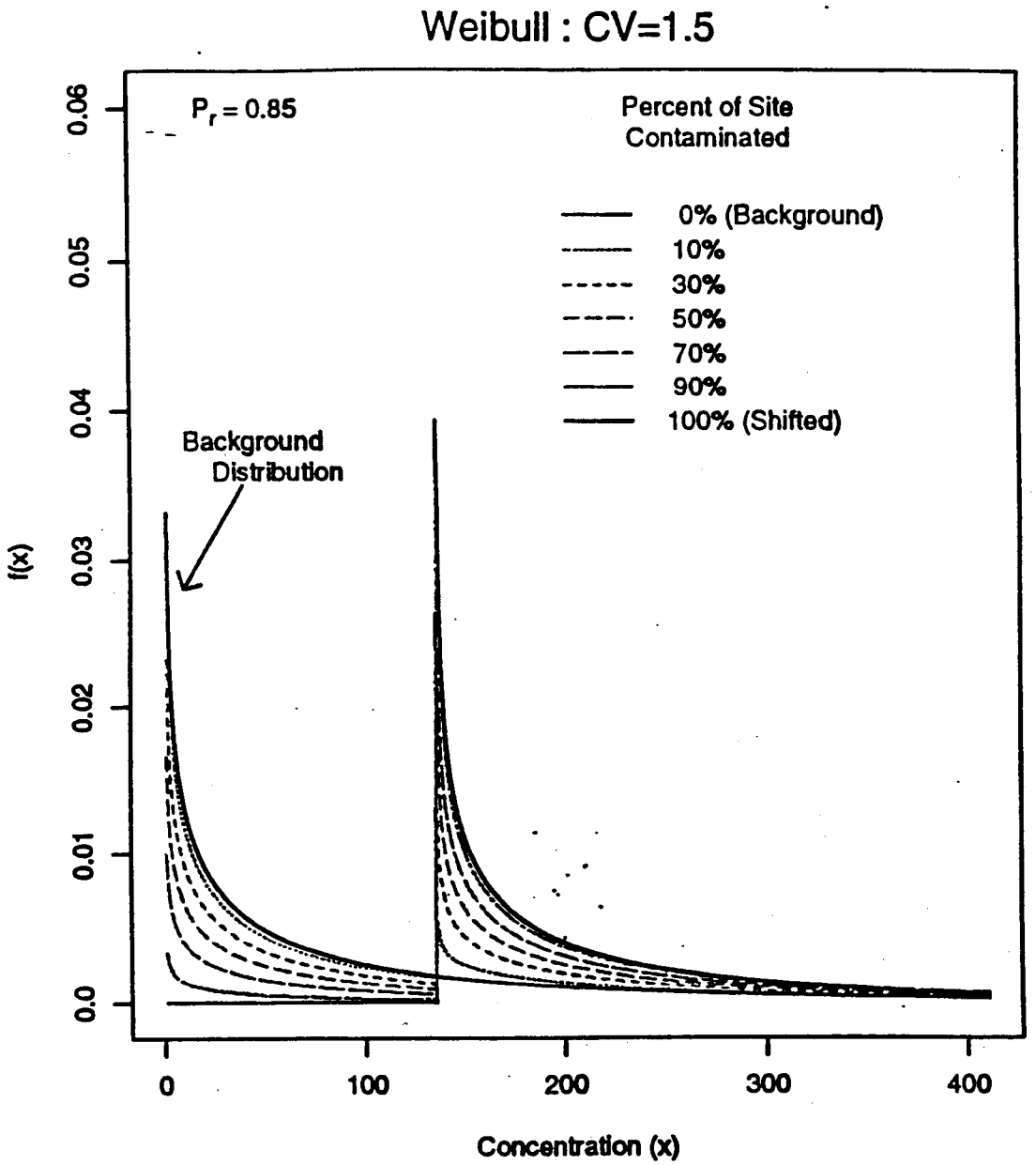


Figure 3.11: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for $\epsilon = 0.1, 0.3, 0.5, 0.7$, and 0.9. The background distribution is Weibull with CV=1.5 and the shifted distribution is characterized by $P_r = 0.85$, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

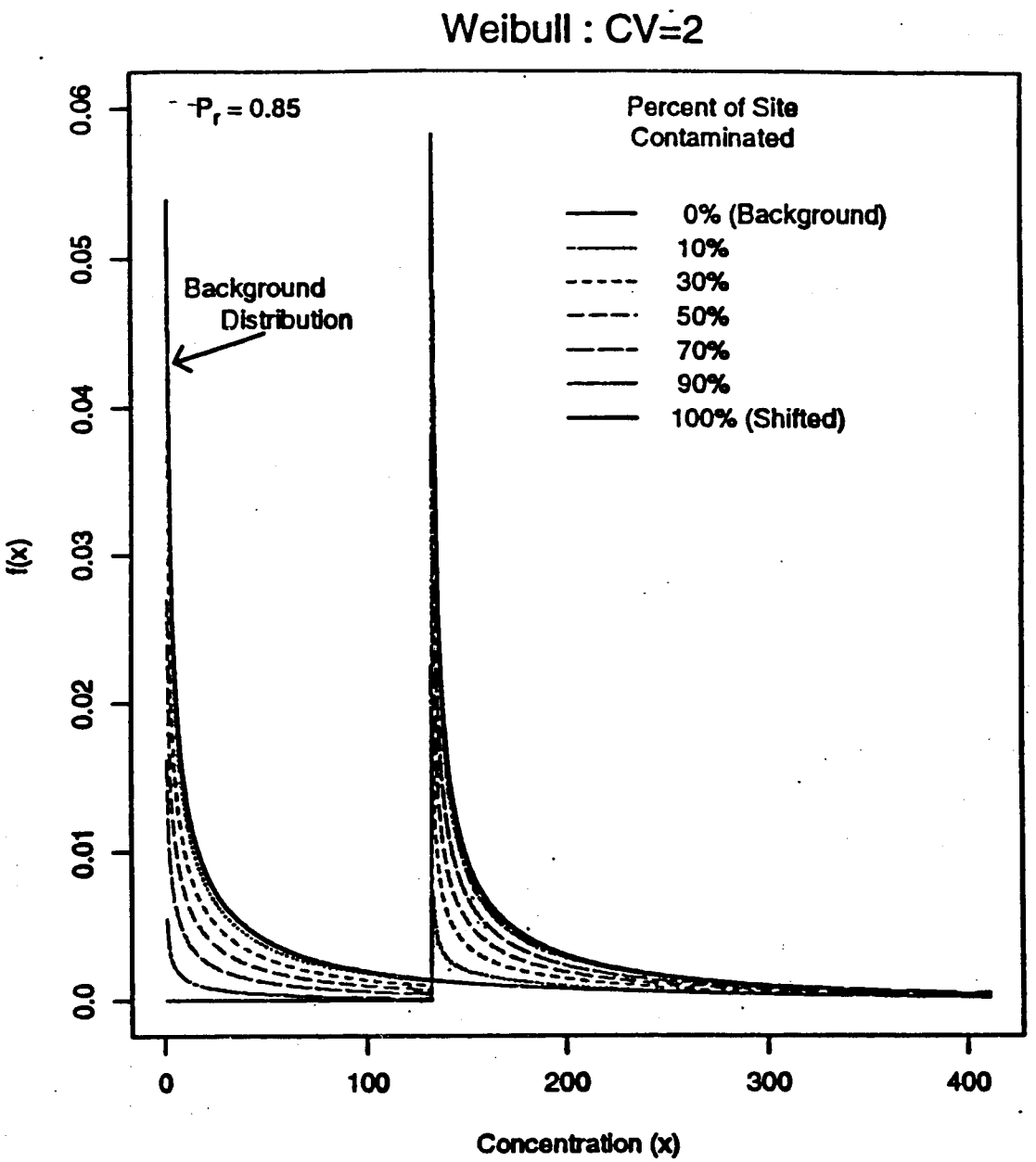


Figure 3.12: Illustration of five site distributions obtained using the mixture model (Equation 2) to mix 100ϵ% of the shifted distribution (solid curve to the right) with 100(1-ϵ)% of the background distribution (solid curve to the left) for $\epsilon = 0.1, 0.3, 0.5, 0.7,$ and 0.9. The background distribution is Weibull with CV=2.0 and the shifted distribution is characterized by $P_r = 0.85$, i.e., the probability is 0.85 that a random measurement from the shifted distribution is greater than a random measurement from the background distribution.

4. RESULTS

This section discusses and illustrates the results of the Monte Carlo simulation study. Our focus is on comparing the power and Type I error rates of the tests outlined in Section 2 for the various background distributions and site contamination scenarios in Table 3.1. The basic tool used to make the comparisons is a set of graphs of the power. As there are 1200 such graphs, only a small percentage of them can be shown here. However, the graphs shown capture the essential information. A copy of the complete set of graphs on microfiche can be obtained from the first author of this report. The graphs show results for the State test and the Wilcoxon Rank Sum test, Quantile test, Savage test and t test. Graphs for the median test, Slippage test and Satterthwaite's t test are not shown because, in most cases, their power is less than that of the Wilcoxon, Quantile, and t tests, respectively. In following discussions, the Wilcoxon Rank Sum test is frequently denoted simply as the Wilcoxon test.

In Section 4.1, test performance is considered for the case of $\epsilon = 1$, i.e., for when 100 percent of the site is contaminated. In other words, the specified background and site distributions are identical in shape, but the site distribution is simply shifted to the right of the background distribution. For this case, each site data set was constructed by randomly drawing all site measurements from the shifted distribution, F_c. Section 4.2 gives results for when only 10 percent of the site is contaminated above background. That is, each site data set was constructed by randomly drawing 10 percent of the site measurements from the shifted distribution and 90 percent from the background distribution. Section 4.3 summarizes the Type I error rates of the tests for the various contamination scenarios. In Section 4.4, the contamination scenarios for which the state test performs better than other tests studied here are summarized.

Some remarks on how to read and interpret the power curves may be helpful. The y axis of each graph is the probability that a test indicates the site is contaminated, i.e., the probability that the test rejects H_o and accepts H_a [Equation (1)]. The x axis gives values of P_r ranging from 0.5 (no

contamination above background) to 0.999 (contamination such that the site and background distributions overlap only a negligible amount). The power curve for each test is labeled to correspond to the number assigned to each test at the bottom of the figure. The Type I error rates for each test are plotted on the graph at the value $P_r = 0.5$ and are also listed at the bottom of the figure. The contamination scenario parameters (type of background distribution, ϵ, CV, and number of background and site measurements) for which the results were obtained are also given at the bottom of each figure.

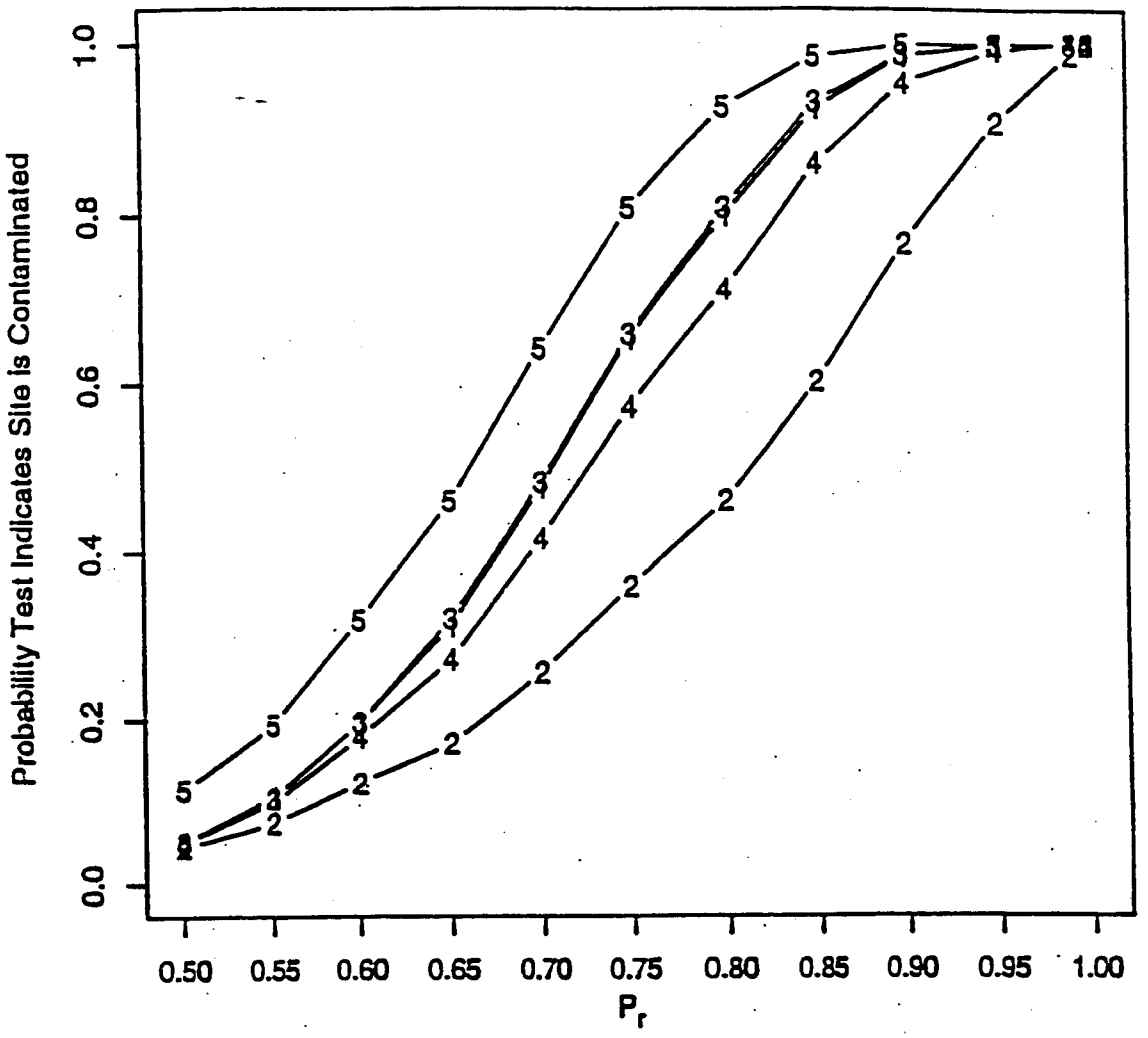
When looking at the power curves, a test is performing well if the curve starts at a y-axis value of 0.05 when $P_r = 0.50$ (which indicates the test has achieved the target Type I error rate of 0.05) and then increases rapidly as P_r increases. The preferred test is the one which has an achieved Type I error rate of 0.05 and which reaches a power of 1 soonest as P_r increases. For any specified value of P_r, one can quickly determine from the graph which test has the most power. The odd numbered figures (Figure 4.1, 4.3, etc) are for a lognormal distributions. The even numbered figures are for Weibull distributions.

4.1 Pervasive Contamination ($\epsilon = 1$)

Recall that for the case of pervasive contamination (for which $\epsilon = 1$) the site and background distributions have the same shape, but the site distribution is shifted to the right of the background distribution. Subsection 4.1.1 presents results obtained when the background and site distributions are only slightly skewed (CV = 0.1), such that they appear to be almost symmetrical. Subsection 4.1.2 considers the case of highly skewed distributions (CV = 1 and 2).

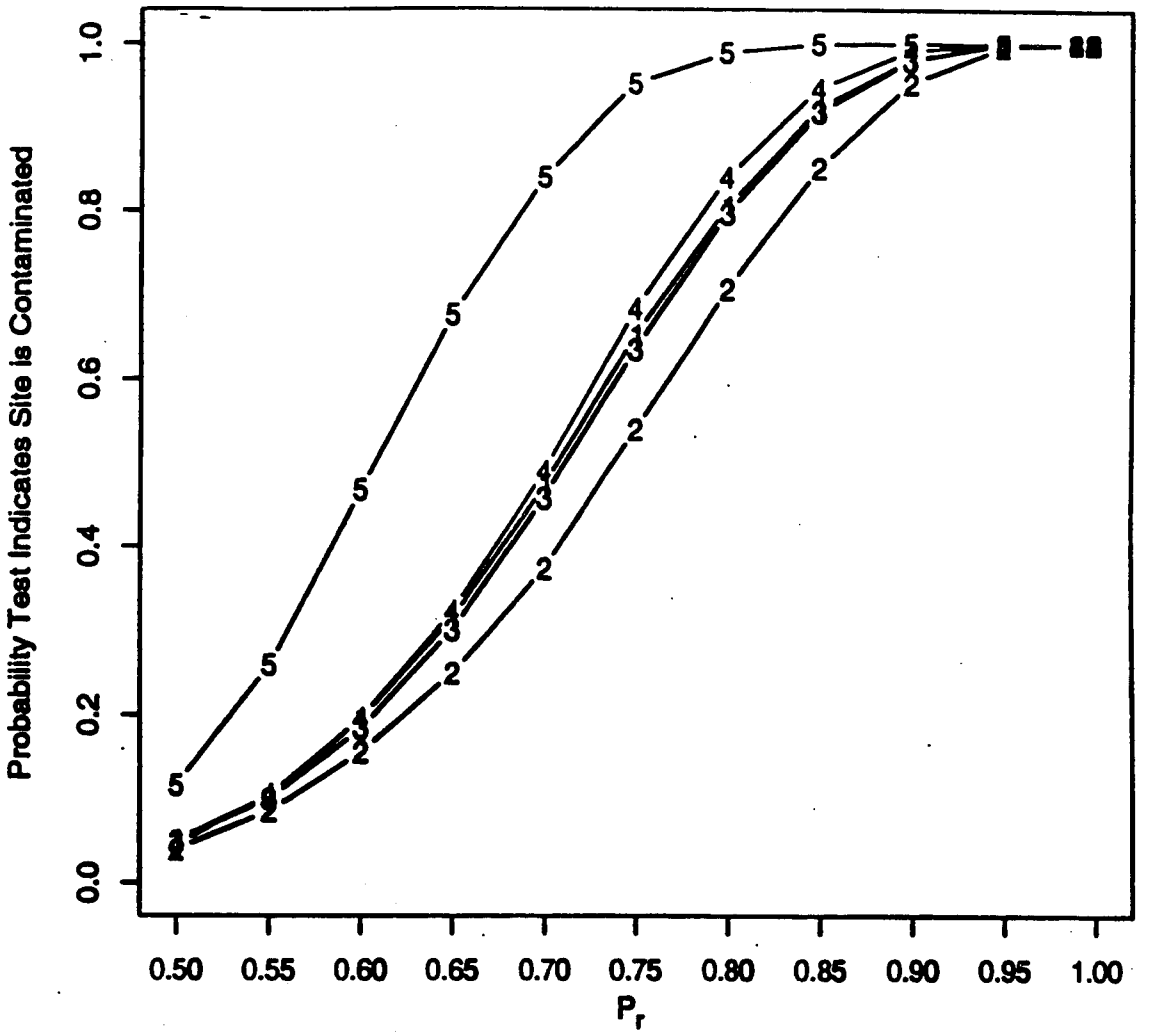
4.1.1 Slightly Skewed Distributions (CV = 0.1)

For the case of pervasive contamination, Figure's 4.1 and 4.2 give results for slightly skewed (CV = 0.1) lognormal and Weibull background distributions, respectively, for when only 10 background and 10 site measurements are available. Note first that the Type I error level for the



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon	0.0526
CV	= 0.10	2	Quantile	0.0451
ϵ	= 1	3	t-Test	0.0508
Samples (backgd.)	= 10	4	Savage	0.0498
Samples (site)	= 10	5	State	0.1138

Figure 4.1: Power curves for five test procedures for a lognormal background distribution, CV = 0.10, $\epsilon = 1$, 10 background samples, and 10 site samples.



PARAMETERS		TEST	TYPE I ERROR
Distribution	= Weibull	1 Wilcoxon	0.0511
CV	= 0.10	2 Quantile	0.0401
ϵ	= 1	3 t-Test	0.0494
Samples (backgd.)	= 10	4 Savage	0.0453
Samples (site)	= 10	5 State	0.1168

Figure 4.2: Power curves for five test procedures for a Weibull background distribution, CV = 0.10, $\epsilon = 1$, 10 background samples, and 10 site samples.

State test is 0.11 rather than the target value of 0.05. In other words, the probability is 0.11 (11 chances out of 100) that the State test would incorrectly indicate that the site is contaminated. The other tests do not have this problem since their Type I error levels cluster very closely around the target value of 0.05.

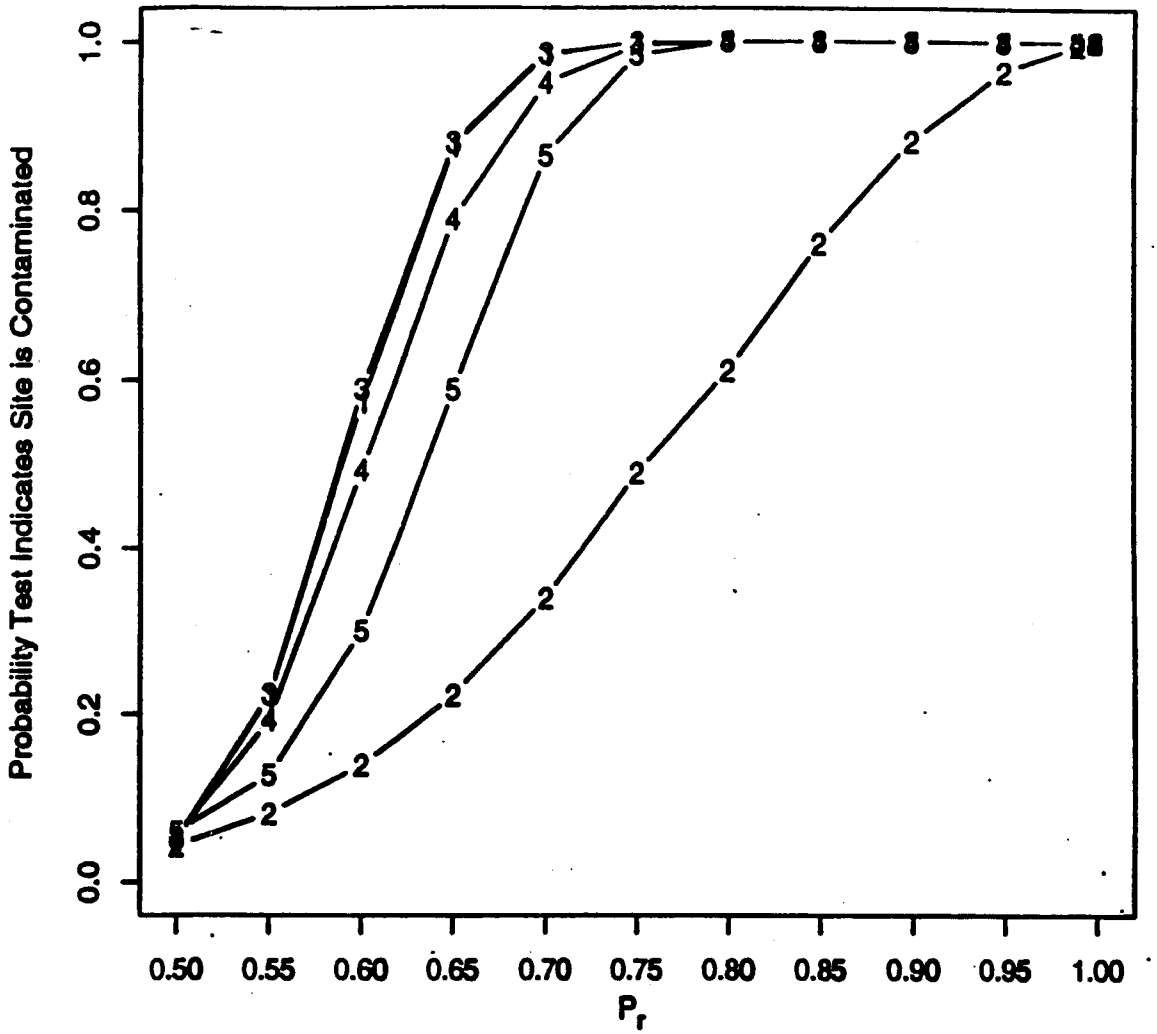
With regard to power, the State test has the most power to detect that the site distribution is shifted to the right of the background distribution. But this power advantage is "purchased" at the expense of an elevated Type I error level. The Wilcoxon and t tests have equal power that is somewhat less than that of the State test. The Savage test is slightly less powerful than the Wilcoxon and t tests when the background distribution is lognormal, and slightly more power than those tests when the background distribution is Weibull. The Quantile test has the least power for both background distributions. Comparing Figure 4.2 with Figure 4.1 indicates that the power of the State, Savage, and Quantile tests is larger for the Weibull distribution than for the lognormal distribution. The Wilcoxon and t tests have about the same power for either distribution.

Summary for Small Sample Case

- If the background and site distributions are only slightly skewed, the site distribution is simply a shifted version of the background distribution, and only a few background and site samples can be collected, then the State test is the preferred test unless the elevated Type I error rate of the State test is not acceptable (Figures 4.1 and 4.2).

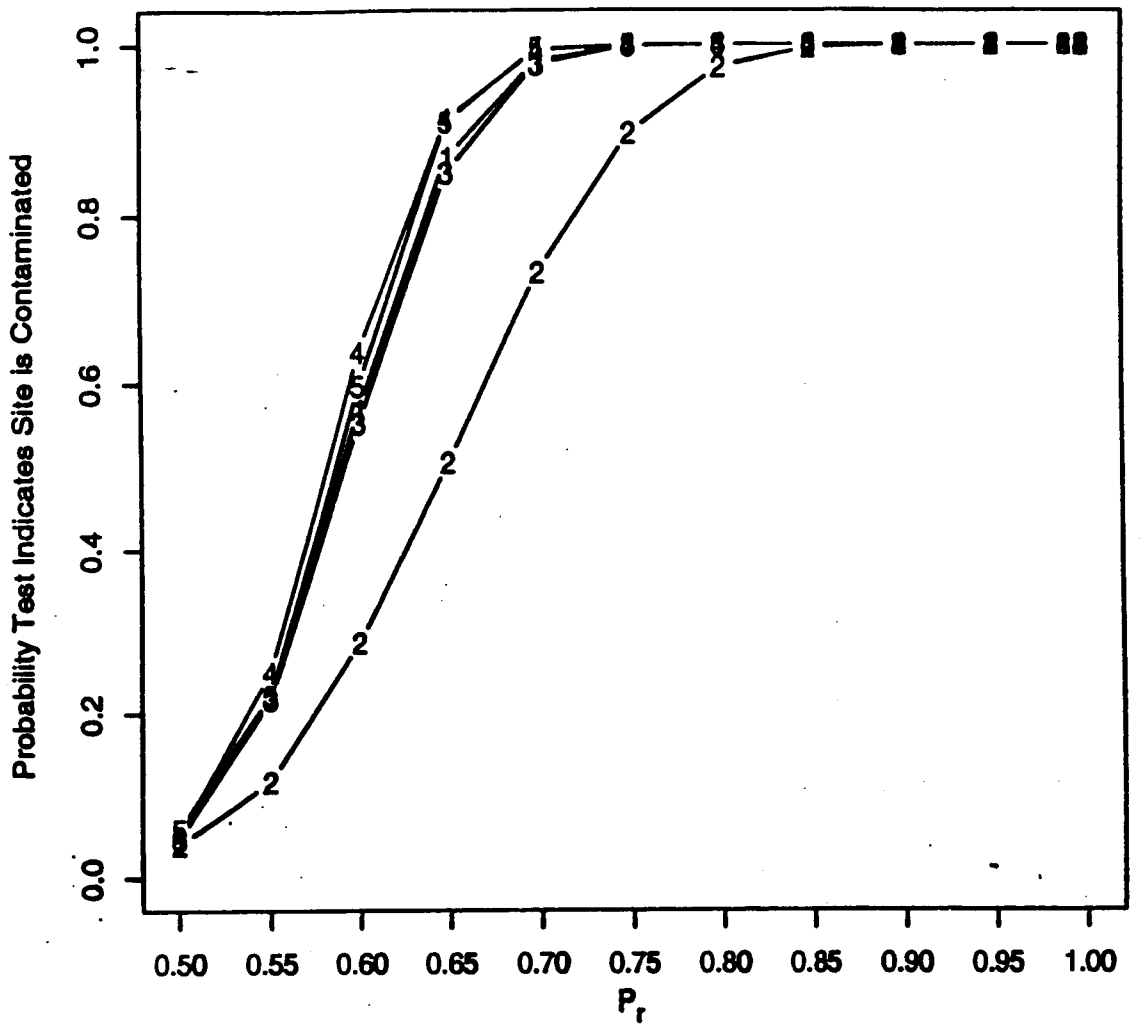
- If the elevated Type I rate of the State test is unacceptable, then among the tests studied in this report, the Wilcoxon, t, or Savage tests are preferred.

Figures 4.3 and 4.4 are for the same contamination scenario as Figures 4.1 and 4.2, except that 60 background and 50 site samples are collected (rather than 10 of each). For this situation the Type I error rate of the



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon	0.0491
CV	= 0.10	2	Quantile	0.0428
ϵ	= 1	3	t-Test	0.0513
Samples (backgd.)	= 60	4	Savage	0.0525
Samples (site)	= 50	5	State	0.0591

Figure 4.3: Power curves for five test procedures for a lognormal background distribution, CV = 0.10, $\epsilon = 1$, 60 background samples, and 50 site samples.



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0488
CV	= 0.10	2	Quantile	0.0399
ε	= 1	3	t-Test	0.0465
Samples (backgd.)	= 60	4	Savage	0.0499
Samples (site)	= 50	5	State	0.0587

Figure 4.4: Power curves for five test procedures for a Weibull background distribution, CV = 0.10, $\epsilon = 1$, 60 background samples, and 50 site samples.

State test is 0.06, which is much closer to the target 0.05. Hence, taking more samples has reduced the size of the elevated Type I error rate of the State test. Also, as expected, increasing the sample sizes increases the power of all tests. However, now the Wilcoxon, t and Savage tests have more power than the State test.

Summary for Large Sample Case

- For slightly skewed lognormal background and site distributions of the same shape, if a relatively large number of samples can be collected, either the Wilcoxon or t test should be used (Figure 4.3).

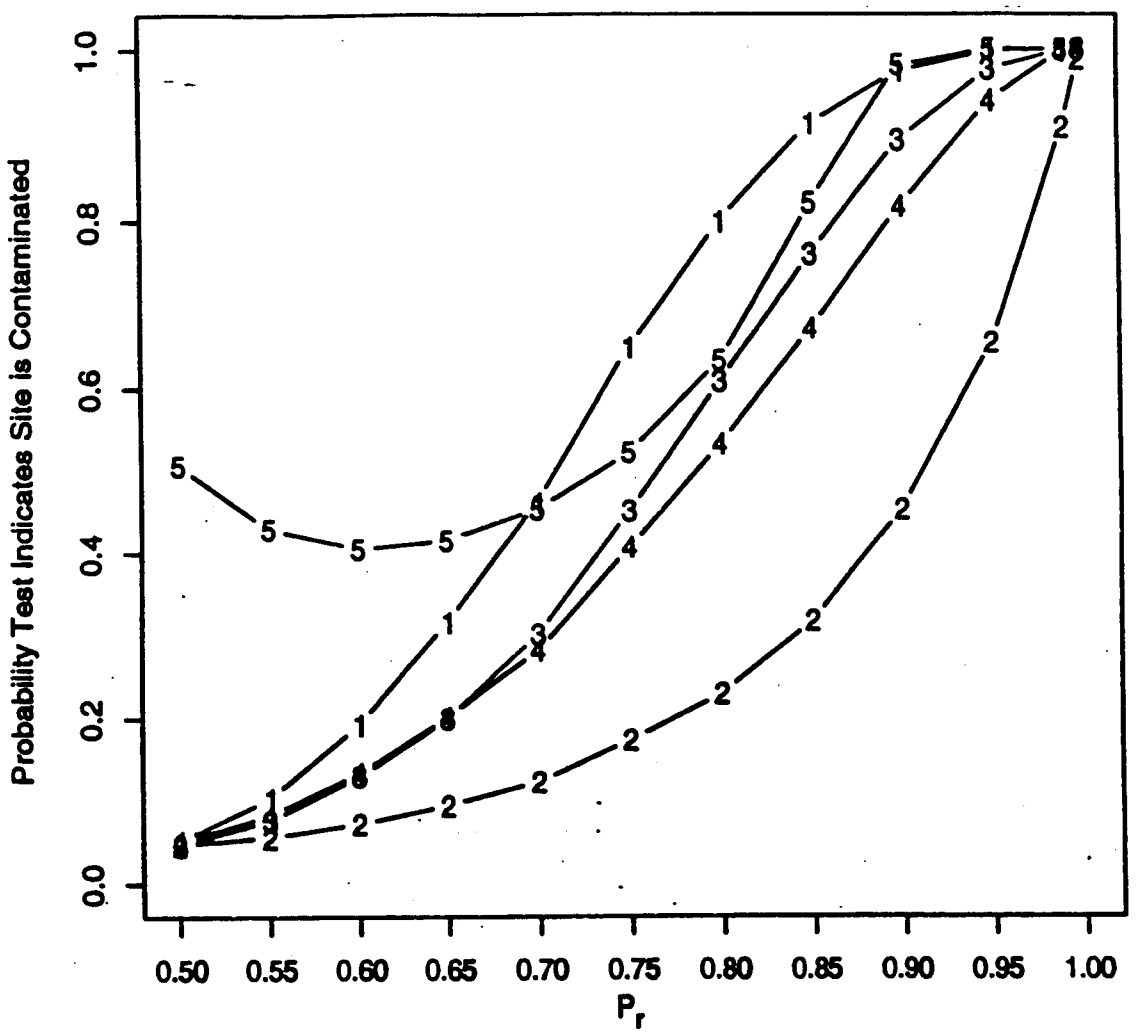
- For Weibull distributions and large sample sizes, the State, Wilcoxon, t and Savage tests have almost identical power so that any one of them could be used (Figure 4.4)

- Since in practice it is difficult to determine whether the distributions are lognormal or Weibull, the Wilcoxon test is preferred when the number of samples is relatively large, since it performs best or very nearly so, for either distribution.

4.1.2 Highly Skewed Distributions (CV = 1 and 2)

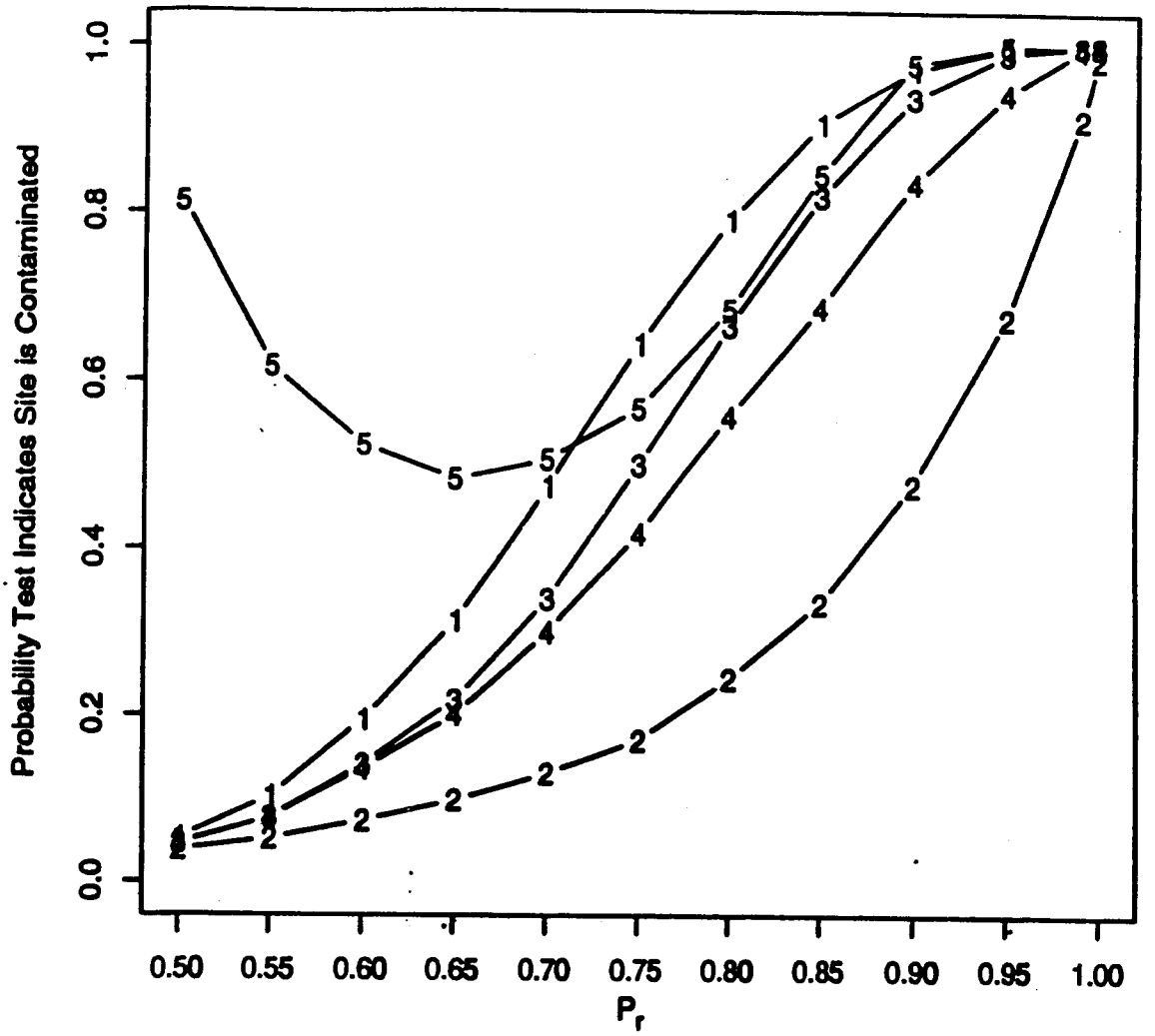
Figures 4.5-4.8 consider the same pervasive contamination scenarios as Figures 4.1-4.4 except that the background and site distributions are more highly skewed, i.e., CV = 1 rather than CV = 0.1. Figures 4.9-4.12 are for an even more highly skewed case in which CV = 2. As the conclusions from Figures 4.5-4.8 are qualitatively the same as those for Figures 4.9-4.12, we shall discuss mainly the latter group of figures.

Figures 4.9 and 4.10 indicate that the State test has an extremely high probability of incorrectly indicating the site is contaminated when the lognormal and Weibull distributions are highly skewed (CV = 2), $\epsilon = 1$, and only 10 background and 10 site measurements are available. Specifically, for the lognormal distribution the Type I error rate of the State test was 0.82



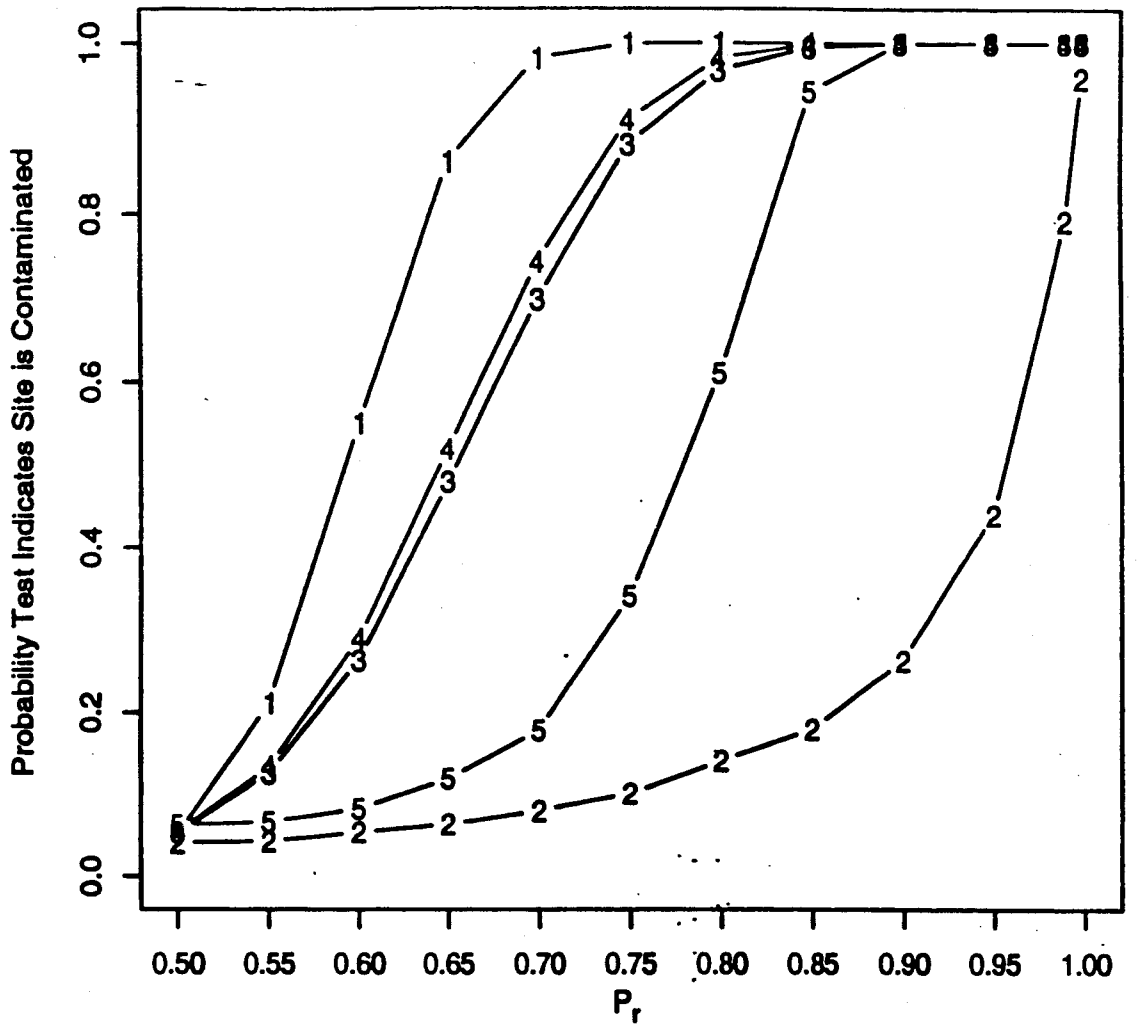
PARAMETERS		TEST	TYPE I ERROR
Distribution	= Lognormal	1 Wilcoxon	0.0513
CV	= 1.00	2 Quantile	0.0466
ϵ	= 1	3 t-Test	0.0458
Samples (backgd.)	= 10	4 Savage	0.0511
Samples (site)	= 10	5 State	0.5063

Figure 4.5: Power curves for five test procedures for a lognormal background distribution, CV = 1.00, $\epsilon = 1$, 10 background samples, and 10 site samples.



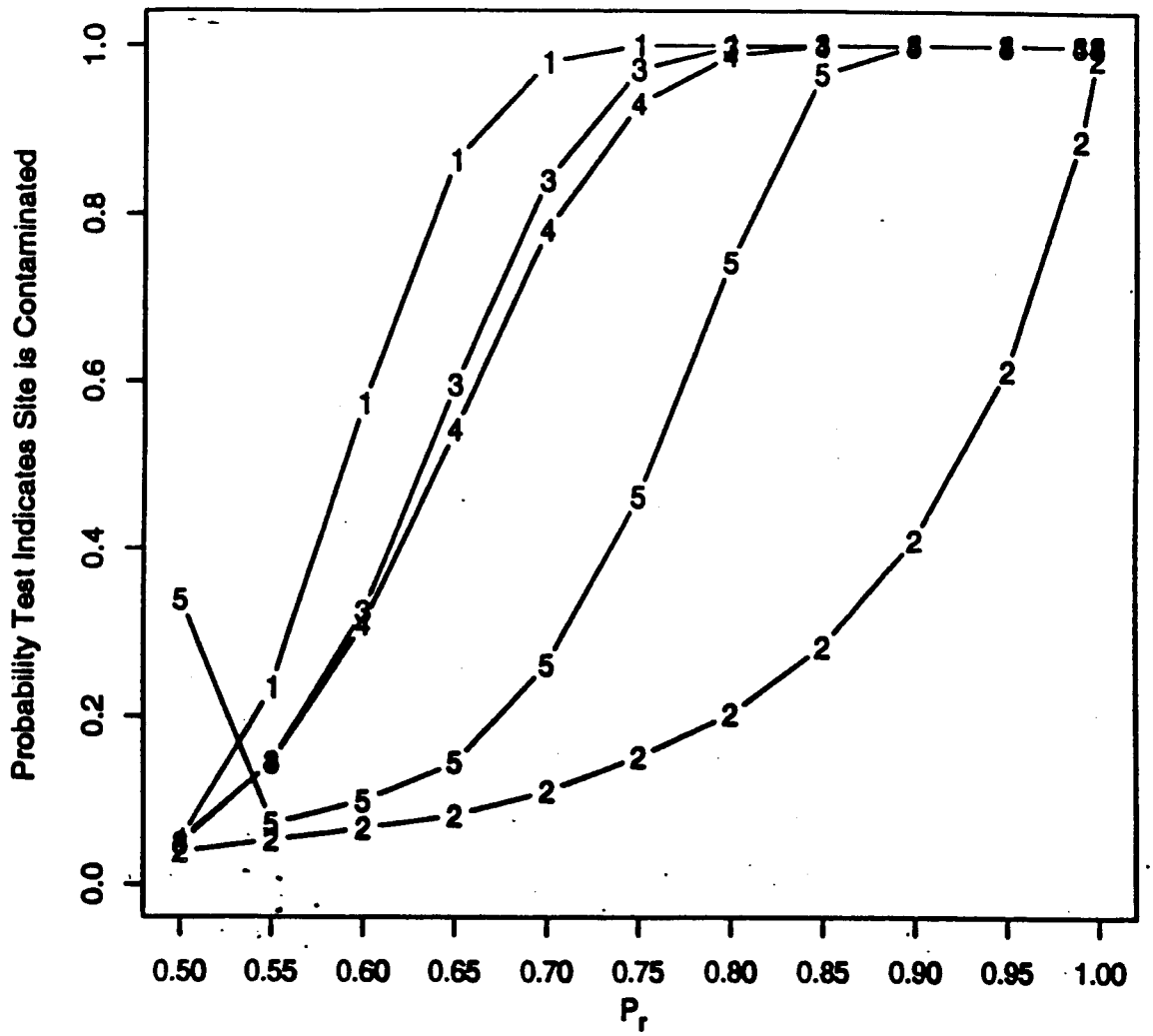
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0532
CV	= 1.00	2	Quantile	0.0391
ε	= 1	3	t-Test	0.0464
Samples (backgd.)	= 10	4	Savage	0.0470
Samples (site)	= 10	5	State	0.8123

Figure 4.6: Power curves for five test procedures for a Weibull background distribution, CV = 1.00, $\epsilon = 1$, 10 background samples, and 10 site samples.



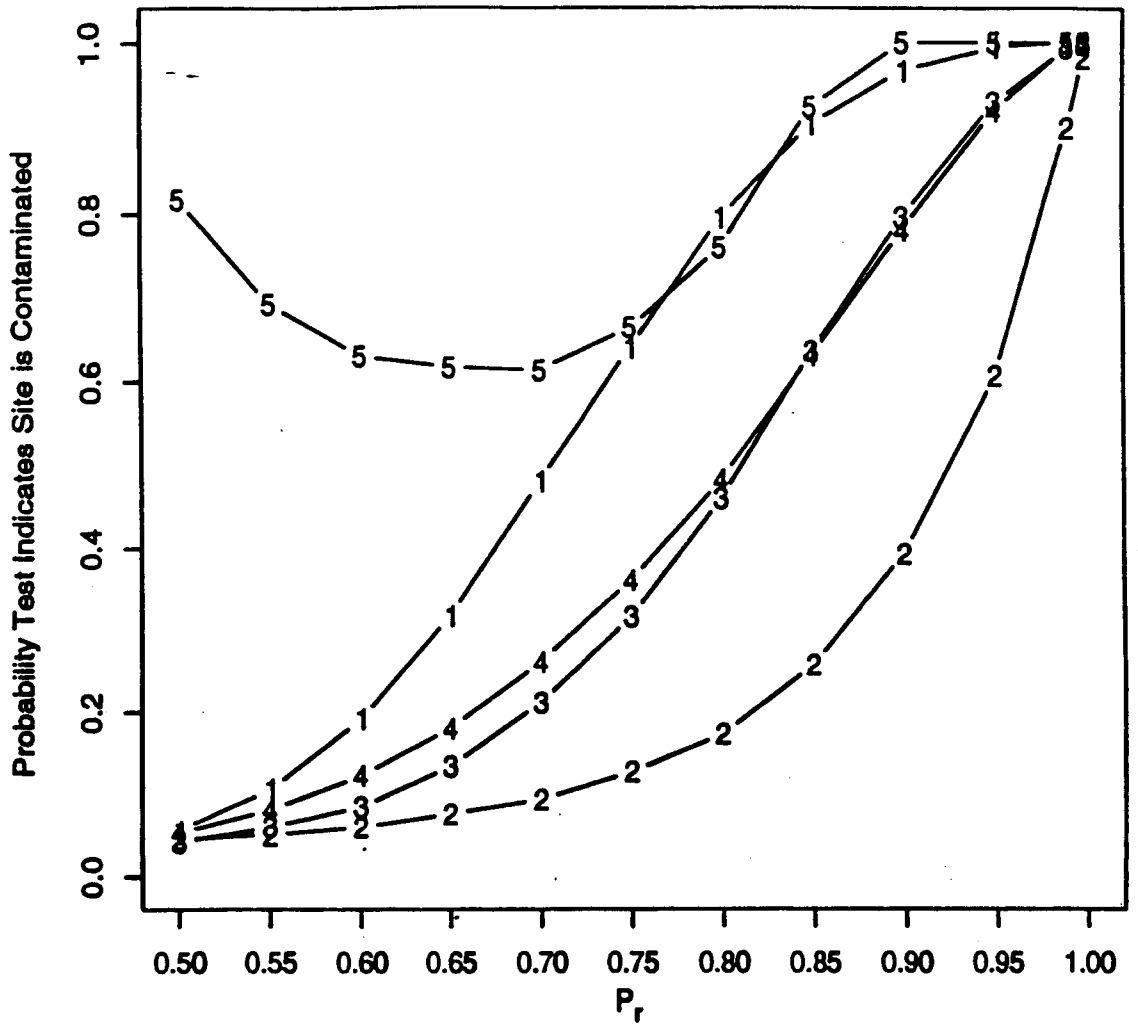
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon	0.0516
CV	= 1.00	2	Quantile	0.0409
ε	= 1	3	t-Test	0.0534
Samples (backgd.)	= 60	4	Savage	0.0541
Samples (site)	= 50	5	State	0.0625

Figure 4.7: Power curves for five test procedures for a lognormal background distribution, CV = 1.00, $\epsilon = 1$, 60 background samples, and 50 site samples.



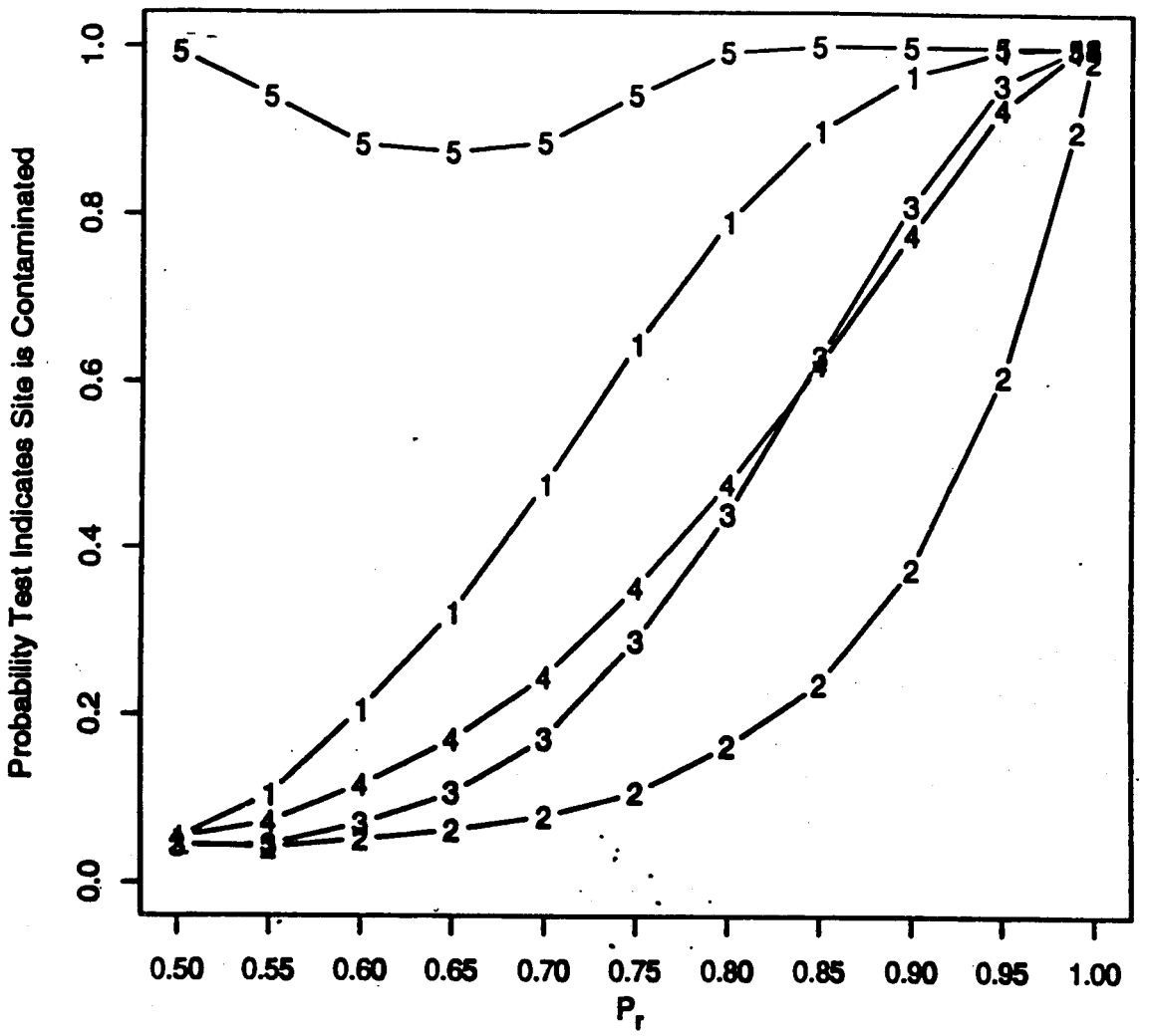
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0525
CV	= 1.00	2	Quantile	0.0388
ϵ	= 1	3	t-Test	0.0492
Samples (backgd.)	= 60	4	Savage	0.0514
Samples (site)	= 50	5	State	0.3385

Figure 4.8: Power curves for five test procedures for a Weibull background distribution, CV = 1.00, $\epsilon = 1$, 60 background samples, and 50 site samples.



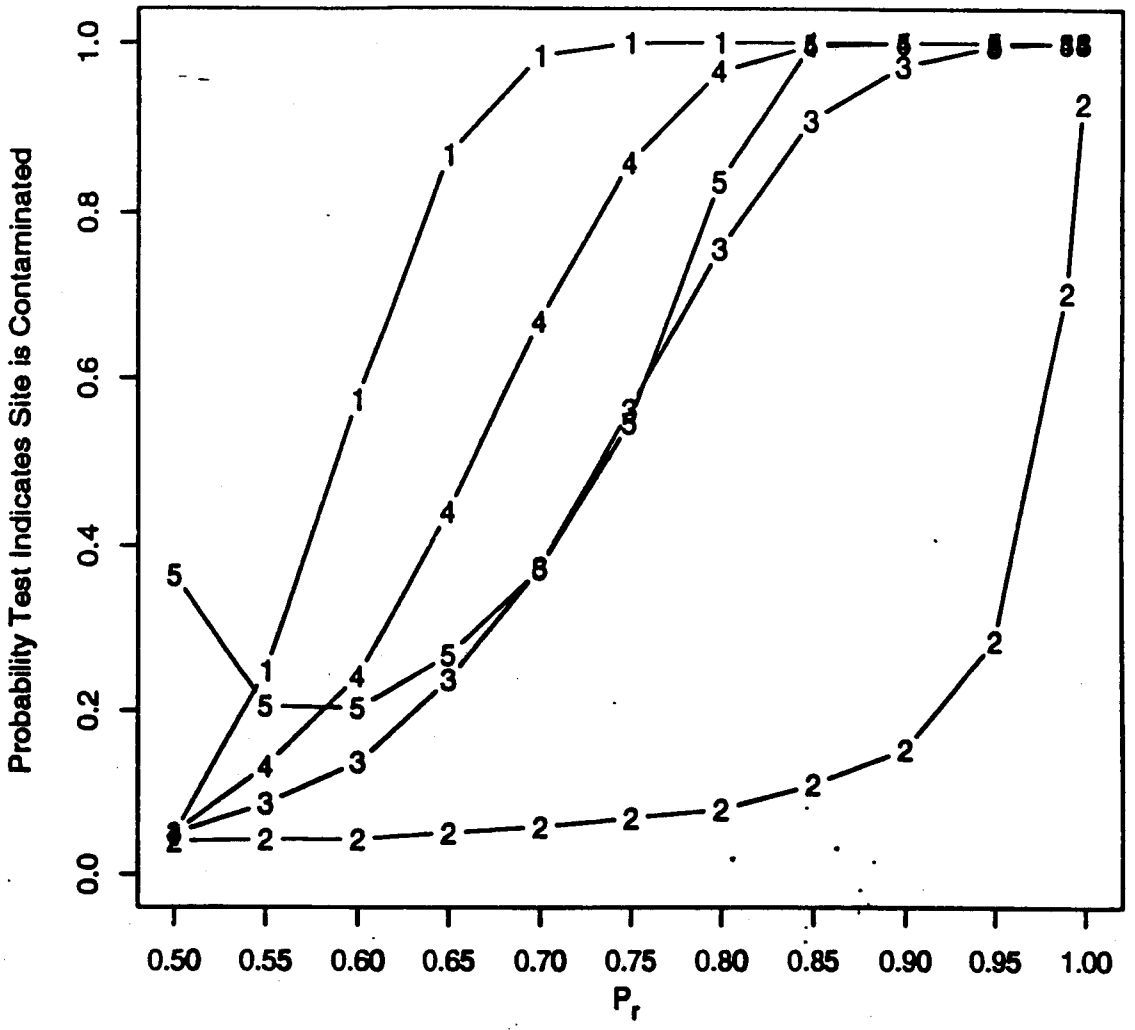
PARAMETERS		TEST	TYPE I ERROR
Distribution	= Lognormal	1 Wilcoxon	0.0555
CV	= 2.00	2 Quantile	0.0444
ε	= 1	3 t-Test	0.0410
Samples (backgd.)	= 10	4 Savage	0.0520
Samples (site)	= 10	5 State	0.8169

Figure 4.9: Power curves for five test procedures for a lognormal background distribution, CV = 2.00, $\epsilon = 1$, 10 background samples, and 10 site samples.



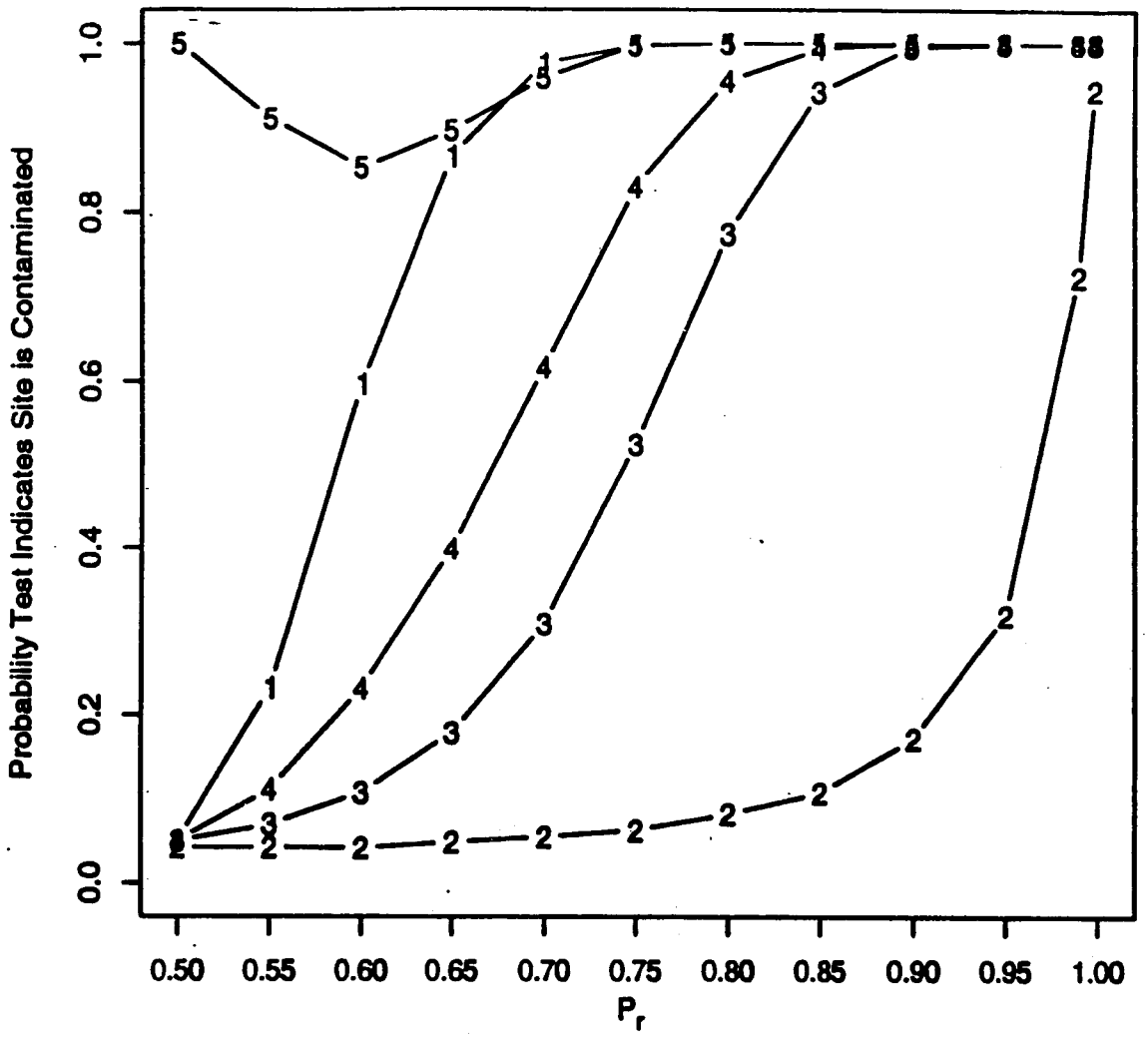
PARAMETERS		TEST	TYPE I ERROR
Distribution	= Weibull	1 Wilcoxon	0.0544
CV	= 2.00	2 Quantile	0.0469
ϵ	= 1	3 t-Test	0.0438
Samples (backgd.)	= 10	4 Savage	0.0544
Samples (site)	= 10	5 State	0.9921

Figure 4.10: Power curves for five test procedures for a Weibull background distribution, CV = 2.00, $\epsilon = 1$, 10 background samples, and 10 site samples.



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon	0.0479
CV	= 2.00	2	Quantile	0.0403
ε	= 1	3	t-Test	0.0502
Samples (backgd.)	= 60	4	Savage	0.0509
Samples (site)	= 50	5	State	0.3643

Figure 4.11: Power curves for five test procedures for a lognormal background distribution, CV = 2.00, $\epsilon = 1$, 60 background samples, and 50 site samples.



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0509
CV	= 2.00	2	Quantile	0.0427
ϵ	= 1	3	t-Test	0.0504
Samples (backgd.)	= 60	4	Savage	0.0509
Samples (site)	= 50	5	State	0.9997

Figure 4.12: Power curves for five test procedures for a Weibull background distribution, CV = 02.0, $\epsilon = 1$, 60 background samples, and 50 site samples.

rather than the target 0.05 (Figure 4.9). That is, the probability is 0.82 (82 times out of 100) that the State test will incorrectly indicate the site is contaminated. This effect was even more extreme for the Weibull distribution, the rate being 0.99 (Figure 4.10). This value implies virtual certainty that the State test will indicate contamination greater than background is present when that is not the case. The degree that the Type I error rate increases as distributions become more skewed (CV increases) when 10 background and 10 site measurements are used can be seen by comparing Figures 4.5 and 4.6 (for which CV = 1) to Figures 4.9 and 4.10 (for which CV = 2). The Type I error rates for the State test are 0.51 and 0.82 when CV = 1 and 2, respectively, for the lognormal distribution. The rates are 0.81 and 0.99 for CV = 1 and 2, respectively, for the Weibull distribution.

Figures 4.9 and 4.10 indicate the superior power of the State test when CV = 2. However, this must be balanced against the extremely elevated Type I error levels of the test. Also, when CV = 1 (Figures 4.5 and 4.6), the State test does not continue to have superior power for all values of P_r. Rather, the Wilcoxon test has equal or more power than the State test when $P_r > 0.7$. It should also be noted from Figures 4.5, 4.6, 4.9 and 4.10 that the t test has considerably less power than the Wilcoxon test when distributions are highly skewed. This difference in power increases as the skewness increases. Overall, the power results suggest that the Wilcoxon test is the preferred test, assuming the State test is not used because of its elevated Type I error rates.

Summary for Small Sample Case

- If the background and site distributions are highly skewed, the site distribution is simply a shifted version of the background distribution, and only a few background and site samples can be collected, then the State test will almost always indicate the site is contaminated greater than background when in fact there are no differences. Clearly, this performance is unacceptable.

- Assuming the State test is not used because of its highly elevated Type I error rate, the Wilcoxon test is preferred.

- The t test should not be used in place of the Wilcoxon test when distributions are highly skewed. These skewed distributions are a severe violation of the normality assumption underlying that test, which results in lower power for the t test than for the Wilcoxon test.

For the large sample case (60 background and 50 site samples) when CV = 1, the Type I error rate for the State test almost achieves the target value (0.06 instead of 0.05) when the data are lognormally distributed (Figure 4.7). For Weibull distributions the Type I error rate is 0.34 (Figure 4.8). When CV = 2 the Type I error rate for the State test is 0.36 for the lognormal distribution (Figure 4.11) and 1.0 for the Weibull distribution (Figure 4.12).

For lognormal and Weibull distributions with CV = 1, the power of the State test for the large sample-size case is much less than that of the Wilcoxon, Savage and t tests (Figures 4.7 and 4.8). A similar result is obtained for the lognormal distribution with CV = 2 (Figure 4.11). However, the power of the State test for the Weibull case with CV = 2 (Figure 4.12) is still superior or equal to that of the Wilcoxon test, but at the expense of the virtual certainty of the State test making a Type I error.

<u>Summary for Large Sample Case</u>

- For the pervasive contamination case and highly skewed lognormal background and site distributions of the same shape, if a relatively large number of samples can be collected, the Wilcoxon test is preferred because it has the best power among all tests that achieve the target Type I error of 0.05. For the Weibull case the Wilcoxon test is also preferred, assuming the State test is considered to be unacceptable because it virtually always indicates the site is contaminated when that is not true.

4.1.3 Tandem Testing

Figures 4.13-4.24 give test performance information for three tandem tests for the same contamination scenarios as Figures 4.1-4.12. The three tandem tests are Wilcoxon-Quantile, Quantile-Savage and Wilcoxon-Savage. The performance of the State test, which is also a tandem test, is also given in the figures for comparison purposes. The power curves for the State test in Figures 4.13-4.24 are repeated from Figures 4.1-4.12.

A comparison of Figures 4.13-4.24 with Figures 4.1-4.12 indicates:

- <u>For slightly skewed lognormal or Weibull distributions (CV = 0.1)</u>:

  - Tandem tests other than the State test have Type I error rates of about 0.06 to 0.08 rather than the target value of 0.05 for both small and large samples sizes.

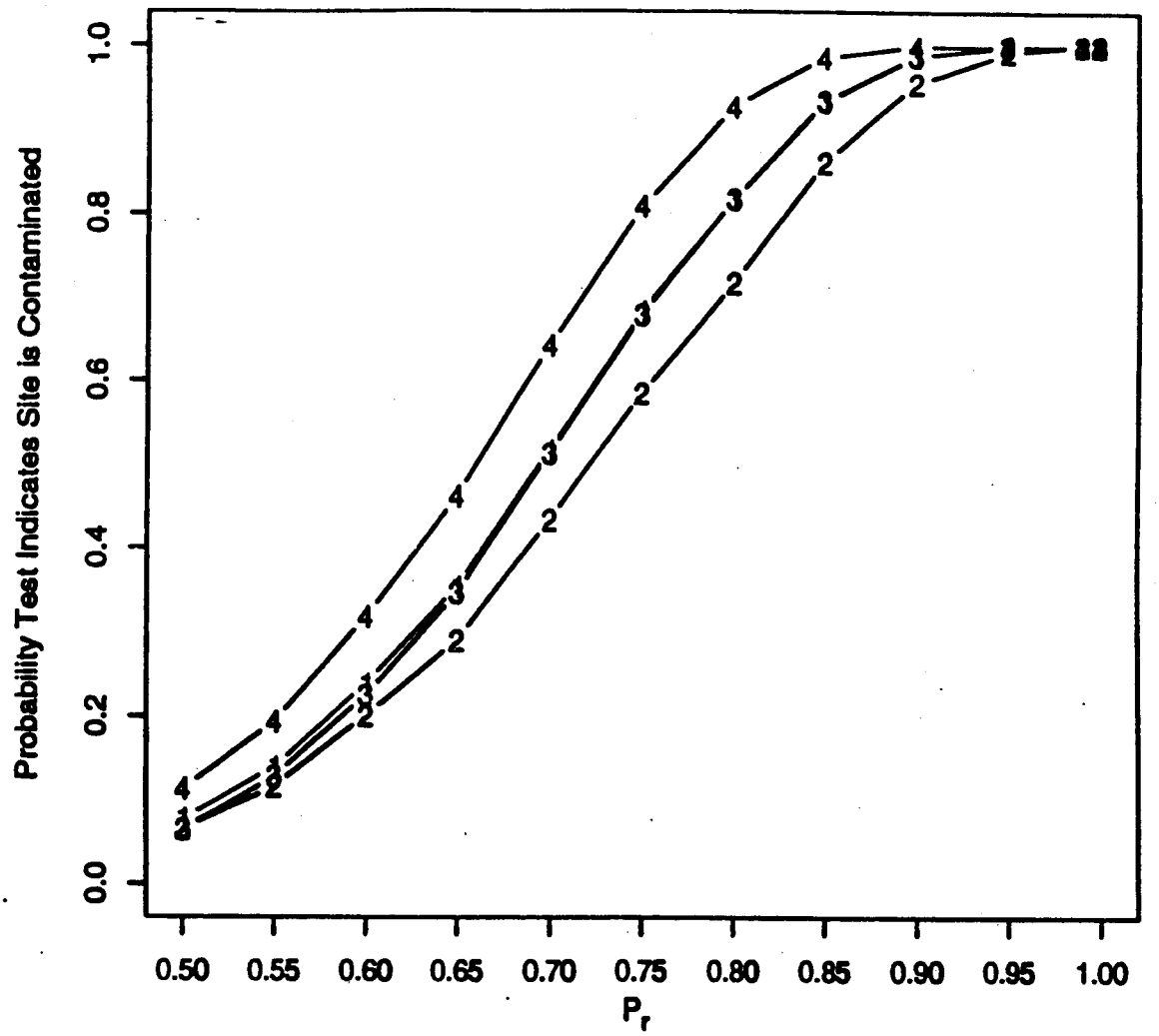
  - Tandem tests give a modest increase in power over individual tests when 10 background and 10 site measurements are used, but no increase in power when 60 background and 50 site samples are used.

  - The Quantile-Savage tandem test usually has less power than the Wilcoxon-Savage or Wilcoxon-Quantile tandem tests.

- <u>For highly skewed distributions (CV = 1 or 2)</u>:

  - Tandem tests other than the State test have Type I error rates of about 0.07 to 0.09 rather than the target value of 0.05 for both small and large sample sizes.

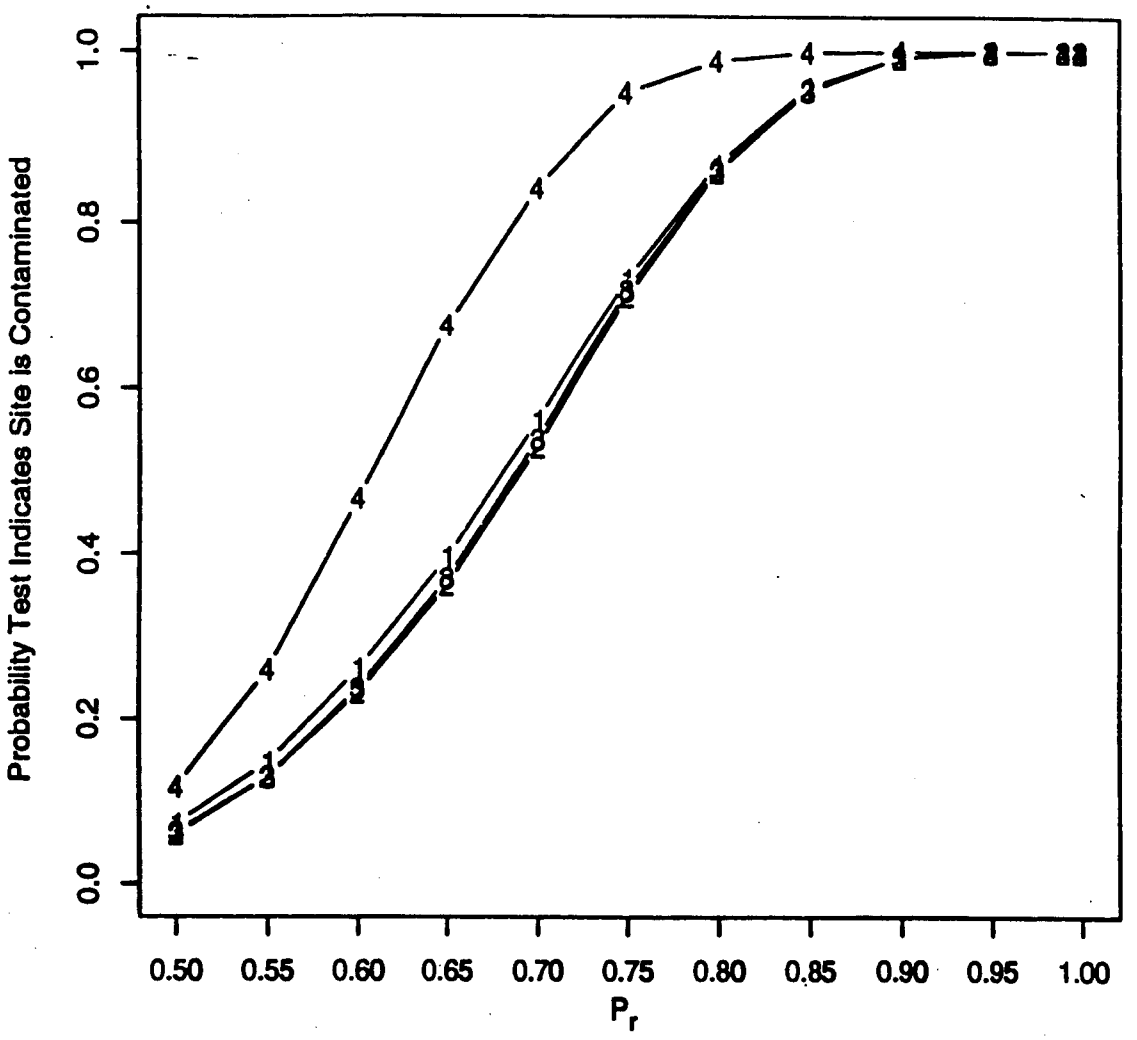
  - The Wilcoxon-Quantile and Wilcoxon-Savage tandem tests have essentially the same power as the Wilcoxon test by itself for either small or large sample sizes. However, these tandem tests have much greater power than the individual Quantile or Savage tests.



PARAMETERS	
Distribution	= Lognormal
CV	= 0.10
ϵ	= 1
Samples (backgd.)	= 10
Samples (site)	= 10

	TEST	TYPE I ERROR
1	Wilcoxon-Quantile	0.0785
2	Quantile-Savage	0.0667
3	Wilcoxon-Savage	0.0655
4	State	0.1138

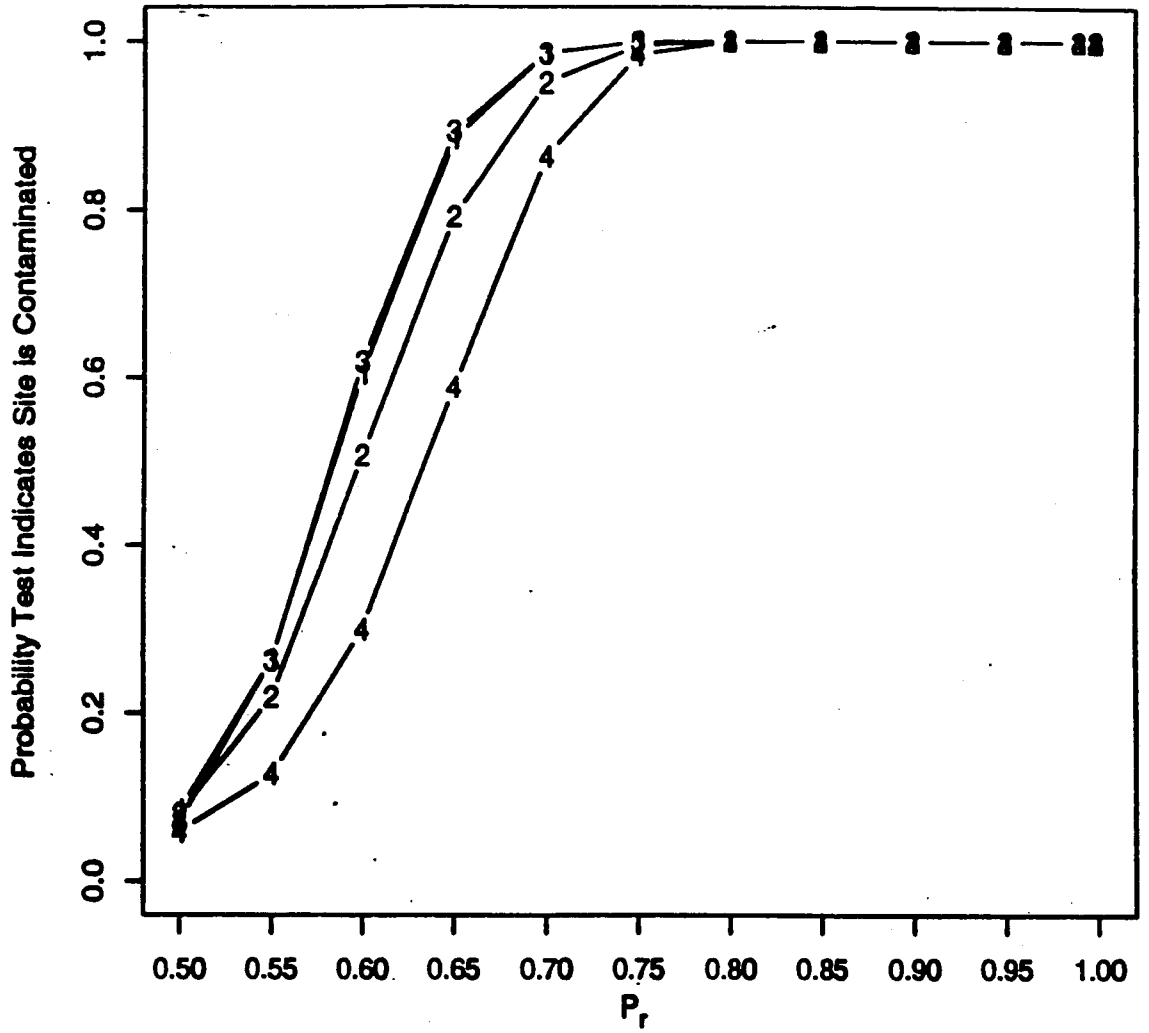
Figure 4.13: Power curves for three tandem tests and the state procedure for a lognormal background distribution, CV = 0.10, $\epsilon = 1$, 10 background samples, and 10 site samples.



PARAMETERS	
Distribution	= Weibull
CV	= 0.10
ε	= 1
Samples (backgd.)	= 10
Samples (site)	= 10

	TEST	TYPE I ERROR
1	Wilcoxon-Quantile	0.0728
2	Quantile-Savage	0.0603
3	Wilcoxon-Savage	0.0619
4	State	0.1168

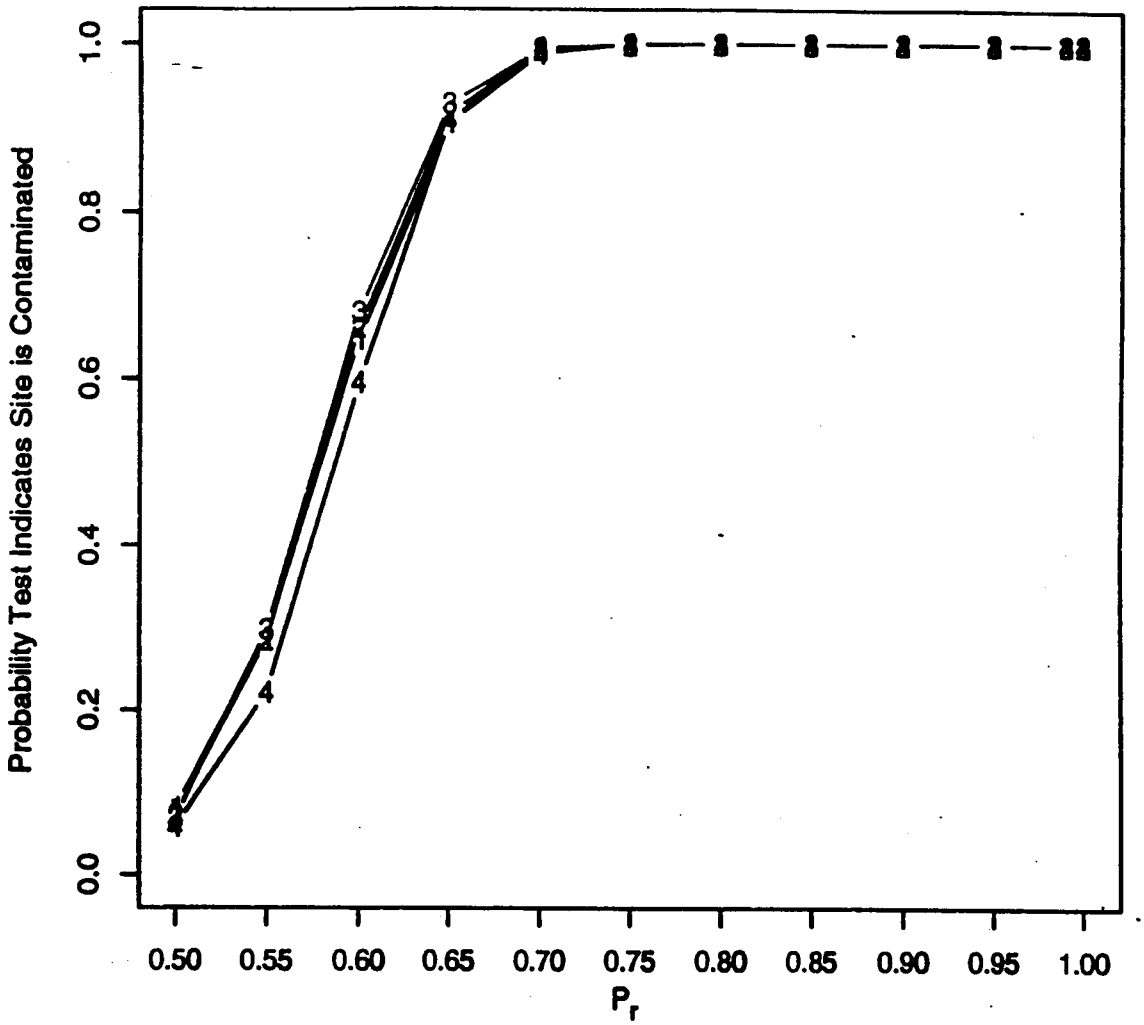
Figure 4.14: Power curves for three tandem tests and the state procedure for a Weibull background distribution, CV = 0.10, $\epsilon = 1$, 10 background samples, and 10 site samples.



PARAMETERS	
Distribution	= Lognormal
CV	= 0.10
ϵ	= 1
Samples (backgd.)	= 60
Samples (site)	= 50

	TEST	TYPE I ERROR
1	Wilcoxon-Quantile	0.0852
2	Quantile-Savage	0.0800
3	Wilcoxon-Savage	0.0707
4	State	0.0591

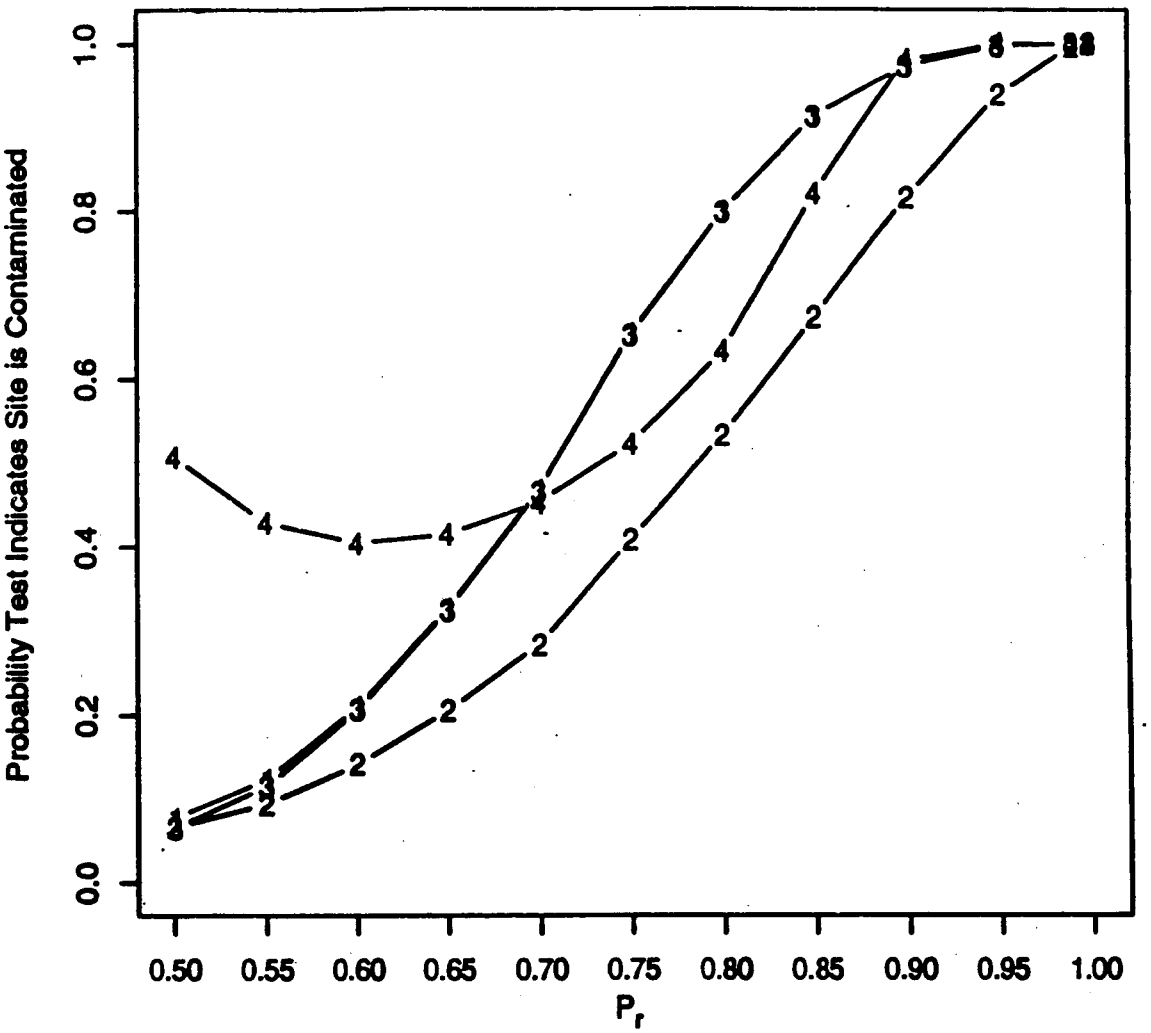
Figure 4.15: Power curves for three tandem tests and the state procedure for a lognormal background distribution, CV = 0.10, $\epsilon = 1$, 60 background samples, and 50 site samples.



PARAMETERS	
Distribution	= Weibull
CV	= 0.10
ε	= 1
Samples (backgd.)	= 60
Samples (site)	= 50

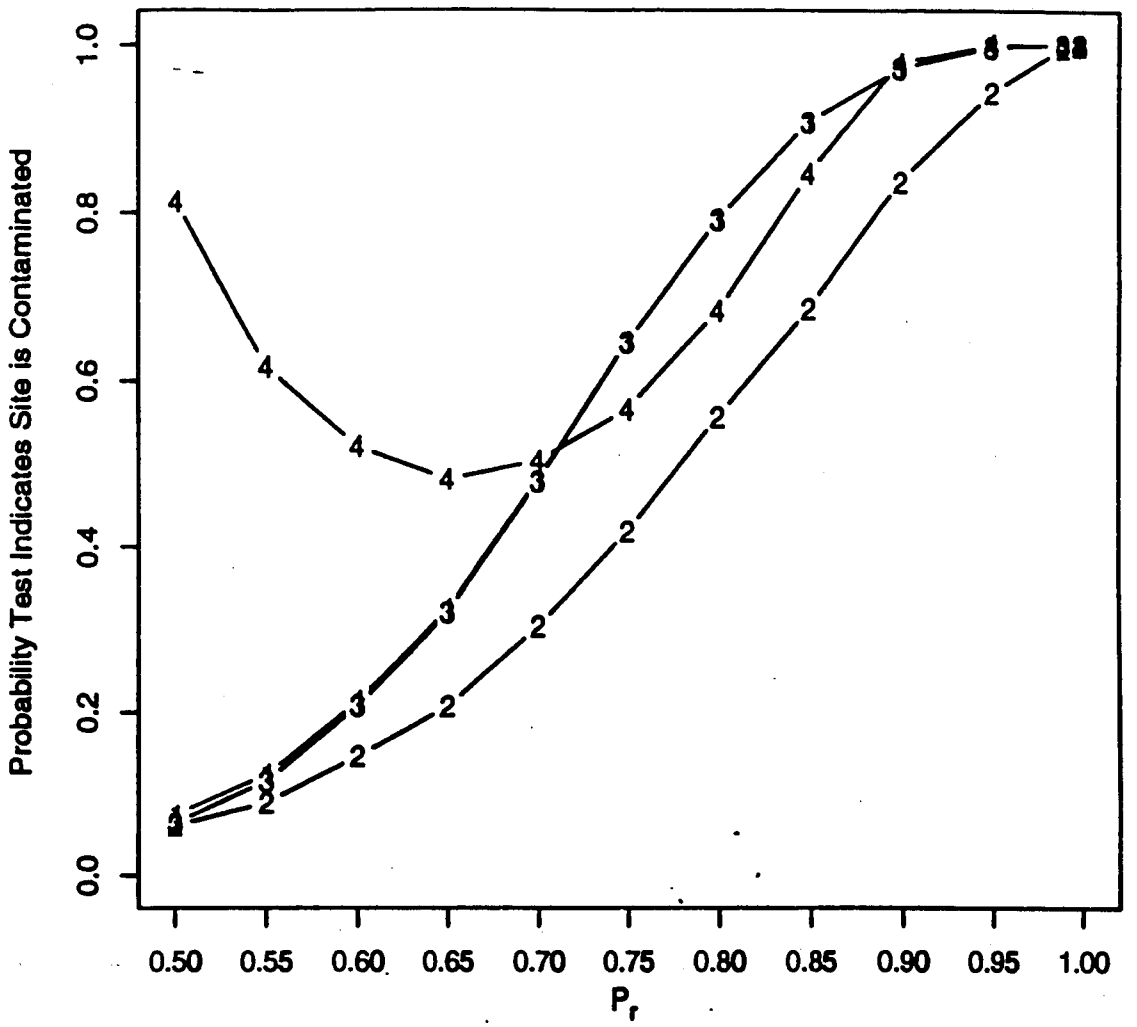
	TEST	TYPE I ERROR
1	Wilcoxon-Quantile	0.0822
2	Quantile-Savage	0.0739
3	Wilcoxon-Savage	0.0692
4	State	0.0587

Figure 4.16: Power curves for three tandem tests and the state procedure for a Weibull background distribution, CV = 0.10, $\epsilon = 1$, 60 background samples, and 50 site samples.



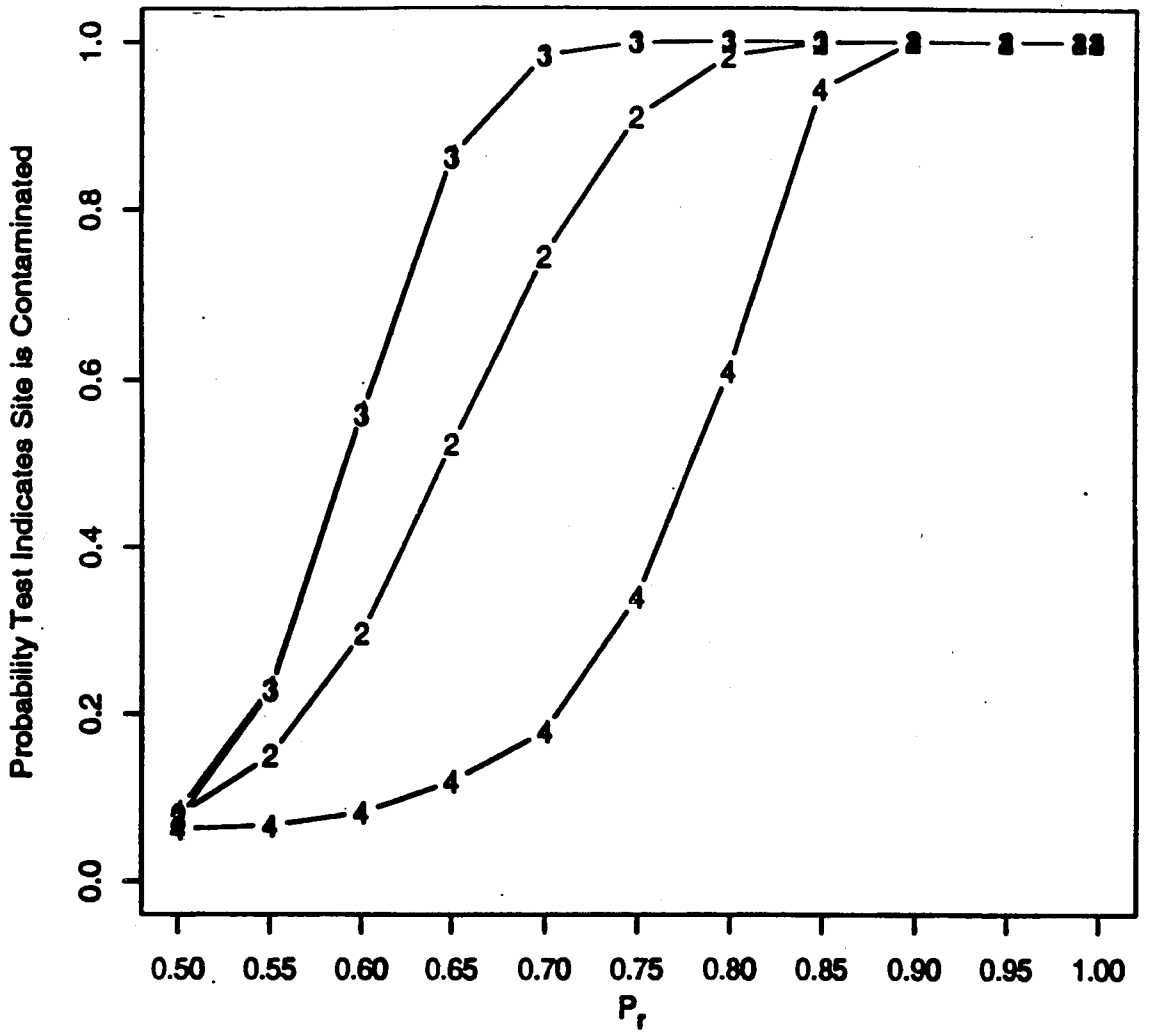
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon-Quantile	0.0770
CV	= 1.00	2	Quantile-Savage	0.0668
ϵ	= 1	3	Wilcoxon-Savage	0.0653
Samples (backgd.)	= 10	4	State	0.5063
Samples (site)	= 10			

Figure 4.17: Power curves for three tandem tests and the state procedure for a lognormal background distribution, CV = 1.00, $\epsilon = 1$, 10 background samples, and 10 site samples.



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon-Quantile	0.0742
CV	= 1.00	2	Quantile-Savage	0.0606
ε	= 1	3	Wilcoxon-Savage	0.0641
Samples (backgd.)	= 10	4	State	0.8123
Samples (site)	= 10			

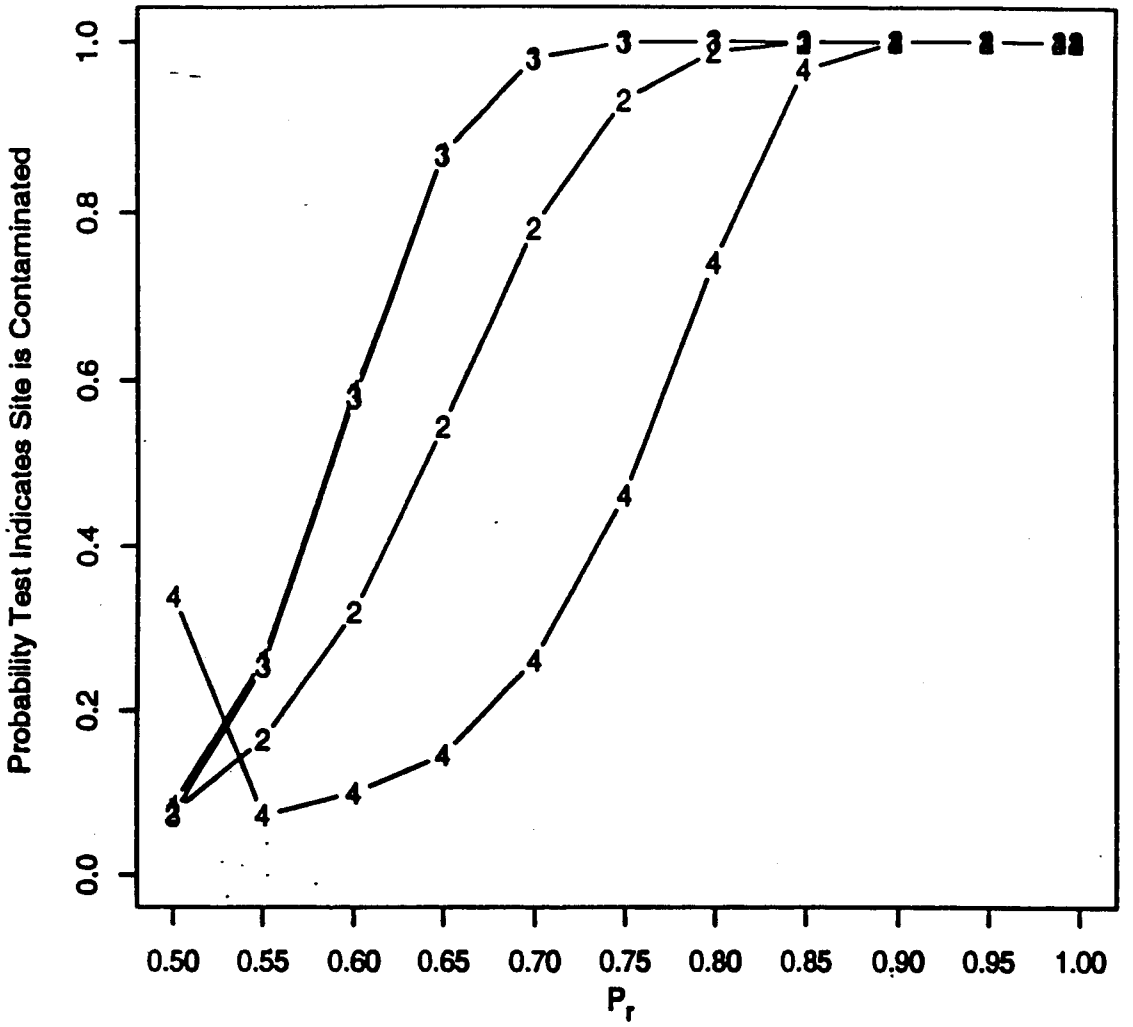
Figure 4.18: Power curves for three tandem tests and the state procedure for a Weibull background distribution, CV = 1.00, $\epsilon = 1$, 10 background samples, and 10 site samples.



PARAMETERS	
Distribution	= Lognormal
CV	= 1.00
ϵ	= 1
Samples (backgd.)	= 60
Samples (site)	= 50

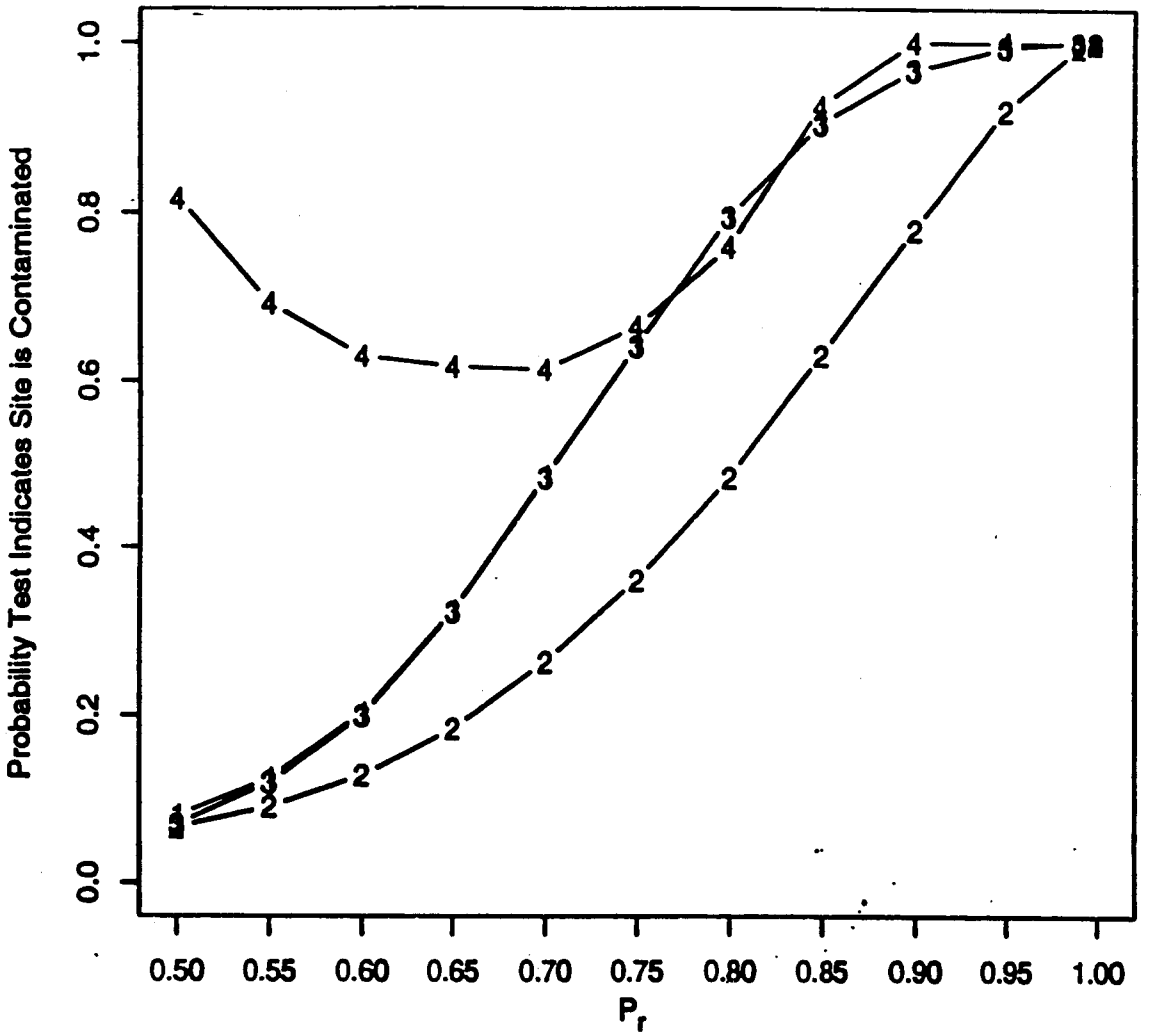
	TEST	TYPE I ERROR
1	Wilcoxon-Quantile	0.0844
2	Quantile-Savage	0.0793
3	Wilcoxon-Savage	0.0728
4	State	0.0625

Figure 4.19: Power curves for three tandem tests and the state procedure for a lognormal background distribution, CV = 1.00, $\epsilon = 1$, 60 background samples, and 50 site samples.



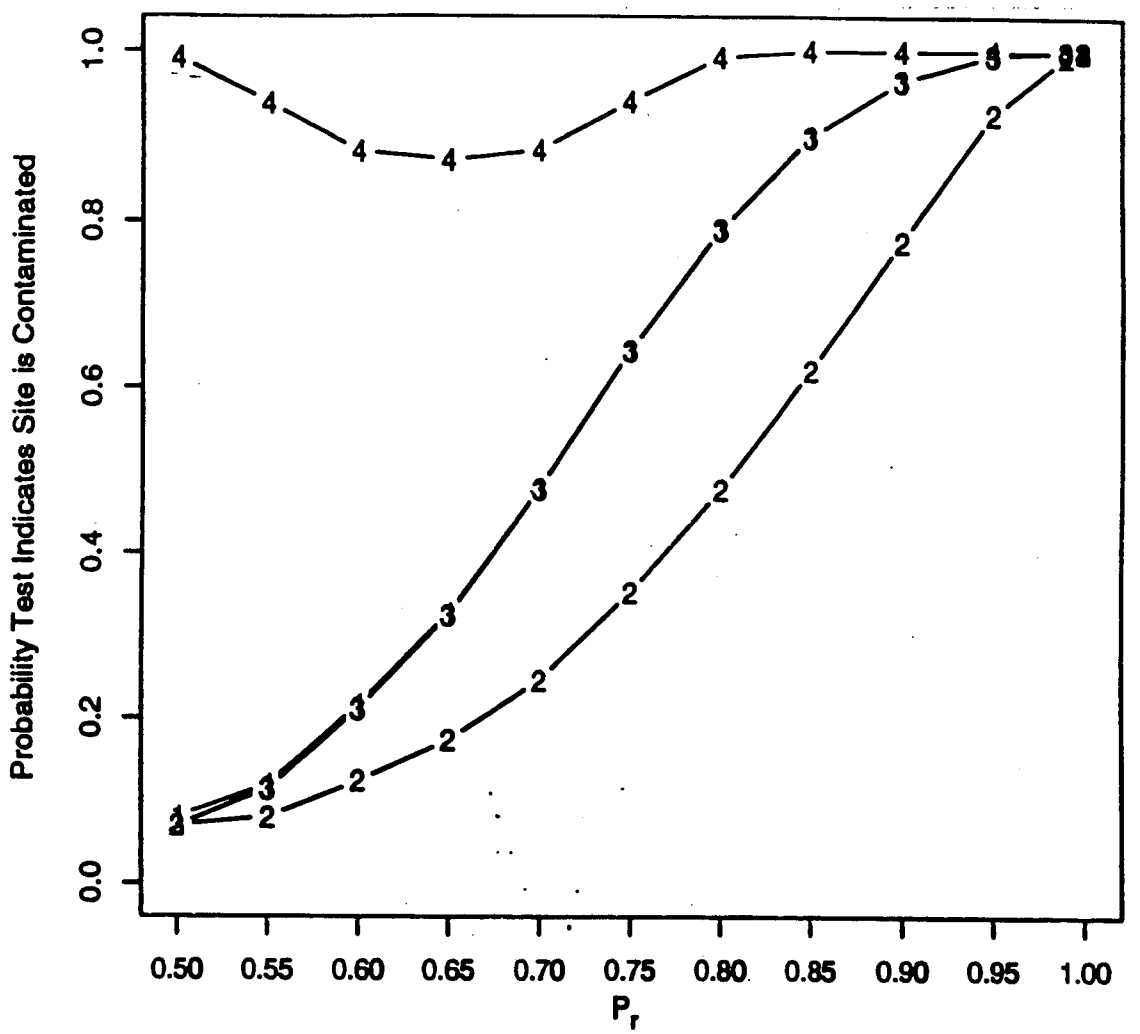
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon-Quantile	0.0842
CV	= 1.00	2	Quantile-Savage	0.0751
ϵ	= 1	3	Wilcoxon-Savage	0.0713
Samples (backgd.)	= 60	4	State	0.3385
Samples (site)	= 50			

Figure 4.20: Power curves for three tandem tests and the state procedure for a Weibull background distribution, CV = 1.00, ϵ = 1, 60 background samples, and 50 site samples.



PARAMETERS		TEST	TYPE I ERROR
Distribution	= Lognormal	1 Wilcoxon-Quantile	0.0795
CV	= 2.00	2 Quantile-Savage	0.0660
ϵ	= 1	3 Wilcoxon-Savage	0.0690
Samples (backgd.)	= 10	4 State	0.8169
Samples (site)	= 10		

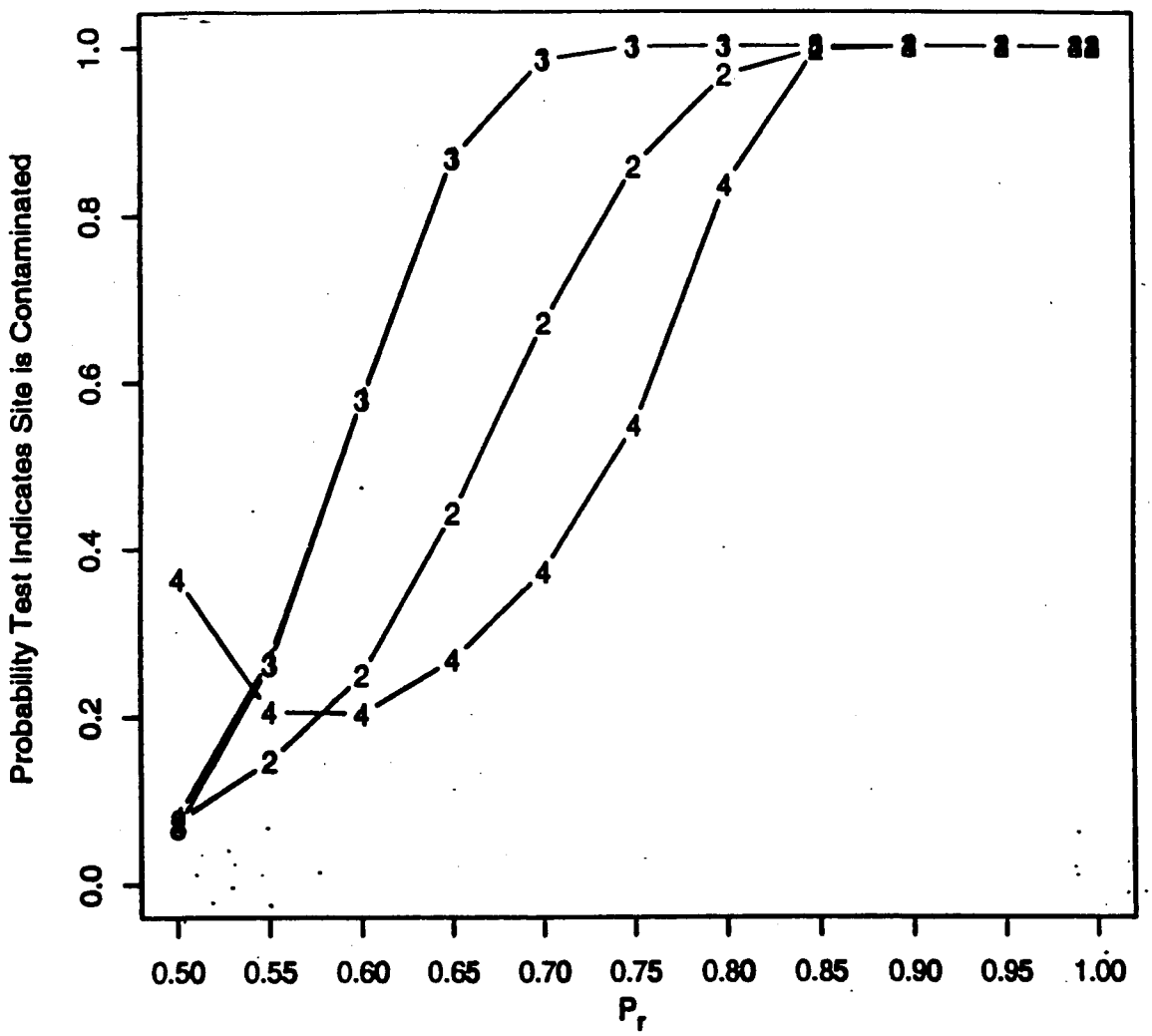
Figure 4.21: Power curves for three tandem tests and the state procedure for a lognormal background distribution, CV = 2.00, $\epsilon = 1$, 10 background samples, and 10 site samples.



PARAMETERS	
Distribution	= Weibull
CV	= 2.00
ε	= 1
Samples (backgd.)	= 10
Samples (site)	= 10

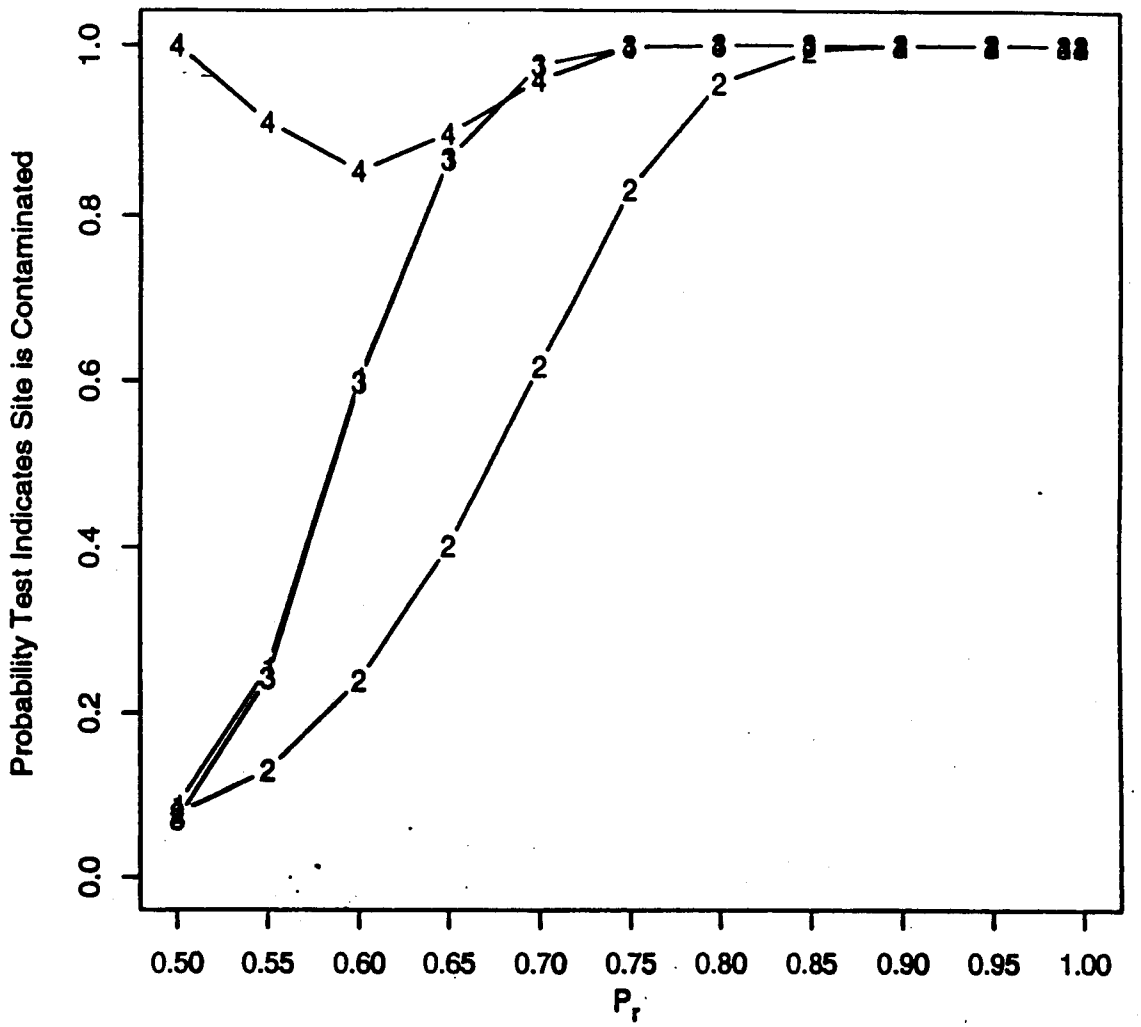
	TEST	TYPE I ERROR
1	Wilcoxon-Quantile	0.0805
2	Quantile-Savage	0.0701
3	Wilcoxon-Savage	0.0693
4	State	0.9921

Figure 4.22: Power curves for three tandem tests and the state procedure for a Weibull background distribution, CV = 2.00, $\epsilon = 1$, 10 background samples, and 10 site samples.



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon-Quantile	0.0810
CV	= 2.00	2	Quantile-Savage	0.0756
ϵ	= 1	3	Wilcoxon-Savage	0.0684
Samples (backgd.)	= 60	4	State	0.3643
Samples (site)	= 50			

Figure 4.23: Power curves for three tandem tests and the state procedure for a lognormal background distribution, CV = 2.00, $\epsilon = 1$, 60 background samples, and 50 site samples.



PARAMETERS	
Distribution	= Weibull
CV	= 2.00
ε	= 1
Samples (backgd.)	= 60
Samples (site)	= 50

	TEST	TYPE I ERROR
1	Wilcoxon-Quantile	0.0875
2	Quantile-Savage	0.0783
3	Wilcoxon-Savage	0.0707
4	State	0.9997

Figure 4.24: Power curves for three tandem tests and the state procedure for a Weibull background distribution, CV = 2.00, $\epsilon = 1$, 60 background samples, and 50 site samples.

- The Quantile-Savage tandem test always has considerably less power than the Wilcoxon-Savage or Wilcoxon-Quantile tests.

- The State test has less power than any of the other tandem tests for the large sample-size case when CV = 1.

In Summary:

- For the pervasive contamination case with slightly or highly skewed lognormal background and site distributions of the same shape, and for small or large numbers of samples, the Wilcoxon-Quantile or Wilcoxon-Savage tandem tests:

  - provide a modest increase in power over using the Wilcoxon test alone and a large increase in power over using the Quantile or Savage tests alone.

  - have Type I error rates of 0.06 to 0.09 compared to the target value of 0.05

It should be noted that Gilbert and Simpson (1992) recommend using the Wilcoxon test and the Quantile test in tandem. Their recommendation is based on computer simulations they conducted to evaluate the performance of each test. They found that the Wilcoxon test had more power than the Quantile test to detect pervasive site contamination. They also found that the Quantile test had more power than the Wilcoxon test when only a small portion of the site was highly contaminated above background. As the two tests have good power for different contamination scenarios, they recommend conducting both tests (i.e., using the Wilcoxon-Quantile tandem test) unless it is known which contamination scenario is most likely. The complementary power of the Wilcoxon and Quantile tests reported by Gilbert and Simpson (1992) is confirmed by the results in the present report.

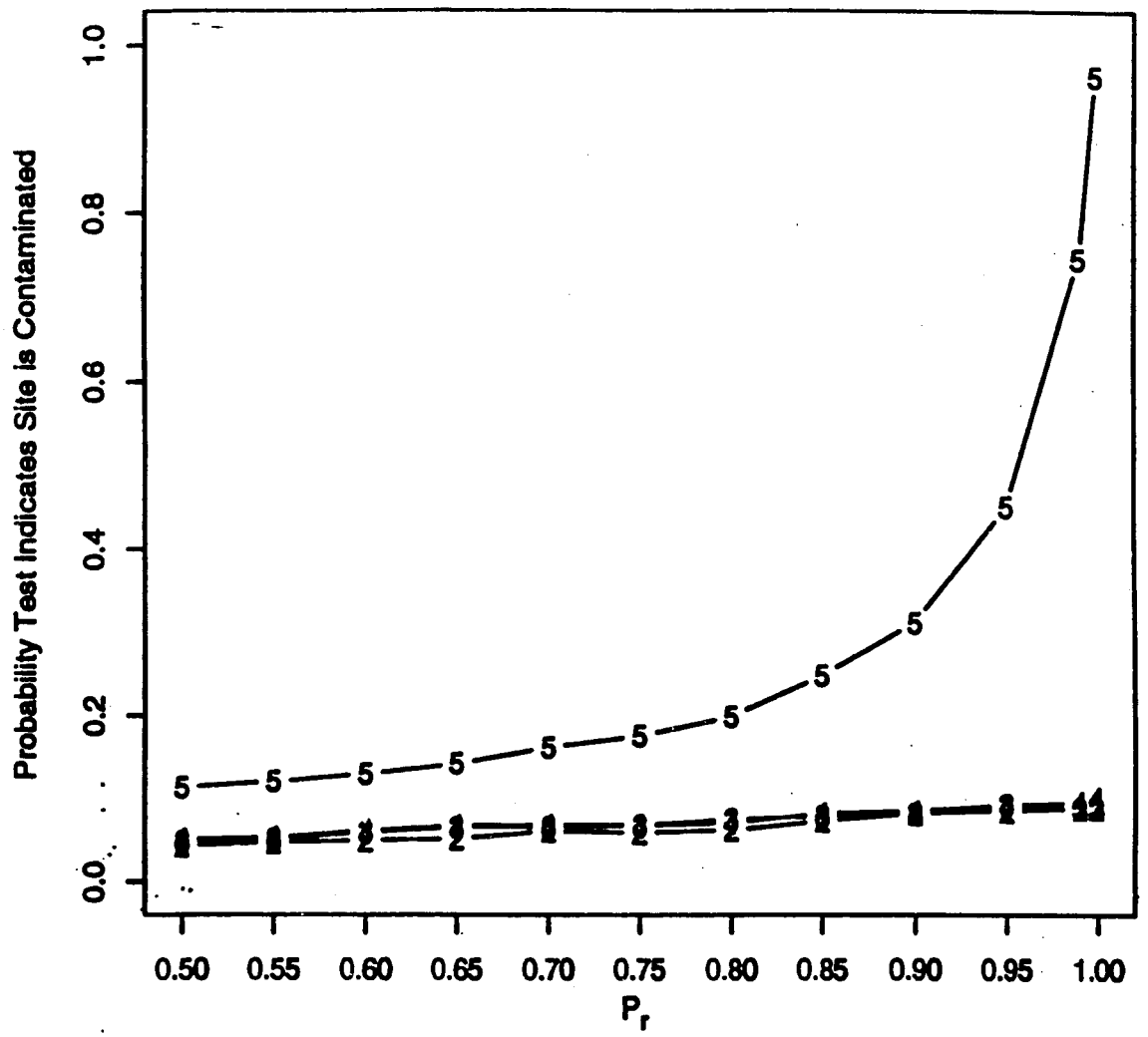
4.2 Ten Percent Contamination at the Site (ϵ = 0.1)

Sections 4.1.1 through 4.1.3 presented results for the case of pervasive contamination. Now, we consider cases for which only 10 percent of the site is contaminated. Note that for this situation the shape of the site distribution is not identical to that of the background distribution because the site is no longer uniformly contaminated a specified amount greater than background. Subsection 4.2.1 presents results for the case of a slightly skewed background distribution (CV = 0.1). Subsection 4.2.2 gives results for a highly skewed case (CV = 2). The shape of site distributions for various degrees of skewness and percent contamination are shown in Figures 3.3-3.12 for a magnitude of contamination P_r = 0.85.

4.2.1 Slightly Skewed Distribution (CV = 0.1)

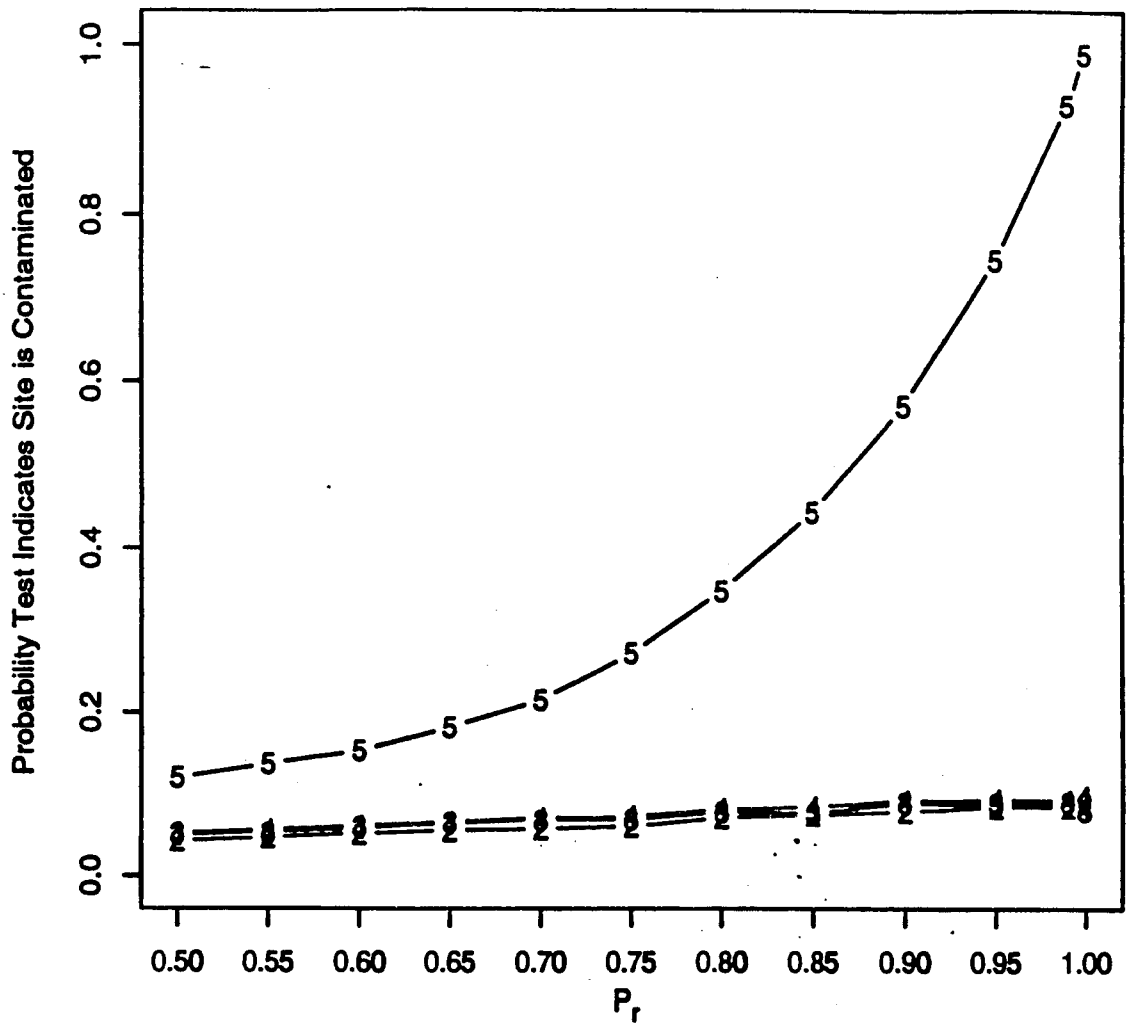
For the case of 10 percent contamination, Figures 4.25 and 4.26 give results for slightly skewed (CV = 0.1) lognormal and Weibull background distributions, respectively, for when only 10 background and 10 site measurements are available. First, the State test has an elevated Type I error rate, 0.11 instead of 0.05, and the other tests do not. These same results were obtained for the pervasive contamination case (Figure 4.1). As regards power, the State test has the best power, but even its power remains very low until the magnitude of contamination in the contaminated areas of the site becomes large. The power is slightly better for the Weibull distribution than for the lognormal. The other tests have essentially no power to detect the 10 percent contamination greater than background for either distribution.

When 60 background and 50 site measurements are available (Figures 4.27 and 4.28) the Type I error rate of the State test (0.06) is much closer to the target value of 0.05, the same conclusion obtained for the pervasive contamination scenario (Figure 4.3). Also, the larger number of samples improves the power. The State test has the most power, and the power of other tests improves somewhat, but all tests have low power except when P_r is close to 1. The power of the State test and the Quantile test is larger for the Weibull than the lognormal distributions. Also, when P_r approaches one, the



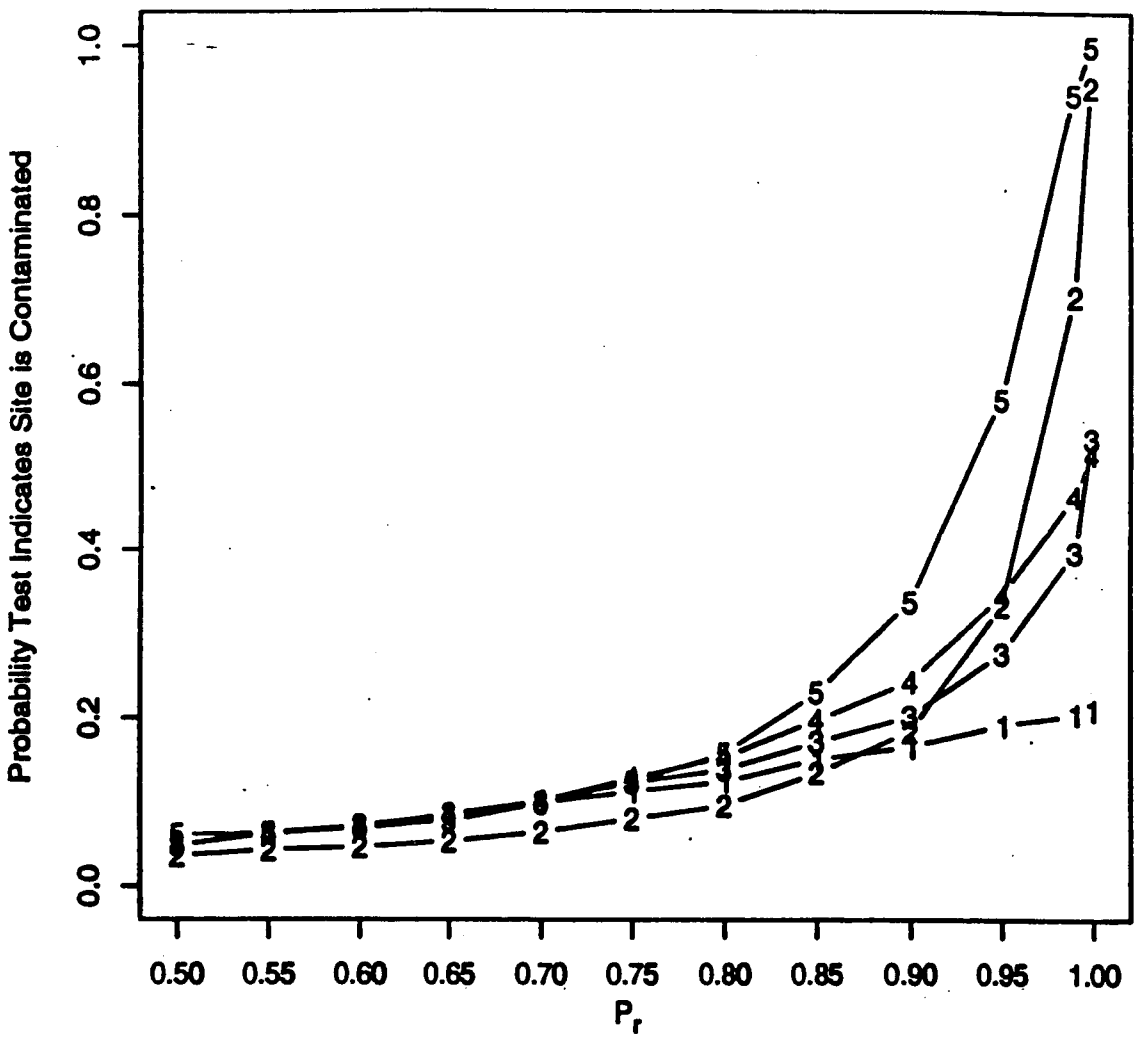
PARAMETERS		TEST	TYPE I ERROR
Distribution	= Lognormal	1 Wilcoxon	0.0509
CV	= 0.10	2 Quantile	0.0429
ϵ	= 0.1	3 t-Test	0.0479
Samples (backgd.)	= 10	4 Savage	0.0499
Samples (site)	= 10	5 State	0.1128

Figure 4.25: Power curves for five test procedures for a lognormal background distribution, CV = 0.10, $\epsilon = 0.1$, 10 background samples, and 10 site samples.



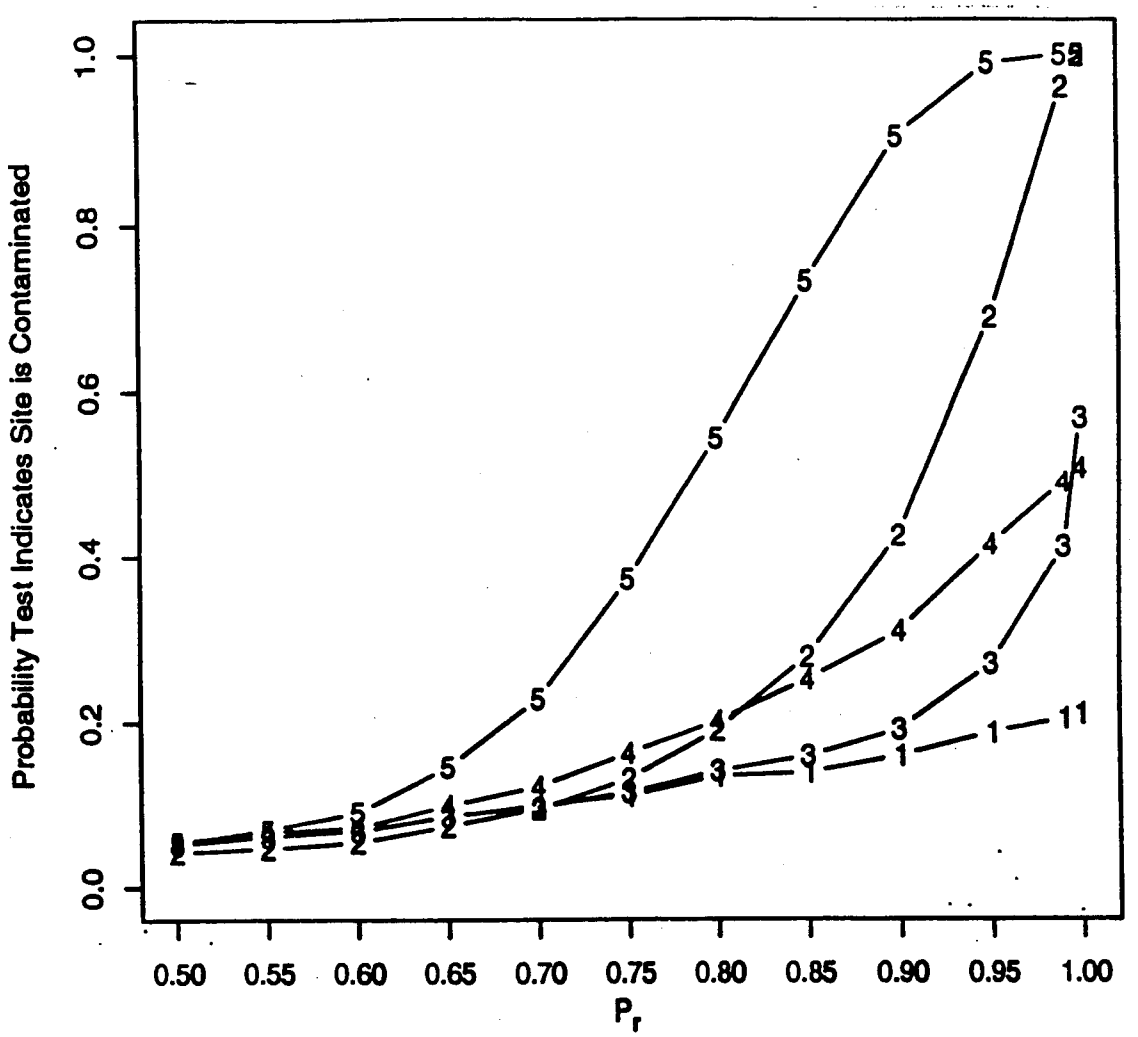
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0527
CV	= 0.10	2	Quantile	0.0428
ε	= 0.1	3	t-Test	0.0515
Samples (backgd.)	= 10	4	Savage	0.0498
Samples (site)	= 10	5	State	0.1195

Figure 4.26: Power curves for five test procedures for a Weibull background distribution, CV = 0.10, $\epsilon = 0.1$, 10 background samples, and 10 site samples.



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon	0.0481
CV	= 0.10	2	Quantile	0.0363
ε	= 0.1	3	t-Test	0.0487
Samples (backgd.)	= 60	4	Savage	0.0487
Samples (site)	= 50	5	State	0.0610

Figure 4.27: Power curves for five test procedures for a lognormal background distribution, CV = 0.10, $\epsilon = 0.1$, 60 background samples, and 50 site samples.



PARAMETERS		TEST	TYPE I ERROR
Distribution	= Weibull	1 Wilcoxon	0.0512
CV	= 0.10	2 Quantile	0.0419
ε	= 0.1	3 t-Test	0.0515
Samples (backgd.)	= 60	4 Savage	0.0518
Samples (site)	= 50	5 State	0.0545

Figure 4.28: Power curves for five test procedures for a Weibull background distribution, CV = 0.10, $\epsilon = 0.1$, 60 background samples, and 50 site samples.

power of the Quantile test becomes larger than that of all other tests except the State test. In particular, the Wilcoxon test has the least power of any test when P_r is greater than about 0.85. This result is in contrast to that for the pervasive contamination case (Figures 4.3 and 4.4). For that case, the Wilcoxon test had the most (or nearly so) power of any test. These results confirm the conclusions of Gilbert and Simpson (1992) that the Wilcoxon test has more power than the Quantile test to detect pervasive contamination, but that the Quantile test has more power to detect "hot spots" or partial contamination.

Summary

- When the background distribution is only slightly skewed and only 10 percent of the site is contaminated greater than background, the State test has the best power. However, the State test has a somewhat elevated Type I error rate when only 10 background and 10 site samples are collected.

- The power of all tests is low.

- For larger sample sizes, the Wilcoxon test has the least power of all tests studied for this 10 percent contamination scenario, and the Quantile test has much greater power than the Wilcoxon test when the magnitude of contamination (value of P_r) is large.

4.2.2 Highly Skewed Distributions (CV = 1 and 2)

Figures 4.29-4.32 consider the case of 10 percent contamination when the background distribution is more highly skewed (CV = 1). Figures 4.33-4.36 are for even more highly skewed distributions (CV = 2).

For the small sample-size case (Figures 4.29, 4.30, 4.33 and 4.34, the State test has an extremely elevated Type I error level, whereas other tests have Type I errors that cluster closely around the target value of 0.05.

However, the power of the State test is far superior to the very low power of all other tests.

For the large sample-size case (Figures 4.31, 4.32, 4.35, 4.36) and CV = 1, the Type I error of the State test for the lognormal background distribution is no longer elevated above the target value (Figure 4.31), but it remains elevated for the Weibull distribution (Figure 4.32). The increased sample sizes improve the power of all tests somewhat, but power remains low in general unless P_r is very near 1. When CV = 2, the Type I error rate of the State test is virtually 1 for the Weibull distribution and remains high (0.36) for the lognormal distribution. The power of all tests remains low.

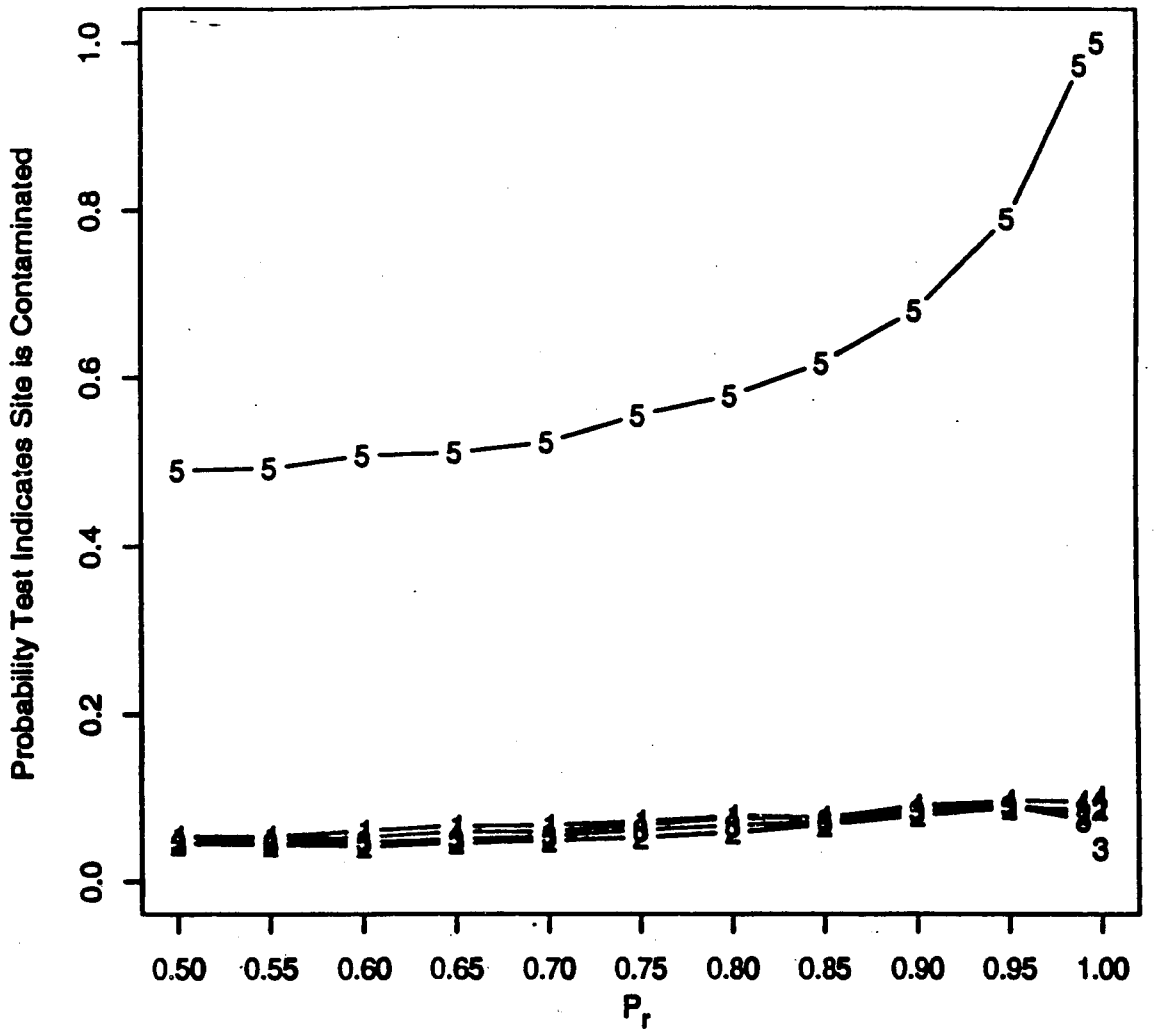
Summary

- When the background distribution is highly skewed and only 10 percent of the site is contaminated, the State test has the best power, but its Type I error level is highly elevated, particularly for the extremely skewed Weibull background distributions. The power of all other tests is low even for large numbers of samples.

- For highly skewed background Weibull distributions and a site with 10 percent contamination, the State test has a probability of 1 of declaring the site is contaminated, regardless of whether the site is identical to background, very slightly contaminated, or heavily contaminated. In other words, for this contamination scenario, the State test will always indicate the site is contaminated, even when it is not.

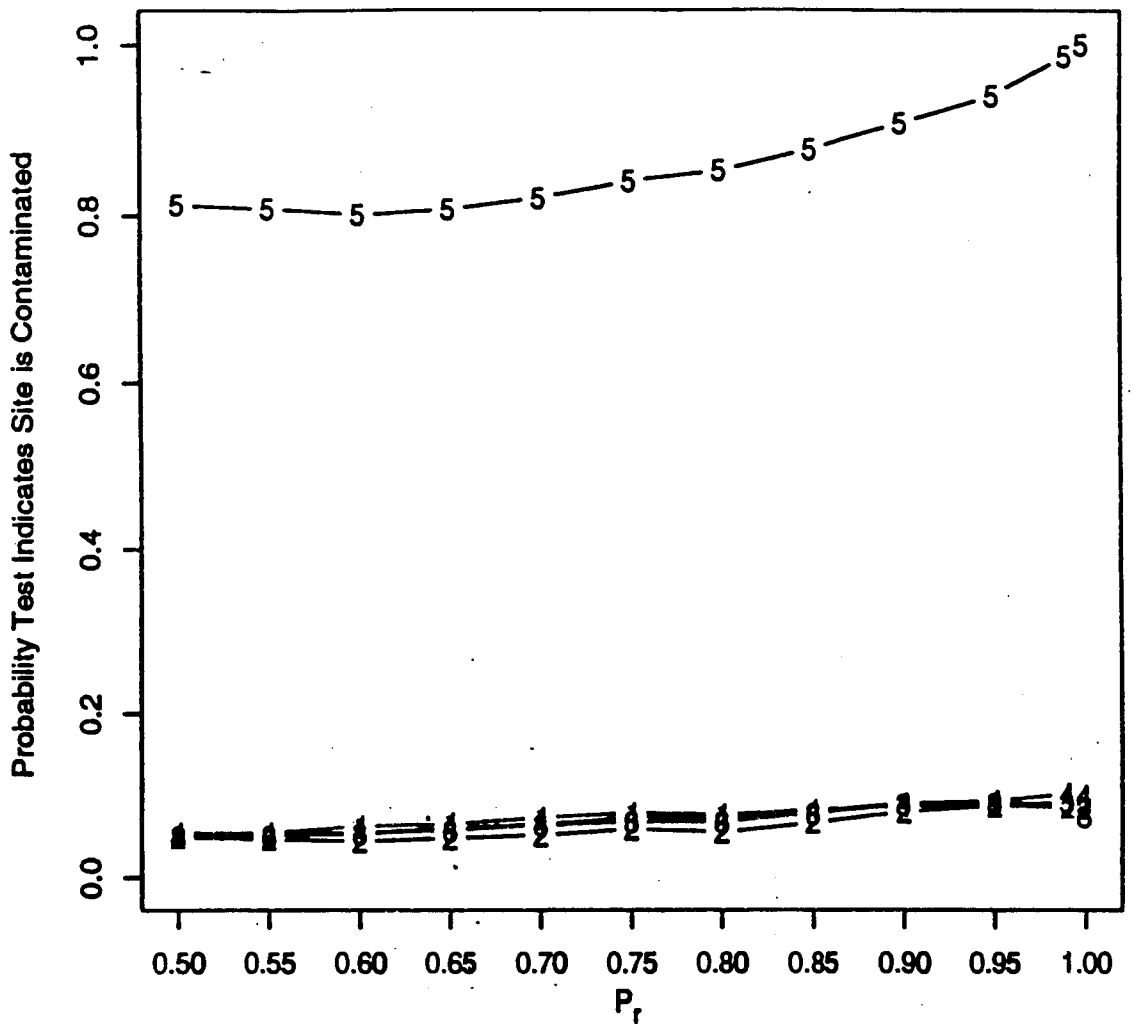
4.2.3 Tandem Testing

Power curves are not shown here for tandem tests applied when only 10 percent of the site is contaminated. However, these curves (available on microfiche) indicate:



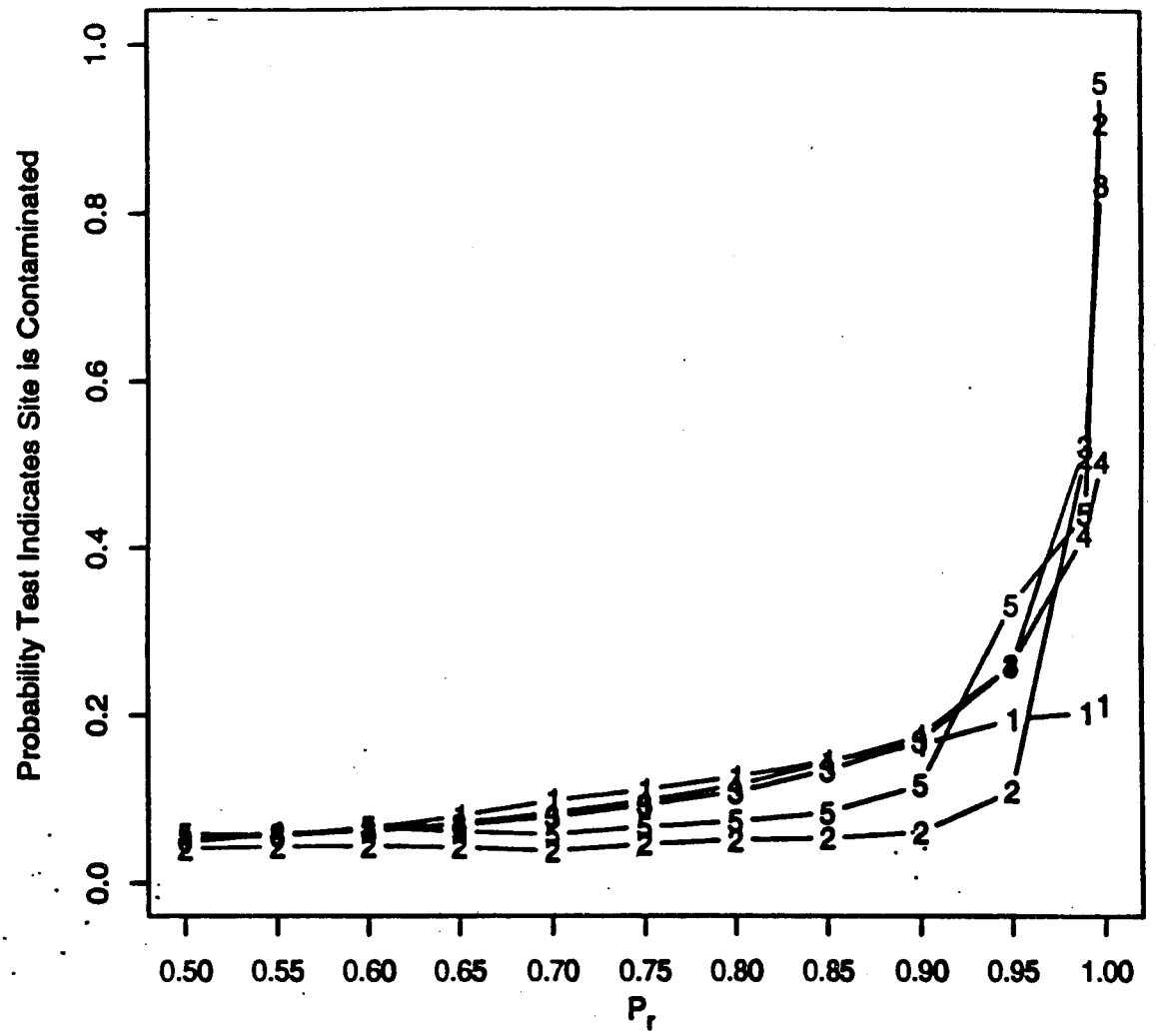
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon	0.0548
CV	= 1.00	2	Quantile	0.0454
ε	= 0.1	3	t-Test	0.0467
Samples (backgd.)	= 10	4	Savage	0.0530
Samples (site)	= 10	5	State	0.4894

Figure 4.29: Power curves for five test procedures for a lognormal background distribution, CV = 1.00, $\epsilon = 0.1$, 10 background samples, and 10 site samples.



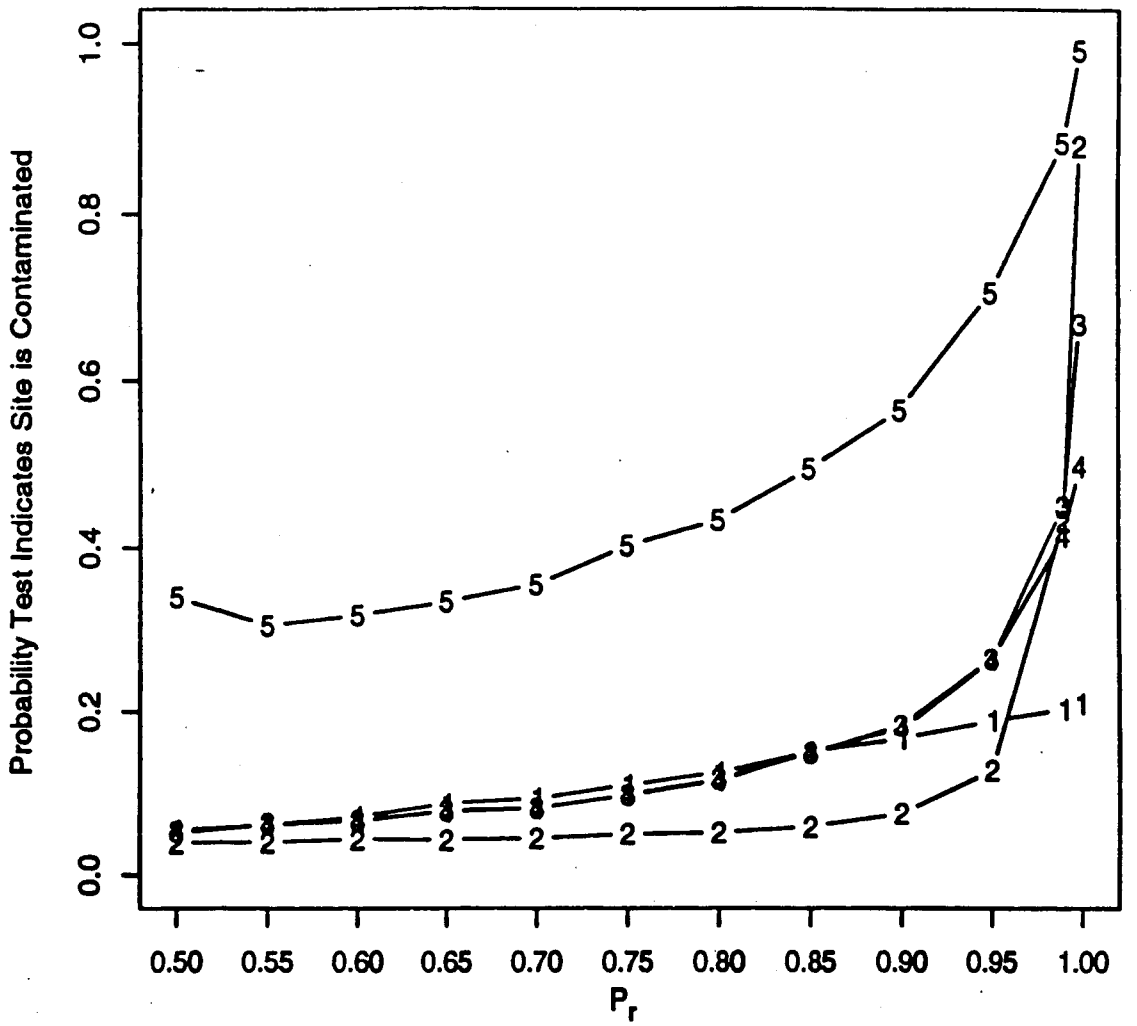
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0543
CV	= 1.00	2	Quantile	0.0481
ε	= 0.1	3	t-Test	0.0493
Samples (backgd.)	= 10	4	Savage	0.0496
Samples (site)	= 10	5	State	0.8116

Figure 4.30: Power curves for five test procedures for a Weibull background distribution, CV = 1.00, $\epsilon = 0.1$, 10 background samples, and 10 site samples.



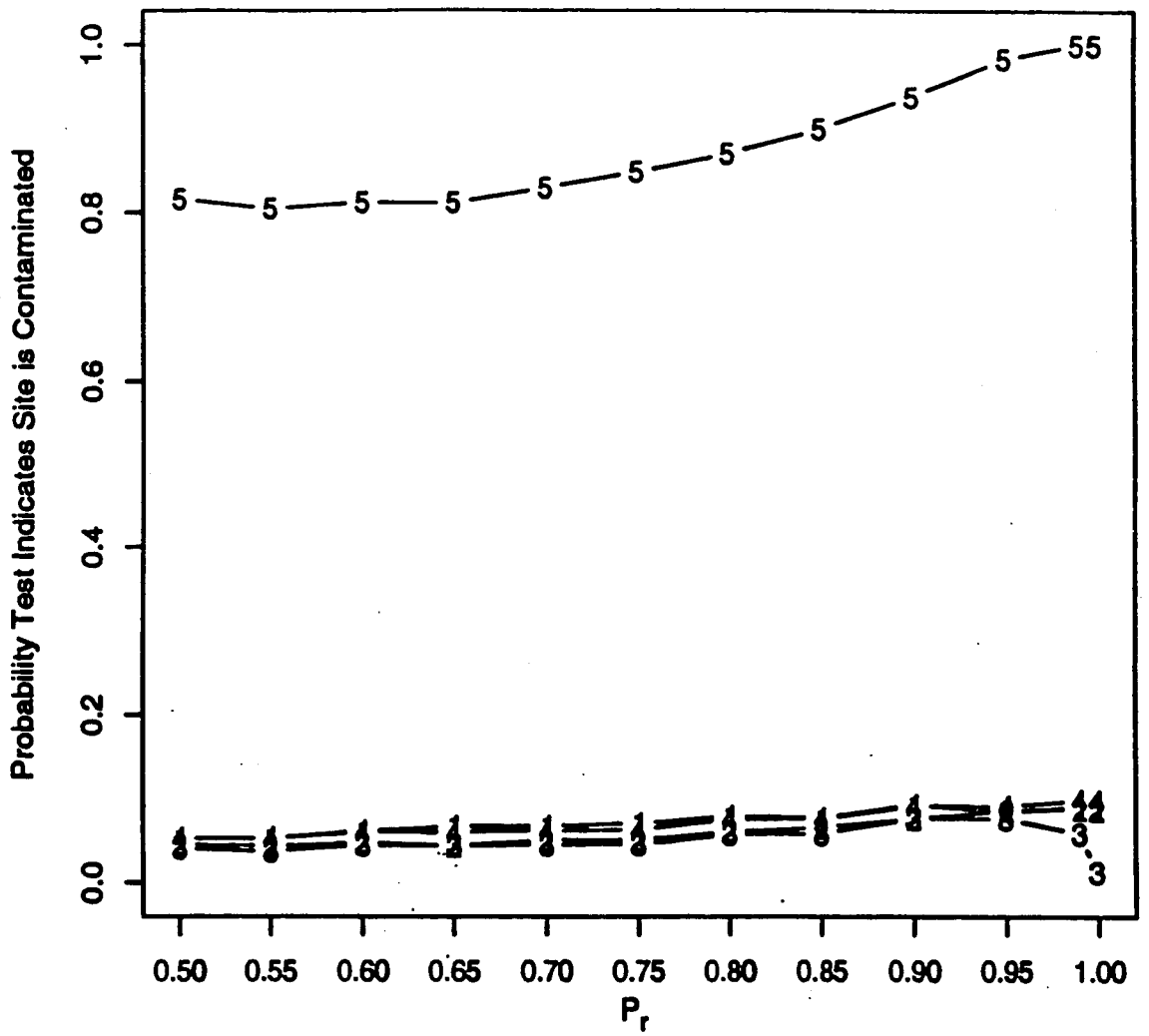
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon	0.0489
CV	= 1.00	2	Quantile	0.0409
ϵ	= 0.1	3	t-Test	0.0511
Samples (backgd.)	= 60	4	Savage	0.0529
Samples (site)	= 50	5	State	0.0583

Figure 4.31: Power curves for five test procedures for a lognormal background distribution, CV = 1.00, $\epsilon = 0.1$, 60 background samples, and 50 site samples.



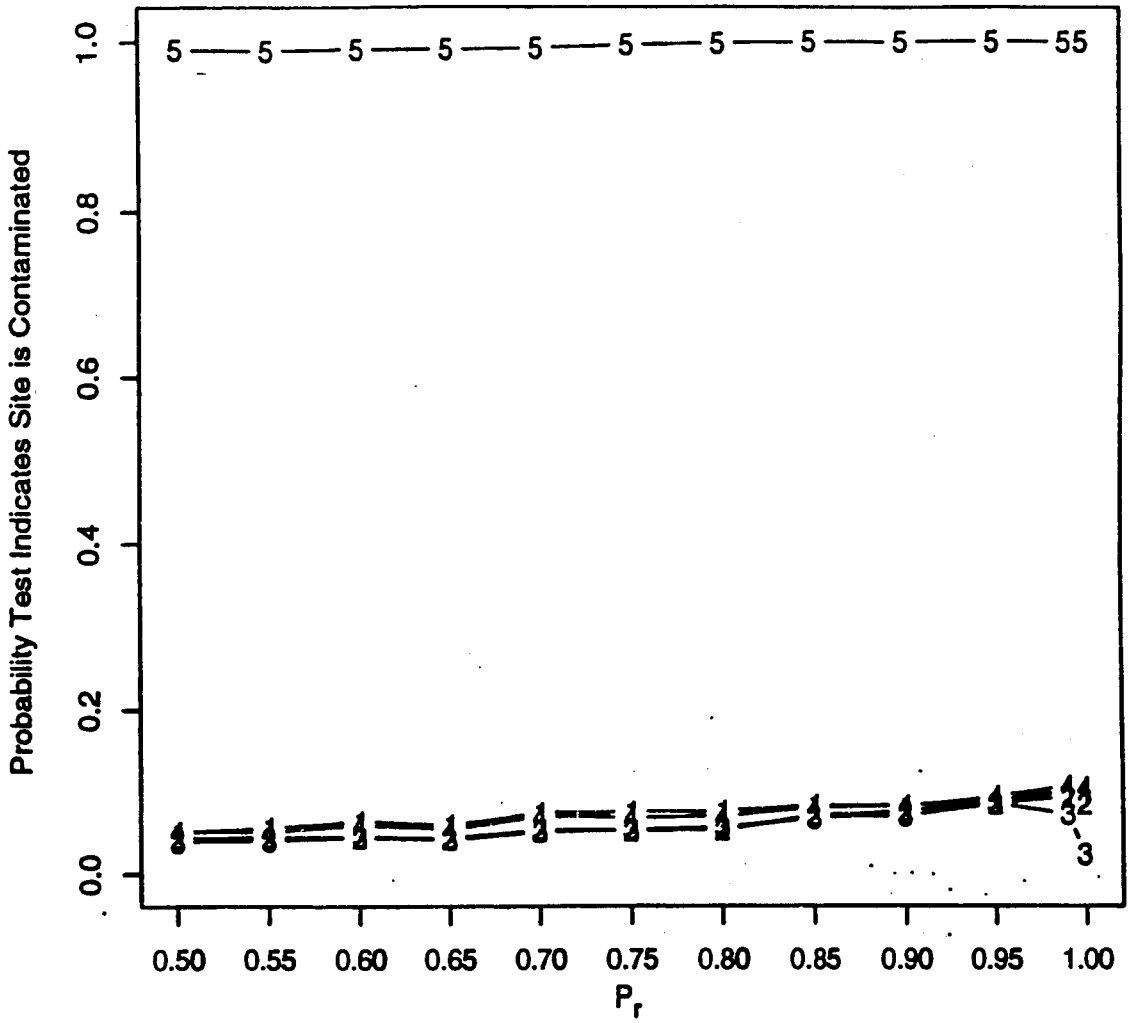
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0517
CV	= 1.00	2	Quantile	0.0394
ε	= 0.1	3	t-Test	0.0533
Samples (backgd.)	= 60	4	Savage	0.0554
Samples (site)	= 50	5	State	0.3401

Figure 4.32: Power curves for five test procedures for a Weibull background distribution, CV = 1.00, $\epsilon = 0.1$, 60 background samples, and 50 site samples.



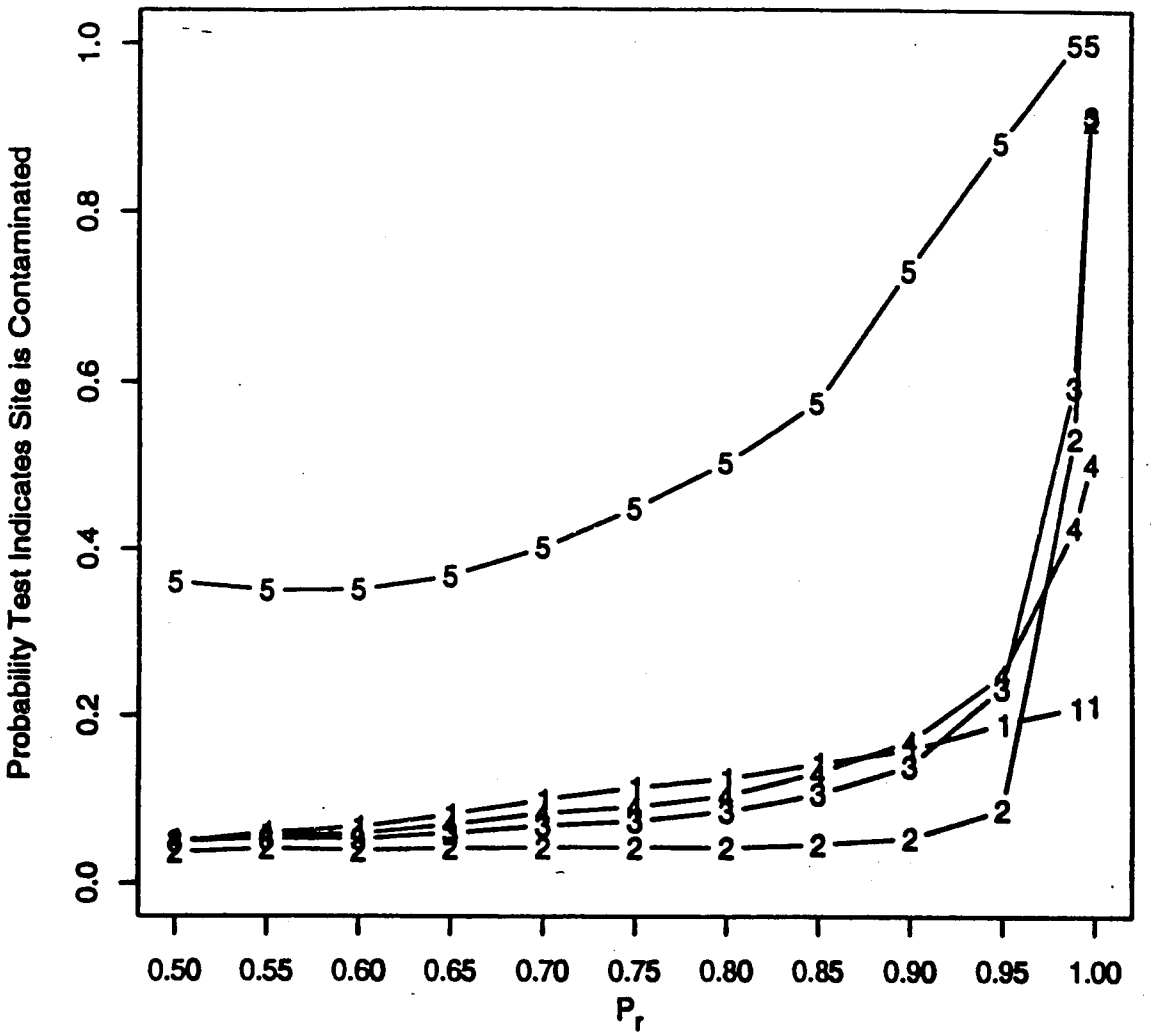
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	Wilcoxon	0.0526
CV	= 2.00	2	Quantile	0.0451
ϵ	= 0.1	3	t-Test	0.0405
Samples (backgd.)	= 10	4	Savage	0.0526
Samples (site)	= 10	5	State	0.8160

Figure 4.33: Power curves for five test procedures for a lognormal background distribution, CV = 2.00, $\epsilon = 0.1$, 10 background samples, and 10 site samples.



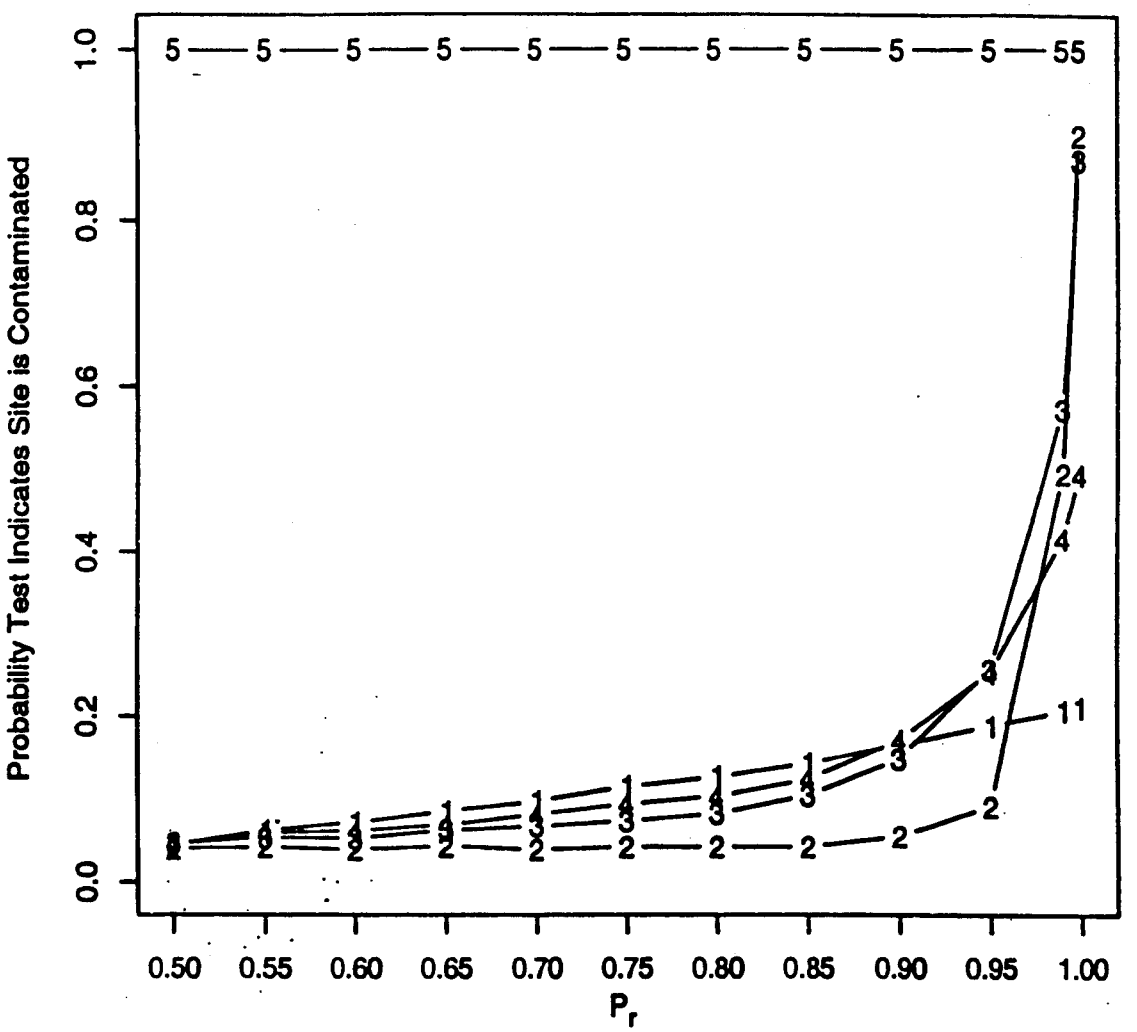
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0514
CV	= 2.00	2	Quantile	0.0438
ε	= 0.1	3	t-Test	0.0391
Samples (backgd.)	= 10	4	Savage	0.0514
Samples (site)	= 10	5	State	0.9906

Figure 4.34: Power curves for five test procedures for a Weibull background distribution, CV = 2.00, $\epsilon = 0.1$, 10 background samples, and 10 site samples.



PARAMETERS		TEST	TYPE I ERROR
Distribution	= Lognormal	1 Wilcoxon	0.0502
CV	= 2.00	2 Quantile	0.0378
ε	= 0.1	3 t-Test	0.0504
Samples (backgd.)	= 60	4 Savage	0.0520
Samples (site)	= 50	5 State	0.3601

Figure 4.35: Power curves for five test procedures for a lognormal background distribution, CV = 2.00, $\epsilon = 0.1$, 60 background samples, and 50 site samples.



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Weibull	1	Wilcoxon	0.0453
CV	= 2.00	2	Quantile	0.0403
ε	= 0.1	3	t-Test	0.0465
Samples (backgd.)	= 60	4	Savage	0.0463
Samples (site)	= 50	5	State	0.9999

Figure 4.36: Power curves for five test procedures for a Weibull background distribution, CV = 2.00, $\epsilon = 0.1$, 60 background samples, and 50 site samples.

- For slightly skewed lognormal or Weibull distributions (CV = 0.1)

  - Tandem tests other than the State test have Type I error rates of about 0.06 to 0.08 rather than the target value of 0.05 for both small and large sample sizes.

  - The power of all tandem tests is very low when only 10 background and 10 site samples are used. The State test has better power than the other tandem tests, but its power is low except when P_r is near 1, and it has an elevated Type I error rate (0.11 versus the target value of 0.05).

  - The Wilcoxon-Quantile tandem test is significantly more powerful than the Wilcoxon test applied alone when 60 background and 50 site measurements are used. This extra power comes from using the Quantile test in addition to the Wilcoxon test. However, very little increase in power occurs by using the Quantile test when 10 background and 10 site samples are used.

  - The Quantile-Savage tandem test has about the same power as the Wilcoxon-Quantile test, and more power than the Wilcoxon-Savage test.

  - For the 10 percent contamination case, any tandem test that includes the Quantile test has more power than a tandem test that does not.

- For highly skewed lognormal or Weibull distributions (CV = 1 or 2)

  - Tandem tests other than the State test have Type I error rates of between 0.06 and 0.10 rather than the target value of 0.05.

- The power of all tandem tests is very low when only 10 background and 10 site samples are used. The power of the State test is very high, but it almost always incorrectly indicates that uncontaminated sites are contaminated.

- The power of tandem tests to detect 10 percent contamination is lower for highly skewed distributions (high CV) than for slightly skewed distributions (low CV).

In Summary:

- For the 10 percent contamination case:

  - The tandem tests have Type I error rates that range between 0.06 and 0.10.

  - The State test has better power than other tandem tests considered here, but its Type I error rate is highly elevated, particularly for highly skewed distributions.

  - All tandem tests have low power unless P_r is near 1.0, particularly when distributions are highly skewed.

4.3 Summary of Type I Error Rates

This section summarizes the Type I error rates for the State test (Section 4.3.1) and for all tests (Section 4.3.2).

4.3.1 State Test

The Type I error rates for the State test are summarized in Table 4.1. The State test has very large Type I error rates when CV ≥ 0.5, particularly when the number of site samples is less than about 30. The Type I error rate

is generally larger for the Weibull than for the lognormal distribution. The rate is less than 0.10 for either distribution when CV < 0.5 and the site sample size is 30 or larger.

These results indicate that if the CV is large enough, the State test will always indicate that a site is contaminated, even when the site has no or very little contamination. If the State test is used when CV is large, and if the decision to take remedial action is based solely on the results of the test, then unnecessary remedial actions will occur.

Table 4.1 Estimated Type I Error Rates for the State Test for Lognormal and Weibull Background Distributions[a]

	Lognormal				Weibull			
	Number of Site Samples				Number of Site Samples			
CV	10	30	50	70	10	30	50	70
0.1	0.11	0.07	0.06	0.055	0.12	0.07	0.06	0.05
0.5	0.24	0.07	0.06	0.055	0.43	0.09	0.06	0.05
1.0	0.50	0.09	0.06	0.055	0.81	0.53	0.33	0.22
1.5	0.65	0.20	0.08	0.060	0.97	0.97	0.97	0.97
2.0	0.81	0.55	0.36	0.230	0.99	1.00	1.00	1.00

[a]The State test does not explicitly take into account the number of background samples. Hence, the results in this table are the same for all background sample sizes in the Monte Carlo study.

Figures 4.37-4.40 show power curves and Type I error rates for the State test as well as for each of the three component parts of the test (upper confidence limit, frequency of exceedance and magnitude of exceedance). The Type I error rates for the two exceedance tests are conditional rates. That is, the Type I error rate for the percent exceedance test is the probability that test rejects H_o given that the first step of the State test (comparing the upper 95 percent confidence limit on the site mean to the background standard) does not reject H_o. Similarly, the Type I error rate for the magnitude of exceedance test is the probability that the test rejects H_o given that neither the first and second steps of the State test (upper limit test and the frequency of exceedance test) rejected H_o. The Type I error for the

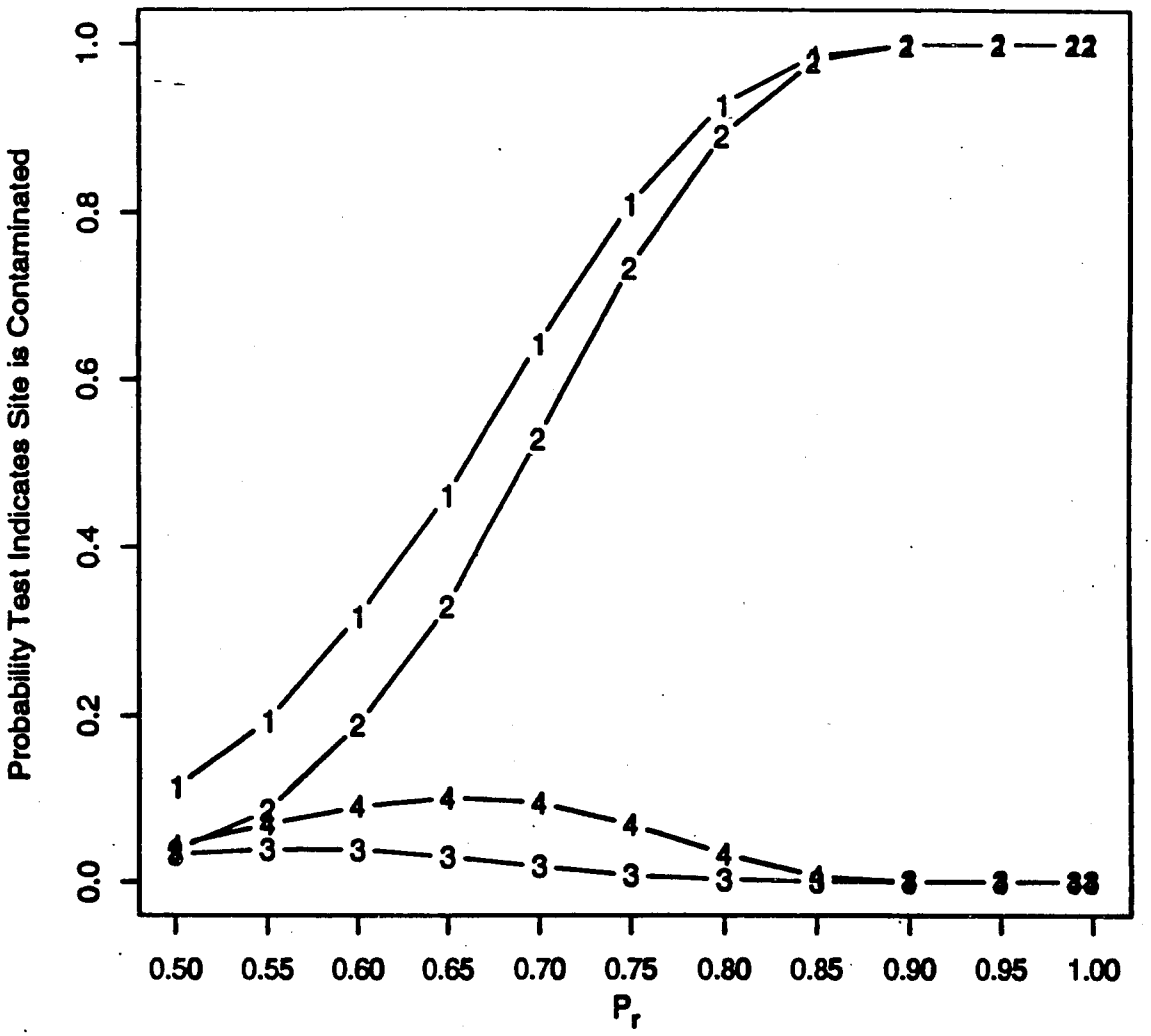
complete 3-step State test was computed as the sum of the Type I error for the confidence limit test and the two conditional Type I errors for the two exceedance tests.

Figure 4.37 indicates that for a slightly skewed distribution (CV = 0.1) when 10 background and 10 site samples are collected, each component part contributes about the same amount to the overall Type I error rate. However, Figure 4.38 indicates that when the number of samples is increased to 60 background and 50 site samples for the slightly skewed case, the overall Type I error rate comes entirely from the two exceedance tests. It is also interesting that the power of the overall test comes entirely from the exceedance test components when the magnitude of contamination, P_r, is relatively small. It isn't until P_r exceeds 0.70 that the upper 95 percent confidence limit test component contributes any power to the State test.

Figures 4.39 and 4.40 indicate that when only 10 background and 10 sites samples are collected from more highly skewed distributions (CV = 0.5 and 1), essentially all the Type I error for the overall test is contributed by the upper confidence limit test.

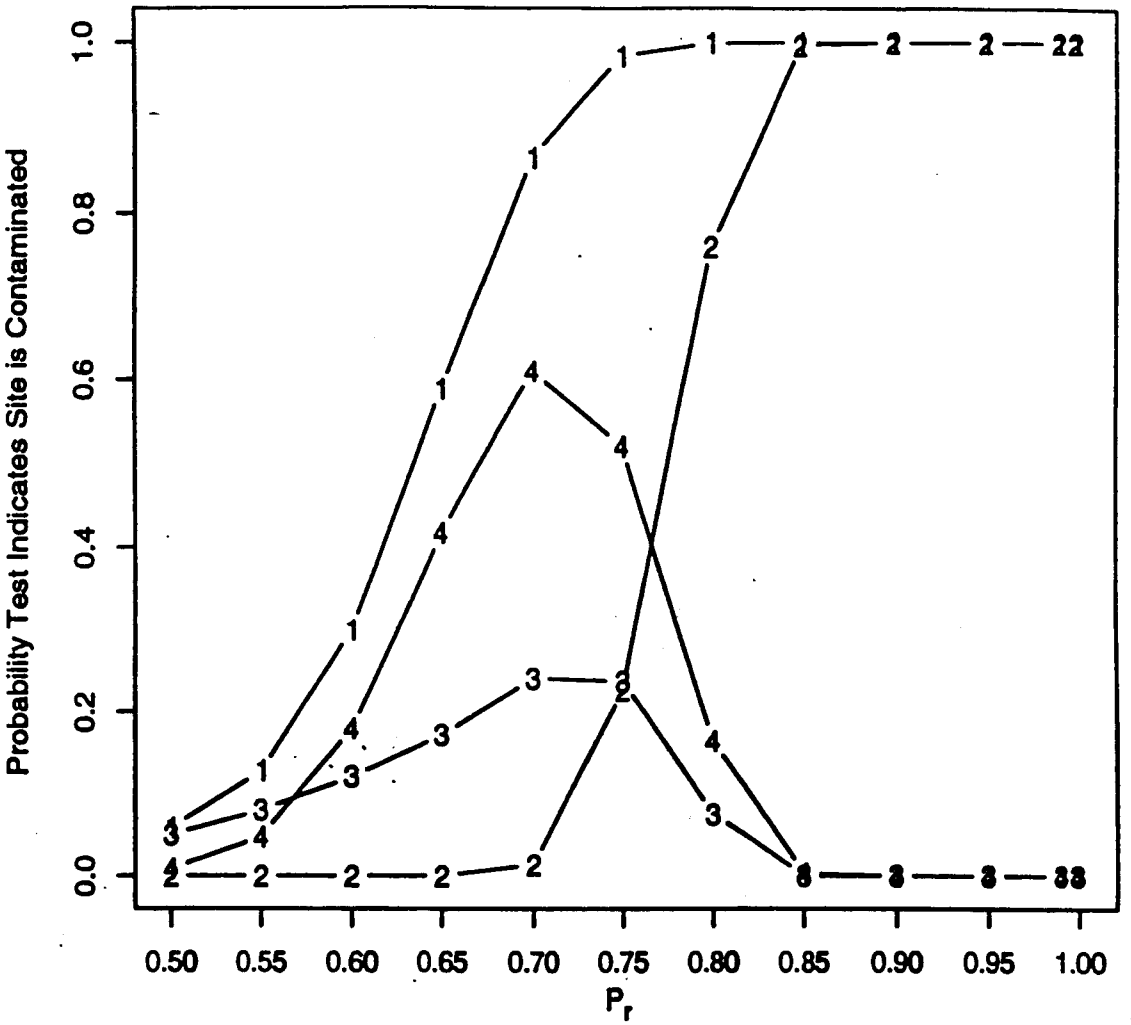
4.3.2 Other Tests

Elevated Type I error rates are not a serious problem for any other test studied in this report except perhaps Satterthwaite's t test. This conclusion is based on Table 4.2, which shows the range of estimated Type I error rates obtained from the Monte Carlo study for all tests. For Satterthwaite's test, the rather wide spread in rate levels (from 0.01 to 0.18) is believed to have occurred because the test was designed for use when data are normally distributed, which is not true for this study. Also the test is approximate in the sense that the achieved Type I error level may not be achieved even if the normality assumption is true. For the t test, the spread from 0.01 to 0.08 is also believed to have occurred because the assumptions of normality and equal variances for the background and site distributions are not true for this study.



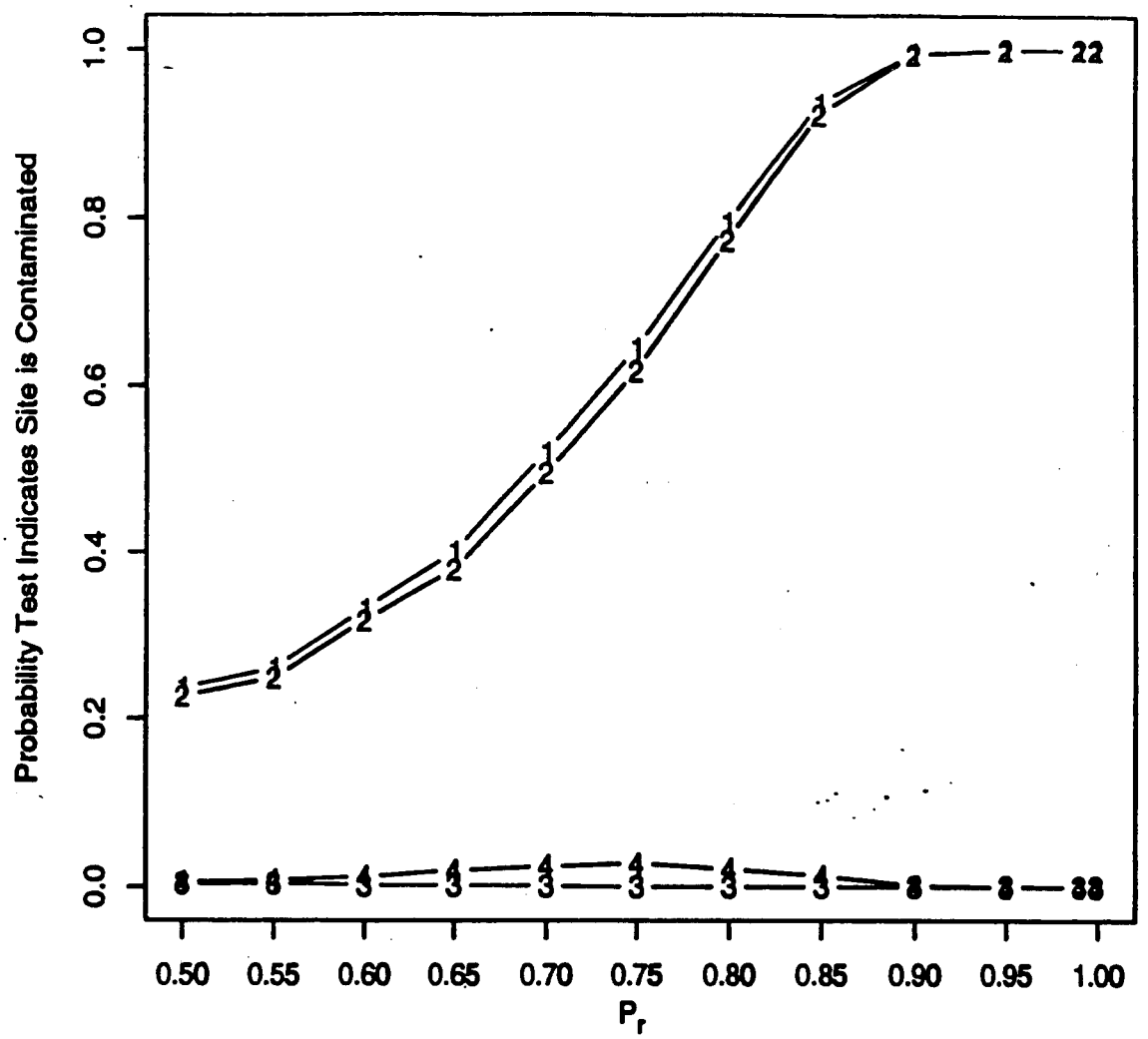
PARAMETERS		TEST	TYPE I ERROR
Distribution	= Lognormal	1 State	0.1138
CV	= 0.10	2 95% CI	0.0378
ε	= 1	3 % Exceedance	0.0325
Size (background)	= 10	4 Max Exceedance	0.0435
Size (site)	= 10		

Figure 4.37: Power curves for the state test and its component parts for the lognormal distribution, CV=0.10, ϵ = 1, 10 background samples, and 10 site samples. The figure illustrates the overall conclusion of the state test and the relative contributions of each of the three factors making up the state test (the overall conclusion, labelled 1, is the sum of the component parts, labelled 2, 3, and 4.



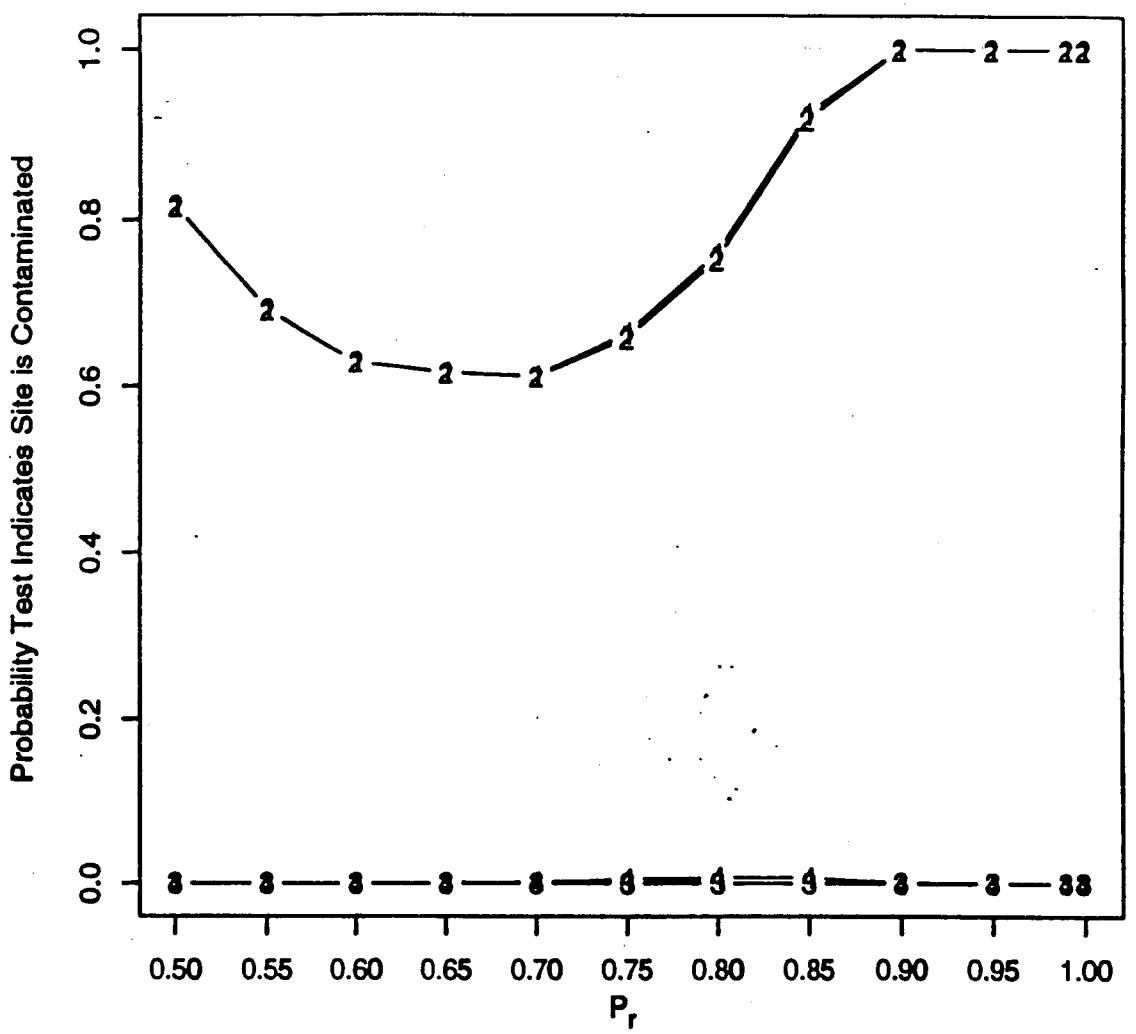
PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	State	0.0591
CV	= 0.10	2	95% CI	0 0000
ε	= 1	3	% Exceedance	0.0503
Size (background)	= 60	4	Max Exceedance	0.0088
Size (site)	= 50			

Figure 4.38: Power curves for the state test and its component parts for the lognormal distribution, CV=0.10, $\epsilon = 1$, 60 background samples, and 50 site samples. The figure illustrates the overall conclusion of the state test and the relative contributions of each of the three factors making up the state test (the overall conclusion, labelled 1, is the sum of the component parts, labelled 2, 3, and 4).



PARAMETERS			TEST	TYPE I ERROR
Distribution	= Lognormal	1	State	0.2366
CV	= 0.50	2	95% CI	0.2269
ε	= 1	3	% Exceedance	0.0040
Size (background)	= 10	4	Max Exceedance	0.0057
Size (site)	= 10			

Figure 4.39: Power curves for the state test and its component parts for the lognormal distribution, CV=0.50, $\epsilon = 1$, 10 background samples, and 10 site samples. The figure illustrates the overall conclusion of the state test and the relative contributions of each of the three factors making up the state test (the overall conclusion, labelled 1, is the sum of the component parts, labelled 2, 3, and 4).



PARAMETERS		TEST		TYPE I ERROR
Distribution	= Lognormal	1	State	0.8169
CV	= 2.00	2	95% CI	0.8169
ε	= 1	3	% Exceedance	0 0000
Size (background)	= 10	4	Max Exceedance	0 0000
Size (site)	= 10			

Figure 4.40: Power curves for the state test and its component parts for the lognormal distribution, CV=2.00, ϵ = 1, 10 background samples, and 10 site samples. The figure illustrates the overall conclusion of the state test and the relative contributions of each of the three factors making up the state test (the overall conclusion, labelled 1, is the sum of the component parts, labelled 2, 3, and 4).

The low value of 0.04 for the Wilcoxon test is believed to be due primarily to sampling error from using a finite (although large; 10,000) number of iterations of the computer code. Of course, sampling error contributes to the uncertainty of the estimated Type I error rates for all the tests. However, as the number of iterations was large, if a test really has a Type I error rate of 0.05, it is very unlikely that the Type I error rates obtained from the simulations would fall outside the range from 0.04 to 0.06.

The estimated Type I error rates for the Savage, Slippage, Quantile and Median tests were also below the target value of 0.05 for some contamination scenarios. The low rate of 0.03 for the Savage test may be an indication that the Type I error rate of the test is slightly sensitive to the degree of skewness of the distributions.

However, the rates of 0.01 or 0.03 for the Slippage, Quantile and Median tests do not necessarily mean the tests are performing poorly. Indeed, the target values for these tests were actually specified to be less than 0.05, the exact value depending on the sample size. This was done because the discrete nature of these tests makes it impossible for them to achieve a target value of exactly 0.05. To avoid the possibility of elevated Type I error rates, the authors decided to conduct these three tests such that the achieved Type I error rates would tend to fall below 0.05 rather than above it.

Table 4.2 Range of Estimated Type I Error Rates for All Tests

Test	Range of Estimated Type I Error Rates
State	0.05 to 1.00
Wilcoxon	0.04 to 0.06
Student's t	0.01 to 0.08
Satterthwaite's t	0.01 to 0.18
Savage	0.03 to 0.06
Slippage	0.01 to 0.05
Quantile	0.03 to 0.06
Median	0.01 to 0.06

4.4 Cases Where State Test Performs Well

This report indicates that the 3-step State test has extremely elevated Type I error rates when the CV is large and the number of site samples is small. However, the State test performs better in other situations. Table 4.3 summarizes the scenarios studied here for which the State test did not have seriously elevated Type I error levels (i.e., the levels are less than about 0.10) and also had the best power (over most values of ϵ and P_r) of any test considered. The results in Table 4.3 indicate that the State test meets these performance criteria when CV $\leq$ 0.5 and the number of site samples is at least 30, although the number of background samples can be as low as 10.

Table 4.3 Contamination Scenarios for which the State Test has Type I Error Levels Less than Approximately 0.10 and Generally Better Power than Other Tests[a]

Background Distribution	CV	Number of Background Samples				Number of Site Samples		
Lognormal	0.1	10	30			30	50	70
	0.5	10				30	50	70
Weibull	0.1	10	30	60	90	30	50	70
	0.1	30				120		
	0.5	10				50	70	

[a] Considers results for whole range of values of ϵ and P_r.

5. DISCUSSION

The results of this study clearly show that no single test or combination of tests can be relied upon to never make a mistake. Moreover, no single test will perform well for all contamination scenarios and sample sizes. For example, the Wilcoxon Rank Sum test is one of the best performers when the site is contaminated throughout its extent, whereas the Quantile test performs better than the Wilcoxon test at detecting sites that have small areas of very high contamination when a large number of samples are taken in both the background and site areas. The easy-to-use slippage test does well at detecting such hot spots if 50 or more background and site samples are taken. In some situations the State test has equal or more power than other tests studied here, but when the background and site measurement distributions are highly skewed and only a few samples are collected, the State test has extremely high probabilities of declaring that an uncontaminated site is contaminated. The t test and Satterthwaite's t test appear to have no performance advantages over the Wilcoxon test. The performance of the Savage test is sometimes better than that of the t test, but it did not perform as well as the Wilcoxon test. The median test has no performance advantages over the Wilcoxon test.

An important caveat on the performance assessments is that the tests evaluated here do not make use of information in the data related to the spatial pattern of contamination or the correlation among measurements. Therefore, it is important that no test should be conducted without also displaying the background and site data graphically to compare visually how those data sets compare and the extent that they overlap. This graphical analysis should include plotting each measurement on a map at its sampling location so that patterns can be visually detected. Also, histograms, box plots, and simple one-dimensional scatter (line) plots (Chambers et al. 1983) of both the site and background data should be constructed and displayed side by side to compare their means, medians, percentiles, extreme values and patterns of variability. These, and other visual aids, are easily constructed using readily available statistical software and graphics packages.

Among the tests considered here, the 3-step State test is perhaps the most complicated test to conduct. Furthermore, its performance behavior is erratic in the sense that it is impossible to know whether the results of the test are valid unless one collects many samples and knows a great deal about the shape of the background and site distributions prior to conducting the test. Specificly, we have seen that this test has great potential for declaring an uncontaminated site as being contaminated when the background and site data distributions are highly skewed and only a few samples are taken. The test should not be used blindly, i.e., without taking sufficient samples to conclude with confidence that the test's Type I error is not excessively large. To obtain such confidence, perhaps 40 or more samples are required in both the background and site areas. Clearly, the use of visual aids as discussed above is particularly important if the State test is used. The option of taking additional samples to confirm the results of the State test should always be available.

To continue our discussion of the State test, suppose, for a moment, that the stakeholders reach consensus that a target Type I error rate larger than 0.05 is acceptable. For example, if a site is very small and the cost of remediating the site is low, it may be acceptable to have a probability of approximately say, 0.30 of declaring an uncontaminated site to be contaminated. In that case, suppose the number of samples collected and the skewness of the distributions are such that the State test is believed to have a Type I error rate of, say 0.30 instead of the nominal 0.05. Hence, one could decide to use the usual State test. However, one could also consider using the Wilcoxon test at a Type I error rate of 0.30 rather than 0.05. The choice of whether the State or Wilcoxon test should be used in this situation depends on the power curves of the two tests when the Type I error rate is specified to be 0.30. This question cannot be addressed rigorously here because this report has only evaluated the power of the Wilcoxon test for a Type I error rate of 0.05. Nevertheless, the results here indicate that when the power curve of the State test is U shaped (see e.g., Figures 4.8 and 4.11), the Wilcoxon test is the preferred test because its power curve is known from theory to be monotonicly non-decreasing as the contamination magnitude parameter P_r increases. Hence, the power curve of the State test

will necessarily be below that of the Wilcoxon test because both curves would start at 0.30 and the State test curve initially declines while the Wilcoxon curve does not decline. This example indicates that if a Type I error level greater than 0.05 is acceptable to the stakeholders, that does not necessarily mean that the State test should be used, i.e., it may be better to use the Wilcoxon test at the higher, specified Type I error rate.

It should also be noted that the State test does not explicitly consider the number of background samples that are used to estimate the background standard, say the 90th percentile of background. Hence, the State test ignores information about the estimation variance of the standard. That variance will decrease if more background measurements are collected. All tests considered here except the State test explicitly take into account the number of background measurements collected.

In addition to using statistical tests and graphical aids, the process of comparing the site to the background area includes determining the geographical boundaries of the background area itself. The sampling and data analysis procedures to define the background area are not discussed in this report. The broader site evaluation process also includes developing site-specific sampling and measurement protocols, sample collection and laboratory analysis procedures, and the field and laboratory quality assurance plan.

It is important that the planning and design aspects be conducted by teams whose members work together and use each others expertise to refine the sampling, analysis, and testing process. Geochemists, geologists, and other professionals must be involved to assure the validity of the background and site data for testing whether background standards have been exceeded. All stakeholders should be involved in planning and implementing the entire process.

Considering the above remarks, the authors believe that a reasonable strategy (process) for comparing a site to the background area is to:

- Specify that field and laboratory sampling methods and designs and procedures should be reviewed by geochemists, geologists, statisticians and other professionals to assure that the measurements used in the statistical tests are valid for the purpose and meet all quality requirements.

- Determine the Type I error rate that is acceptable to all stakeholders. Higher Type I error rates than were used in this simulation study may be acceptable if the monetary, ecological, and human costs of cleaning up an uncontaminated site are not large.

- Always use graphical aids to visually display the data and to help interpret the statistical test results

- Routinely use the Wilcoxon Rank Sum Test in preference to the State test

- Use both the Wilcoxon and Quantile tests in preference to the State test if an overall Type I error rate (for both tests combined) of about 0.08 or 0.09 is acceptable when both tests are conducted at the 0.05 Type I error rate. If either test is significant, this is evidence that the site may be contaminated.

- Use the State test only if sufficient data have been collected and prior analysis has been done to assure with the required level of confidence (specified by the stakeholders) that:

    - the distributions of the background and site measurements have CVs less than about 0.50

    - the background and site data can be modeled sufficiently well by lognormal distributions

    - enough samples are collected in both the background and site areas so that the Type I error rate of the State test does not exceed the established acceptable level

- the power of the State test is greater than the Wilcoxon test for the same Type I error rate.

Although the present study has provided useful information on test performance, there is much that remains to be learned. One area where additional evaluations could help is in developing tests or data analysis procedures that work well for censored data sets, i.e., for data sets that contain one or more values less than the detection limit or the limit of quantitation. WSDE (1993) makes recommendations for this situation and some Monte Carlo studies have been conducted (e.g., Millard and Deverel 1988 and Helsel 1990). However, to the authors' knowledge, the performance of these available tools in practice has not been thoroughly evaluated. Additional evaluations might also be done for a wider range of distribution shapes than are considered in this report.

It should also be noted that the performance assessments in this report are for the case of simple random sampling from both the background and site areas. This sampling plan is used because the assumption is made that there is no information available about any contamination patterns that exist (or might exist) at the site. The performance of tests reported in this document can be considered as a worst case that applies when there is no prior information about where contamination may exist at the site.

In practice, one may use statistical analysis methods such as geostatistics (Cressie 1991) on data collected previously at the site to search for possible spatial patterns of contamination. If information is available on possible patterns, that information can be used to design a field sampling plan that improves the probability of detecting or mapping contamination. For example, using a systematic grid of samples of different densities in different areas within the site may work better than simple random sampling. Adaptive sampling designs (Thompson 1992) should also be considered, wherein the procedure for selecting sample locations depends on measurements made on previously collected samples.

6. REFERENCES

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. 1983. *Graphical Methods for Data Analysis*, Duxbury Press, Boston, MA.

Conover, W.J. 1980. *Practical Nonparametric Statistics*, 2nd edition, John Wiley, NY.

Cressie, N. 1991. *Statistics for Spatial Data*, Wiley, NY.

DOE. 1993. *Hanford Site Background: Soil Background for Nonradioactive Analytes, Revision 1 of Volumes 1 and 2*, DOE/RL-92-24. U.S. Department of Energy, Richland, WA.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, NY.

Gilbert, R.O. and Simpson, J.C. 1990. "Statistical Sampling and Analysis Issues and Needs for Testing Attainment of Background-Based Standards at Superfund Sites," In *Proceedings of the Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs*, pp. 1-16. Pennsylvania State University, University Park, PA.

Gilbert, R.O. and Simpson, J.C. 1992. *Statistical Methods for Evaluating the Attainment of Cleanup Standards, Volume 3: Reference-Based Standards for Soils and Solid Media*, PNL-7409 Vol. 3, Rev. 1, Pacific Northwest Laboratory, Richland, WA.

Helsel, D.R. 1990. "Less Than Obvious, Statistical Treatment of Data Below the Detection Limit," *Environmental Science & Technology* 24(12):1766-1774.

Hsieh, H.K. 1988. "Savage Test," In *Encyclopedia of Statistical Sciences*, Volume 8, pp. 267-270, Wiley, NY.

Iman, R.L. and Conover, W.J. 1983. *A Modern Approach to Statistics*, Wiley, NY.

Johnson, N.L. and Kotz, S. 1970. *Continuous Univariate Distributions - 1 Distributions in Statistics*, Wiley, NY.

Johnson, R.A., S. Verrill, and D.H. Moore II. 1987. "Two-Sample Rank Test for Detecting Changes that Occur in a Small Proportion of the Treated Population," *Biometrics* 43:641-655.

Kalos, M.H., and P.A. Whitlock. 1986. *Monte Carlo Methods*, Wiley, NY.

Knuth, D.E. 1981. *Seminumerical Algorithms*, 2nd edition, Addison-Wesley, Reading, MA.

Land, C.E. 1975. "Tables of Confidence Limits for Linear Functions of the Normal Mean and Variance," In *Selected Tables in Mathematical Statistics*, pp. 385-419, American Mathematical Society, Providence, RI.

Lehmann, E.L. 1975. *Nonparametrics, Statistical Methods Based on Ranks*, McGraw-Hill, New York.

Millard, S.P. and Deverel, S.J. 1988. "Nonparametric Statistical Methods for Comparing Two Sites Based on Data with Multiple Nondetect Limits," *Water Resources Research* 42:2087-2098.

Mood, A.M., F.A. Graybill, and D.C. Boes. 1974. *Introduction to the Theory of Statistics, Third Edition*, McGraw-Hill, New York.

Press, W.H., B.P. Flannery, S.A. Teukalsky, and W.T. Vetterling, 1988. *Numerical Recipes in C*, Cambridge University Press, NY.

Rosenbaum, S. 1954. "Tables for a Nonparametric Test of Location," *Annals of Mathematical Statistics* 24:146-150.

Savage, I.R. 1956. "Contributions to the Theory of Rank Order Statistics-The Two-Sample Case," *Annals of Mathematical Statistics* 27:590-616.

Thompson, S.K. 1992. *Sampling*, Wiley, NY.

WSDE August, 1992. *Statistical Guidance for Ecology Site Managers*, Washington State Department of Ecology, Olympia, WA.

WSDE August, 1993. *Statistical Guidance for Ecology Site Managers, Supplement S-6, Analyzing Site or Background Data with Below-Detection Limit or Below-P08L Values (Censored Data Sets)*, Washington State Department of Ecology, Olympia, WA.

APPENDIX

SIMULATION METHODS

APPENDIX

SIMULATION METHODS

A.1 Determining Distribution Parameters

The computer simulation code generated background and site measurement values from lognormal and Weibull distributions. The shape and location of these distributions is determined by the distribution parameters, which are defined below. These parameters were determined algebraically using specified values of the mean and coefficient of variation, CV, of the distribution. The mean was arbitrarily set equal to 100 in all cases because it was determined empirically that the performance assessment results do not depend on the value of the mean. Specifically, simulations were conducted using mean values of 10, 1000 and 10,000. The performance assessment results were the same as when the mean was 100.

A.1.1. Lognormal Distribution

The lognormal distribution may be written as

$$f(x) = \delta[(x-\theta)(2\pi)^{1/2}]^{-1}\exp(-[\nu+\delta \log_e(x-\theta)]^2/2)$$

where $x \geq \theta$. For the background distribution, θ was set equal to zero, i.e. the two-parameter lognormal distribution was used. Using formulas found in Johnson and Kotz (1970, pp. 112-117) the other two parameters, ν and δ, can be determined once the mean and CV have been specified. The formulas for the mean and CV of the lognormal distribution are

$$\text{mean} = \exp[-\nu/\delta + 1/(2\delta^2)]$$

$$\text{CV} = [\exp(\delta^{-2})-1]^{1/2}$$

The values of δ and ν were determined using specified values of the CV and a mean value of 100 using the equations

$$\delta = [1/\log_e(CV^2+1)]^{1/2}$$

$$\nu = -\delta\log_e(100)+1/(2\delta)$$

A.1.2 Weibull Distribution

The Weibull distribution can be written as

$$f(x) = c\alpha^{-1}[(x-\xi_0)/\alpha]^{c-1} \exp\{-[(x-\xi_0)/\alpha]^c\}$$

where $x > \xi_0$. For this simulation study, the value of ξ_0 for the background distribution was set to zero.

To determine values for parameters α and c, we first specified a value for the CV and a mean of 100, then solved numerically the following two equations (from Johnson and Kotz 1970, pp. 250-255):

$$\text{mean} = \alpha\Gamma(1+1/c)$$
$$CV = [\alpha\Gamma(1+2/c) - \Gamma^2(1+1/c)]^{1/2}/[\alpha\Gamma(1+1/c)]$$

where Γ is the Gamma function (Press et al. 1988, page 167).

A.2 Simulating Random Normal Values

Random normal values were obtained using the Box-Muller method (Kalos and Whitlock 1986, page 48). This method uses two random values from a Uniform (0,1) distribution to generate two independent standard normal values. Random uniform values between zero and one were obtained using a subtractive method proposed by Knuth (1981) and described by Press at al. (1988).

The procedure for generating normal values is not defined for just any pair of random uniform values. If we consider the two uniform values as defining a point in the x-y plane, then that point must lie within a circle centered at (0,0) with a radius of 1. The transformation between the two uniform values, x_1 and x_2, given by

$$y_1 = (-2 \log_e x_1)^{1/2} \cos 2\pi x_2$$

$$y_2 = (-2 \log_e x_1)^{1/2} \sin 2\pi x_2$$

is used to generate two random normal values, y_1 and y_2, for each pair of random uniform values, x_1 and x_2.

In summary, the procedure is to obtain a pair of uniform values, check that the point defined by the pair is in the unit circle, generate a replacement pair if necessary, and apply the above transformation to obtain the pair of standard normal values.

A.3 Simulating Random Lognormal Values

To obtain random lognormal values, we made use of the cumulative distribution function (cdf) of the lognormal distribution:

$$F(x) = \alpha + \beta \log_e(z-\gamma)$$

where z is a random normal value. This cdf was used to obtain the quantile function, which is the inverse of the cdf. The quantile function for the lognormal distribution is

$$x = \exp[(z-\alpha)/\beta] + \gamma \qquad \text{(A.1)}$$

Hence, a lognormal value x was obtained by first generating a random normal value z using the method described in Section A.2 and then using Equation (A.1).

A.4 Simulating Random Weibull Values

The probability integral transform method (Mood et al. 1974, page 202) was used to obtain random Weibull values. The cumulative Weibull distribution function is

$$F(x) = 1.0 - \exp\{-[(x-\alpha)/\beta]^{1/\gamma}\}$$

The quantile function for the Weibull distribution is

$$x = \alpha + \beta[-\log_e(1.0-u)]^{\gamma} \tag{A.2}$$

which was used to obtain Weibull values. In summary, a random Weibull value, x, was determined by first generating a random uniform value, u, and then using Equation (A.2).

DISTRIBUTION

No. of Copies

OFFSITE

12 DOE/Office of Scientific and Technical Information

2 A. Wallo, III*
U.S. Department of Energy
Environment, Safety & Health
Forrestal Building (EH-232)
Washington, DC 20585

B.K. Thatcher*
U.S. Department of Energy
Rocky Flats Office
P.O. Box 928
Golden, CO 80402-0928

R.J. Schassburger
U.S. Department of Energy
Rocky Flats Office
P.O. Box 928
Golden, CO 80402-0928

N. Blakley*
State of Washington Department of Ecology
Toxics Cleanup Program
P.O. Box 47600
Olympia, WA 98504-7600

W. Busby
EG&G Rocky Flats Inc.
Rocky Flats Plant
P.O. Box 464
Golden, CO 80402-0464

K. Campbell*
Los Alamos National Laboratory
MS/F600
Los Alamos, NM 87545

E.Y. Chai
Shell Development Company
Westhollow Research Center
P.O. Box 1380
Houston, TX 77251-1380

No. of Copies

C. Cline
State of Washington Department of Ecology
P.O. Box 47600
Olympia, WA 98504-7600

D. Delistraty
State of Washington Department of Ecology
North 4601 Monroe, Suite 202
Spokane, WA 99205

D.A. Edwards
Exxon Biomedical Sciences, Inc.
Mettlers Road
East Millstone, NJ 08875-2350

D.R. Elle
U.S. Department of Energy
Nevada Operations Office
P.O. Box 98518
Las Vegas, NV 89193-8518

J. Erickson
State of Washington Department of Health
P.O. Box 47827
Olympia, WA 98504

M. Gansecki
U.S. Environmental Protection Agency
999 18th Street, Suite 500
Mail Code HWM-HW
Denver, Colorado 80202

3 L. Gadbois*
U.S. Environmental Protection Agency
712 Swift Blvd., Suite 5
Richland, WA 99352

No. of
Copies

S. Holtzman
Brookhaven National Laboratory
Associated Universities, Inc.
Building 490A
Upton, L.I., NY 11973

P.S. Innis
U.S. Environmental Protection Agency
712 Swift Blvd., Suite 5
Richland, WA 99352

K.M. Keer
EnecoTech Inc.
303 Camp Craft Road, Suite 100
Austin, Texas 78746

R.G. Knowlton, Jr
Sandia National Laboratories
Albuquerque, NM 87185-5800

B. Lavelle
U.S. Environmental Protection Agency
999 18th Street, Suite 500
Mail Code HWM-FF
Denver, Colorado 80202

G.E. Powers
U.S. Nuclear Regulatory Commission
NLS-139
Washington, DC 20555

B. Ramsey
Systematic Management Services, Inc.
20201 Century Boulevard
Germantown, MY 20874

N.P. Ross*
U.S. Environmental Protection Agency
401 M Street, SW
Washington, DC 20460

No. of
Copies

2 J. Schieffelin*
Hazardous Waste Facility Unit
Colorado Department of Health
4300 Cherry Creek Drive South
Denver, CO 80222-1530

M.A. Siders
EG&G Rocky Flats
P.O. Box 464
Golden, CO 80402-0464

W. Staubitz
U.S. Geological Survey
1201 Pacific Avenue
Suite 600
Tacoma, WA 98402

RuAnne Thomas*
Westinghouse Savannah River Company
P.O. Box 616
Aiken, SC 29802

R.C. Tuckfield*
Westinghouse Savannah River Company
P.O. Box 616
Aiken, SC 29802

J. Warren*
U.S. Environmental Protection Agency
401 M Street, SW
Washington, DC 20460

G.E. Watkin
Harding Lawson Associates
105 Digital Drive
Novato, CA 94949

D.R. Weier*
Statistical Applications
EG&G Rocky Flats Inc.
Rocky Flats Plant
P.O. Box 464
Golden, CO 80402-0464

No. of Copies	
	D. Wells State of Washington Department of Health P.O. Box 47827 Olympia, WA 98504
4	DOE Richland Field Office
	C.E. Clark J.K. Erickson A. Foote R.K. Stewart*
11	Westinghouse Hanford Company
	C.J. Chou* T. LeGore (5*) D.J. Hoff J.D. Hoover N.K. Lane S.W. Petersen F.A. Ruck

No. of Copies		
36	Pacific Northwest Laboratory	
	D.J. Bates	K7-34
	P.G. Doctor	K6-98
	R.O. Gilbert (10*)	K7-34
	N.L. Hassig	K7-34
	J.W. Hardin (10*)	K7-34
	R.E. Jaquish	K1-30
	J.P. McKinley	K3-61
	R.F. O'Brien	K7-34
	B.A. Pulsipher	K7-34
	F.M. Ryan	K7-70
	A.J. Schilk	P8-01
	J.M. Thomas	K6-63
	Publishing Coordination Technical Report Files (5)	

*Indicates one copy with microfiche attachment.