United States
Environmental Protection
Agency

Performance of Statistical Tests for Site Versus Background Soil Comparisons When Distributional Assumptions Are Not Met

Technology Support Center Issue

RESEARCH AND DEVELOPMENT

EPA/600/R-07/020
March 2007
www.epa.gov

Performance of Statistical Tests for Site Versus Background Soil Comparisons When Distributional Assumptions Are Not Met

Technology Support Center Issue

by

Evan J. Englund

U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Environmental Sciences Division
Characterization and Monitoring Branch
Las Vegas, NV 89119

U.S. Environmental Protection Agency
Office of Research and Development
Washington, DC 20460

141pos07

Technology Support Center Issue

Performance of statistical tests for site versus background soil comparisons when distributional assumptions are not met

Evan J. Englund

Abstract

Statistical distributions of site and background soil samples often do not meet the assumptions of statistical tests. This is true even of "non-parametric" tests. This paper evaluates several statistical tests over a variety of cases involving realistic population distribution scenarios and sampling schemes. Over the range of cases, performance was erratic for most tests. When planning a project, the sampling scheme must be designed together with the statistical test, and the choice of test may vary depending on which scenario best matches the conceptual model for the site.

Introduction

This report began as an inquiry into the ability of the Wilcoxon Rank Sum (WRS) test to distinguish between background and contaminated soils in the "Choccolocco Corridor" (CC) area of Fort McClellan Army Base in Alabama. The WRS test was performed under the null hypothesis that there is no difference between the concentrations in the site and concentrations in the background reference area. The alternative hypothesis was that the site concentrations are greater than those in the background reference area. A positive test result for any analyte indicated the presence of site contamination that would require further evaluation for possible cleanup. Concerns about the performance of the WRS test led to an investigation of the test under conditions similar to those in the Choccolocco Corridor.

Poor WRS performance under some conditions in the initial tests from the CC raised the obvious question: Are there better alternatives? This question prompted the expanded and more generalized investigation presented here, which compares the WRS test directly with several other tests.

Student's t is the most well-known and widely-used statistical test for comparing two samples sets. However, Student's t is based on very specific assumptions about the populations from which the samples are drawn. These include the assumption that the population variances are equal, and the assumption that the shape of both populations is normal. Welch's test (also called Satterthwaite's t or the unequal-variance t) is a modified Student's t test that attempts to correct for unequal variances, though it still requires the assumption of normality. The Wilcoxon Rank Sum test (WRS) is a non-parametric alternative to the Student's t test that assumes that the population shapes are identical, though not necessarily normal. The WRS test and the Welch's t

test are often suggested when their assumptions appear to be more accurate than those of the Student's t test. (EPA 2002, EPA 2006).

The literature is confusing at best regarding the relative merits of the WRS test and the Student's t test. Hodges and Lehman (1956) argue on theoretical grounds in favor of the WRS, noting that it is never much less efficient than Student's t (where less efficient means needing more samples to get the same performance) but can be infinitely more efficient.

A sampling of papers that compare the tests in the context of specific applications or specific distributions produces mixed conclusions. Bridge and Sawilowsky (1999) conclude that WRS is better for small samples of skewed distributions; they recommend WRS when distributions are unknown. Blair and Higgins (1980) found that WRS generally held power advantages over the t-test for various non-normal distributions. Potvin and Roff (1993) conclude that the WRS test is more powerful than the t-test for skewed distributions.

Johnson (1995), however, criticizes Potvin and Roff, pointing out that The Central Limit Theorem works in favor of Student's t, making it insensitive to modest departures from normality. Johnson also notes that the WRS test requires a strict assumption that the two distributions are identical in both shape and variance – something that is rarely tested. He concludes that if investigators use random sampling properly, "…parametric methods will ordinarily be adequate; if they are not, nonparametric methods will not protect them from sailing off course." Modarres et al. (2005) likewise note the strict distributional requirements of the WRS test (and rank-based nonparametric tests in general), claiming results of these tests are not valid when the assumptions are violated. For log-normal distributions, Zhou et al. (1997) reject both WRS and t, and propose an alternative test that calculates a z-statistic using maximum likelihood estimators.

The disparity among these papers is likely due to differences in the scenarios being investigated. No two of the papers compare statistical tests on identical distributions and distribution shifts; the conclusion to be drawn from the literature is that t-tests are superior in some cases, WRS in others. None of the papers deals with the kinds of distributions that we would expect to encounter when testing to distinguish between contaminated and uncontaminated soils.

Hypotheses

Statistical tests are framed in the form of a null hypothesis which is assumed to be true and an alternative hypothesis which is accepted as true only if the data strongly indicate that the null hypothesis is actually false.

Hypothesis tests favor the null hypothesis by setting stringent limits on the probability that the null hypothesis will be rejected simply by chance, when it is, in fact, true. This is called a false positive, or Type I error, and the limit on its probability is called the significance level of the test, or alpha. Most hypothesis tests are designed so that when the assumptions of the test are met, the test will perform as specified. In practical terms, that means in a large number of repeated trials, the fraction of false positive decisions should be very close to alpha.

Two alternative null hypotheses are considered in this investigation. Adopting the terminology from EPA (2002a), we have:

- Test Form 1 – The null hypothesis is that the distribution of concentrations in the site population is identical to that of the background. In this case, no further action will be taken unless the statistical test indicates that the site measurements are sufficiently higher than the background measurements that they are unlikely to have occurred by chance. In formal terms, the p-value of the test, which is the estimated probability that the result could have occurred by chance, must be less than alpha. This test form directly controls the costs of taking unnecessary action, while the environmental risks from failure to detect contamination are determined by the sample size and the population variability.
- Test Form 2 – The null hypothesis is that the site mean concentration exceeds the background mean concentration by a specified threshold amount or more. In this case, action will be taken unless the difference between site and background measurements is significantly less than the threshold. This test form controls environmental risks by setting an upper limit on the amount of contamination that can be missed by sampling, while the costs of unnecessary action are a function of sample size and variability.

Statistical Tests and Assumptions

The statistical tests evaluated here include Wilcoxon Rank Sum, Student's t, Welch's t, the quantile test, the quantile test combined with the WRS test, and the difference in sample means compared to a specified threshold difference. In addition, Welch's t was performed on log-transformed data. The test descriptions in italics below are taken from EPA (2006).

The Two-Sample Student's t Test

Purpose: Test for a difference or estimate the difference between two population means when it is suspected the population variances are not equal.

Data: A simple or systematic random sample $x_1, x_2, \ldots, x_m$ from the one population, and an independent simple or systematic random sample $y_1, y_2, \ldots, y_n$ from the second population.

Assumptions: The two populations are independent. If not, then it is possible that a paired method could be used. Both are approximately normally distributed or the sample sizes are large (m and n both at least 30). If this is not the case, then a nonparametric procedure is an alternative.

Limitations and Robustness: The two-sample t-test with unequal variances is robust to moderate violations of the assumption of normality. The t-test is also not robust to outliers because sample means and standard deviations are sensitive to outliers.
(U.S.EPA 2006, Section 3.3.1.1.1)

The Two-Sample t-Test (Welch-Satterthwaite: Unequal Variances)

Purpose: Test for a difference or estimate the difference between two population means when it is suspected the population variances are not equal.

Data: A simple or systematic random sample x1, x2, . . . , xm from the one population, and an independent simple or systematic random sample y1, y2, . . . , yn from the second population.

Assumptions: The two populations are independent. If not, then it is possible that a paired method could be used. Both are approximately normally distributed or the sample sizes are large (m and n both at least 30). If this is not the case, then a nonparametric procedure is an alternative,

Limitations and Robustness: The two-sample t-test with unequal variances is robust to moderate violations of the assumption of normality. The t-test is also not robust to outliers because sample means and standard deviations are sensitive to outliers.
(U.S.EPA 2006, Section 3.3.1.1.2)

The Wilcoxon Rank Sum Test

Purpose: Test for a difference between two population means. The Wilcoxon Rank Sum test, applied with the Quantile test, provides a powerful combination for detecting true differences between two population distributions.

Data: A random sample $x_1, x_2, \ldots, x_m$ from one population, and an independent random sample $y_1, y, \ldots, y_n$ from the second population.

Assumptions: The validity of the random sampling and independence assumptions should be verified by review of the procedures used to select the sampling points. The two underlying distributions are assumed to have approximately the same shape (variance) and that the only difference between them is a shift in location. A qualitative test of this assumption can be done by comparing histograms.

Limitations and Robustness: The Wilcoxon signed rank test may produce misleading results if there are many tied data values. When many ties are present, their relative ranks are the same, and this has the effect of diluting the statistical power of the Wilcoxon test. If possible, results should be recorded with sufficient accuracy so that a large number of tied values do not occur. Estimated concentrations should be reported for data below the detection limit, even if these estimates are negative, as their relative magnitude to the rest of the data is of importance. If this is not possible, substitute the value DL/2 for each value below the detection limit providing all the data have the same detection limit. When different detection limits are present, all data could be censored at the highest detection limit but this will substantially weaken the test. A statistician should be consulted on the potential use of Gehan ranking.
(U.S.EPA 2006, Section 3.3.2.1.1)

The Quantile Test

Purpose: Test for a shift to the right in the right-tail of population 1 versus population 2 This may be regarded as being equivalent to detecting if the values in the right-tail of population 1 distribution are generally larger than the values in the right-tail of the population 2 distribution.

Data: A simple or systematic random sample, $x_1, x_2, \ldots, x_n$, from the site population and an independent simple or systematic random sample, $y_1, y_2, \ldots, y_m$, from the background population.

Assumptions: The validity of the random sampling and independence assumptions is assured by using proper randomization procedures, which can be verified by reviewing the procedures used to select the sampling points.

Limitations and Robustness: Since the Quantile test focuses on the right-tail, large outliers will bias results. Also, the Quantile test says nothing about the center of the two distributions.

Therefore, this test should be used in combination with a location test like the t-test (if the data are normally distributed) or the Wilcoxon Rank Sum test.
(U.S.EPA 2006, Section 3.3.2.1.2)

The Sample Means Test

The difference in arithmetic means of the two samples is tested against a threshold value. If the site mean minus the background mean is greater than the threshold, then the site is considered contaminated.

The arithmetic means test is not a true statistical test in the sense that it does not attempt to control either alpha or beta. It is perhaps better described as a decision rule. The test is error-neutral. It is the same whether using test form 1 or 2. If the error distributions are symmetrical, or the sample sizes are large, then false positive and false negative error rates will be equal. It has some potential advantages: it is simple; it directly tests the parameter of concern; and because it uses no distributional information, it can be performed equally well with composite data at lower analytical costs.

Alternately, the sample means test could be considered a degenerate form of the t-test. It is equivalent to setting the alpha value of the t-test at the sample mean threshold to 0.5. In this case the value of t becomes zero, and the critical value of the test is simply the threshold.

Tests on Transformed Data

Log transforms are not generally recommended when analyzing environmental data (EPA, 1997). Two problems are common. Upper confidence limits calculated by Land's method, while theoretically correct if the true distribution of the population is exactly log-normal, may be much too high when the population is only approximately log-normal. Also, it can easily be demonstrated that in a one-sample test, comparing a statistic calculated on log-transformed data against a log-transformed threshold limit can produce biased, non-protective decisions. To test whether the same is true for a two-sample test, Welch's t on log-transformed data is included in this study for comparison. The Welch's t test is chosen here instead of the Student's t test because its assumptions are less stringent.

For test form 1, the approach is straight-forward: take the logarithms of the two data sets and perform the Welch's t test. However, test form 2 presents complications when the significant difference being tested is greater than zero. With untransformed data, the significant difference is tested by either subtracting the difference from the site data or by adding the difference to the background data, and then performing the test. Either way the result is the same. But the results are not the same if the data are transformed before the test. Subtracting the significant difference from the site data yields negative values that cannot be transformed, so the only real option is to add the significant difference to the background values, transform the data, and perform the test.

An alternative approach is to transform the original data before adding the significant difference. There are a wrong way and two right ways to do this. The wrong way is to log-transform the sample data and then to add the log of the significant difference (equal to the width of the gray region – in this case, 50) to the transformed background values. That is the equivalent of multiplying the background values by 50. The right ways differ in the interpretation of the performance objectives (see Figure 1 below). We might be testing for a 50% increase, in which case the log of 1.5 would be added to the transformed background data. Or we might be testing for an increase of 50 units over an unknown background mean; then the added value would be the log of the ratio: background mean plus 50 over the background mean. The latter is used in this paper.

Combined Tests

The U.S. EPA (EPA, 1992, 1996) recommends using the quantile test in conjunction with the Wilcoxon Rank Sum test. The reasoning is that the WRS test is robust to the presence of outliers, which also makes it insensitive to localized contamination that would show up in the site distribution as an increase or shift in the upper tail. The quantile test looks only at the upper tail and should detect such an increase if it has occurred. The decision logic is that if either test indicates further action, then further action will be taken.

Evaluation of the performance of a combination of two tests is beyond the scope of this paper. To perform a quantile test requires choosing two parameters – the significance level alpha and the quantile to test (i.e., the upper 10% of the distribution). A full evaluation would require testing various combinations of alpha and quantile for the quantile test, along with various levels of alpha for the WRS. This paper shows performance for a single combination as an example. Both alpha values are set at 0.05, and the quantile is the upper 10%.

It has been suggested (EPA 1992) that a "hot sample" test be included in combination with the WRS and quantile tests. This involves choosing an upper threshold value such that the site is declared contaminated if any site measurement exceeds the threshold. That reference, however, was not explicit as to how to choose or calculate a hot sample threshold. This paper does not evaluate hot sample tests, alone or in combination. That is left, along with a full evaluation of the combined WRS - quantile test, as a subject for future research.

Decision Objectives

Comparing statistical tests can be an "apples and oranges" exercise, because the tests operate on different characteristics of the samples. Both the Student's t and the WRS tests set out to test exactly the same thing. Both assume that the two populations being sampled have identical shapes and variances, and that the populations differ only by a location shift. "Location shift" usually denotes a change in mean, but the shape and variance constraints imply something stronger – not only the mean, but every quantile of the population, including the median, are all shifted by exactly the same amount.

When the tests are performed, they tell us different things. The t-tests actually estimate the population means and might conclude that the mean of population A is significantly higher than the mean of population B. The WRS test does not conclude anything about means – only that the

relative rankings of population A are greater than the relative rankings of population B. As Modarres et al. (2005) noted, when the WRS assumptions are violated, a difference in rankings might be detecting a change in shape or variance instead of the intended location shift.

Similarly, the t-test performed on log transformed data compares the means of the logs, which tells us nothing about the arithmetic population means. If the log-transformed populations are strictly normal, then the t-test on log-transformed data is a test of medians because the mean of a normal distribution of logs is the median of the corresponding arithmetic values. The simulated site populations that will be used in this investigation are bimodal log-normal distributions – the background and contaminant components of the site distribution are each log-normal, but the combination is only log-normal if the entire site is contaminated.

As pointed out earlier, this investigation began as an evaluation of the performance of the WRS test and expanded to include comparisons of the WRS test with some alternative tests. For site-background comparisons, the WRS test is used (in the author's experience) only as a non-parametric substitute for a t test when the hypothesis of normality is rejected; therefore it is reasonable to assume in such cases that the purpose of performing either test is to compare the population means. In this investigation, we choose to define the objective of performing a test to be the detection of a significant difference in population mean, regardless of what the test itself may actually be measuring. There may be good reasons to compare other population parameters, but such comparisons are beyond the scope of this paper.

We adopt a pragmatic approach toward hypothesis testing. Contaminated sites are not controlled experiments. In the real world, if we evaluate sets of sample data and find that there are no available statistical tests whose assumptions are perfectly met, we do not have the luxury of not making a decision. We do not abandon tests simply because the underlying assumptions are not met. After all, t-tests and the WRS test are widely used at least in part because they are robust to moderate violations of the assumptions. However, we do abandon the notion of formal statistical inference. P-values can not be interpreted as exact probabilities; therefore, alpha values are not true significance levels. The tests in effect become mechanical decision algorithms or "decision rules" – the data go in and a decision comes out. Our concern is not about which test is formally "correct", but about which test performs best under the circumstances.

Performance Evaluation

Performance was evaluated by computer simulations designed to mimic the sampling and decision process as closely as possible. For this investigation, the site and reference area are assumed to be sampled randomly, and that the comparison of samples is intended to answer the question: Is the mean concentration of the site significantly higher than the mean concentration of background? The approach is to produce realistic simulated "populations" for the background and the site, to re-sample and test the populations numerous times, and to record the decision outcomes. Each decision is classified as "action" or "no action," and the decimal fraction of "action" decisions over a number of trials is plotted as an estimate of the true performance probability for the test. A performance curve is constructed by repeating the process with a series of simulated site populations with increasing mean concentrations.

Hypothesis tests are run at specified levels of alpha, which control (or attempt to control) the false rejection rate at a specified threshold value. Here, the threshold for test form 1 is set at zero difference. An alpha value of 0.05 would represent a 0.05 probability of cleanup when the site

mean concentration is actually identical to the background mean concentration. A DQO diagram (Figure 1) is used during project planning to quantify the desired performance of a sampling/decision-making effort. The alpha "control point" for test form 1 is indicated by the lower corner on the line bounding the left side of the gray region. The alpha control point for test form 2 is the upper corner on the right boundary. For this study, the threshold difference for test form two is set at 50, or 50% higher than the nominal background mean of 100. An alpha of 0.05 would represent a 0.95 probability of cleanup at this threshold.

A statistical test that is performing correctly because all of the assumptions are met should always pass through the alpha control point. Whether it passes through the other control point is a function of the variability of the populations and the number of samples.

For this investigation, the threshold value for the sample mean test is set at the midpoint of the gray region, or 25. This is just a first approximation to make the performance of this test comparable to the other tests. This is basically a Central Limit theorem approximation - if the sample size is large enough to make the distribution of errors about the mean look approximately normal, then the mean will approximately equal the median. When the true value equals the threshold, the decision probability will be 0.5 either way; thus the performance curve will pass through the center of the gray region.

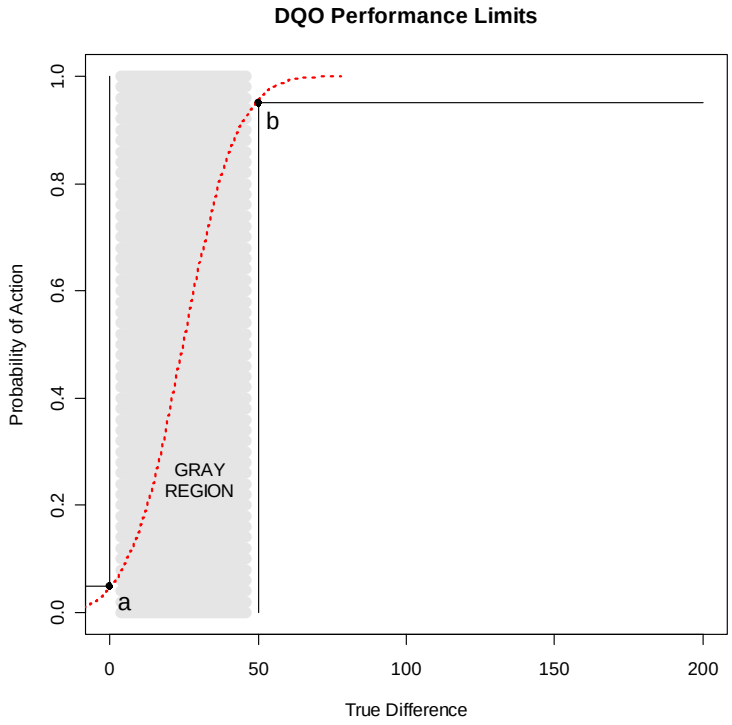


Figure 1. Example DQO diagram specifying performance requirements for a two-sample test. (a) The alpha control point for test form 1. (b) The alpha control point for test form 2. The dotted red line is a performance curve for an acceptable DQO design; "acceptable" because it stays within the boundaries of the gray region, or equivalently, within the control points. Different combinations of sample size and method, analytical method, and statistical test can produce numerous acceptable designs. In the typical DQO process, the lowest cost of these would be "optimal."

Performance results for the simulations in this study are presented in the form of simplified DQO diagrams. To reduce visual clutter when comparing several performance curves on the same graph, the gray region and most of its boundaries are removed. Only the corner sections of the bounding lines remain to indicate the design control points. The alpha control point for the test form being illustrated is shown in black; the other in gray, as in Figure 2. All results will be shown in this fashion; as pairs of plots with test form 1 on the left and test form 2 on the right.

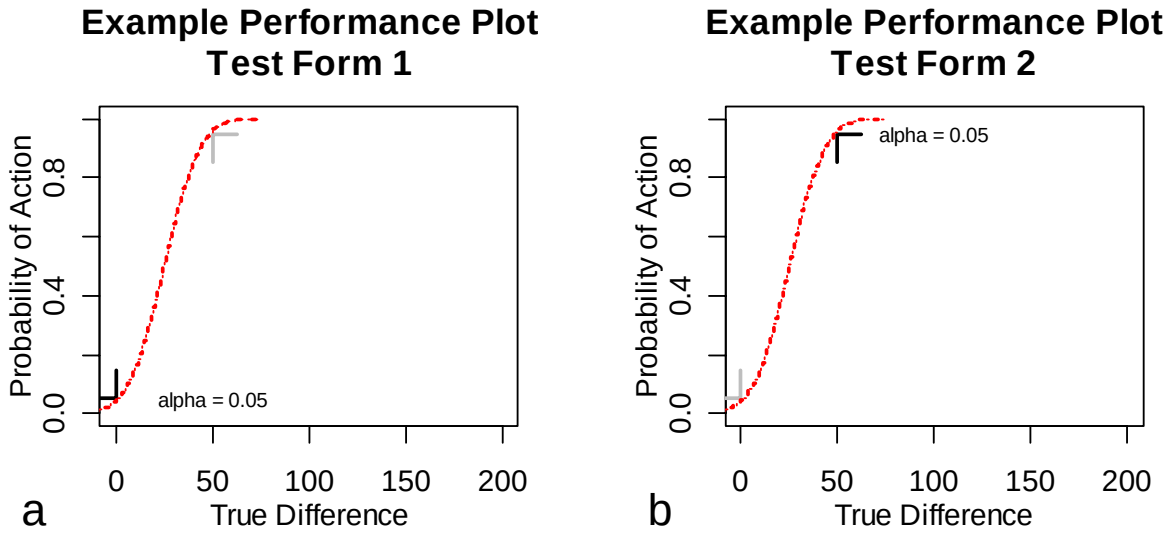


Figure 2. Simplified DQO diagram used to present results in this paper. The gray region and most of the gray region boundaries in Figure 1 have been removed to reduce visual clutter. The gray and black "ticks" remain to show the target points. The alpha point for formal statistical tests is shown in black for the appropriate test form.

The Simulation Algorithm

1. Generate a simulated background population: 100,000 random values log-normally distributed with arithmetic mean = 100 and a specified log standard deviation.
2. Generate a simulated contaminant population of 100,000 values containing both zero values and contaminated values in a specified proportion. The non-zero contaminated values are log-normally distributed with a specified log standard deviation and an arithmetic mean such that the arithmetic mean of the entire contaminant population, including the zeros, is equal to D, a user-specified arithmetic "true difference" value. D is initially set to zero—no contamination.
3. Add the two distributions to create the site population.
4. Draw a simple random background sample of size nb and a simple random site sample of size ns. The sample sizes, nb and ns, are user-specified and may differ.
5. Perform a Wilcoxon Rank Sum test with the null hypothesis: site = background, versus the alternative hypothesis: site > background. Record the p-value.
6. Perform a two-sample Student's t test as in step 5.
7. Repeat steps 4 through 6. The number of repetitions is user-specified, and is determined by trial-and-error to produce acceptably smooth performance curves. Evaluate the performance of the tests by computing the fraction of repetitions with positive action decisions (p-value<alpha for test form 1, or p-value>alpha for test form 2).
8. Repeat steps 2 through 7, for a series of specified D values.
9. Plot the performance for each test as a function of the true difference in means.

Simulations and graphics were done with R software (R Development Core Team, 2005).

The Background Population

Trace metal concentrations in natural soils are usually positively skewed and are often approximately log-normal. Background sample data from the Fort McClellan Choccolocco Corridor were used as a basis for choosing a reasonable background population distribution for this study. Normal probability plots of the log concentrations for six metals in the Choccolocco Corridor background samples are shown in Figure 3. A perfectly normal, or in this, case log-normal, distribution would plot on a straight line. All but mercury (Hg) appear approximately log-normal. Mercury is anomalous – possibly bimodal. For purposes of this investigation, we will assume that "typical" background distributions are log-normal.

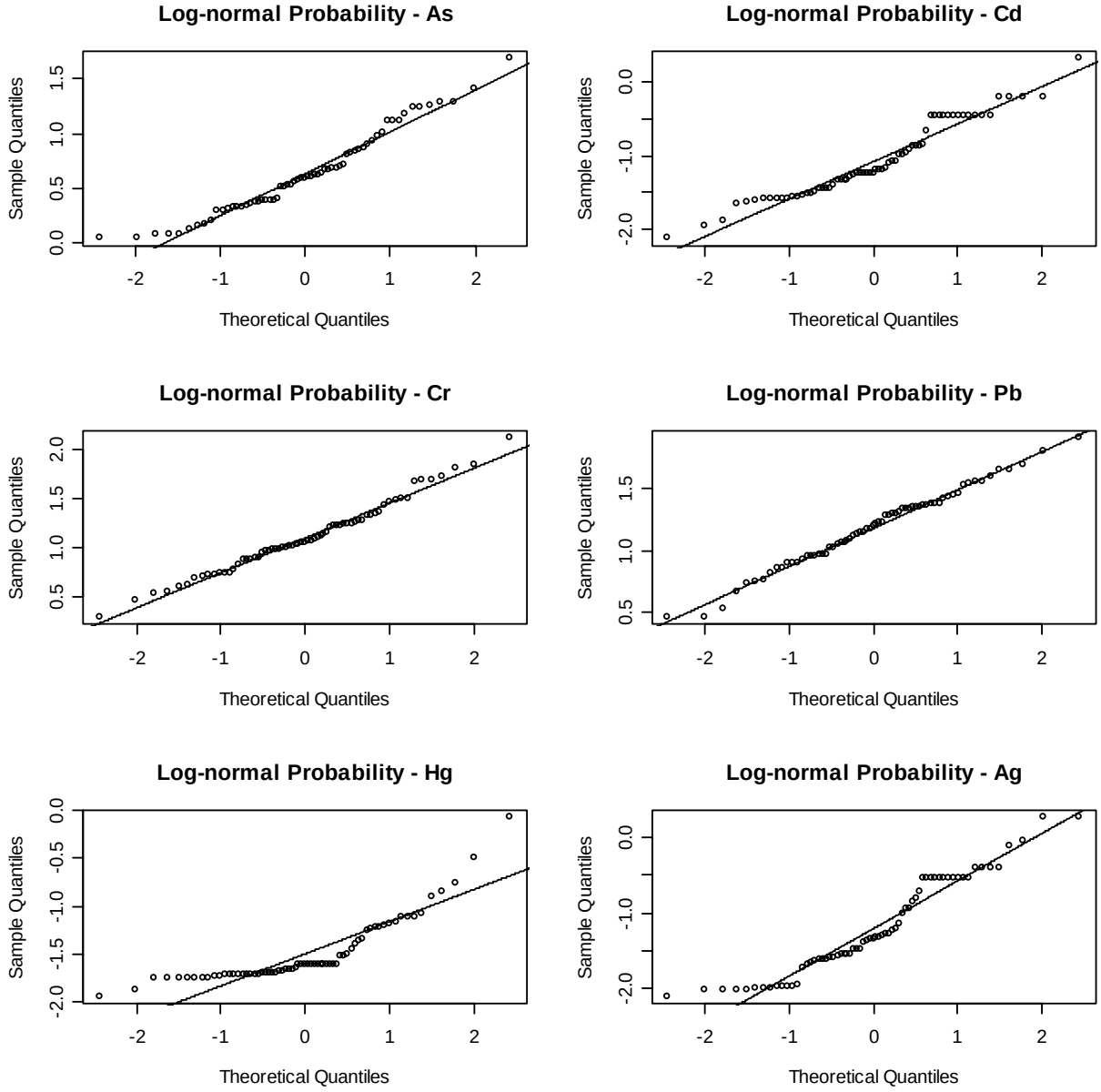


Figure 3. Probability plots of log concentrations of trace metals in Choccolocco Corridor background samples. The solid lines are the theoretical log-normal distributions matching the sample parameters.

Table 1 contains natural log standard deviations for the six CC background distributions in Figure 2. Four corresponding values from a USGS soil survey (USGS, 1984) covering the contiguous 48 States are also shown. Normally, we would expect to see lower variability in a local sample than in the USGS sample; but soil variability can be very sensitive to the soil sampling method, the total mass of a sample, and sub-sampling procedures. In any event, Table 1 provides what is needed for this investigation – a range of values for real-world sample variability. For this study, we will assume a lognormal background distribution, and choose a middle-of-the-pack natural log standard deviation of 0.80.

Table 1. Natural log standard deviations of Choccolocco Corridor and USGS background samples

	CC	USGS
Arsenic	0.87	0.80
Cadmium	1.17	---
Chromium	0.83	0.86
Lead	0.71	0.52
Mercury	0.67	0.92
Silver	1.45	---

The Site Populations

Four scenarios for the distribution of the site population will be evaluated:

1. 100% of the site is contaminated.
2. 50% of the site is contaminated.
3. 20% of the site is contaminated.
4. 10% of the site is contaminated.

In these scenarios, the site population is initially set equal to the log-normal background population and a lognormal contaminant distribution is added to the background distribution in the contaminated fraction. The contaminant distribution is assumed to be more variable than the background distribution - a natural log standard deviation of 1.5 is assumed for the contaminant distribution in the simulations.

Figures 4 – 7 show site and background population distributions for the four site scenarios. The distributions are shown as density plots, which can be thought of as smoothed histograms. Each case is plotted with arithmetic concentration on the x-axis, and then plotted again to the right with log concentration on the x-axis. In each figure, the upper pair of plots (a and b) shows the case when the site concentration exceeds the background concentration by 50; that is, the target threshold or significant difference that was chosen for test form 2. In the lower pair of plots (c and d) the difference between the site and the background is 200. The background population, of course, remains constant through all of these scenarios and cases.

As the fraction of contaminated soil decreases, the visual differences between the site and background populations become less obvious. Intuitively, we might expect these low-fraction scenarios to be more difficult to distinguish by statistical testing as well. Scenario 4 (Figure 7), with only 10% of the soil contaminated, is statistically equivalent to a hot-spot scenario.

Background and site populations are largely identical, with all of the differences occurring in the upper tail.

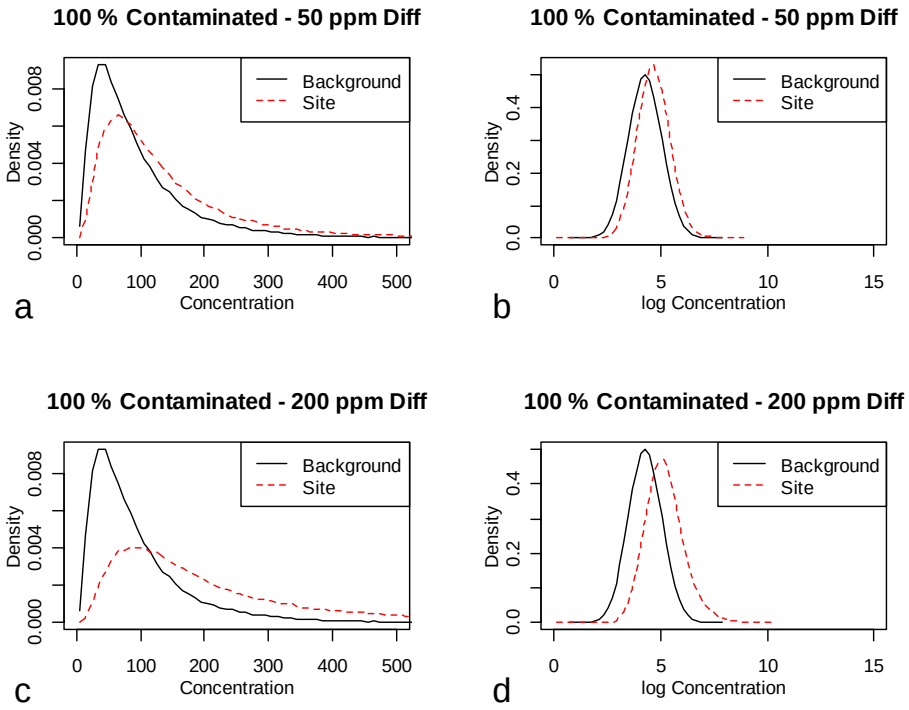


Figure 4. Distributions from Scenario 1.

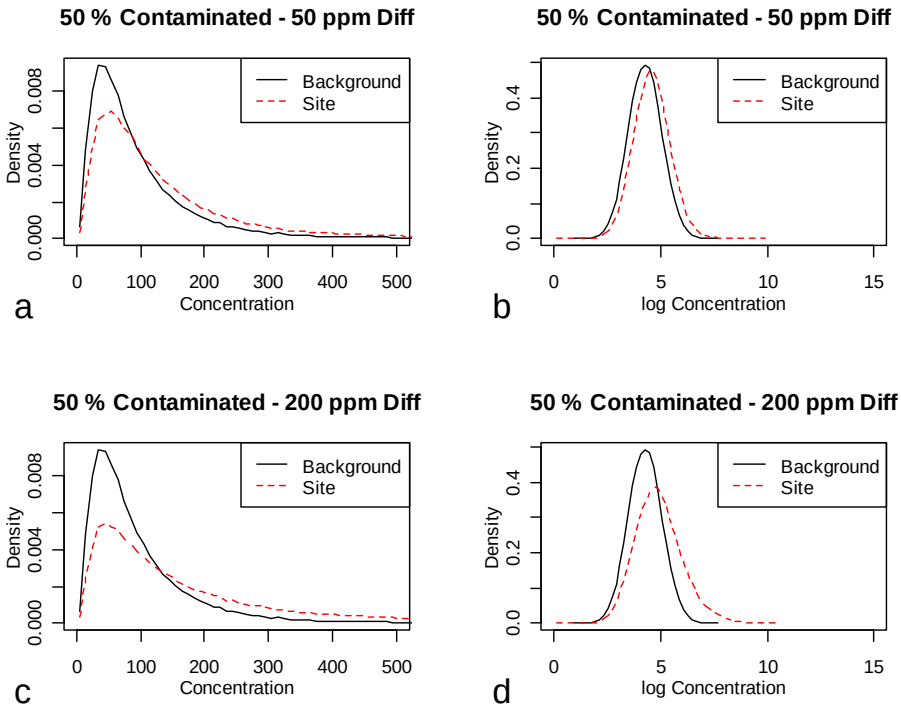


Figure 5. Distributions from Scenario 2.

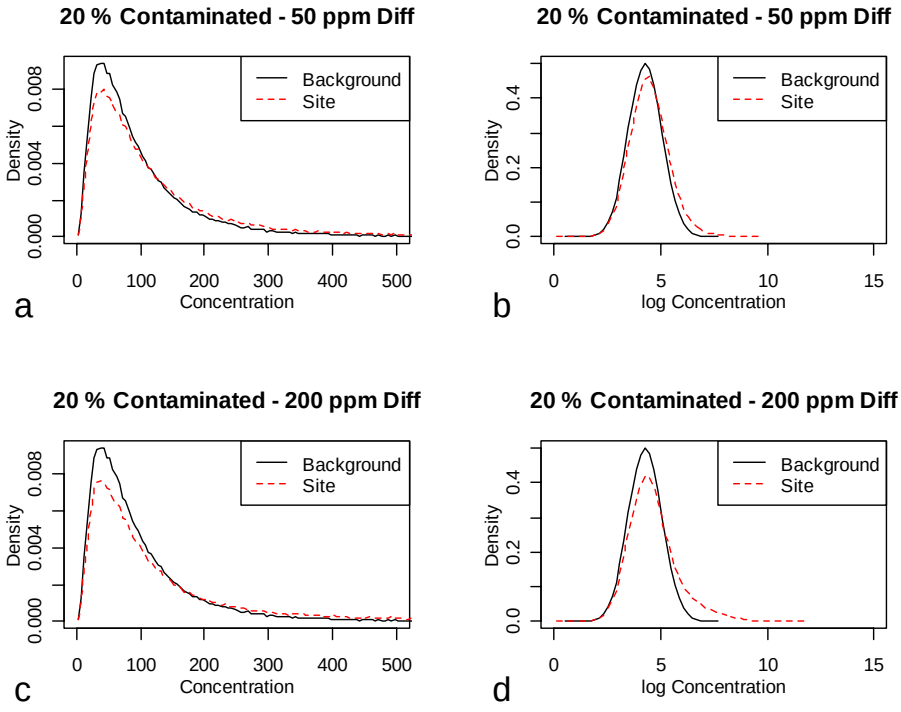


Figure 6. Distributions from Scenario 3.

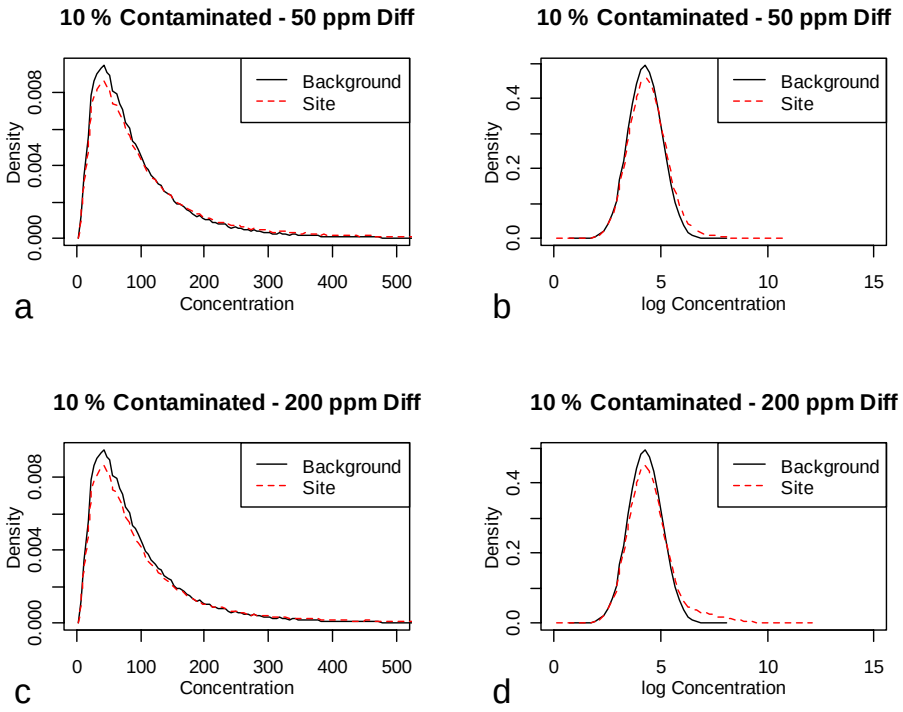


Figure 7. Distributions from Scenario 4.

Null Hypotheses

The performance of a formal statistical test of two populations depends on whether we begin by assuming that the two populations are the same, or that they are different. EPA (2002) refers to the former as "test form 1" and the latter as "test form 2." In statistical terms, the null hypothesis for test form 1 is that the site mean is less than or equal to the background mean. When the test rejects the null hypothesis, the site mean concentration is deemed significantly elevated above the background mean concentration, thus requiring a cleanup operation, or at least a further evaluation phase. When test form 2 is used the null hypothesis is that the site is contaminated by at least a specified amount above background. In this case, cleanup or further evaluation will proceed unless the null hypothesis is rejected in favor of the alternative.

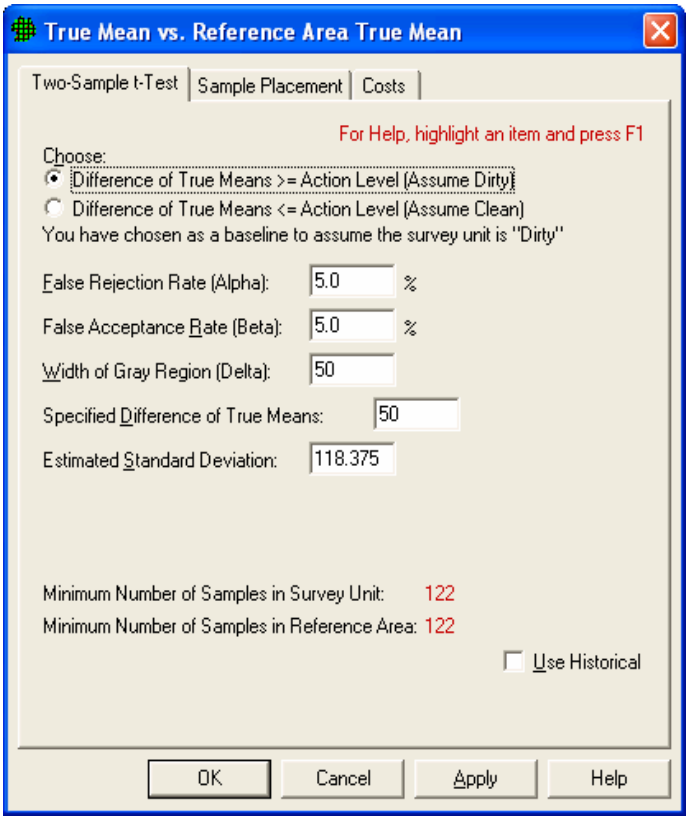
The Student's t, Welch's t, and WRS tests are commonly used for both test forms. When test form 1 is used, the sample data are used directly; the tests assume the two data sets are equal, and determine whether or not the site data are higher than background. When test form 2 is used, one of the data sets is shifted by the significant difference – the width of the gray region. The difference can either be added to the background measurements or subtracted from the site measurements. Statistical software packages usually have an option to test for a significant difference between populations, making it unnecessary for the user to modify the data. The tests assume that the modified data sets are equal, and then determine whether or not the background data are lower than the site.

The quantile test is performed only using test form 1. The quantile test only looks at the upper tails of the distributions; so, it does not make sense to shift an entire data set.

Sample Sizes

Two sample sizes were used in the simulation: n = 30 and n = 150. N = 30 is a frequently used (and misused) rule-of-thumb sample size. From the Central Limit Theorem, we know that for any population distribution, the distribution of the sample mean approaches the normal distribution as the sample size increases. An example of this is the Student's t distribution, where the distribution of the mean for sample sizes of 30 or more is very close to normal. It is sometimes mistakenly assumed that 30 samples are adequate for any distribution.

The sample size of 150 comes from an *ad hoc* design approach intended to provide a "ballpark" number. The theoretical sample size required to achieve the performance objectives in Figure 1 was calculated by assuming a normally distributed population. An arbitrary "safety factor" of 20% was added, and the result rounded to the nearest 10 samples. The theoretical part of the design was done using the *Visual Sample Plan, v.4.6D* freeware package (http://dqo.pnl.gov/) developed by the Pacific Northwest National Laboratory, as shown in the screen shot below.



The first four inputs are taken directly from Figure 1, and are obvious. The arithmetic estimated standard deviation (ESD) was calculated from the log standard deviation of the background population by first calculating the coefficient of variation (CV):

$CV^2 = e^{sy\text{^}2} - 1 = e^{0.8\text{^}2} - 1 = 0.896$, where sy is the log standard deviation of the background population. Then,

$$CV = 0.947$$

By definition,

$$CV = ESD/mean, \text{ so}$$
$$ESD = CV * mean.$$

The appropriate mean in this case is the midpoint of the gray region, or 125, so

$$ESD = 0.947 * 125 = 118.375.$$

The theoretical sample size (Nt) from VSP is 122, and our *ad hoc* adjustment gives

$$N = 1.2 * Nt = 146.4 \approx 150.$$

Measurement Errors

Measurement errors were not treated separately in the simulations. The simulated populations were assumed to be the sets of all possible soil sample measurements rather than being the sets of all possible true soil sample concentrations. All measurements were assumed to be above the detection limit.

Simulation Results

Figures 8-23 illustrate performance curves for the various statistical tests. Each figure shows one test case: one contamination scenario sampled by a particular combination of site and background sample sizes.

Each figure contains three pairs of plots as described earlier; each pair comparing several tests for the two test forms. WRS and Welch's t results are repeated in each pair for reference.

- Upper pair (a and b) – WRS, Student's t, and Welch's t.
- Middle pair (c and d) – WRS, Welch's t, Quantile test, and WRS + Quantile test.
- Lower Pair (e and f) – WRS, Welch's t, Welch's t (log transformed data), and Sample Mean test.

Overall test performance is determined by both false positive and false negative rates. When evaluating test performance with performance plots, there are two critical features to look at. First, as indicated above, the curve should pass through the specified alpha control point. Second, the curve should rise steeply to the right of the control point in the case of test form 1 or fall steeply to the left of the control point in the case of test form 2. In general the best overall indicator of performance is the steepness of the curve. Steeper curves indicate a sharper separation between populations.

Table 2 presents some of the performance results in quantitative terms. Each of the 16 cases shown in figures 8-23 has two rows of performance data, with columns representing specific tests. The first row, with text in italics, shows the observed performance when the true

concentration difference is zero. This is the false positive rate for test form 1, or for any test, the false action rate. The second row contains the value of the true difference in concentration at which the test achieved the desired 0.95 action rate. If the test did not achieve a 0.95 action rate, the observed rate at the maximum true difference (200) is shown instead.

The focus on differences in row 2 is because it provides an intuitive evaluation of relative risk. The true difference between the mean site concentration and the background concentration is proportional to the increase in risk at the site. The upper performance target – a 0.95 action decision rate at a threshold value of 50 – sets a *de facto* upper limit on the increase in risk that can be allowed to go undetected. If a test meets this target, it can be considered "protective" against a risk increase of 50 (or 50%, where the true background mean is 100). If the observed performance of a test achieves the 0.95 action rate at a true difference of 70, that test can be said to be protective against a 70% increase in risk.

Performance rates are used in row 1 because decision errors of this type increase action costs unnecessarily. These (expected) increases are directly proportional to the error rate.

Table 2 highlights the "better" performance results. In row 1, error rates less than or equal to 0.10 are underlined. In row 2, distances less than or equal to 100 are shown in bold. Instances where both criteria are met are shaded. These ranking criteria are arbitrary.

Equal Sample Sizes: 30 Background Samples, 30 Site Samples

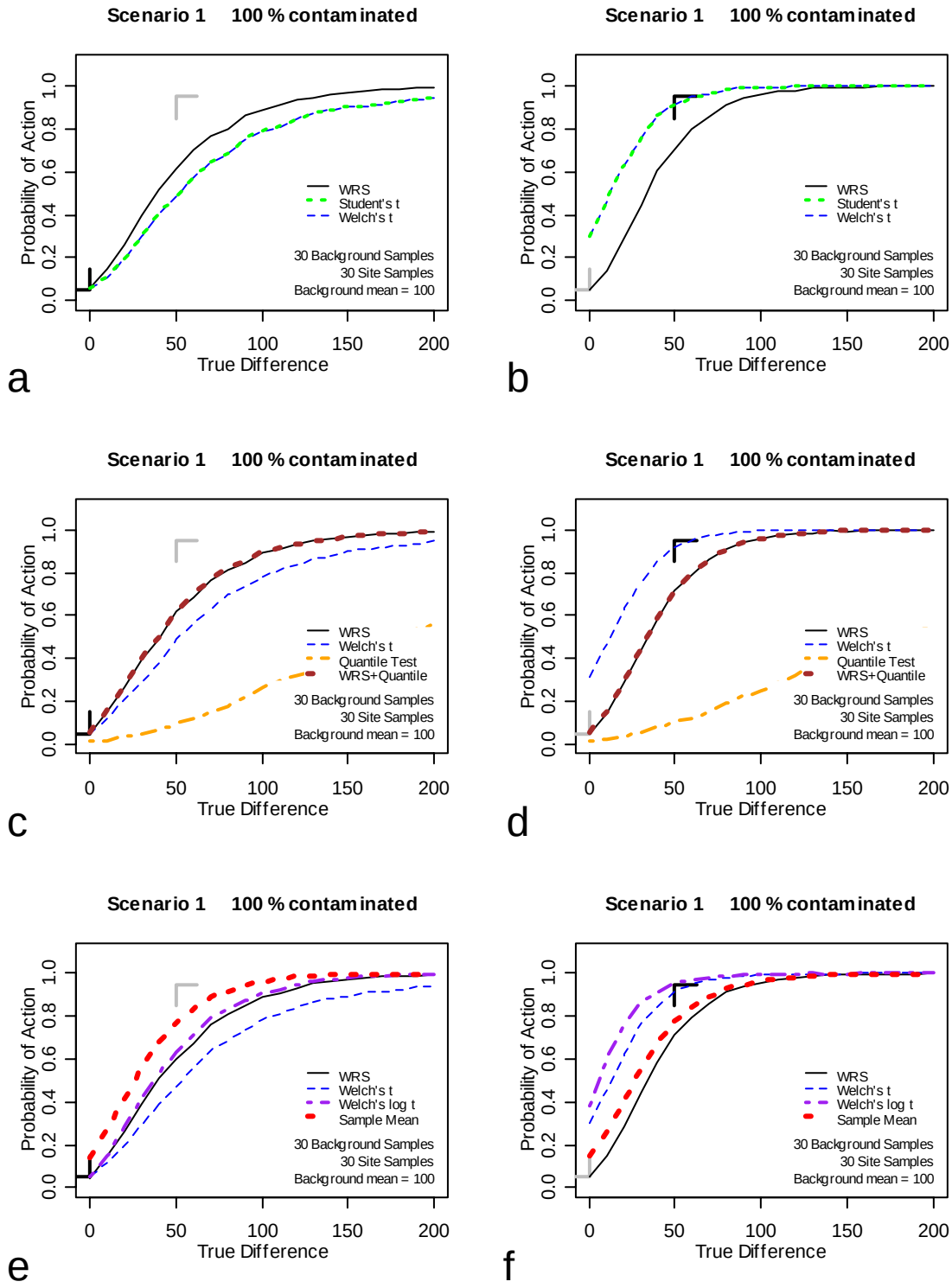


Figure 8. Scenario 1. Performance of several statistical tests. 30 background samples, 30 site samples. Left: Test Form 1. Right: Test Form 2.

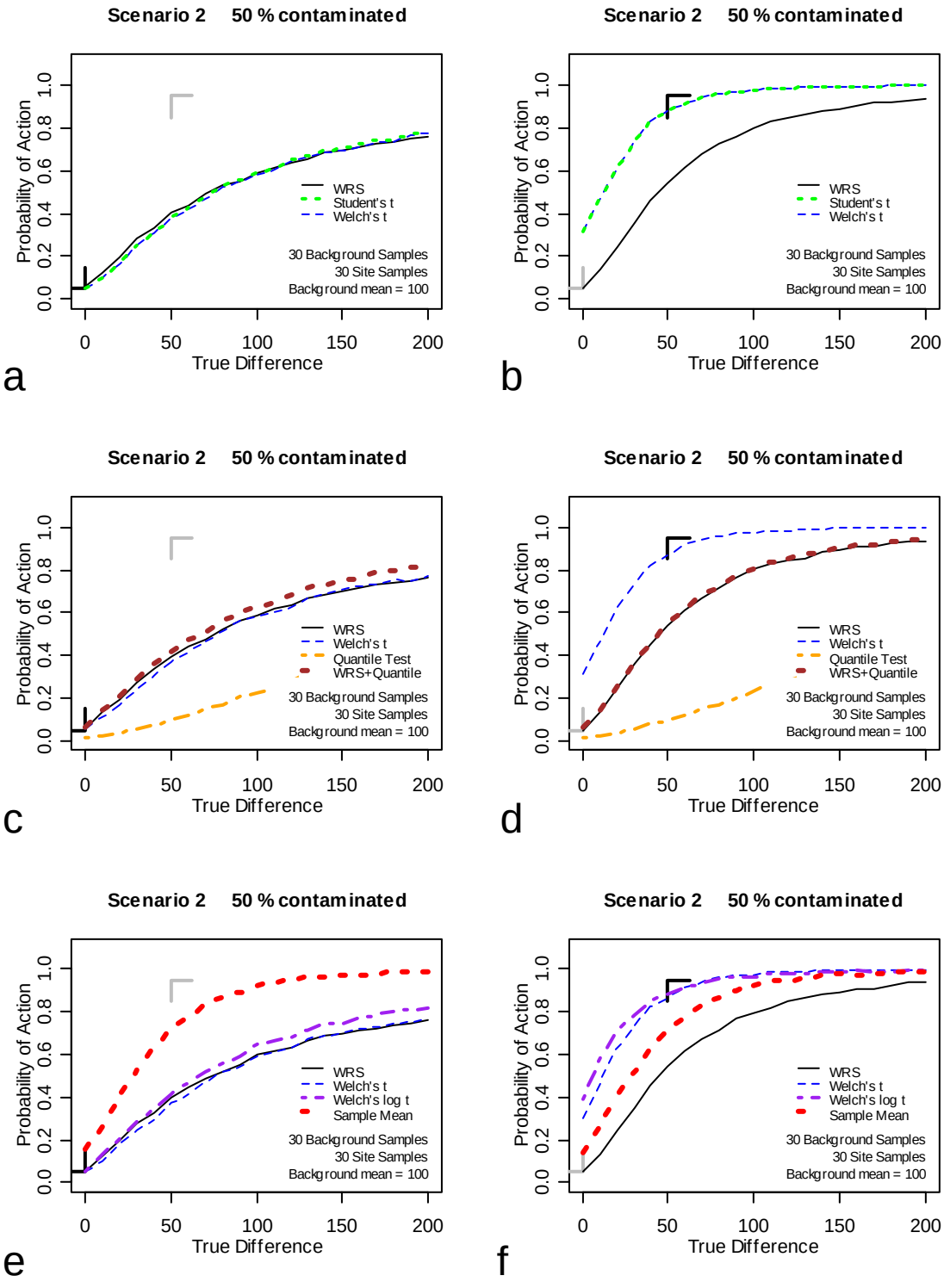


Figure 9 Scenario 2. Performance of several statistical tests. 30 background samples, 30 site samples. Left: Test Form 1. Right: Test Form 2.

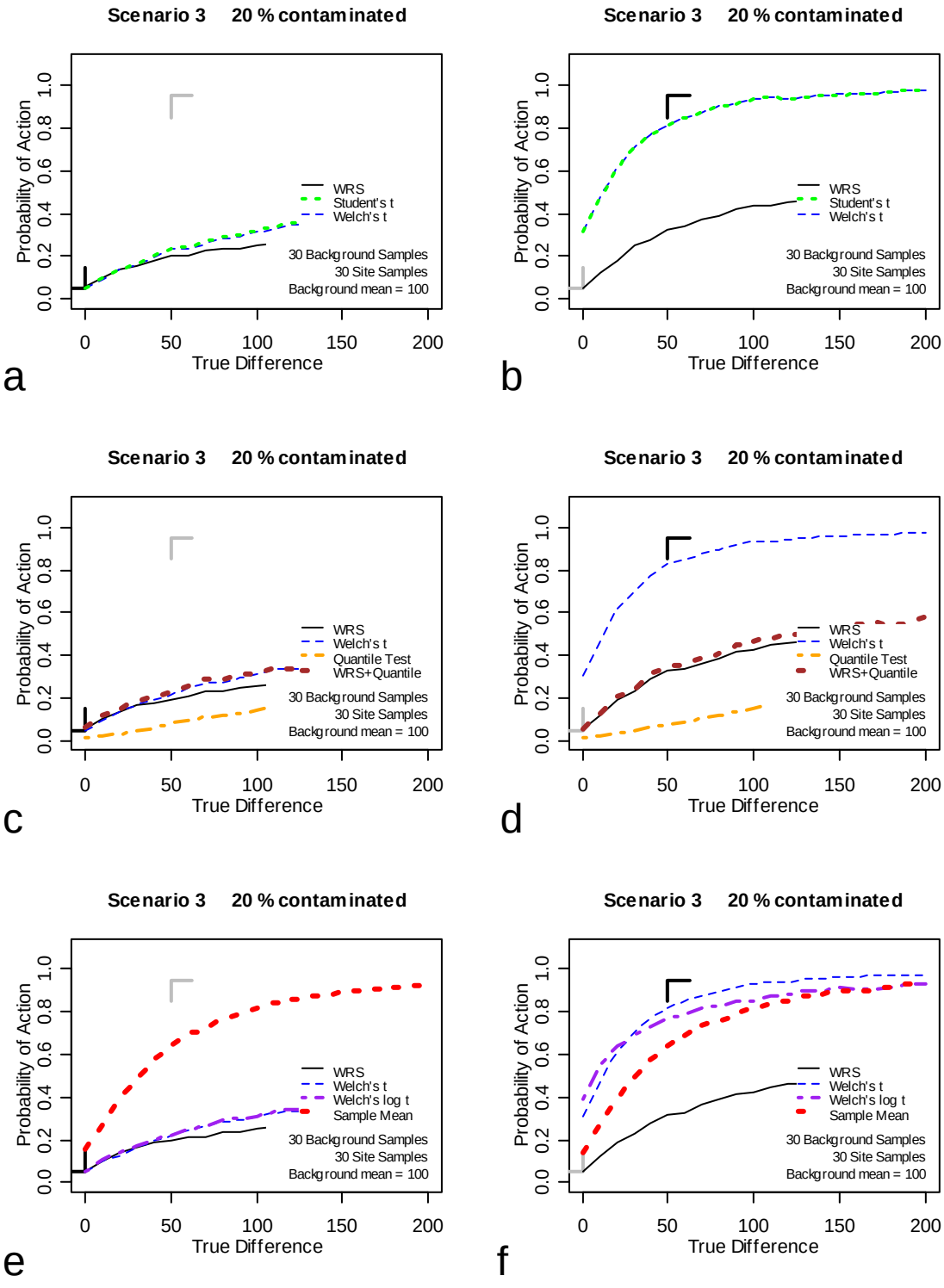


Figure 10 Scenario 3. Performance of several statistical tests. 30 background samples, 30 site samples. Left: Test Form 1. Right: Test Form 2.

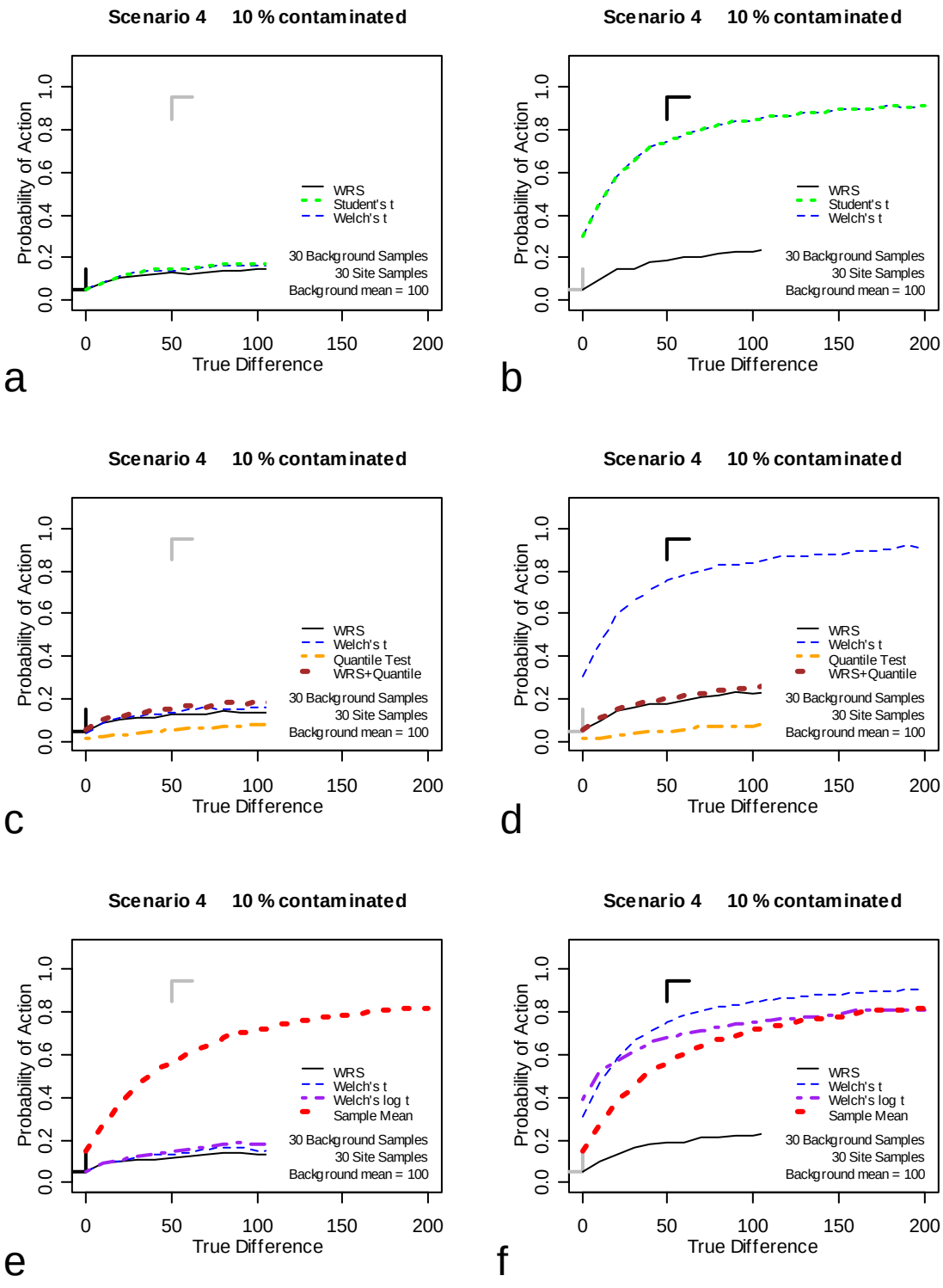


Figure 11. Scenario 4. Performance of several statistical tests. 30 background samples, 30 site samples. Left: Test Form 1. Right: Test Form 2.

Equal Sample Sizes: 150 Background Samples, 150 Site Samples

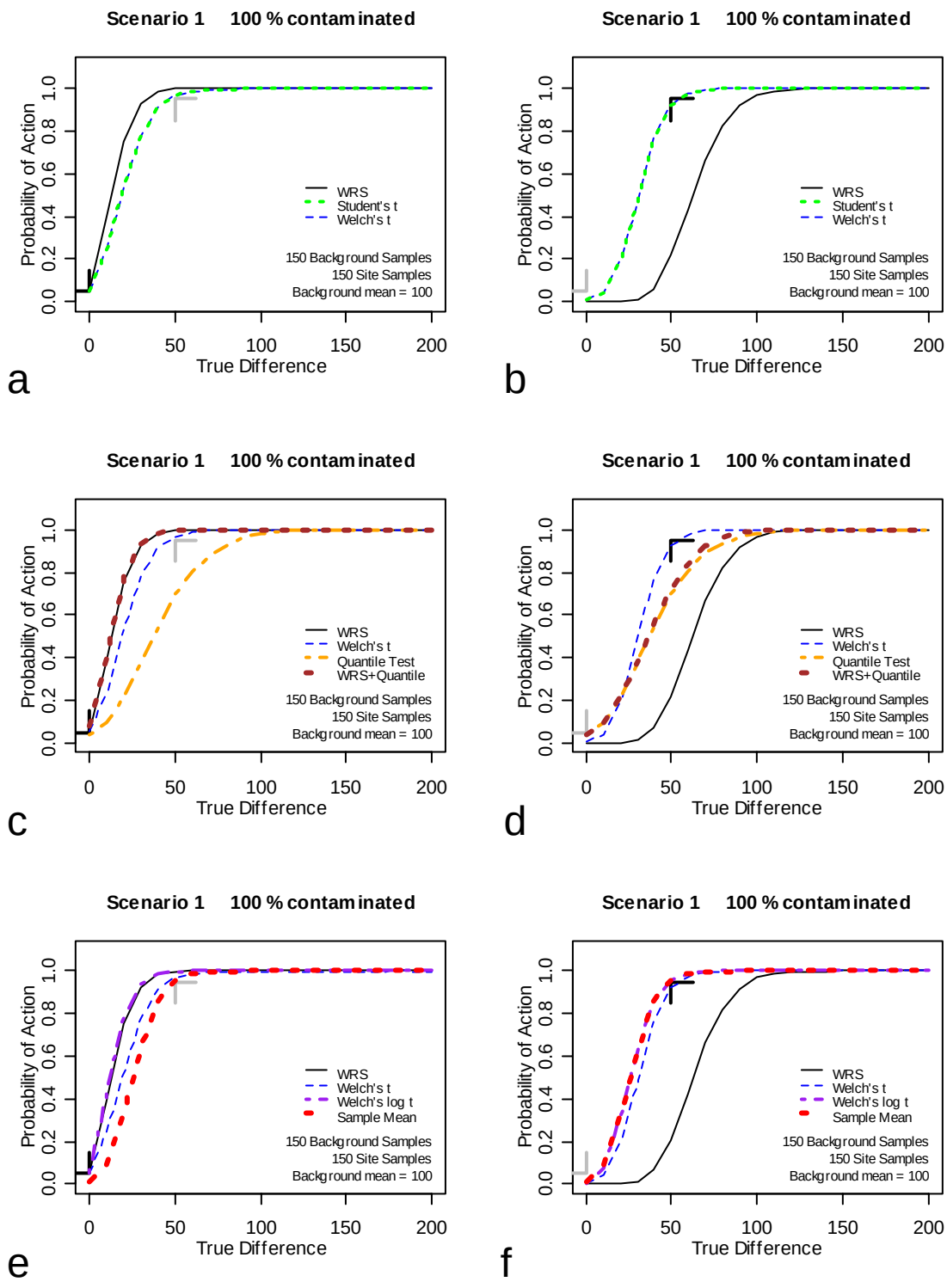


Figure 12. Scenario 1. Performance of several statistical tests. 150 background samples, 150 site samples. Left: Test Form 1. Right: Test Form 2.

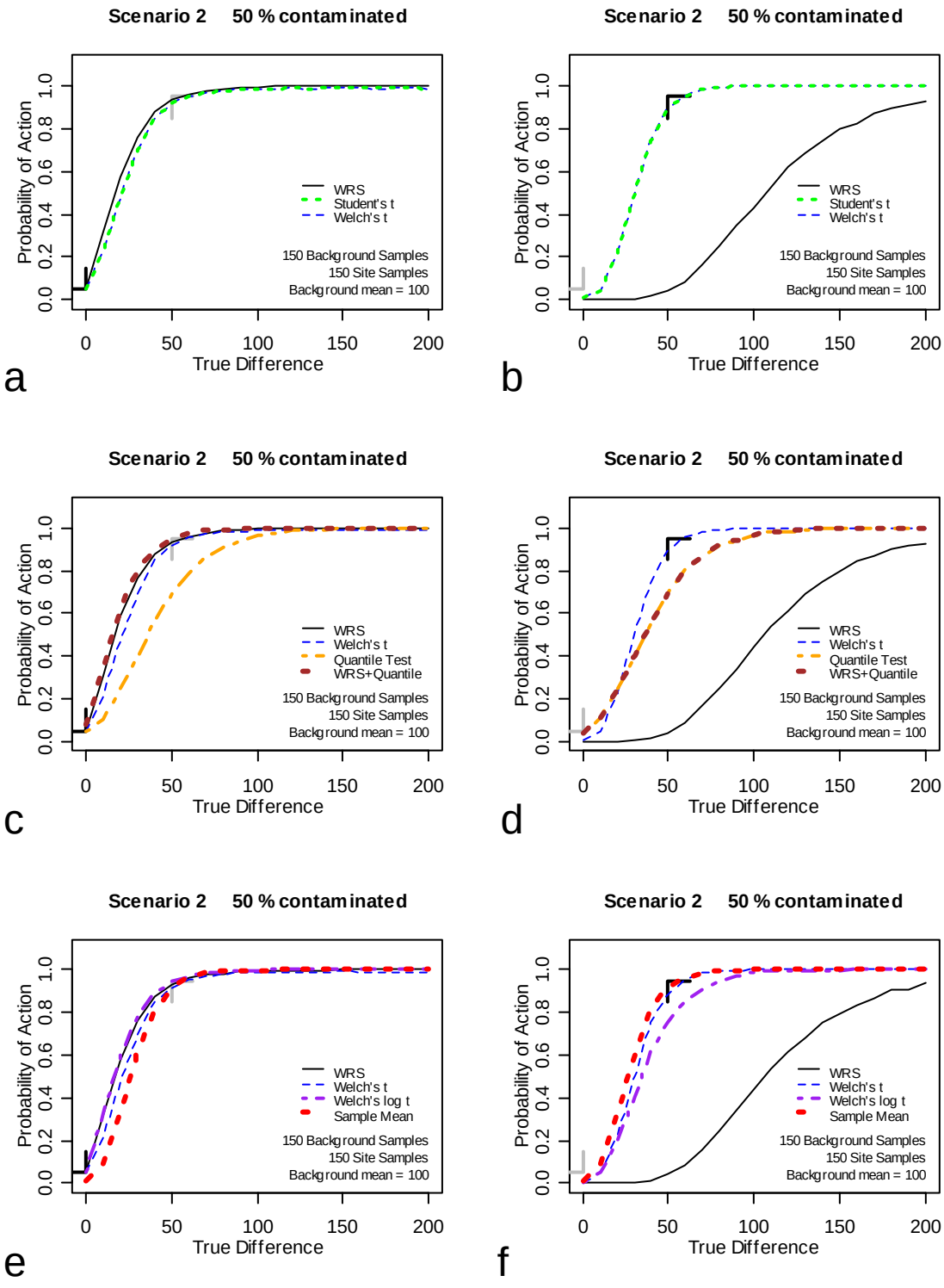


Figure 13. Scenario 2. Performance of several statistical tests. 150 background samples, 150 site samples. Left: Test Form 1. Right: Test Form 2.

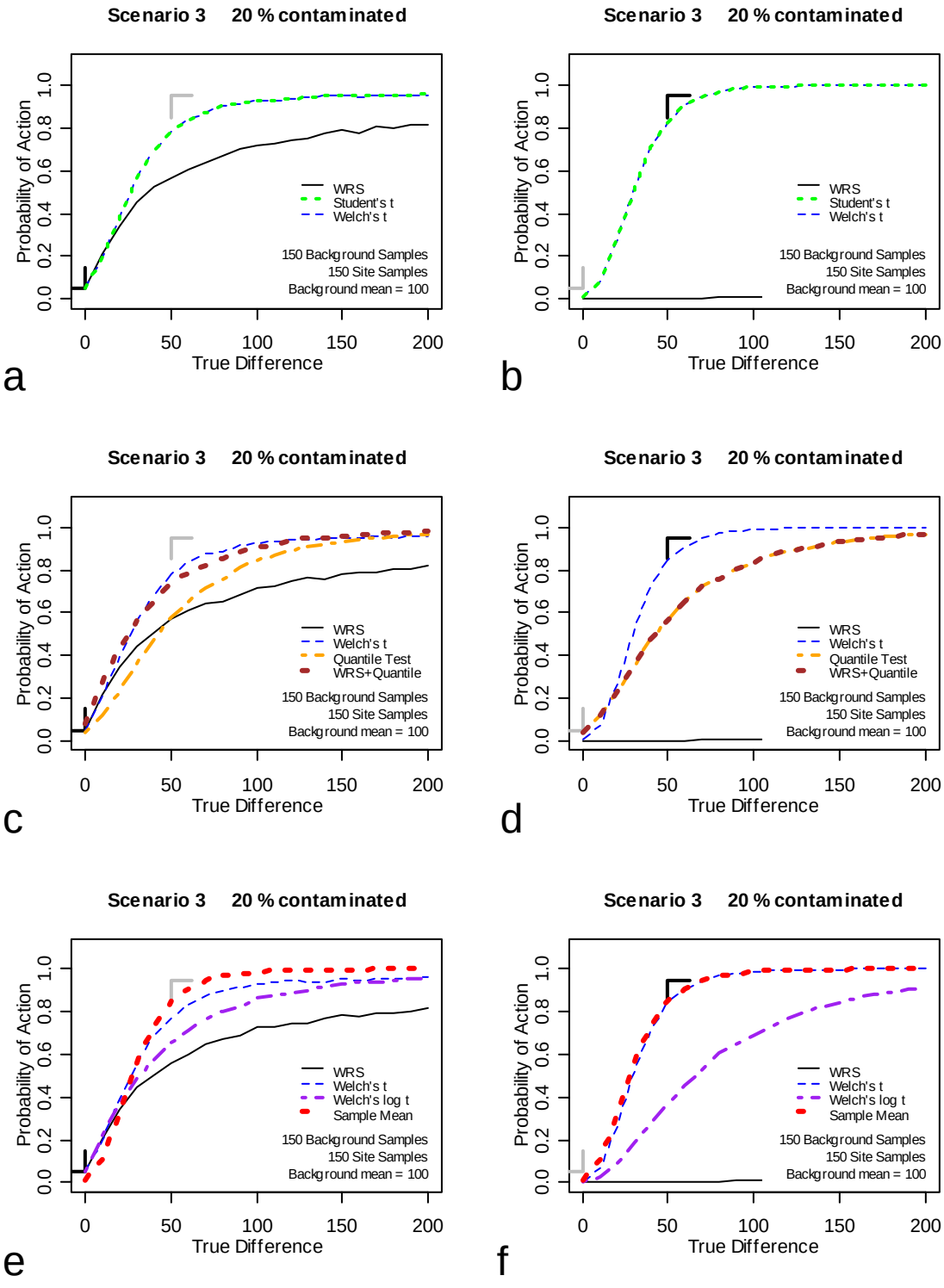


Figure 14. Scenario 3. Performance of several statistical tests. 150 background samples, 150 site samples. Left: Test Form 1. Right: Test Form 2.

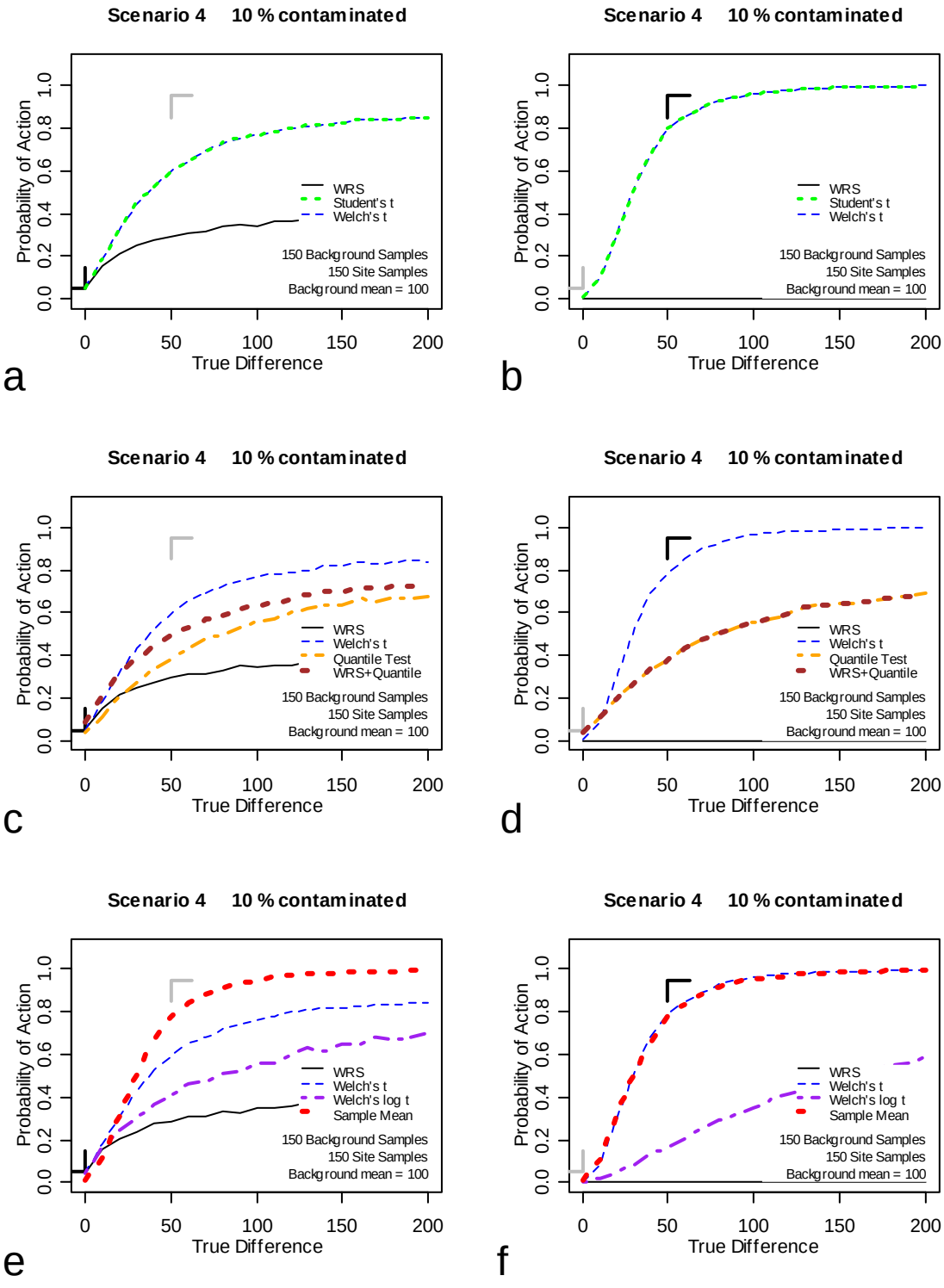


Figure 15. Scenario 4. Performance of several statistical tests. 150 background samples, 150 site samples. Left: Test Form 1. Right: Test Form 2.

Unequal Sample Sizes: 30 Background Samples, 150 Site Samples

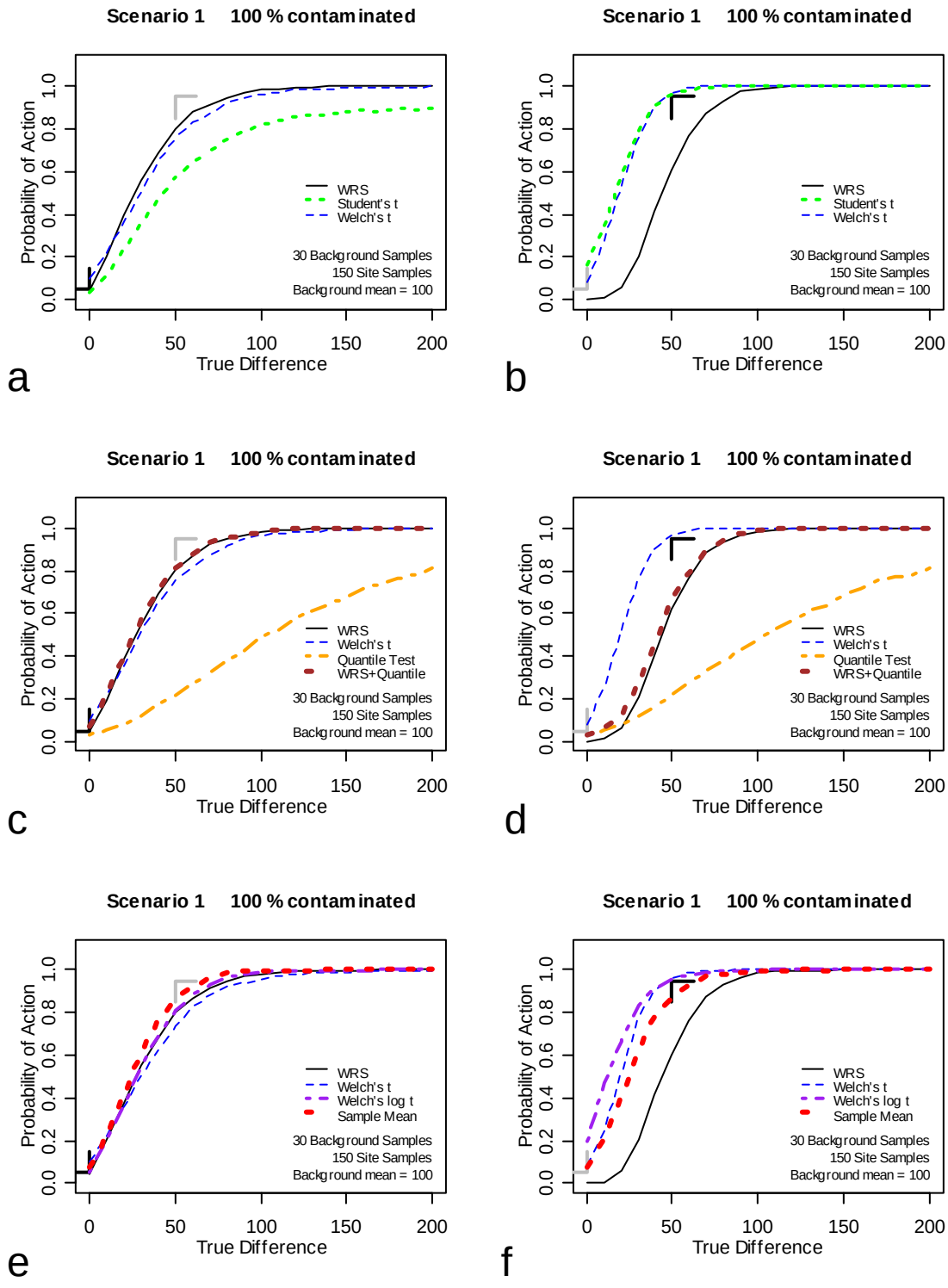


Figure 16. Scenario 1. Performance of several statistical tests. 30 background samples, 150 site samples. Left: Test Form 1. Right: Test Form 2.

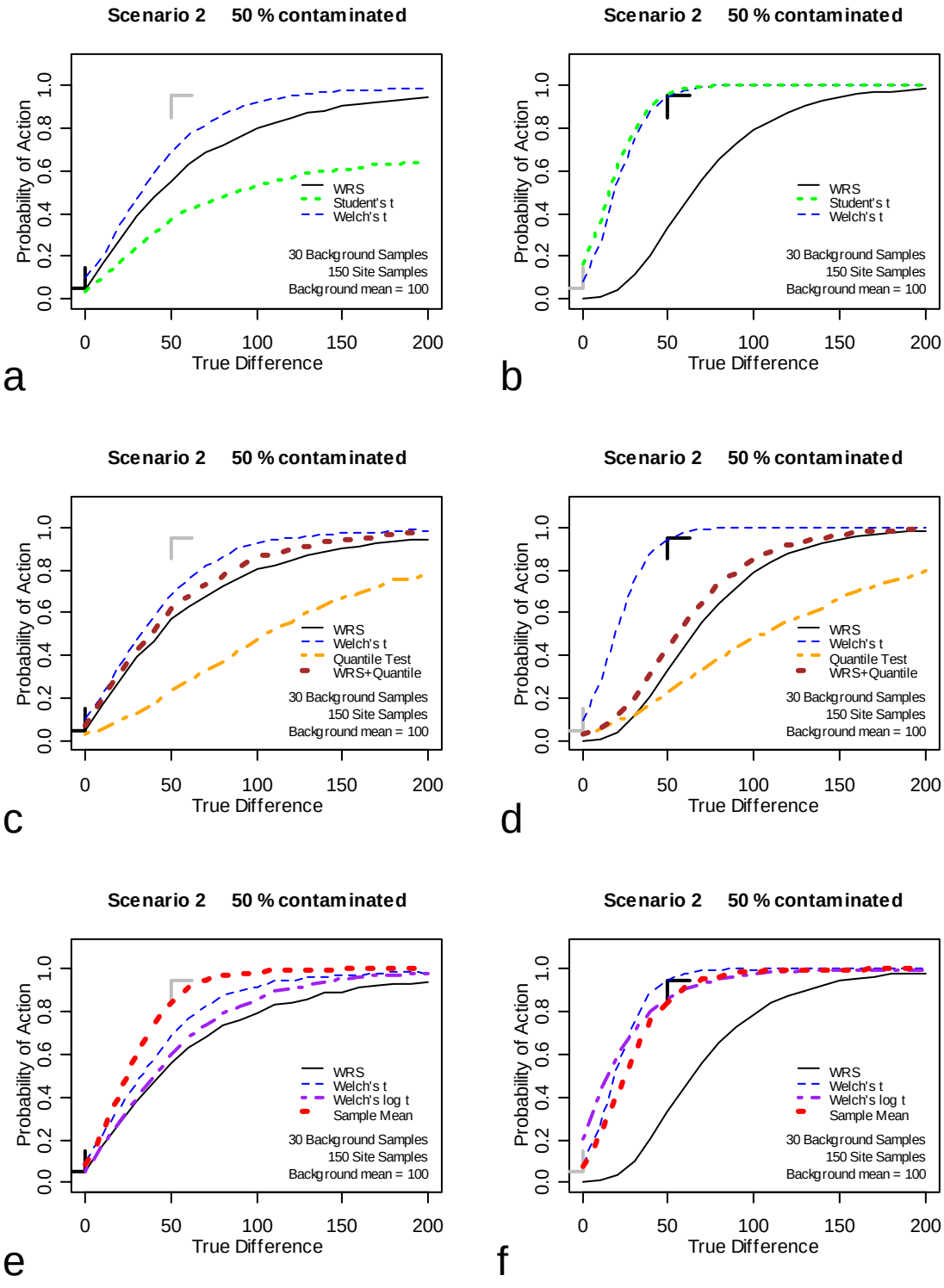


Figure 17. Scenario 2. Performance of several statistical tests. 30 background samples, 150 site samples. Left: Test Form 1. Right: Test Form 2.

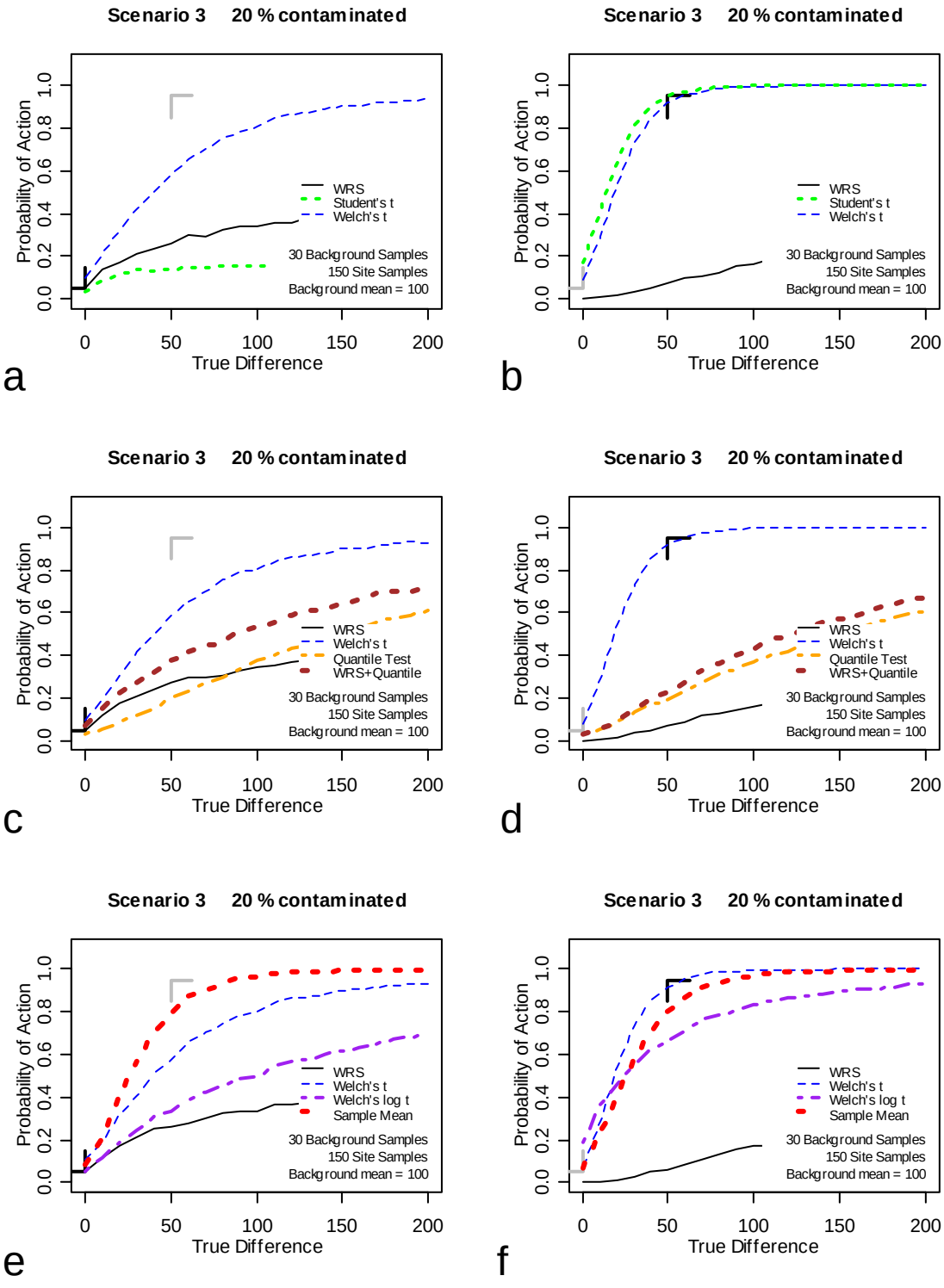


Figure 18. Scenario 3. Performance of several statistical tests. 30 background samples, 150 site samples. Left: Test Form 1. Right: Test Form 2.

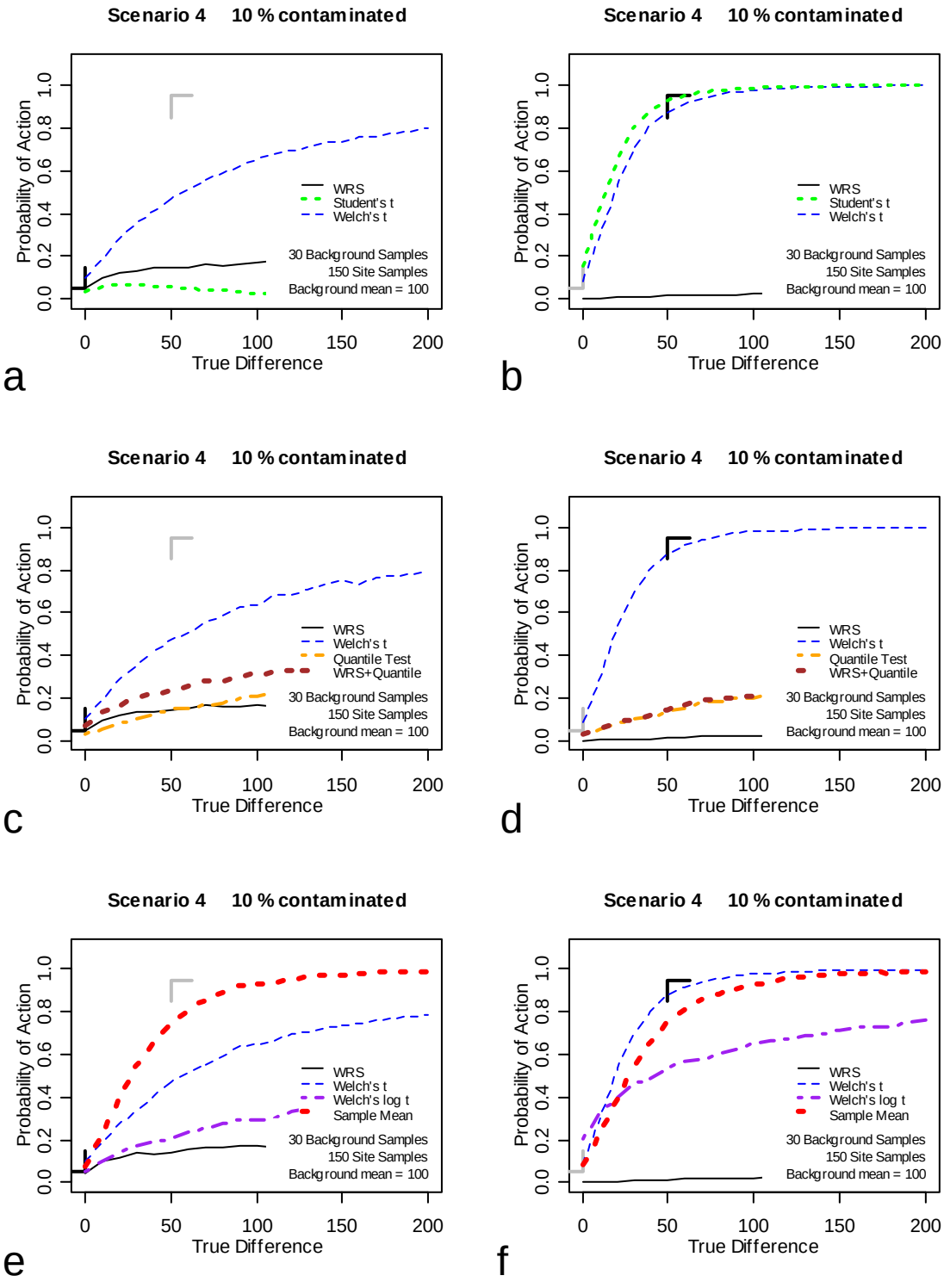


Figure 19. Scenario 4. Performance of several statistical tests. 30 background samples, 150 site samples. Left: Test Form 1. Right: Test Form 2.

Unequal Sample Sizes: 150 Background Samples, 30 Site Samples

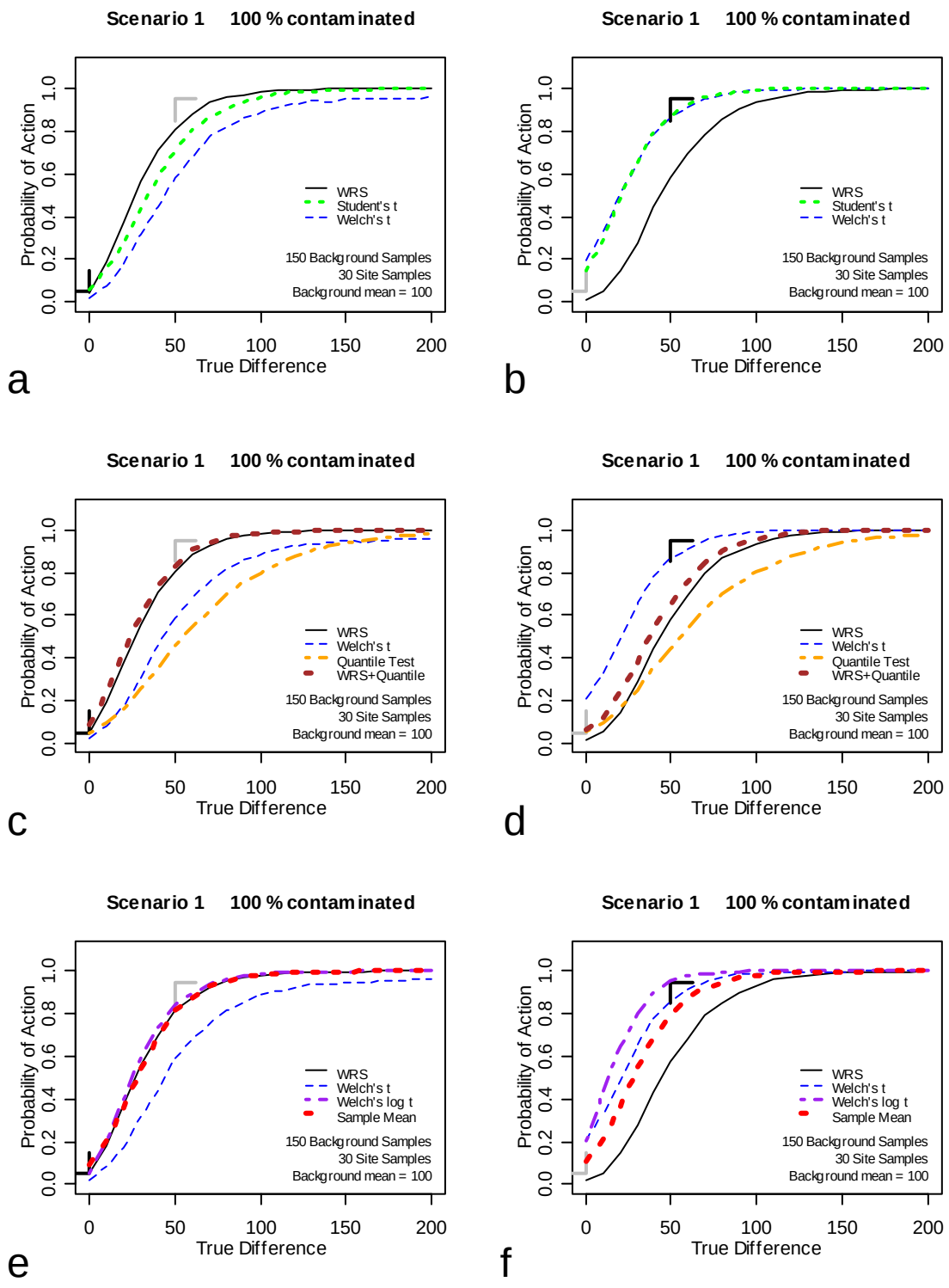


Figure 20. Scenario 1. Performance of several statistical tests. 150 background samples, 30 site samples. Left: Test Form 1. Right: Test Form 2.

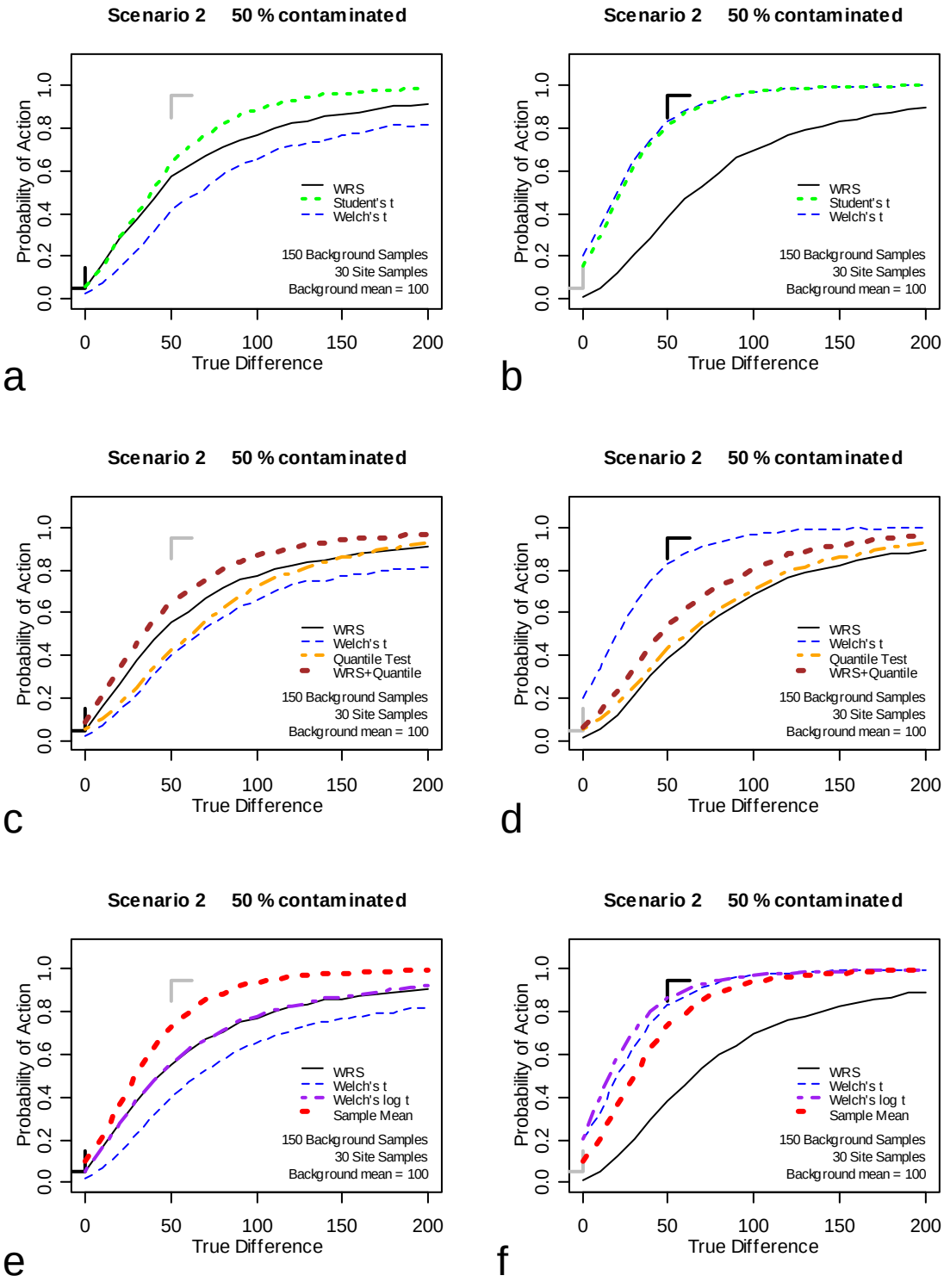


Figure 21. Scenario 2. Performance of several statistical tests. 150 background samples, 30 site samples. Left: Test Form 1. Right: Test Form 2.

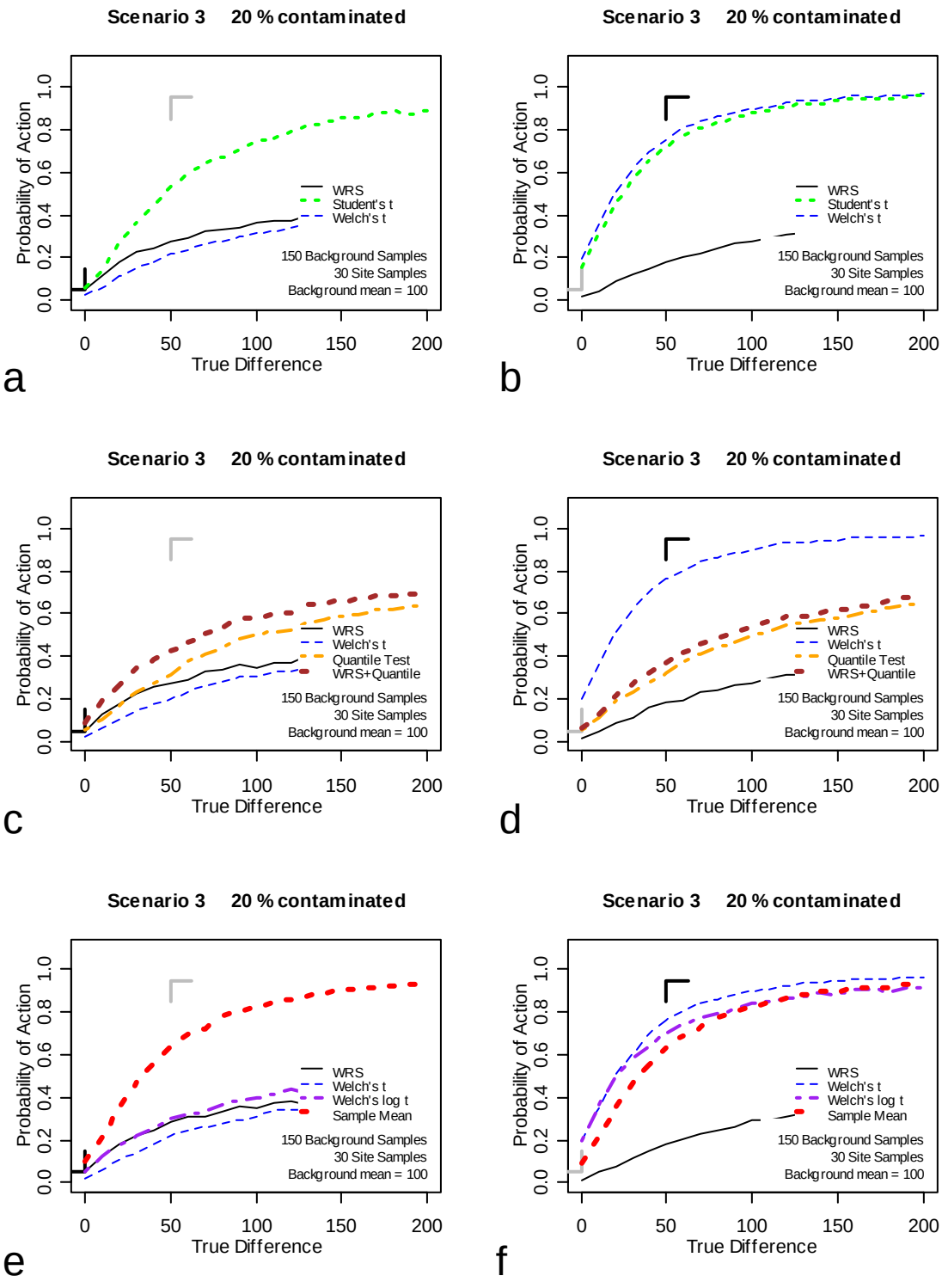


Figure 22. Scenario 3. Performance of several statistical tests. 150 background samples, 30 site samples. Left: Test Form 1. Right: Test Form 2.

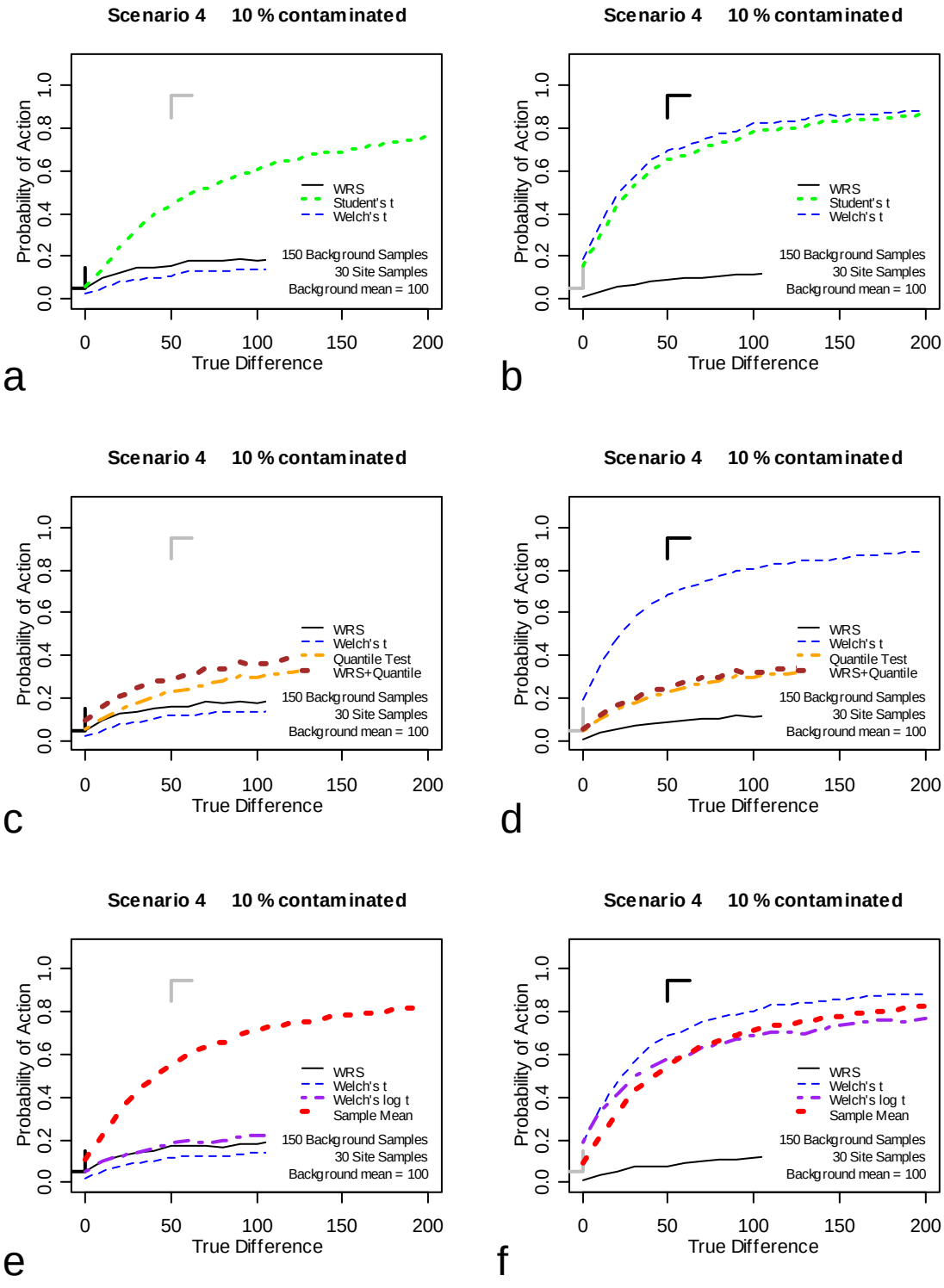


Figure 23. Scenario 4. Performance of several statistical tests. 150 background samples, 30 site samples. Left: Test Form 1. Right: Test Form 2.

Table 2. Performance of Two-Sample Comparison Tests

% Contaminated	Background N	Site N		WRS_1	WRSQ_1	Student's T_1	Welch's T_1	Quantile	Welch log T_1		WRS_2	WRSQ_2	Student's T_2	Welch's T_2	Quantile	Welch log T_2		Mean
100	30	30		*0.046*	*0.055*	*0.049*	*0.047*	*0.013*	*0.048*		*0.043*	*0.048*	*0.295*	*0.296*	*0.01*	*0.387*		*0.14*
				140	130	0.93	0.93	0.55	130		**100**	**100**	**70**	**70**	0.56	**50**		**90**
50	30	30		*0.051*	*0.057*	*0.051*	*0.051*	*0.01*	*0.051*		*0.05*	*0.058*	*0.308*	*0.308*	*0.012*	*0.379*		*0.145*
				0.76	0.82	0.78	0.77	0.47	0.82		0.94	0.94	**80**	**80**	0.48	90		130
20	30	30		*0.054*	*0.061*	*0.052*	*0.051*	*0.011*	*0.054*		*0.052*	*0.058*	*0.312*	*0.313*	*0.01*	*0.392*		*0.152*
				0.3	0.42	0.41	0.39	0.25	0.41		0.53	0.57	130	130	0.26	0.93		0.93
10	30	30		*0.05*	*0.056*	*0.05*	*0.049*	*0.011*	*0.052*		*0.046*	*0.052*	*0.298*	*0.3*	*0.011*	*0.368*		*0.141*
				0.14	0.21	0.18	0.17	0.11	0.22		0.26	0.3	0.91	0.91	0.11	0.82		0.82
100	150	150		*0.049*	*0.083*	*0.049*	*0.049*	*0.044*	*0.051*		*0*	*0.042*	*0.006*	*0.006*	*0.042*	*0.003*		*0.011*
				40	**40**	**50**	**50**	**90**	**40**		**100**	**80**	**60**	**60**	**90**	**50**		**50**
50	150	150		*0.047*	*0.081*	*0.049*	*0.049*	*0.045*	*0.048*		*0*	*0.041*	*0.007*	*0.007*	*0.041*	*0.001*		*0.01*
				60	**60**	**60**	**60**	**100**	**60**		0.93	**100**	**60**	**60**	**100**	**90**		**60**
20	150	150		*0.052*	*0.08*	*0.052*	*0.052*	*0.04*	*0.055*		*0*	*0.044*	*0.007*	*0.007*	*0.044*	*0.003*		*0.012*
				0.81	140	150	150	170	190		0.04	170	**80**	**80**	170	0.92		**80**
10	150	150		*0.058*	*0.084*	*0.051*	*0.051*	*0.038*	*0.054*		*0*	*0.038*	*0.006*	*0.006*	*0.038*	*0.004*		*0.011*
				0.4	0.74	0.85	0.85	0.68	0.71		0	0.69	**100**	**100**	0.69	0.57		110
100	30	150		*0.049*	*0.074*	*0.035*	*0.102*	*0.033*	*0.051*		*0.001*	*0.036*	*0.155*	*0.08*	*0.035*	*0.189*		*0.084*
				90	**80**	0.9	**100**	0.81	**80**		**90**	**80**	**50**	**50**	0.8	**50**		**70**
50	30	150		*0.046*	*0.07*	*0.031*	*0.093*	*0.029*	*0.047*		*0*	*0.027*	*0.15*	*0.077*	*0.027*	*0.192*		*0.073*
				0.94	160	0.65	130	0.78	150		160	140	**50**	**50**	0.77	80		**80**
20	30	150		*0.047*	*0.073*	*0.035*	*0.103*	*0.033*	*0.047*		*0*	*0.027*	*0.157*	*0.083*	*0.027*	*0.193*		*0.085*
				0.43	0.73	0.14	0.93	0.61	0.7		0.3	0.68	**60**	**70**	0.61	0.93		**90**
10	30	150		*0.048*	*0.07*	*0.035*	*0.095*	*0.029*	*0.051*		*0*	*0.034*	*0.158*	*0.079*	*0.034*	*0.205*		*0.078*
				0.18	0.38	0.01	0.78	0.3	0.39		0.03	0.33	**60**	**80**	0.32	0.77		120
100	150	30		*0.049*	*0.082*	*0.06*	*0.022*	*0.049*	*0.049*		*0.013*	*0.061*	*0.157*	*0.202*	*0.054*	*0.202*		*0.09*
				80	**80**	**100**	170	170	**80**		110	**100**	**70**	**80**	170	**60**		**80**
50	150	30		*0.049*	*0.092*	*0.065*	*0.021*	*0.059*	*0.052*		*0.016*	*0.06*	*0.155*	*0.199*	*0.054*	*0.199*		*0.1*
				0.91	170	140	0.81	0.93	0.92		0.89	180	**90**	**90**	0.92	90		110
20	150	30		*0.048*	*0.088*	*0.058*	*0.02*	*0.054*	*0.052*		*0.013*	*0.066*	*0.159*	*0.206*	*0.06*	*0.196*		*0.097*
				0.43	0.71	0.9	0.39	0.65	0.52		0.38	0.68	190	160	0.65	0.91		0.94
10	150	30		*0.047*	*0.087*	*0.062*	*0.019*	*0.054*	*0.048*		*0.012*	*0.058*	*0.153*	*0.196*	*0.052*	*0.192*		*0.1*
				0.21	0.43	0.75	0.16	0.38	0.28		0.13	0.38	0.86	0.89	0.36	0.77		0.83

The first three columns identify a test case -- a combination of contaminated fraction and sample sizes

Each case has two rows of results:

1. Fraction of unnecessary "action" decisions. (The target is 0.05)
2. True difference in population means above which the fraction of "action" decisions is >0.95

(the target difference is 50) … or, if the result is a number less than 1:

2. The maximum "action " fraction at a true difference of 200

Underscore: Row 1 results less than twice the target fraction

Bold: Row 2 results less than twice the target difference

Shaded: Both underscore and **bold**

"1" or "2" in column heading indicates test form

Design Considerations

The purpose of this paper was to compare the performance of statistical tests, not sampling designs. But perhaps the most important message to be gleaned from the results presented here is that decision performance is determined by a combination of the sampling design and the statistical test. If one wants to optimize a decision making process, then both factors must be considered together.

To design sampling plans for actual sites, it is recommended that case-specific simulations be performed, similar to those used in this report. Programs for running such simulations, or "scripts" as they are called in the R language, are included in Appendix 1. These scripts will permit a user to reproduce, within the limits of simulation variability, the results presented in this paper, and to experiment with different sample sizes, alpha levels, etc.

A cautionary note to anyone wishing to use these scripts as a design tool: although the scenarios used in the simulations are fairly realistic and present difficult sampling problems, they do not cover the full range of difficulties that the real world has to offer. In particular, these scripts do not provide for superimposing a regional anthropogenic background component over both site and reference areas. Nor do they allow for any real differences between the site and reference area background distributions due simply to natural variability; or for measurement errors or non-detects. Choosing appropriate background and site scenarios is a critical part of developing the "conceptual model" in the early stages of project planning.

The decision quality objectives used in this investigation (Figure 1) were totally arbitrary. They are used only to provide reference points to assist comparisons of the statistical tests and should not be considered an example to emulate at actual sites. Nevertheless, the objectives used here provide a useful starting point for discussion when setting objectives for a real site. The objectives from Figure 1 protect against missing an increase in mean concentration of 50% over the background mean level. Is a 50% increase too high, so that we need to aim to detect a 25% or even a 10% increase? Or is it too low, allowing us to relax the detection threshold to 100%, 200%, or more?

Test Recommendations

The tests recommended below generally performed well over the range of scenarios and sampling schemes evaluated. The recommendations should be considered tentative, and challenged when developing case-specific designs, especially if the design objectives or the site scenarios differ significantly from those assumed in this paper.

Perhaps the best approach is to start with the recommended tests, and then experiment to try to find a better alternative. Alternatives are not limited to the tests that have been evaluated in this paper. Any of the tests evaluated here can be modified by choosing different values for test parameters, such as alpha, significant difference, or action threshold. Many different combination tests are possible other than the WRS – quantile combination tested here.

Use Student's t or Welch's t with Test Form 2

Although the Wilcoxon Rank Sum test with Test Form 1 performs somewhat better than the t tests with Test Form 2 in the specific case of 100% contamination, WRS performance roughly equals the t tests at 50% contamination and rapidly becomes much less protective as the contaminated fraction drops. The t tests are consistently protective over the range of scenarios. If the site conceptual model suggests a real possibility that the site may be less than 50% contaminated, then t tests are the safer choice. When site and background sample sizes are equal, it makes no difference whether Student's t or Welch's t is used. Although the two tests may differ for any particular sampling event, over many simulations their overall performances become indistinguishable. There is a difference, however, when the sample sizes are unequal. The Welch's t test is superior to the Student's t test when the site sample size is larger than the background sample size and inferior when the site sample is smaller. (The opposite relationship holds for Test Form 1, but Test Form 1 is not recommended).

Consider composite sampling with a sample means test.

The sample means test did not perform as well as the t tests, but it was not far behind. The only statistic used is the sample mean, which can be obtained as easily by analyzing a composite sample as by averaging analyses of individual samples. This has considerable potential for reducing analytical costs when the required sample size is large and there are many target analytes requiring individual and costly analyses.

Observations and Discussion

Quantile test results are shown in Table 2 for both test forms, though the test was run the same way in both cases. Comparing the two columns provides an indication of the precision of the simulation process.

Within the limits of the precision of the simulation method, the WRS test and the Welch's log t test achieved the specified alpha requirement for Test Form 1 over all scenarios and sample sizes. This is as expected because when there is no difference between the site and background distributions, the assumptions for the WRS test are met; and when there is no difference between the distributions and they are log-normal, the assumptions of the Welch's log t test are met. Unfortunately this is of limited practical value for two reasons. Real reference area populations are inevitably different from true site background populations, so actual decision performance can differ from theoretical performance. More importantly, neither test performs well for Test Form 1 on scenarios 3 and 4 (20% and 10% contamination, respectively).

The WRS test fails completely for Test Form 2 on scenarios 3 and 4 when sample sizes are large. This is a rather unusual case where test performance actually gets worse with more data. This happens because Test form 2 subtracts 50 from each site sample. For example, in Figure 7c, the downshift would move approximately 30% of the site population below the background population, while only 10% of the site population was shifted upward by contamination. With enough data, the ranks will always show this net downward shift, causing rejection of the null. With few data, the test sometimes gets it wrong, which ironically gives the appearance of better performance with respect to our mean-oriented objectives.

The quantile test is combined with the WRS to address the WRS insensitivity to shifts in the upper tail. The quantile test is also a rank test, but looks only at the upper tail. However, neither the quantile test by itself, nor the quantile test in combination with the WRS, performed

consistently well enough to be recommended. This conclusion only applies to the particular combination of parameters used in this paper. For Test Form 1, the WRS test parameters were: alpha = 0.05 and significant difference = 0; the quantile test parameters were: alpha = 0.05 and quantile = 0.9 (meaning that the test looks only at the upper 10% tail.) For Test Form 2, the significant difference was 50 while the other three parameters did not change. As suggested earlier, all four of the parameters could be varied, and it would be a substantial project to search for the optimal combination. With the right set of parameters, a combination of the WRS test and the quantile test might prove to be the best overall performer, but he parameter sets used in this paper are not it.

The Welch's log t test performed similarly to the WRS test for Test Form 1. For those specific scenarios where the WRS test with Test Form 1 was equal to or better than the untransformed t tests with Test Form 2, the Welch's log t test equaled or slightly outperformed the WRS test.

References

Blair, C.R., and Higgins, J.J, 1980. *A comparison of the Power of Wilcoxon's Rank-Sum Statistic to That of Student's t Statistic under Various Nonnormal Distribution.* J Educ Statistics, 5:4, 309-335.

Bridge, P D., and Sawilowsky, S.S.,1999, *Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research.* J Clin Epidemiol 52:3, 229-235.

Hodges, J.L. Jr. and Lehman, E.L., 1956. *The Efficiency of some Nonparametric Competitors of the "t"-Test.* Annals Math Statistics 27:2,324-335.

Modarres, R., Gastwirth, J.L. and Ewens, W., 2005.*A cautionary note on the use of non-parametric tests in the Analysis of Environmental Data.* Environmetrics 16, 319-326.

Potvin, C. and Roff, D.A., 1993. *Distribution-Free and Robust Statistical Methods: Viable Alternatives to Parametric Statistics?* Ecology 74:6, 1617-1628.

R Development Core Team , 2005. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Shacklette, H.T. and Boerngen, J.G., 1984. *Element Concentrations in Soils and Other Surficial Materials of the Conterminous United States*, U.S.G.S. Prof. Paper 1270.

Singh, A.K., Singh, A. and Engelhardt, M., 1997. *The Lognormal Distribution In Environmental Applications*, EPA/600/R-97/006.

U.S.EPA, 2002. *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites,* EPA 540-R-01-003, OSWER 9285.7-41.

U.S.EPA, 1992. *Statistical Methods for Evaluating the Attainment of Cleanup Standards, Volume 3: Reference-Based Standards for Soils and Solid Media*, EPA 230-R-94-004.

U.S.EPA, 2006. *Data Quality Assessment: Statistical Methods for Practitioners*, EPA QA/G-9S, EPA/240/B-06/003.

Zhou, X-H., Gao, S. and Hui, S.L.,1997. M*ethods for Comparing the Means of Two Independent Log-Normal Samples.* Biometrics 53, 1129-1135.

Notice: The information in this document has been prepared by the United States Environmental Protection Agency. It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

APPENDIX 1 – Evaluating statistical test performance with R

R programs for simulating statistical test performance are included in the next three appendices.

The scripts presented here have been somewhat modified from the original that was used to produce the results in this report: loops were removed that produced six graphs on a single figure, and that ran all four scenarios in a single long run. Parameters that can be modified by the user were moved to a separate file so that they can be more easily located and changed without inadvertently altering the operational code. Graphical output has been reduced to a single plot; the user can select any or all of the performance curves to be plotted. Text previously written inside the performance plot has been moved outside. Performance results for all tests are automatically written to an output file, along with a copy of the parameter file used during the run. The user has been given the option of changing the arithmetic mean and the log standard deviation of the background population, and changing the log standard deviation of the contaminant population. Numerous comments have been added to the scripts.

The instructions below are not intended as an R tutorial. The intent is that an interested reader who is not an R user (and who has no desire to become one) will be able to run simulations and to be able to experiment with different input parameters. Although it would be possible to perform a rudimentary trial-and-error design optimization using these scripts, the scripts themselves have not been subjected to independent validation, and should be used with appropriate caution.

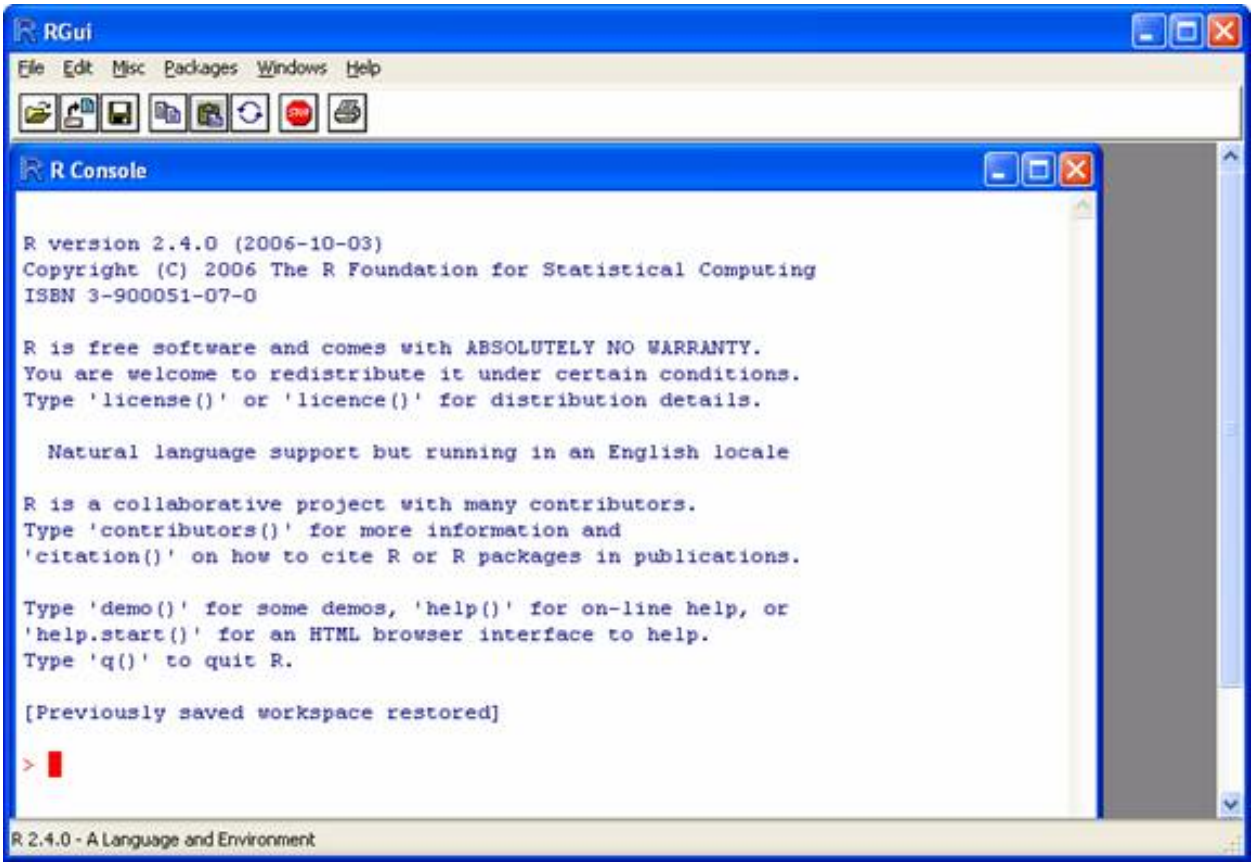
Instructions

Download the R Windows installation program from: http://www.R-project.org. (The current version at this writing is *R-2.4.1-Win32.exe.)*

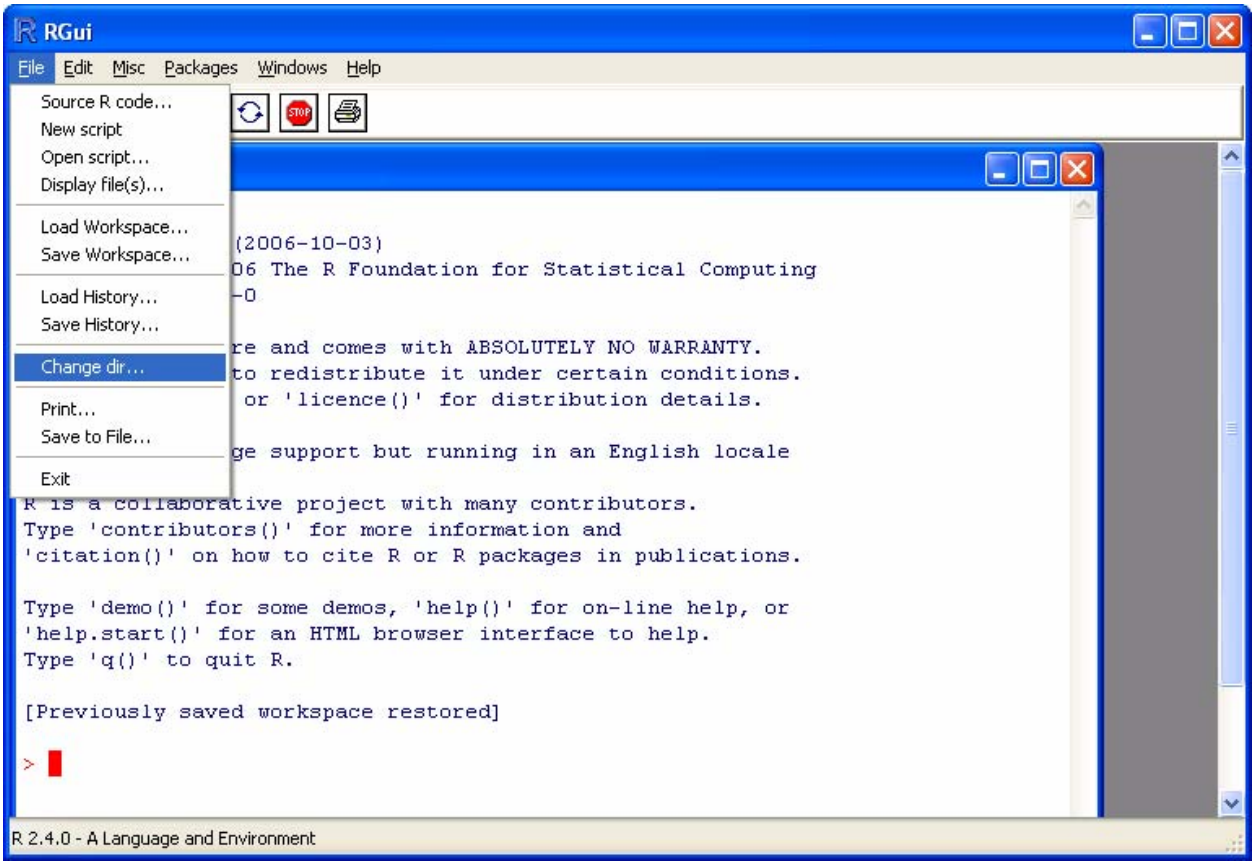
Run the install program.

Create a work folder and copy the scripts below into the folder (use the file names in italics, e.g., *parameters.r*. Make copies in another folder as backup files.

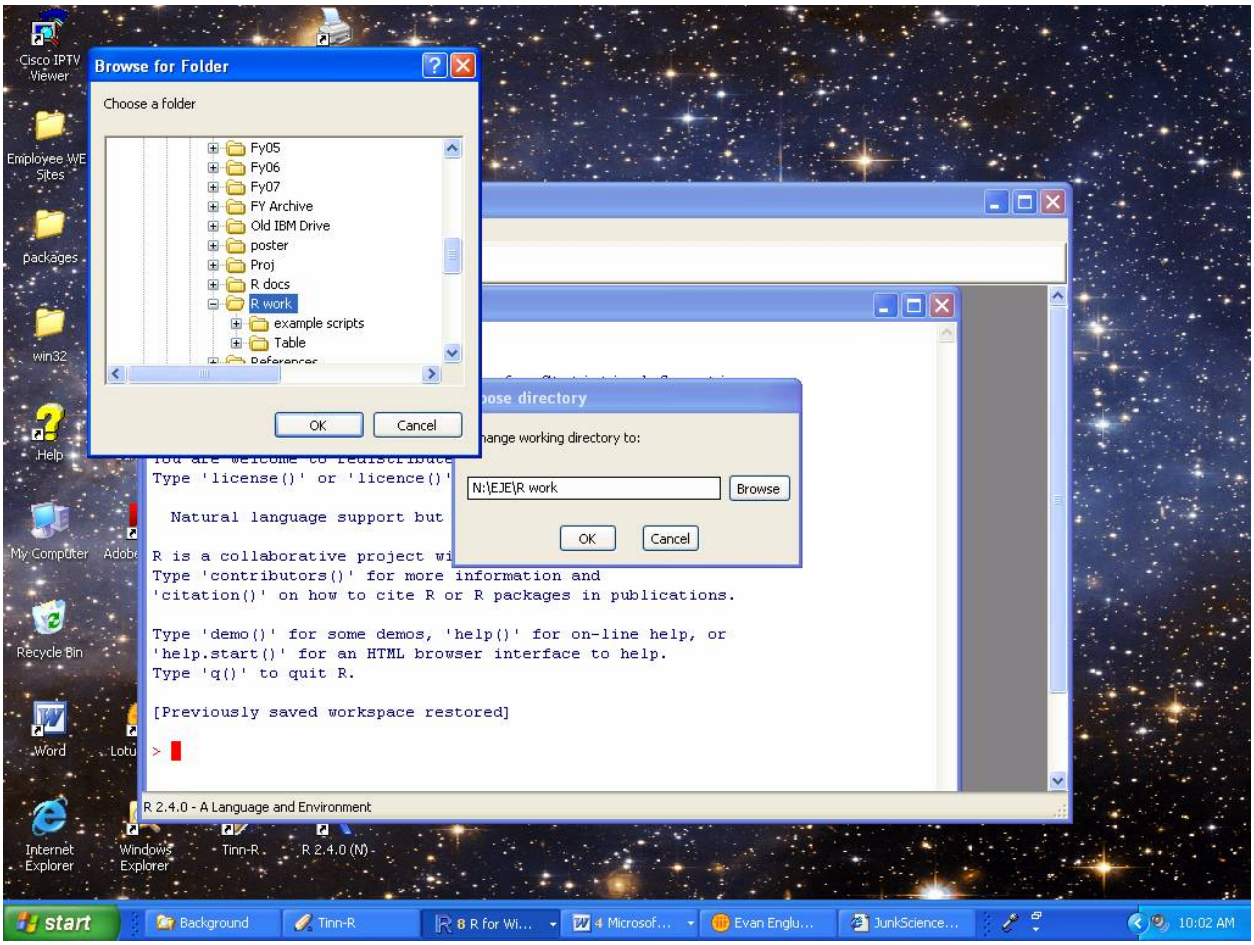
Run R. R will open with a main (RGui) window and a console window:



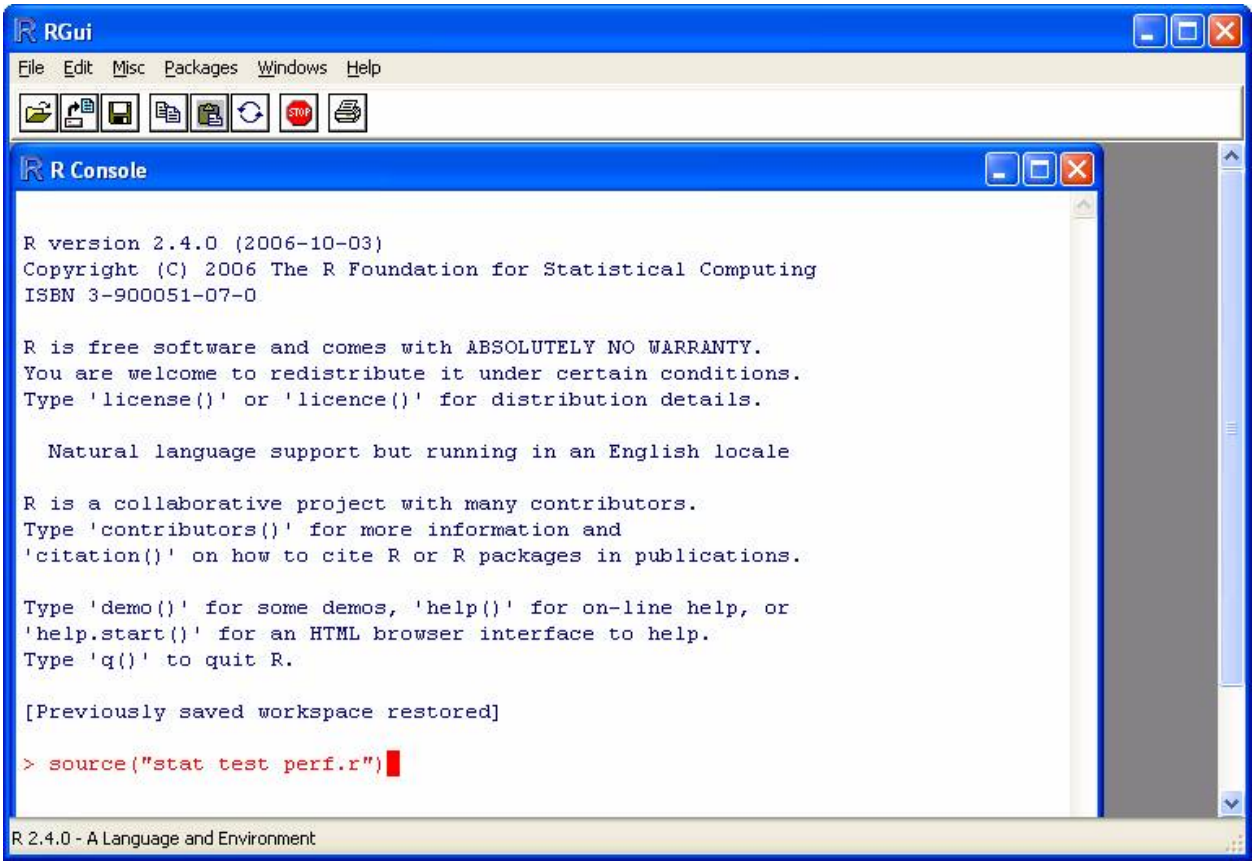
Click <File> on the toolbar, then <Change dir…> on the dropdown menu:



Browse and select your working directory (folder):

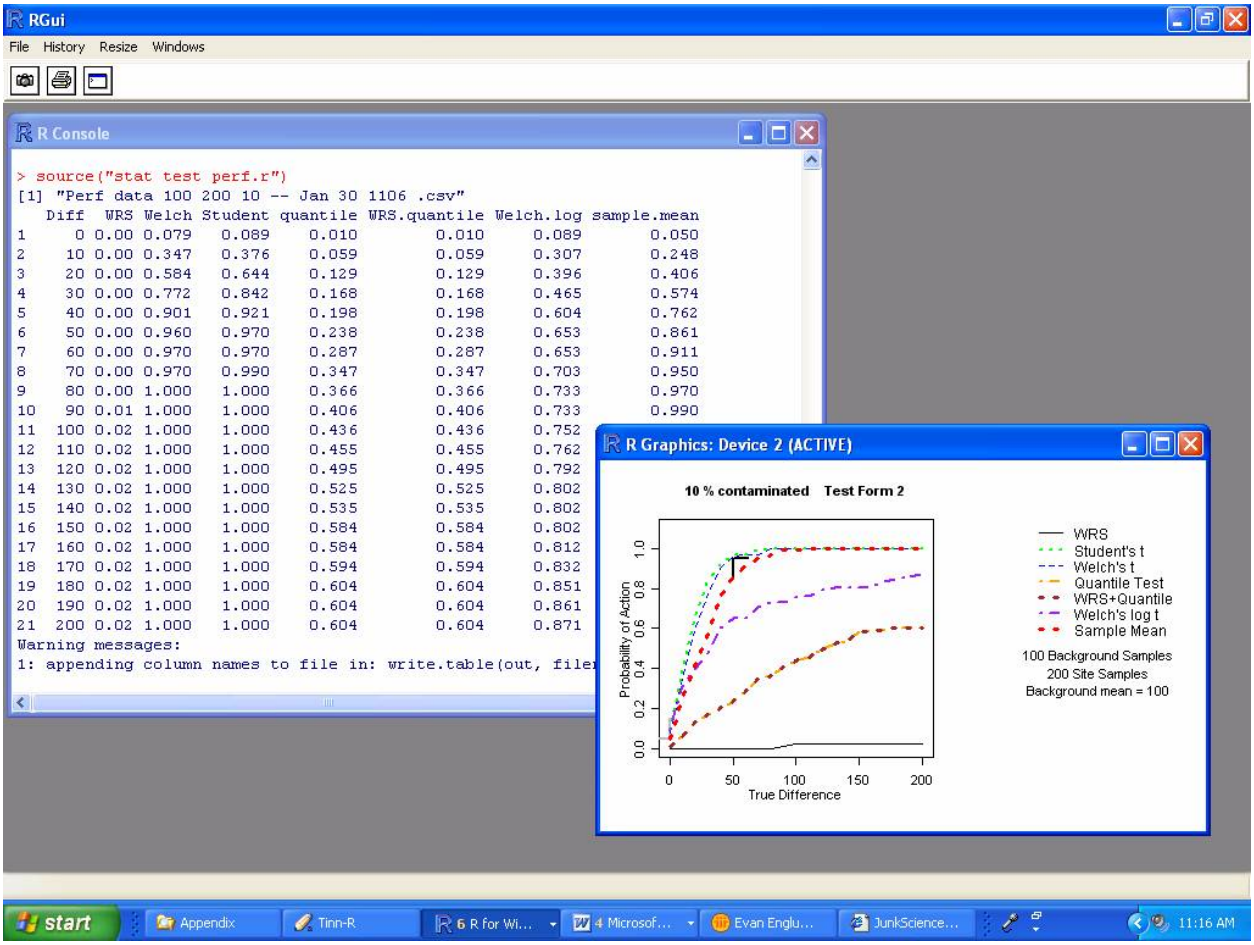


On the R console window, type the command **source("stat test perf.r")**:



R is a "command line interpreter" similar to Basic, rather than a compiled language, like Fortran or C++. When you hit <Enter>, R will execute the **source** command. R then reads and executes the R commands in the *stat test perf.r* script, which in turn reads and executes the R commands in the *parameters.r* and *sort.data.frame.r* script files. The parameters.r file contains the input parameters that control the simulation and statistical tests. These parameters are kept in a separate script file to make it easy for a user to make changes.

When R has finished, it will display a performance plot in a new graphics window, and print output on the console:



The graphics window can be saved in any of six common graphics formats.

The first line on the console after the **source** command is a file name into which R has written the output data, as well as a copy of the input parameters. The output file is a comma separated values file (.csv) that can be opened directly by a spreadsheet program. The file name is generated by the script, and includes the background sample size, site sample size, contamination percentage, and the date and time to make the name unique.

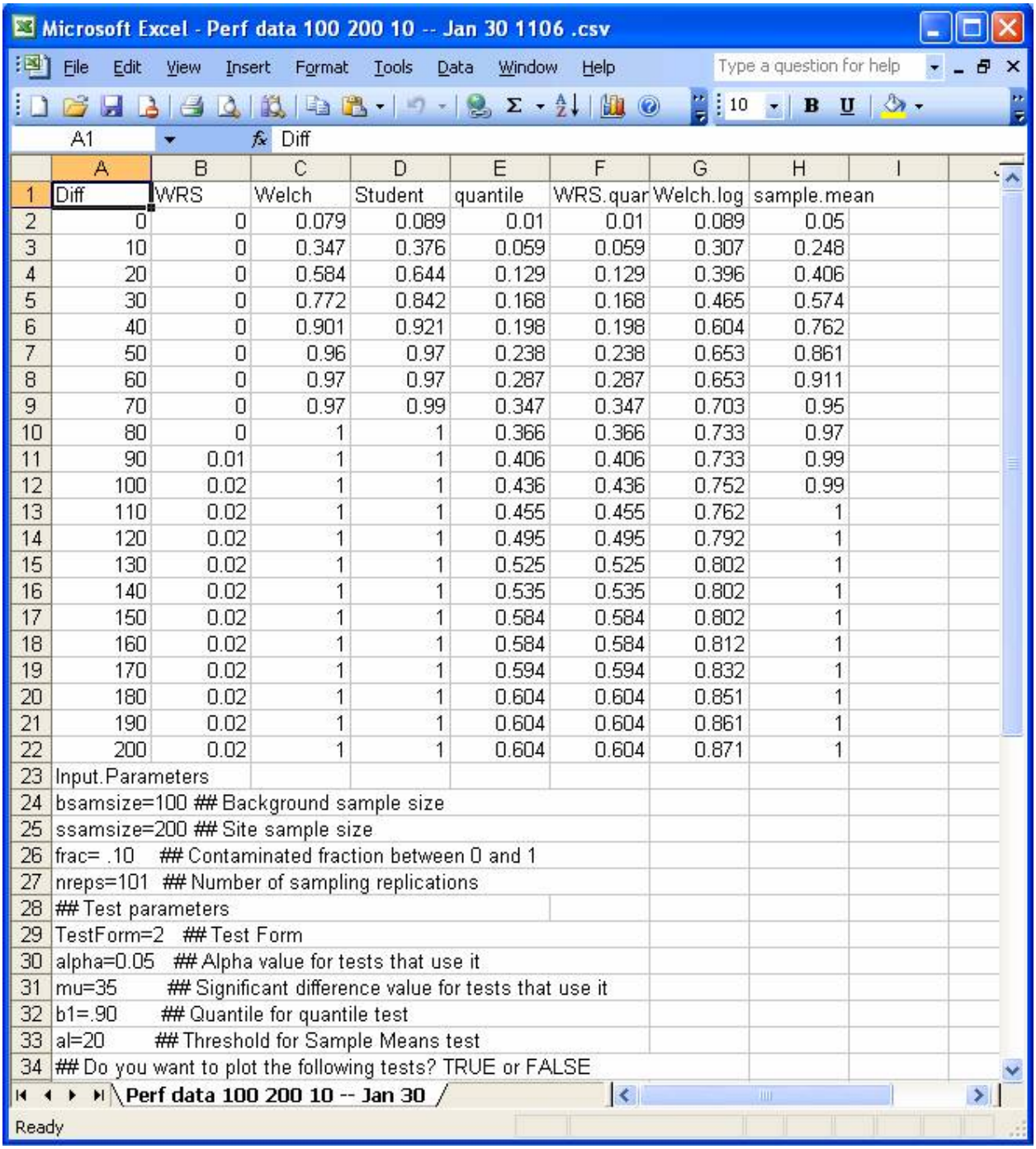
The data table on the console contains the data shown in the performance plot. Plotting the Diff column on the x axis versus any other column on the y axis will reproduce one of the performance curves.

The warning messages can be ignored. R is explaining – not very clearly – that it reformatted the data table in order to create the csv format output file.

Note that the sample sizes in the example above are not the same as any of those in the body of this report. They illustrate a step in the author's own rudimentary trial-and-error optimization attempt. Compare these results with Figure 15. This example uses the same total number of samples, but improves the performance of the t tests by taking fewer background and more site samples. Although the performance targets are not changed, the tests were "tweaked" by lowering the significant difference value ("mu" in the parameter file) from 50 to 35. This has the

effect of shifting the performance curve a little to the left. Searching for a better design is left as an exercise for the reader.

The screen shot below shows the output file after being opened in a spreadsheet.



	A	B	C	D	E	F	G	H	I
1	Diff	WRS	Welch	Student	quantile	WRS.quar	Welch.log	sample.mean	
2	0	0	0.079	0.089	0.01	0.01	0.089	0.05	
3	10	0	0.347	0.376	0.059	0.059	0.307	0.248	
4	20	0	0.584	0.644	0.129	0.129	0.396	0.406	
5	30	0	0.772	0.842	0.168	0.168	0.465	0.574	
6	40	0	0.901	0.921	0.198	0.198	0.604	0.762	
7	50	0	0.96	0.97	0.238	0.238	0.653	0.861	
8	60	0	0.97	0.97	0.287	0.287	0.653	0.911	
9	70	0	0.97	0.99	0.347	0.347	0.703	0.95	
10	80	0	1	1	0.366	0.366	0.733	0.97	
11	90	0.01	1	1	0.406	0.406	0.733	0.99	
12	100	0.02	1	1	0.436	0.436	0.752	0.99	
13	110	0.02	1	1	0.455	0.455	0.762	1	
14	120	0.02	1	1	0.495	0.495	0.792	1	
15	130	0.02	1	1	0.525	0.525	0.802	1	
16	140	0.02	1	1	0.535	0.535	0.802	1	
17	150	0.02	1	1	0.584	0.584	0.802	1	
18	160	0.02	1	1	0.584	0.584	0.812	1	
19	170	0.02	1	1	0.594	0.594	0.832	1	
20	180	0.02	1	1	0.604	0.604	0.851	1	
21	190	0.02	1	1	0.604	0.604	0.861	1	
22	200	0.02	1	1	0.604	0.604	0.871	1	
23	Input.Parameters								
24	bsamsize=100 ## Background sample size								
25	ssamsize=200 ## Site sample size								
26	frac= .10 ## Contaminated fraction between 0 and 1								
27	nreps=101 ## Number of sampling replications								
28	## Test parameters								
29	TestForm=2 ## Test Form								
30	alpha=0.05 ## Alpha value for tests that use it								
31	mu=35 ## Significant difference value for tests that use it								
32	b1=.90 ## Quantile for quantile test								
33	al=20 ## Threshold for Sample Means test								
34	## Do you want to plot the following tests? TRUE or FALSE								

Perf data 100 200 10 -- Jan 30

Ready

To change parameters, open the *parameters.r* file using Notepad. Edit the parameter values as desired, being careful to change only the numerical values or the TRUE-FALSE logical values.

Save the revised file under the same file name, return to the R console, and execute the **source("stat test perf.r")** command again.

If the scripts fail to run properly, replace them with copies of your backup scripts and try again.

Note. Each time the scripts are executed, they open another graphics window and add another output file to the working directory. Periodic housekeeping is necessary.

APPENDIX 2 - Script file: *parameters.r*

```
########### Begin Script ##############################
## parameters for statistical test performance simulation            |#
bsamsize=100 ## Background sample size                               |#
ssamsize=200 ## Site sample size                                     |#
frac= .10    ## Contaminated fraction between 0 and 1                |#
nreps=101  ## Number of sampling replications                        |#
## Test parameters                                                   |#
TestForm=2   ## Test Form                                            |#
alpha=0.05   ## Alpha value for tests that use it                    |#
mu=35        ## Significant difference value for tests that use it   |#
b1=.90       ## Quantile for quantile test                           |#
al=20        ## Threshold for Sample Means test                      |#
## Do you want to plot the following tests? TRUE or FALSE            |#
plotWRS=TRUE                # WRS                                    |#
plotStudentst=TRUE          # Students t                             |#
plotWelchst=TRUE            # Welch's t                              |#
plotQuantile=TRUE           # Quantile                               |#
plotWRSplusQuantile=TRUE    # WRS plus Quantile                      |#
plotWelchslogt=TRUE         # Welch's log t                          |#
plotSampleMean=TRUE         # Sample mean                            |#
## DQO targets (just for putting the ticks on, not used in tests)    |#
lft=0    # left bound of the gray region                             |#
rgt=50   # right bound of the gray region                            |#
bot=.05  # the alpha you want for test form 1                        |#
top=.95  # 1-(the alpha you want) for test form 2                    |#
## Defining the population parameters                                |#
bsd=0.8       # background log standard deviation (natural logs)     |#
bmean=100     # background arithmetic mean                           |#
csd=1.5       # contaminant log standard deviation                   |#
## Defining the True Difference points for the performance curve     |#
## (step=10 and numsteps=21 calculates 21 points:  0,10,20,...,200)  |#
step=10 ##                                                           |#
numsteps=21 ##
########### End Script ############################
```

APPENDIX 3 – Script file: *stat test perf.r*

```
########### Begin Script ##############################
## Evaluate performance of several statistical tests
##-------------------------------------------------
## get the parameters and a sort utility script
source("parameters.r")
source("sort.data.frame.r")
## initialize variables
```

```
stpvalue=numeric()
wtpvalue=numeric()
wpvalue=numeric()
probcleanw=numeric()
probcleanst=numeric()
probcleanwt=numeric()
qpvalue=numeric()
probcleanq=numeric()
cpvalue=numeric()
probcleanc=numeric()
lpvalue=numeric()
mpvalue=numeric()
probcleanl=numeric()
probcleanm=numeric()
diff=numeric()
bsam=numeric(nreps*bsamsize)
sbsam=numeric(nreps*ssamsize)
ssam=numeric(nreps*ssamsize)
csam=numeric(nreps*ssamsize)
## open the graphics window and set the layout
windows(width=5,height=3)
mx=c(1,1,1,2,2,1,1,1,2,2)
layout(matrix(mx, 2, 5, byrow = TRUE))
par(mgp=c(2,1,0)) ## reset at end to 3,1,0
## preliminary calculation for the quantile test
qnsite=qn=ssamsize
qmbkg=qm=bsamsize
qc=qmbkg+qnsite-floor((qmbkg+qnsite-1)*b1)-1
## set test parameters for thr R test functions
alt="l"
if(TestForm==1)alt="g"
mew=0
if(alt=="l")mew=mu
## create sets of background samples; set to specified mean
bsam=exp(rnorm(nreps*bsamsize,0,bsd))
bsam=bsam/mean(bsam)
bsam=bsam*bmean
dim(bsam)=c(nreps,bsamsize)
## create sets of site background samples before contamination
sbsam=exp(rnorm(nreps*ssamsize,0,bsd))
sbsam=sbsam/mean(sbsam)
sbsam=sbsam*bmean
dim(sbsam)=c(nreps,ssamsize)
## create sets of contanimant values
csam=exp(rnorm(nreps*ssamsize,0,csd))
csam[1:((1-frac)*length(csam))]=0 ## zero out the uncontaminated fraction
csam=csam/mean(csam)  ## make site contaminant mean = 1.0
csam=sample(csam) ## shuffle the data
dim(csam)=c(nreps,ssamsize)

## calculate fraction of samplings that result in a cleanup decision

for (m in (1:numsteps))  # test a range of differences between site and
background
{
## add site-related contaminant to background
diff[m]=step*(m-1)
ssam=sbsam+csam*diff[m]

for(i in 1:nreps)
```

```
  {
  ## generate p-values for the tests
  ## WRS
  wout=wilcox.test(ssam[i,],bsam[i,],alternative=alt,mu=mew)
  wpvalue[i]=wout$p.value
  ## Welch's t
  wtout=t.test(ssam[i,],bsam[i,],alternative=alt,mu=mew)
  wtpvalue[i]=wtout$p.value
  ## Student's t
  stout=t.test(ssam[i,],bsam[i,],alternative=alt,mu=mew,var.equal=TRUE)
  stpvalue[i]=stout$p.value
  ## quantile test
  qsam=c(bsam[i,],ssam[i,])
  code=c(rep(0,qm),rep(1,qn))
  first=length(code)-qc+1
  last=length(code)
  qdata=data.frame(code,qsam)
  qdata=sort.data.frame(~qsam,qdata)
  qs=sum(qdata$code[first:last])
    qmew=(qn*qc)/(qm+qn)
  qsigma=sqrt(qn*(qc/(qm+qn))*(1-(qc/(qm+qn)))*(qm/(qm+qn-1)))
  qpvalue[i]=1-pnorm((qs-0.5-qmew)/qsigma)
  ## combined quantile and WRS
  if(alt=="g")
    {
    cpvalue[i]=1
    if(wpvalue[i]<alpha)cpvalue[i]=0
    if(qpvalue[i]<alpha)cpvalue[i]=0
    }
  else
    {
    cpvalue[i]=0
    if(wpvalue[i]>alpha)cpvalue[i]=1
    if(qpvalue[i]<alpha)cpvalue[i]=1
    }
  ## Welch's log t

if(alt=="l"){lout=t.test(log(ssam[i,]),log(bsam[i,])+log((bmean+mew)/bmean),
    alternative=alt,mu=0,var.equal=FALSE)}
  else{lout=t.test(log(ssam[i,]),log(bsam[i,]),alternative=alt,
    mu=0,var.equal=FALSE)}
  lpvalue[i]=lout$p.value
  ## sample mean test
  ms=mean(ssam[i,])
  mpvalue[i]=1
  if((ms-mean(bsam[i,]))>al){mpvalue[i]=0}

} ## next i

## rejection probability (equals action probability for test form 1)
probcleanw[m]=length(wpvalue[wpvalue<(alpha)])/length(wpvalue)  #WRS
probcleanst[m]=length(stpvalue[stpvalue<(alpha)])/length(stpvalue)#Students t
probcleanwt[m]=length(wtpvalue[wtpvalue<(alpha)])/length(wtpvalue)#Welchs t
probcleanc[m]=length(cpvalue[cpvalue<(alpha)])/length(cpvalue) #combo: WRS &
quantile
probcleanq[m]=length(qpvalue[qpvalue<(alpha)])/length(qpvalue) #quantile test
probcleanl[m]=length(lpvalue[lpvalue<(alpha)])/length(lpvalue) #log Welch's t
probcleanm[m]=length(mpvalue[mpvalue<(alpha)])/length(mpvalue) #mean

}# next m diff step
```

```
md=max(diff)

## 1-rejection probability needed for test form 2
if(alt=="l")
  {
  probcleanst=1-probcleanst
  probcleanwt=1-probcleanwt
  probcleanw=1-probcleanw
  probcleanc=1-probcleanc
  probcleanl=1-probcleanl
  }
## plot empty performance diagram with title
plot(diff,probcleanw,xlim=range(0,diff[m]),type="n",lwd=1,
    ylim=c(0,1.1),cex=.4,xlab="",ylab="")
title(main=paste((frac*100),"% contaminated","   Test
Form",TestForm),cex.main=1,
     xlab="True Difference",
     ylab="Probability of Action")
## plot gray region tick marks
if(lft!=rgt)
{
cp=c("black","gray")
if(alt=="l"){cp=c("gray","black")}
lines(c(lft-.0625*md,lft,lft),c(bot,bot,bot+.1),col=cp[1],lwd=2)
lines(c(rgt,rgt,rgt+.0625*md),c(top-.1,top,top),col=cp[2],lwd=2)
}
## performance curves
if(plotWRS==TRUE)lines(diff,probcleanw,col="black",lwd=1)
if(plotWelchst==TRUE)lines(diff,probcleanwt,col="blue",lty=2)
if(plotStudentst==TRUE)lines(diff,probcleanst,col="green",lty=3,lwd=2)
if(plotQuantile==TRUE)lines(diff,probcleanq,lwd=2,col="orange",lty=4)
if(plotWRSplusQuantile==TRUE)lines(diff,probcleanc,col="brown",lty=3,lwd=3)
if(plotWelchslogt==TRUE)lines(diff,probcleanl,lwd=2,col="purple",lty=4)
if(plotSampleMean==TRUE)lines(diff,probcleanm,col="red",lty=3,lwd=3)

## plot legend
plot(0:1,0:1,type="n",xaxt="n",yaxt="n",xlab="",ylab="",axes=FALSE)
legend(0,1.05,
  legend=c("WRS","Student's t","Welch's t","Quantile Test",
    "WRS+Quantile","Welch's log t","Sample Mean"),
  lwd=c(1,2,1,2,3,2,3),
  col=c("black","green","blue","orange","brown","purple","red"),
  bty="n",cex=1.1,lty=c(1,3,2,4,3,4,3))

text(.5,.42,paste(bsamsize,"Background Samples"))
text(.5,.34,paste(ssamsize,"Site Samples"))
text(.5,.26,paste("Background mean =",bmean))

## reset graphics parameters and layout
par(mgp=c(3,1,0))
layout(c(1,1))

## prepare data for output
filename=paste("Perf data",bsamsize,ssamsize,frac*100,"--
",format(Sys.time(),"%b %d %H%M"),.”csv")
pw=round(probcleanw,3)
pwt=round(probcleanwt,3)
pst=round(probcleanst,3)
pq=round(probcleanq,3)
pc=round(probcleanc,3)
```

```
pl=round(probcleanl,3)
pm=round(probcleanm,3)
out=data.frame(diff,pw,pwt,pst,pq,pc,pl,pm)
names(out)=c("Diff","WRS","Welch","Student","quantile","WRS.quantile","Welch.
log","sample.mean")
print(filename)
print(out)
a=read.delim("parameters.r",sep="|",header=TRUE)
a=data.frame(a[,1])
names(a)="Input.Parameters"
## Write output file

write.csv(out,filename,append=TRUE,row.names=FALSE)
write.csv(a,filename,append=TRUE,row.names=FALSE)
########### End Script ##########################
```

APPENDIX 4 – Script file: *sort.data.frame*.r.

This script was written by Kevin Wright and kindly made available on the internet (http://tolstoy.newcastle.edu.au/R/help/04/09/4300.html). Mr. Wright's script has been evaluated by the author, who is solely responsible for its validity in this application.

This script looks different than the scripts in the two previous appendices because R recognizes two alternative symbols as the assignment operator: "=" and "<-." The command **a = b + c** has the identical meaning as **a <- b + c**, namely, to replace the value of the variable (**a)** to the left of the assignment operator symbol with the result obtained by evaluating the expression **(b+c)** to the right of the symbol. The choice of symbol is just a matter of personal preference.

```
########### Begin Script ###############################
sort.data.frame <- function(form,dat){
  # Author: Kevin Wright
  # Some ideas from Andy Liaw
  # http://tolstoy.newcastle.edu.au/R/help/04/07/1076.html


  # Use + for ascending, - for decending.
  # Sorting is left to right in the formula

  # Useage is either of the following:
  # sort.data.frame(~Block-Variety,Oats)
  # sort.data.frame(Oats,~-Variety+Block)

  # If dat is the formula, then switch form and dat
  if(inherits(dat,"formula")){
    f=dat
    dat=form
    form=f
  }
  if(form[[1]] != "~")
    stop("Formula must be one-sided.")


  # Make the formula into character and remove spaces
  formc <- as.character(form[2])
  formc <- gsub(" ","",formc)
```

```
  # If the first character is not + or -, add +
  if(!is.element(substring(formc,1,1),c("+","-")))
    formc <- paste("+",formc,sep="")
  # Extract the variables from the formula
  vars <- unlist(strsplit(formc, "[\\+\\-]"))
  vars <- vars[vars!=""] # Remove spurious "" terms


  # Build a list of arguments to pass to "order" function
  calllist <- list()
  pos=1 # Position of + or -
  for(i in 1:length(vars)){
    varsign <- substring(formc,pos,pos)
    pos <- pos+1+nchar(vars[i])
    if(is.factor(dat[,vars[i]])){
      if(varsign=="-")
        calllist[[i]] <- -rank(dat[,vars[i]])
      else
        calllist[[i]] <- rank(dat[,vars[i]])
    }
    else {
      if(varsign=="-")
        calllist[[i]] <- -dat[,vars[i]]
      else
        calllist[[i]] <- dat[,vars[i]]
    }
  }
  dat[do.call("order",calllist),]


}



d = data.frame(b=factor(c("Hi","Med","Hi","Low"),levels=c("Low","Med","Hi"),
               ordered=TRUE),
               x=c("A","D","A","C"),y=c(8,3,9,9),z=c(1,1,1,2))
sort.data.frame(~-z-b,d)
sort.data.frame(~x+y+z,d)
sort.data.frame(~-x+y+z,d)
sort.data.frame(d,~x-y+z)
############ End Script ############################
```

United States
Environmental Protection
Agency

Office of Research
and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
$300

EPA/600/R-07/020
March 2007
www.epa.gov

Please make all necessary changes on the below label, detach or copy, and return to the address in the upper left-hand corner.

If you do not wish to receive these reports CHECK HERE ☐; detach, or copy this cover, and return to the address in the upper left-hand corner.

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT No. G-35