APPENDIX M
Hypothesis Testing—Two-Population and General Cases

M-1. Introduction. A two-sample test is used when a data user is interested in making inferences about two independent populations, comparing some parameter from one population to the corresponding parameter from a second population. For example, a common environmental application entails comparing the population mean or median of the study area data set to the population mean or median of the background data set. EPA 600/R-96/084, QA/G-9 contains additional examples of the basic statistical tests presented here. Lehmann (1975) is a good resource for nonparametric tests. Montgomery (1997) contains a fuller treatment of two-sample *t*-tests, matched pairs *t*-tests, ANOVA, and multiple comparison tests.

M-2. Comparing Two Means. Two-sample tests do not require equal sample sizes, though equal sample sizes are recommended. The accuracy of estimating summary statistics from each sample is based on the number of samples available; data sets with many samples can provide more accurate estimates of the mean and standard deviation than those with only a few. When sample sizes are not equal, it may mean that one population is not defined as well as the other. If sample sizes are grossly unequal, the result of the two-sample test may produce an incorrect conclusion.

M-2.1. *Student's Two-Sample* t-*Test.* Student's two-sample *t*-test is a parametric statistical test that can be used to compare two population means based on the independent random samples x_1, x_2,..., x_m from the first population, and samples y_1, y_2,..., y_n from the second population. This test assumes the variances of the two populations are approximately equal. This supposition can be verified using an *F*-test or Levene's test (Appendix N, Paragraph N-4). However, the *F*-test is not recommended because it is not robust to deviations from normality. A positively skewed distribution tends to give rise to higher values of *F* and false rejection of the null hypothesis that the variances of two distributions are equal. If the two variances are not equal, the Satterthwaite's *t*-test is recommended (See Paragraph M-2.1.2 for directions and Paragraph M-2.1.3 for an example).

M-2.1.1. *Introduction*. The principal assumption required for the two-sample *t*-test is that a random sample of size *m* (x_1, x_2,..., x_m) is drawn from population 1, and an independent random sample of size *n* (y_1, y_2,..., y_n) is drawn from population 2. The second assumption required for the two-sample *t*-test is that the sample means, $\bar{x}$ (sample 1) and $\bar{y}$ (sample 2), are approximately normally distributed (if *X* and *Y* are normal, the sample means $\bar{x}$ and $\bar{y}$ will be also be normally distributed).

M-2.1.1.1. The two-sample *t*-test is commonly used to compare site contaminant concentrations to background concentrations:

$$H_0 : \mu_S - \mu_B \le \delta_0 ,\ H_A : \mu_S - \mu_B > \delta_0 \ .$$

The "true" mean site concentration and "true" mean background concentrations are denoted by μ_S and μ_B, respectively. When the above null hypothesis is selected, often $\delta_0 = 0$ and $\alpha = 0.2$ or 0.1. For this situation, the value of α tends to be somewhat higher than that used for other statistical applications (e.g., where α may be 0.05 or 0.01). This occurs to avoid a large Type II error (in this case, concluding the site is "clean" when it is "dirty" relative to background). As α decreases, the value of $\bar{x} > \bar{y}$ required to reject $H_0 : \mu_X \le \mu_Y$ increases. The following null and alternative hypotheses are also frequently used:

$$H_0 : \mu_S - \mu_B \ge \delta_0 ,\ \ H_A : \mu_S - \mu_B < \delta_0 \ .$$

M-2.1.1.2. In this situation, a common value for α is 0.05. However, the value for δ_0 depends greatly on the project. To reject H_0, that is, to demonstrate that the site is "clean" relative to background, the site mean must be significantly less than the background plus δ_0 (e.g., $\bar{x} << \bar{y} + \delta_0$). When there is actually no difference between the site and background populations (i.e., $\mu_S = \mu_B$), rejecting the null hypothesis in favor of the alternative hypothesis (i.e., the site is "clean" relative to background), becomes less probable as the selected value of δ_0 decreases. In general, a small value of δ_0 is undesirable from a cost perspective as a larger than budgeted number of samples may be required to determine if the means differ by δ_0. However, an extremely large value of δ_0 is undesirable from an environmental risk perspective as H_0 may be rejected even when the site mean is much larger than the background mean. Occasionally, δ_0 is equal to one or two standard deviations of the background data set. *The selection of an appropriate value of* δ_0 *is a critical component of the DQO process during project planning; the value should be established only after input is obtained from all users and stake holders.*

M-2.1.2. *Directions to Apply the Two-sample* t*-test for Differences Between the Population Means*. Steps to apply the two-sample *t*-test for differences between the population means for Case 1 and Case 2 are as follows: Case 1: $H_o : \mu_X - \mu_Y \le \delta_o$, $H_A : \mu_x - \mu_y \ge \delta_0$; and Case 2: $H_0 : \mu_x - \mu_y \ge \delta_0$, $H_A : \mu_x - \mu_y \le \delta_0$, which is given in braces { }.

M-2.1.2.1. Verify that both data sets are normal, using procedures in Appendices F and J, such as the Shapiro-Wilk test (Paragraphs F-3.2 and F-3.3) and a normal probability plot (Paragraphs J-5.5 and J-5.6).

M-2.1.2.2. Calculate the sample mean, $\bar{x}$, and the sample variance, s_X^2 (Appendix D), for the first data set (containing *m* points) and compute the sample mean, $\bar{y}$, and the sample variance, s_Y^2, for the second data set (containing *n* points).

M-2.1.2.3. Determine if the variances of the two populations are equal. If the variances of the two populations are not equal, use Satterthwaite's *t*-test (presented below). Otherwise, compute the pooled standard deviation:

$$s_E = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{(m-1) + (n-1)}}\ .$$

M-2.1.2.4. Calculate

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s_E\sqrt{1/n + 1/m}}\ .$$

M-2.1.2.5. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,m+n-2}$, such that $(1-\alpha)100\,\%$ of the *t*-distribution with $(m + n - 2)$ degrees of freedom is below $t_{1-\alpha,m+n-2}$.

M-2.1.2.5.1. If $t > t_{1-\alpha,m+n-2}\{t < -t_{1-\alpha,m+n-2}\}$, reject H_0. Go to step M-2.1.2.7.

M-2.1.2.5.2. If $t \le t_{1-\alpha,m+n-2}\{t \ge -t_{1-\alpha,m+n-2}\}$, there is not enough evidence to reject H_0. Therefore, the false acceptance error rate will need to be verified. Go to M-2.2.6.

M-2.1.2.6. To calculate the power of the test, assume that the true values for the mean and standard deviation are those obtained in the sample and use a statistical software package like DEFT (EPA QA/G-4D) or DataQUEST (EPA QA/G-9D) to generate the power curve of the two-sample *t*-test. If only one false acceptance error rate (β) has been specified (at δ_1), it is possible to calculate the sample size that achieves the DQOs, assuming the true mean and standard deviation are equal to the values estimated from the sample, instead of calculating the power of the test.

M-2.1.2.7. Calculate:

$$m^* = n^* = \frac{2s_E^2\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{(\delta_1 - \delta_0)^2} + (0.25)\,z_{1-\alpha}^2\ .$$

If $m^* \le m$ and $n^* \le n$, the false acceptance error rate has been satisfied. Otherwise, the false acceptance error rate has not been satisfied.

M-2.1.2.8. The results of the test could be:

M-2.1.2.8.1. H_0 is rejected; $\mu_x - \mu_y > \delta_0\{\mu_x - \mu_y < \delta_0\}$.

M-2.1.2.8.2. H_0 is not rejected and the false acceptance error rate is satisfied; $\mu_x - \mu_y \le \delta_0\{\mu_x - \mu_y \ge \delta_0\}$.

M-2.1.2.8.3. H_0 is not rejected and the false acceptance error rate was not satisfied; $\mu_x - \mu_y \le \delta_0 \{\mu_x - \mu_y \ge \delta_0\}$, but this conclusion is uncertain because the sample size was too small.

M-2.1.3. *Example of the Student's Two-Sample* t-*Test (Equal Variances) for Simple and Systematic Random Samples*. Consider the case where nickel (Ni) surface soil concentrations are compared between Site A and Background using the test:

$$H_o : \mu_X - \mu_Y \le \delta_o, \;\; H_A : \mu_x - \mu_y > \delta_0 \; .$$

Let *X* refer to the site Ni concentrations and *Y* to the background Ni concentrations. Let $\delta_0 = 0$.

M-2.1.3.1. The following Ni concentrations are obtained for the site soil (m = 6): 2.665, 3.610, 5.470, 7.150, 8.340, and 7.960 mg/kg.

M-2.1.3.2. The following Ni concentrations are obtained for the background soil (n = 10): 5.140, 7.460, 5.990, 3.360, 3.190, 2.870, 5.950, 1.720, 4.770, and 5.605 mg/kg.

M-2.1.3.3. In this example, the Shapiro-Wilk test was used to test the assumption of normality and an *F*-test was used to test the assumption of equal variances. Because the data have equal variances at a significance level of 0.05, the Student's two-sample *t*-test is more appropriate.

	Sample Mean	Sample Variance	Sample Size
Site data (*X*)	5.87	5.53	6
Background data (*Y*)	4.61	3.12	10

M-2.1.3.4. Using methods presented above in Paragraph M-2.1, determine if the variances of the two populations are equal. If the variances of the two populations are not equal, use Satterthwaite's *t*-test (Paragraph M-2.2). Otherwise, compute the pooled standard deviation:

$$s_E = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{(m-1) + (n-1)}} = \sqrt{\frac{(6-1)\times 5.53 + (10-1)\times 3.12}{(6-1) + (10-1)}} = 1.995 .$$

M-2.1.3.5. Calculate

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s_E\sqrt{1/n + 1/m}} = \frac{5.87 - 4.61 - 0}{1.995\sqrt{1/10 + 1/6}} = 1.22 .$$

M-2.1.3.6. Because we want an 80% level of confidence, $\alpha = 0.20$. So, $t_{0.80,14} = 0.8681$. Now compare the calculated value, *t*, with the critical value, $t_{0.80,14}$: $1.22 > 0.8681$. Therefore, reject H_0. At the 80% level of confidence, the mean concentration of Ni at Site A is greater than the mean background concentration of Ni.

M-2.2. *Satterthwaite's* t-*Test (Unequal Variances).* If the two variances are not equal, the use of Satterthwaite's *t*-test is recommended. Directions are provided below in Paragraph M-2.2.1, followed by an example in Paragraph M-2.2.2.

M-2.2.1. *Directions for Applying Satterthwaite's* t-*Test to Unequal Variances.* This describes the steps for applying the two-sample *t*-test for differences between the population means for: Case 1: $H_0 : \mu_X - \mu_Y \le \delta_o$ vs. $H_A : \mu_x - \mu_y > \delta_0$; and Case 2: $H_0 : \mu_x - \mu_y \ge \delta_0$ vs. $H_A : \mu_x - \mu_y < \delta_0$, which is given in braces { }.

M-2.2.1.1. Verify that both data sets come from a normal distribution, using the tests presented in Appendices F and J, such as the Shapiro-Wilk test (Paragraph F-3.2) and a normal probability plot (Paragraph J-5.5).

M-2.2.1.2. Calculate the sample mean, $\bar{x}$, and the sample variance, s_X^2 (Appendix C), for sample 1 and compute the sample mean, $\bar{y}$, and the sample variance, s_Y^2, for sample 2.

M-2.2.1.3. Test for equal variances, using tests presented in Appendix N, such as Bartlett's test (Paragraph N-3). If the variances are approximately equal, use the two-sample *t*-test (presented in Paragraph M-2.2.2). Otherwise, compute the standard deviation for unequal variances:

$$s_{NE} = \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}\ .$$

M-2.2.1.4. Calculate

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s_{NE}}.$$

M-2.2.1.5. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,\nu}$, such that $100(1-\alpha)$ % of the *t*-distribution with *v* degrees of freedom is below $t_{1-\alpha,\nu}$, and

$$\nu = \frac{\left[\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right]^2}{\frac{s_x^2}{m^2(m-1)} + \frac{s_y^2}{n^2(n-1)}}\ .$$

Round down the degrees of freedom to the nearest integer. Compare *t* to the critical value:

M-2.2.1.5.1. If $t > t_{1-\alpha,v} \{t < -t_{1-\alpha,v}\}$, H_0 may be rejected.

M-2.2.1.5.2. If $t \le t_{1-\alpha,v} \{t \ge -t_{1-\alpha,v}\}$, there is not enough evidence to reject H_0. Therefore, the false acceptance error rate will need to be verified. Go to M-2.2.1.6.

M-2.2.1.6. If H_0 was not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. To calculate the power, assume that the true values for the mean and standard deviation are those obtained in the sample and use a statistical software package to generate the power curve of the two-sample *t*-test. A simple method to check on statistical power *does not exist.*

M-2.2.1.7. The results of the test could be:

M-2.2.1.7.1. H_0 is rejected: $\mu_x - \mu_y > \delta_0 \{\mu_x - \mu_y < \delta_0\}$.

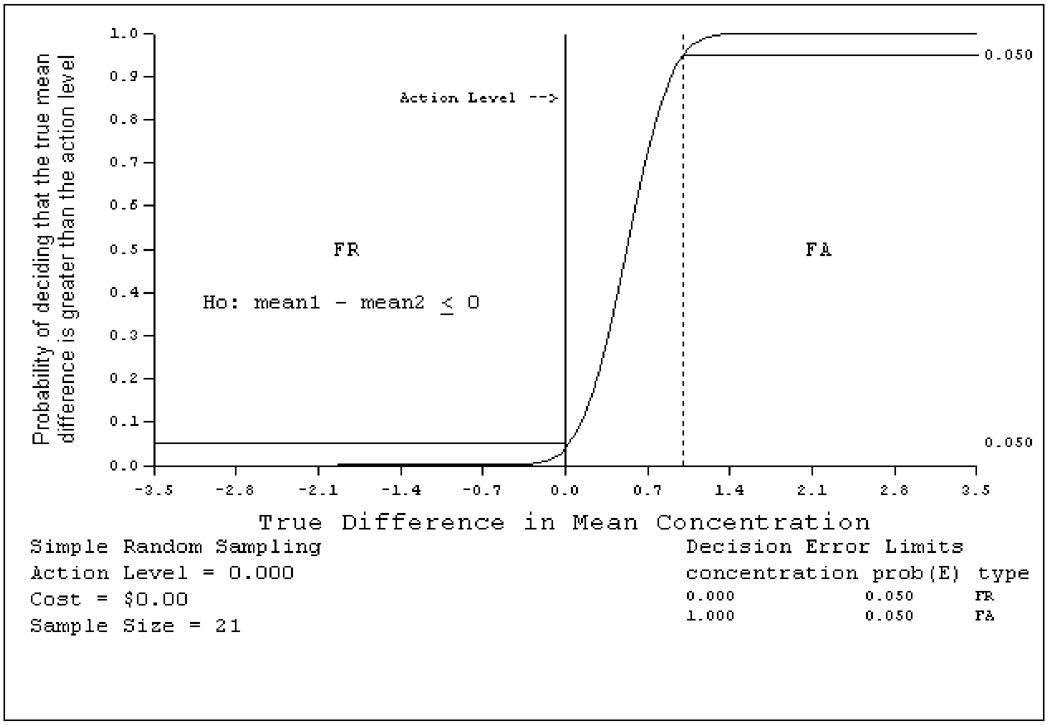
M-2.2.1.7.2. H_0 is not rejected and the false acceptance error rate is satisfied, $\mu_x - \mu_y \le \delta_0 \{\mu_x - \mu_y \ge \delta_0\}$.

M-2.2.1.7.3. H_0 is not rejected but the false acceptance error rate is not satisfied; H_0 is uncertain because the sample size was too small.

M-2.2.2. *Example of Applying Satterthwaite's* t*-test to Unequal Variances*. Because we want a 95% level of confidence, $\alpha = 0.05$ and v = 6 (round down to the nearest integer). So, $t_{0.95,6} = 1.943$. Now compare the calculated value (*t*) with the critical value, $t_{0.95,6}$. Because $-1.031 \le 1.943$, there is not enough evidence to reject H_0.

M-2.2.2.1. As a result of not having enough evidence to reject the null hypothesis, it is necessary to calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. DEFT can be used to evaluate power and sample size and is presented in this example. To calculate the power of the test, one must consider what an acceptable difference among the means is before concluding H_0 should be rejected. The difference that one is willing to accept depends on the detection limits achieved, the range of concentrations from each data set, and what is considered to have practical significance vs. statistical significance.

M-2.2.2.2. The power curve (Figure M-1) shows where a statistically significant difference between the means was assumed to be 1 mg/kg (the region between the vertical dashed and solid lines). According to DEFT, 21 samples are needed for the estimated performance curve. In the above example, the site data have 36 samples and the background data only have 8. Therefore, there may be a need to take more background samples. It is important to note that the true difference in the mean (4.619 – 4.925 = –0.31) is to the left of the action level.



Simple Random Sampling
Action Level = 0.000
Cost = $0.00
Sample Size = 21

Decision Error Limits

concentration	prob(E)	type
0.000	0.050	FR
1.000	0.050	FA

Figure M-1. Estimated power performance curve.

M-2.3. *Matched Pairs* t-*Test.*

M-2.3.1. *Introduction.* Sometimes, the two populations of interest represent different measurements on the same homogenous group. For example, contaminant concentration in groundwater before and after a certain remediation treatment may need to be compared. If measurements are taken from the same set of wells both before and after treatment, we can match the results by well. That is, each well will have a result from before the treatment and a result from after the treatment. Under this experimental design, the observed differences for each well before and after treatment become the sample data because we expect the two results from each well to be more homogeneous than the results among wells.

M-2.3.1.1. The differences are then analyzed using the one-sample *t*-test if the assumptions for that test are met. Namely, the one-sample *t*-test assumes the differences represent a random sample. It also assumes that the average difference follows a normal distribution. If the normal assumption is not valid, Paragraph M-4.1.6 discusses a non-parametric alternative for matched pairs designs. In addition to matched pairs, one would ideally assign the order of the treatments randomly to each subject, although that would not be possible in the groundwater remediation example. Matching can also occur between subjects that are closely alike in all respects except the treatment that is applied.

M-2.3.1.2. The matched pairs *t*-test is commonly used to compare site contaminant concentrations before and after a treatment:

$$H_0 : \mu_A \geq \mu_B \, , \, H_A : \mu_A < \mu_B \ .$$

M-2.3.1.3. The “true” mean concentration *before* treatment and the “true” mean concentration *after* treatment are denoted by μ_B and μ_A, respectively. The before treatment mean is often referred to as the “baseline” mean. Directions are provided below in Paragraph M-2.3.2, followed by an example in Paragraph M-2.3.3.

M-2.3.2. *Directions to Apply the Matched Pairs* t-*test for Differences Between the Means Before and After a Treatment*. Steps to apply the Matched Pairs *t*-test for differences between the means for Case 1 and Case 2 are as follows: Case 1: $H_0 : \mu_A \geq \mu_B$, $H_A : \mu_A < \mu_B$; and Case 2: $H_0 : \mu_A \leq \mu_B$, $H_A : \mu_A > \mu_B$, which is given in braces { }.

M-2.3.2.1. Subtract the before treatment concentration (B_i) from the corresponding after treatment concentration (A_i) for each pair of results (B_i, A_i) to obtain the differences:

$$d_i = A_i - B_i \ .$$

M-2.3.2.2. Verify that the differences, $d_1, d_2, d_3 ... d_n$, are normal, using procedures in Appendices F and J, such as the Shapiro-Wilk test (Paragraphs F-3.2 and F-3.3) and a normal probability plot (Paragraphs J-5.5 and J-5.6).

M-2.3.2.3. Calculate the sample mean, $\bar{d}$, and the sample variance, s_d^2 (Appendix D).

M-2.3.2.4. Calculate

$$t = \frac{\bar{d}}{s_d \big/ \sqrt{n}} \, .$$

M-2.3.2.5. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,n-1}$, such that $(1-\alpha)100\,\%$ of the *t* distribution with (n – 1) degrees of freedom is below $t_{1-\alpha,n-1}$.

M-2.3.2.5.1. If $t < -t_{1-\alpha,n-1} \{t > t_{1-\alpha,n-1}\}$, reject H_0. Go to M-2.3.2.7.

M-2.3.2.5.2. If $t \geq -t_{1-\alpha,n-1} \{t \leq t_{1-\alpha,n-1}\}$, there is not enough evidence to reject H_0. Therefore, the false acceptance error rate will need to be verified. Go to M-2.3.2.6.

M-2.3.2.6. To calculate the power of the test, assume that the true values for the mean and standard deviation are those obtained in the sample and use a statistical software package like the DEFT software (EPA QA/G-4D) or the DataQUEST software (EPA QA/G-9D) to generate the power curve of the matched pairs *t*-test. If only one false acceptance error rate (β) has been specified (at μ_1), it is possible to approximately calculate the sample size that achieves the DQOs, assuming the true mean and standard deviation are equal to the values estimated from the sample, instead of calculating the power of the test. A derivation of the following formula is given in Appendix A of EPA 600/R-96/055, QA/G-4.

M-2.3.2.7. Calculate:

$$m = \frac{s_d^2 \left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\left(\bar{d}\right)^2} + (0.5)Z_{1-\alpha}^2$$

where Z_p is the $p100^{\text{th}}$ percentile of the standard normal distribution (Table B-15 of Appendix B). Round m up to the next integer. If $m \le n$, the false acceptance error rate has been satisfied. If $m > n$, the false acceptance error rate has not been satisfied.

M-2.3.2.8. The results of the test could be:

M-2.3.2.8.1. H_0 is rejected; $\mu_A < \mu_B \; \{\mu_A > \mu_B\}$.

M-2.3.2.8.2. H_0 is not rejected and the false acceptance error rate is satisfied; $\mu_A \ge \mu_B \; \{\mu_A \le \mu_B\}$.

M-2.3.2.8.3. H_0 is not rejected and the false acceptance error rate was not satisfied; $\mu_A \ge \mu_B \; \{\mu_A \le \mu_B\}$, but this conclusion is uncertain because the sample size was too small.

M-2.3.3. *Example of the Matched Pairs* t-*Test for the Difference Between Means Before and After Treatment*. Consider the case where the results of a groundwater remediation procedure are compared before and after treatment to determine if the remediation has decreased the concentration of the contaminant. Test the null hypothesis that the treatment had no lowering effect at the 95% level of confidence:

$$H_0 : \mu_A \ge \mu_B , \; H_A : \mu_A < \mu_B \; .$$

M-2.3.3.1. The data consist of measured TCE concentrations (mg/L) at monitoring wells before and after a treatment-test, given in Table M-1.

M-2.3.3.2. Determine if the differences follow a normal distribution. A Shapiro-Wilk test for normality does not reject the hypothesis that the differences are normal (p = 0.4248). So, assuming normality is reasonable.

M-2.3.3.3. Calculate

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{-18.0}{13.9/\sqrt{10}} = -4.10 \, .$$

M-2.3.3.4. Assume that we want a 95% level of confidence, $\alpha = 0.05$. So, $t_{0.95,9} = 1.833$. Now compare the calculated value, *t*, with the critical value $-t_{0.95,9}$: $-4.10 < -1.833$. Therefore, reject H_0. This means that there is a lower mean concentration of TCE after remediation.

Table M-1.
Measured TCE Concentrations (mg/L) at Monitoring Wells Before and After a Treatment Test

Sample ID	Baseline (01/2000)	Post–Test (12/2000)	Difference
Well 1	20.9	0.917	–20.0
Well 2	9.17	8.77	–0.400
Well 3	5.96	4.37	–1.59
Well 4	41.5	4.34	–37.2
Well 5	34.3	10.7	–23.6
Well 6	19.7	1.48	–18.2
Well 7	38.9	0.272	–38.6
Well 8	8.18	0.520	–7.66
Well 9	9.13	3.06	–6.07
Well 10	28.5	1.90	–26.6

M-3. Comparing Proportions and Percentiles: Two-Sample Test for Proportions. This Paragraph considers hypotheses concerning two population proportions (or percentiles). The two-sample test for proportions can be used to compare two population percentiles or proportions and is based on an independent random sample of *m* ($x_1, x_2, \ldots, x_m$) from the first population and an independent random sample of size *n* ($y_1, y_2, \ldots, y_n$) from the second population. The sample proportion for the first population is represented by p_1 and the sample proportion for the second population is represented by p_2.

M-3.1. *Introduction*. The principal assumption for this non-parametric test is that of random sampling from the two populations. The two-sample test for proportions is valid (robust) for any underlying distributional shape and is robust to outliers, providing they are not pure data errors. Directions for a two-sample test for proportions for a simple random sample and a systematic simple random sample are given below in Paragraph M-3.2, followed by an example in Paragraph M-3.3.

M-3.2. *Directions for Applying the Two-Sample Test for Proportions.* Directions for applying the two-sample test for proportions are presented for Case 1: $H_0 : P_1 - P_2 \le \delta_0$ and $H_A : P_1 - P_2 > \delta_0$; and Case 2: $H_0 : P_1 - P_2 \ge \delta_0$ and $H_A : P_1 - P_2 < \delta_0$, which is given in braces { }. Given *m* random samples $x_1, x_2, \ldots, x_m$ from the first population, and *n* samples from the second population, $y_1, y_2, \ldots, y_n$, let k_1 be the number of points from sample 1 which exceed some concentration *C*, and let k_2 be the number of points from sample 2 that exceed *C*.

M-3.2.1. Calculate the sample proportions: $p_1 = k_1 / m$, $p_2 = k_2 / n$.

M-3.2.2. Calculate the pooled proportion: $p = (k_1 + k_2)/(m + n)$.

M-3.2.3. Compute:

mp_1, $m(1-p_1)$, np_2, $n(1-p_2)$.

If all of the above values are greater than or equal to 5, continue. Otherwise, seek assistance from a statistician as analysis is complicated.

M-3.2.4. Calculate:

$$z = (p_1 - p_2)/\sqrt{p(1-p)(1/m + 1/n)}$$

M-3.2.5. Use Table B-15 of Appendix B to find the critical value, $Z_{1-\alpha}$, such that $(1-\alpha)100\%$ of the normal distribution is below $Z_{1-\alpha}$. For example, if α = 0.05 then $Z_{1-\alpha}$ = 1.645.

M-3.2.5.1. If $z > Z_{1-\alpha}$ $\{z < -Z_{1-\alpha}\}$, reject H_0.

M-3.2.5.2. If $z \le Z_{1-\alpha}$ $\{z \ge -Z_{1-\alpha}\}$, do not reject H_0. Proceed to M-3.2.6 to calculate the false acceptance error rate.

M-3.2.6. If H_0 is not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. If only one false acceptance error rate (β) has been specified at $P_1 - P_2$, it is possible to calculate the sample sizes that achieve the DQOs (assuming the proportions are equal to the values estimated from the sample) instead of calculating the power of the test. To do this, calculate:

$$m^* = n^* = \frac{2\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 \overline{P}(1-\overline{P})}{\left(P_2 - P_1\right)^2}$$

$$\overline{P} = \frac{P_1 + P_2}{2}\ .$$

Z_p is the $p100^{\text{th}}$ percentile of the standard normal distribution (Table B-15 of Appendix B).

M-3.2.6.1. If $m > m^*$ and $n > m^*$, then the false acceptance error rate has been satisfied.

M-3.2.6.2. If both m and n are below m^*, the false acceptance error rate has not been satisfied.

M-3.2.6.3. If m^* is between m and n, use a software package like the DEFT or DataQUEST to calculate the power of the test, assuming that the true values for the proportions P_1 and P_2 are those obtained in the sample.

M-3.2.6.4. If the estimated power is below $1 - \beta$, the false acceptance error rate has not been satisfied.

M-3.2.7. The results of the test could be:

M-3.2.7.1. H_0 is rejected; $P_1 - P_2 > \delta_0$ $\{P_1 - P_2 < \delta_0\}$.

M-3.2.7.2. H_0 was not rejected, the false acceptance error rate was satisfied, and it seems $P_1 - P_2 \le \delta_0$ $\{P_1 - P_2 \ge \delta_0\}$.

M-3.2.7.3. H_0 was not rejected, the false acceptance error rate was not satisfied, and it seems $P_1 - P_2 \le \delta_0$ $\{P_1 - P_2 \ge \delta_0\}$, but this outcome is uncertain because the sample size was probably too small.

M-3.3. *Example of Two-Sample Test for Proportions for Simple and Systematic Random Samples*. Gasoline groundwater concentrations at Site A are compared to background concentrations:

$$H_0 : P_1 - P_2 \le \delta_0,\ H_A : P_1 - P_2 > \delta_0 .$$

M-3.3.1. The groundwater site data are following (m = 15): 243, 700, 781, 385, 642, 97.2, 233, 11.1, 10.60, 14.90, 14.90, 12.70, 9.57, 6.04, and 7.32 µg/L.

M-3.3.2. The groundwater background data are following (n = 45): 177.0, 4.27, 10.60, 10.60, 14.90, 14.60, 12.70, 9.57, 95.70, 7.32, 7.32, 7.32, 6.58, 6.90, 6.90, 39.5, 4.27, 10.60, 10.60, 14.90, 14.60, 12.70, 9.57, 6.04, 7.32, 7.32, 7.32, 146.00, 6.90, 6.90, 44.5, 4.27, 10.60, 10.60, 14.90, 14.60, 12.70, 9.57, 6.04, 7.32, 7.32, 7.32, 111.00, 6.90, and 6.90 µg/L.

	k_i	Sample Size
Site data (i = 1)	7	15
Background data (i = 2)	6	45

where k_i is the number of detected concentrations above the regulatory threshold (35 µg/L).

M-3.3.3. Determine whether or not mp_1, $m(1 - p_1)$, np_2, $n(1 - p_2)$ are all greater than 5:

$$p_1 = k_1 / m = 7/15 = 0.467$$

$$p_2 = k_2 / n = 6/45 = 0.133$$

$$mp_1 = 15(0.467) = 7 > 5$$

$m(1-p_1) = 15(1-0.467) = 8 > 5$

$np_2 = 45(0.133) = 6 > 5$

$n(1-p_2) = 45(1-0.133) = 39 > 5$.

M-3.3.4. Calculate the following:

$p = (k_1 + k_2)/(m+n) = (7+6)/(15+45) = 0.217$

$$\begin{aligned} z &= (p_1 - p_2)/\sqrt{p(1-p)(1/m + 1/n)} \\ &= (0.467-0.133)/\sqrt{0.217(1-0.217)(1/15 + 1/45)} = 2.72 \end{aligned} .$$

M-3.3.5. Because the level of confidence is 95%, $\alpha = 0.05$. Using Table B-15, we find that $Z_{1-0.05} = 1.645$. Now compare the calculated value, *z*, with the critical value, $Z_{1-0.05}$: $2.74 > 1.645$.

M-3.3.6. Therefore, there is enough evidence to reject H_0 (i.e., the results suggest that the proportion of samples with gasoline levels above the regulatory threshold in the site well samples is greater than the proportion above the regulatory threshold in the background well samples).

M-4. Nonparametric Comparisons of Two Populations

M-4.1. *The Wilcoxon Rank Sum Test.* The Wilcoxon rank sum test is a nonparametric test that can be used to compare two population distributions based on *n* independent random samples ($x_1, x_2, \ldots, x_n$) from the first population, and *m* independent random samples ($y_1, y_2, \ldots, y_m$) from the second population. The most general form of the hypotheses for a one-tailed Wilcoxon rank sum test can be stated in terms of the probability that an observation from distribution *Y* exceeds a value from distribution *X*, such as:

$$H_0 : P(X < Y) \geq 0.5, \;\; H_A : P(X < Y) < 0.5 \; .$$

M-4.1.2. *Introduction*. Hypotheses on the relative rank of the mean of each population can also be formulated with the additional assumption that the two underlying distributions have the same shape and dispersion (Conover, 1980). That is, one distribution differs by some fixed amount (or is increased by a constant) when compared to the other distribution. An important advantage of the Wilcoxon rank sum test is its partial robustness to outliers, because the analysis is conducted on rankings of the observations. This limits the influence of outliers because a given observation can be no more extreme than the first or last rank. Directions and an example for the Wilcoxon rank sum test are given in Paragraphs M-4.1.3 and M-4.1.4, respectively. If a relatively large number of samples have been taken, it is more efficient to use the large sample approximation to the Wilcoxon rank sum test (Paragraph M-4.1.6) to perform the hypothesis test.

M-4.1.3. *Directions for the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples*.

M-4.1.3.1. Let $x_1, x_2, \ldots, x_n$ represent the *n* observations from population 1 and $y_1, y_2, \ldots, y_m$ represent the *m* observations from population 2, where both *n* and *m* are less than or equal to 20.

M-4.1.3.1.1. *Case 1*:

$H_0 : P(X < Y) \geq 0.5$: Values of *X* tend to be smaller than or equal to values of *Y*.

$H_A : P(X < Y) < 0.5$: Values of *X* tend to be larger than values of *Y*.

M-4.1.3.1.2. *Case 2*:

$H_0 : P(X < Y) \leq 0.5$: Values of *X* tend to be larger than or equal to values of *Y*.

$H_A : P(X < Y) > 0.5$: Values of *X* tend to be smaller than values of *Y*.

M-4.1.3.1.3. *Case 3*:

$H_0 : P(X < Y) = 0.5$: Values of *X* tend to be equal to values of *Y*.

$H_A : P(X < Y) \neq 0.5$: Values of *X* tend to be smaller than or greater than values of *Y*.

M-4.1.3.2. If either *m* or *n* is larger than 20 and the smaller of the two is at least 4 (Lehmann, 1975), use the large sample approximation described in Paragraph M-4.1.5.

M-4.1.3.3. Combine the two data sets and rank the measurements (from both data sets) from smallest to largest, keeping track of which population contributed each measurement.

M-4.1.3.3.1. Assign the rank of 1 to the smallest value of the combined data sets and note whether the smallest value is from population 1 or 2.

M-4.1.3.3.2. Assign the rank of 2 to the second smallest value of the combined data sets (noting the population), and so forth.

M-4.1.3.3.3. If there are ties, assign the average of the ranks that would otherwise have been assigned to the tied observations.

M-4.1.3.4. Calculate *R,* the sum of the ranks of the data from population 1, and then calculate:

$$W = R - \frac{n(n+1)}{2}.$$

M-4.1.3.5. Use Table B-17 of Appendix B to find the critical value, W_α (or $W_{\alpha/2}$ for Case 3).

M-4.1.3.6. Compare W to the critical value W_α.

M-4.1.3.6.1. For Case 1, reject H_0 if $W > nm - W_\alpha$.

M-4.1.3.6.2. For Case 2, reject H_0 if $W < W_\alpha$.

M-4.1.3.6.3. For Case 3, reject H_0 if $W > nm - W_{\alpha/2}$ or $W < W_{\alpha/2}$.

M-4.1.3.7. The results of the test could be:

M-4.1.3.7.1. H_0 was rejected and it seems values from population 1 tend to be greater than (Case 1), smaller than (Case 2), or different from (Case 3) values from population 2.

M-4.1.3.7.2. H_0 was not rejected, and it seems that values from population 1 tend to be smaller than or equal to (Case 1), greater than or equal to (Case 2), or not different from (Case 3) values from population 2.

M-4.1.3.7.3. If H_0 is not rejected, it should be determined whether adequate power was achieved. However, as power calculations tend to be complex and difficult to do manually, it is recommended that a statistician be consulted.

M-4.1.4. *Example of the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples*.

M-4.1.4.1. Consider the Case 1 (Paragraph M-4.1.3), where lead (Pb) surface soil concentrations are compared between Site A and background at a significance level of $\alpha = 0.05$ using the test.

M-4.1.4.1.1. H_0: Site A Pb concentrations tend to be less than or equal to background Pb concentrations.

M-4.1.4.1.2. H_A: Site A Pb concentrations tend to be greater than background Pb concentrations.

M-4.1.4.2. Suppose the Pb surface site concentrations (X) are as follows (n = 20): 8.24, 6.57, 4.48, 4.34, 16.00, 3.83, 4.11, 3.48, 3.66, 5.01, 93.80, 3.70, 129.00, 4.92, 91.80, 3.86, 4.21, 4.32, 10.00, and 9.38 mg/kg.

M-4.1.4.3. Suppose the Pb surface background concentrations (Y) are as follows (m = 16): 3.81, 3.68, 3.72, 3.68, 5.97, 4.12, 6.42, 4.13, 8.88, 3.01, 5.34, 3.74, 10.70, 3.86, 10.80, and 4.40 mg/kg.

Table M-2.
Example M-4.1.4 Pb Concentrations

Location	Result	Rank	Location	Result	Rank
background	3.01	1	background	4.4	19
Site	3.48	2	site	4.48	20
Site	3.66	3	site	4.92	21
background	3.68	4.5	site	5.01	22
background	3.68	4.5	background	5.34	23
Site	3.70	6	background	5.97	24
background	3.72	7	background	6.42	25
background	3.74	8	site	6.57	26
background	3.81	9	site	8.24	27
Site	3.83	10	background	8.88	28
background	3.86	11.5	site	9.38	29
Site	3.86	11.5	site	10.0	30
Site	4.11	13	background	10.7	31
background	4.12	14	background	10.8	32
background	4.13	15	site	16.0	33
Site	4.21	16	site	91.8	34
Site	4.32	17	site	93.8	35
Site	4.34	18	site	129.0	36

$$W = R - \frac{n(n+1)}{2} = 409.5 - \frac{20(20+1)}{2} = 199.5$$

$$W_\alpha = W_{0.05} = 108$$

$$nm - W_\alpha = (20)(16) - 108 = 212 \ .$$

M-4.1.4.4. Because $199.5 \leq 212$, H_0 cannot be rejected. There is insufficient evidence to conclude that the lead concentrations from Site A are greater than background lead concentrations.

M-4.1.5. *Large Sample Approximation of the Wilcoxon Rank Sum Test.* When a relatively large number of samples has been taken, it is more efficient to use a large sample approximation of the Wilcoxon rank sum test to obtain the critical value of W. Directions and an example are presented in Paragraphs M-4.1.5.1 and M-4.1.5.2, respectively. Required sample size to achieve a specified power is explored in Paragraphs M-4.1.4.3 and M-4.1.4.4.

M-4.1.5.1. *Directions for a Large Sample Approximation of the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples.*

M-4.1.5.1.1. Let $x_1, x_2, \ldots, x_n$ represent the *n* observations from population 1 and $y_1, y_2, \ldots, y_m$ represent the *m* observations from population 2 where either *n* or *m* is greater than 20 and the smaller of *n* and *m* is at least 4 (Lehmann, 1975). The following hypothesis tests are considered:

M-4.1.5.1.1.1. *Case 1.* $H_0 : P(X < Y) \geq 0.5$, $H_A : P(X < Y) < 0.5$.

M-4.1.5.1.1.2. *Case 2.* $H_0 : P(X < Y) \leq 0.5$, $H_A : P(X < Y) > 0.5$.

M-4.1.5.1.1.3. *Case 3.* $H_0 : P(X < Y) = 0.5$, $H_A : P(X < Y) \neq 0.5$.

M-4.1.5.1.2. List and rank the measurements from both populations from smallest to largest, keeping track of which population contributed each measurement.

M-4.1.5.1.2.1. The rank of 1 is assigned to the smallest value of the combined data sets, the rank of 2 to the second smallest value of the combined data sets, and so forth.

M-4.1.5.1.2.2. If there are ties, assign the average of the ranks that would otherwise have been assigned to the tied observations.

M-4.1.5.1.3. Calculate *R,* the sum of the ranks of the data from population 1, and then calculate:

$$W = R - \frac{n(n+1)}{2}.$$

M-4.1.5.1.4. Calculate:

$$w_p = \frac{mn}{2} + Z_p \sqrt{mn(n+m+1)/12}\ .$$

M-4.1.5.1.4.1. *Case 1.* $p = 1 - \alpha$

M-4.1.5.1.4.2. *Case 2*: $p = \alpha$

M-4.1.5.1.4.3. *Case 3*. Calculate both $w_{\alpha/2} (p = \alpha/2)$ and $w_{1-\alpha/2} (p = 1 - \alpha/2)$ (Lehmann, 1975).

M-4.1.5.1.5. Note that Z_p is the $p100^{th}$ percentile of the standard normal distribution (Table B-15 of Appendix B).

M-4.1.5.1.5.1. For Case 1, reject H_0 if $W > w_{1-\alpha}$.

M-4.1.5.1.5.2. For Case 2, reject H_0 if $W < w_{\alpha}$.

M-4.1.5.1.5.3. For Case 3, reject H_0 if $W > w_{1-\alpha/2}$ or $W < w_{\alpha/2}$.

M-4.1.5.1.6. The results of the test could be as follows.

M-4.1.5.1.6.1. H_0 was rejected and it seems values from population 1 tend to be greater than (Case 1), smaller than (Case 2), or different from (Case 3) values from population 2.

M-4.1.5.1.6.2. H_0 was not rejected, and it seems that values from population 1 tend to be smaller than or equal to (Case 1), greater than or equal to (Case 2), or not different from (Case 3) values from population 2.

M-4.1.5.2. *Example of the Large Sample Approximation to the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples.*

M-4.1.5.2.1. Consider the case where lead (Pb) surface soil concentrations are compared between Site A and background at a significance level of 0.05 using the test (Case 1 in Paragraph M-4.1.5.1) (Table M-3).

M-4.1.5.2.1.1. H_0: Site A Pb concentrations tend to be less than or equal to background Pb concentrations.

M-4.1.5.2.1.2. H_A: Site A Pb concentrations tend to be larger than background lead concentrations.

M-4.1.5.2.2. Suppose the surface soil Pb concentrations for Site A (X) are: 8.24, 6.57, 4.48, 4.34, 16.00, 3.83, 4.11, 3.48, 3.66, 5.01, 93.80, 3.70, 129.00, 4.92, 91.80, 3.86, 4.21, 4.32, 10.00, and 9.38 mg/kg.

M-4.1.5.2.3. Suppose the background surface soil Pb concentrations (Y) are: 3.05, 3.81, 3.68, 3.72, 4.20, 3.68, 5.97, 4.12, 6.42, 6.20, 4.13, 8.88, 3.01, 15.5, 5.34, 3.74, 20.6, 10.70, 3.86, 10.80, and 4.40 mg/kg.

Table M-3.
Example M-4.1.5.2 Pb Concentrations

Location	Result	Rank	Location	Result	Rank
Background	3.01	1	site	4.48	22
Background	3.05	2	site	4.92	23
Site	3.48	3	site	5.01	24
Site	3.66	4	background	5.34	25
Background	3.68	5.5	background	5.97	26
Background	3.68	5.5	background	6.2	27
Site	3.7	7	background	6.42	28
Background	3.72	8	site	6.57	29
Background	3.74	9	site	8.24	30
Background	3.81	10	background	8.88	31
Site	3.83	11	site	9.38	32
Site	3.86	12.5	site	10	33
Background	3.86	12.5	background	10.7	34
Site	4.11	14	background	10.8	35
Background	4.12	15	background	15.5	36
Background	4.13	16	site	16	37
Background	4.2	17	background	20.6	38
Site	4.21	18	site	91.8	39
Site	4.32	19	site	93.8	40
Site	4.34	20	site	129	41
Background	4.4	21	—	—	—

M-4.1.5.2.4. Note that tied values occur at for concentrations 3.68 and 3.86. These ties are assigned the average of the ranks they would otherwise have been assigned. The rank of 3.68 is 5.5, which is the average of ranks 5 and 6, and the rank of 3.86 is 12.5, which is the average of ranks 12 and 13.

M-4.1.5.2.5. Population 1 is the lead surface site data (n = 20), and population 2 is the background lead data (m = 21). Calculate W as:

$$W = R - \frac{n(n+1)}{2} = 458.5 - \frac{20(20+1)}{2} = 248.5 .$$

M-4.1.5.2.6. Calculate

$$w_p = \frac{mn}{2} + Z_p \sqrt{mn(n+m+1)/12} = \frac{21 \times 20}{2} + 1.645\sqrt{21 \times 20(20+21+1)/12} = 273.1$$

$$Z_p = Z_{1-\alpha} = Z_{0.95} = 1.645 .$$

M-4.1.5.2.6. Compare the calculated statistic W to the critical value $w_{1-\alpha}$, (248.5 < 273.1). Because $W \le w_{1-\alpha}$, do not reject the null hypothesis. Lead concentrations from Site A may be less than or equal to background lead concentrations. The power of the test needs to be determined (refer to Paragraph M-4.1.5.3).

M-4.1.5.3. *Directions for Calculating Sample Size to Achieve a Specified Power for the Wilcoxon Rank Sum Test.*

M-4.1.5.3.1. Noether (1987) discusses the determination of an adequate sample size based on a predefined level of power to apply the Wilcoxon rank sum test for the following hypothesis test. The *n* values of $X\,(x_1, x_2, \ldots, x_n)$ compared to *m* values of $Y\,(y_1, y_2, \ldots, y_m)$:

M-4.1.5.3.1.1. *Case 1.* $H_0 : P(X < Y) \geq 0.5$, $H_A : P(X < Y) < 0.5$.

M-4.1.5.3.1.2. *Case 2.* $H_0 : P(X < Y) \leq 0.5$, $H_A : P(X < Y) > 0.5$.

M-4.1.5.3.1.3. *Case 3*: $H_0 : P(X < Y) = 0.5$, $H_A : P(X < Y) \neq 0.5$.

M-4.1.5.3.2. The total number of samples collected, $N = n + m$, is compared with a conservative estimate of the number of samples N' required to achieve some desired power $1 - \beta$ Under the assumption that the test statistic (in this case, the large sample approximation for the Wilcoxon rank sum statistic in Paragraph M-4.1.5.1) is normally distributed, N' is determined as follows. For Cases 1 and 2:

$$N' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{12c\left(1-c\right)\left(p'' - \frac{1}{2}\right)^2}$$

and for Case 3:

$$N' = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{12c\left(1-c\right)\left(p'' - \frac{1}{2}\right)^2}$$

where Z_q = q quantile of the standard normal distribution (from Table B-15)

α = significance level of the test

$1-\beta$ = desired power for the test

$c = \frac{n}{N}$

p'' = $P(X < Y)$.

M-4.1.5.3.4. Setting *c* equal to 0.5 will be best unless there are reasons to sample more heavily from one of the populations. The value of p'' can be taken from past information, a pilot sample, or chosen to represent a meaningful shift in the data (Noether, 1987). The normality of the test statistic under the null hypothesis is generally valid if either *n* or *m* exceeds 20 and the smaller of the two is at least 4. If the suggested sample size does not meet these requirements, consult a statistician.

M-4.1.5.4. *Example of Calculating Sample Size to Achieve a Specified Power for the Wilcoxon Rank Sum Test.* Suppose Pb surface soil concentrations at a site are to be compared to background concentrations using a 95% level of confidence ($\alpha = 0.05$) using the following hypothesis test (Case 1).

M-4.1.5.4.1. H_0: Site A Pb concentrations tend to be less than or equal to background concentrations.

M-4.1.5.4.2. H_A: Site A Pb concentrations tend to be higher than background concentrations.

M-4.1.5.4.3. We wish to ensure that the sample size is large enough to find a meaningful elevation of lead concentrations with 80% probability ($\beta = 0.20$). Suppose historical information indicates that the probability of site lead concentration being less than background lead concentration is about 1/3. We decide to use this as our estimate of p''. We wish to take an equal number of samples from the site and background, so that *c* = 0.5. The required sample size to meet the power requirement is:

$$N' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{12c\left(1-c\right)\left(p'' - \frac{1}{2}\right)^2} = \frac{\left(1.645 + 0.842\right)^2}{12\left(0.5\right)\left(1-0.5\right)\left(0.333-0.5\right)^2} = 74.2 \ .$$

M-4.1.5.4.4. As we wish to collect and equal number of samples from the site and background, the calculated required total sample size is rounded up to the next largest even whole number, 76 (an even number is required because it is being assumed that the required sample size is equal to the sum of an equal number of site and background samples). If it is assumed that 38 site plus 38 background samples are required to achieve adequate power for the test performed in Paragraph M-4.1.5.2, it follows that, though the null hypothesis was not rejected, the result is not conclusive (as only 20 site and 21 background samples were collected).

M-4.1.6. *Matched Pairs Wilcoxon Signed Ranks Test.* As discussed in Paragraph M-2.3, matching subjects can lead to efficient comparisons between two populations. However, the observed differences between treatments will not always appear to come from a normal distribution. In that case, the one-sample Wilcoxon signed ranks test that was discussed in Appendix L can be used to test whether the mean or median difference differs significantly from zero. Directions for applying the Wilcoxon signed ranks test to a matched pairs design are presented in Paragraph M-4.1.6.1 and an example is presented in Paragraph M-4.1.6.2. See the discussion in Appendix L for more details on applying the Wilcoxon signed ranks test.

M-4.1.6.1. *Directions for the Wilcoxon Signed Ranks Test for Matched Pairs*. The following describes the steps for applying the Wilcoxon signed ranks test for a matched pairs design when the sample size, *n*, is less than 20 for: Case 1: $H_0 : \mu_A \geq \mu_B$, $H_A : \mu_A < \mu_B$; and Case 2: $H_0 : \mu_A \leq \mu_B$, $H_A : \mu_A > \mu_B$, which is given in braces { }.

M-4.1.6.1.1. Subtract each before concentration (B_i) from the after concentration (A_i) to get the difference:

$$d_i = A_i - B_i \ .$$

If any of the differences are zero, delete them and correspondingly reduce the sample size (*n*).

M-4.1.6.1.2. Assign ranks from 1 to *n* based on ordering the absolute deviations $|d_i|$ (i.e., magnitude of differences ignoring the sign) from smallest to largest. The rank 1 is assigned to the smallest value, the rank 2 to the second smallest value, and so forth. If there are ties, assign the average of the ranks that would otherwise have been assigned to the tied observations.

M-4.1.6.1.3. Assign the sign for each observation to create the signed rank. The sign is positive if the deviation d_i is positive and the sign is negative if the deviation d_i is negative. Calculate *R*, the sum of the ranks with a positive sign.

M-4.1.6.1.4. Use Table B-24 of Appendix B to find the critical value $w_{\alpha,n}$.

M-4.1.6.1.5. Compare the calculated test statistic, *R*, to the critical value:

M-4.1.6.1.5.1. If $R \leq \{n(n+1)/2\} - w_{\alpha,n} \ \{R \geq w_{\alpha,n}\}$, H_0 may be rejected.

M-4.1.6.1.5.2. If $R > \{n(n+1)/2\} - w_{\alpha,n} \ \{R < w_{\alpha,n}\}$, there is not enough evidence to reject H_0; verify the false acceptance error rate.

M-4.1.6.1.6. If H_0 was not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates using a software package like DEFT (EPA QA/G-4D).

M-4.1.6.1.7. The results of the test may be:

M-4.1.6.1.7.1. H_0 is rejected; $\mu_A < \mu_B \ \{\mu_A > \mu_B\}$.

M-4.1.6.1.7.2. H_0 is not rejected and the false acceptance error rate is satisfied; $\mu_A \geq \mu_B \ \{\mu_A \leq \mu_B\}$.

M-4.1.6.1.7.3. H_0 is not rejected and the false acceptance error rate was not satisfied; $\mu_A \geq \mu_B \{\mu_A \leq \mu_B\}$, but this conclusion is uncertain because the sample size was too small.

M-4.1.6.2. *Example of the Matched Pairs Wilcoxon Signed Ranks Test for the Difference Between Means Before and After Treatment*. Consider the case where the results of a groundwater remediation procedure are compared before and after treatment to see if the remediation has lowered the concentration of the contaminant. Test the hypothesis that the treatment had no lowering effect at the 95% level of confidence:

$$H_0 : \mu_A \geq \mu_B,\ H_A : \mu_A < \mu_B\ .$$

M-4.1.6.2.1. The data consist of measured TCE concentrations (mg/L) at monitoring wells before and after treatment (Table M-4). Negative values of the difference support the alternative hypothesis.

Table M-4.
Measured TCE Concentrations (mg/L) at Monitoring Wells Before and After Treatment for Example M-4.1.6.2

Sample	Baseline (01/2000)	Post–Test (12/2000)	Difference	Signed Rank
Well 1	20.9	0.917	–20.0	–6
Well 2	9.17	8.77	–0.400	–1
Well 3	5.96	4.37	–1.59	–2
Well 4	41.5	4.34	–37.2	–9
Well 5	34.3	10.7	–23.6	–7
Well 6	19.7	1.48	–18.2	–5
Well 7	38.9	0.272	–38.6	–10
Well 8	8.18	0.520	–7.66	–4
Well 9	9.13	3.06	–6.07	–3
Well 10	28.5	1.90	–26.6	–8

M-4.1.6.2.2. The differences are roughly symmetrical so the Wilcoxon signed ranks test can be applied.

M-4.1.6.2.3. Because the sign ranks are all negative, $R = 0$.

M-4.1.6.2.4. Using Table B-24 of Appendix B, we find the critical value $w_{0.05,10} = 11$.

M-4.1.6.2.5. Recall that negative values of the difference support the alternative hypothesis. Therefore we reject H_0 if R is smaller than the critical value. Comparing the calculated test statistic and the critical value, $R = 0 \leq \{n(n+1)/2\} - w_{\alpha,n} = \{10(11)/2\} - 11 = 44$, so H_0 is rejected. The treatment appears to have lowered TCE concentration in groundwater.

M-4.1.6.2.6. If the differences do not meet the symmetry assumption of the Wilcoxon signed ranks test, the one-sample sign test could be used for the analysis. However, a specific example will not be presented here.

M-4.2. *The Quantile Test.* The quantile test is used to compare two populations using m independent random samples (x_1, x_2,..., x_m) from the first population and n independent random samples (y_1, y_2,..., y_n) from the second population. The quantile test is useful in detecting instances where only parts of the data are different rather than a complete shift in the data. It looks at a certain number of the largest data values to determine if too many data values from one population are present to be accounted for by pure chance. When the quantile test and the Wilcoxon rank sum test (discussed above) are applied together, the combined tests are the most powerful at detecting true differences between two populations.

M-4.2.1. *Introduction*. The quantile test assumes a set of random samples from population 1 and an independent set of random samples from population 2. The quantile test is not robust to outliers, and assumes either a systematic (e.g., a triangular grid) or simple random sampling design. The quantile test may not be used for stratified designs. In addition, exact false rejection error rates are not available, only approximate rates. The quantile test is difficult to do by hand so directions are not included in this guidance, but the DataQUEST software (EPA QA/G-9D) can be used. Directions for a modified quantile test that can be done by hand are contained below in Paragraph M-4.2.2, followed by an example in Paragraph M-4.2.3.

M-4.2.2. *Directions for a Modified Quantile Test Done by Hand.* Let there be m measurements from population 1 (the reference area or background group) and n measurements from population 2 (the test area). The modified quantile test can be used to detect differences in shape and location of the two distributions. For this test, the significance level, α , can either be approximately 0.10 or approximately 0.05.

M-4.2.2.1 H_0 : population 1 = population 2.

M-4.2.2.2. H_A : population 2 > population 1.

M-4.2.2.3. Combine the two samples and order them from smallest to largest, keeping track of which sample a value came from.

M-4.2.2.4. Using Table B-25 of Appendix B, determine the critical number (C) for a sample size n from the reference area and sample size m from the test area using the significance level α . If the C^{th} largest measurement of the combined population is the same as others, increase C to include all of these tied values.

M-4.2.2.4.1. If the largest C measurements from the combined samples are all from population 2 (the test area), then reject the null hypothesis and conclude that there are differences between the two populations.

M-4.2.2.4.2. Otherwise, the null hypothesis is not rejected and it appears that there is no difference between the two populations.

M-4.2.3. *Example of a Modified Quantile Test Done by Hand.* Consider the case where nickel surface soil concentrations are compared between Site A and background using the test (Table M-5).

M-4.2.3.1. H_0: population 1 = population 2.

M-4.2.3.2. H_A: population 1 > population 2.

M-4.2.3.3. Suppose data for nickel surface site data (population 1) are the m = 6 values: 2.67, 3.61, 5.47, 7.15, 8.34, and 7.96 mg/kg.

M-4.2.3.4. Suppose data for nickel surface background data (population 2) are the n = 10 values: 5.14, 7.46, 5.99, 3.36, 3.19, 2.87, 5.95, 1.72, 4.77, and 5.61 mg/kg.

Table M-5.
Nickel Surface Soil Concentrations for Example M-4.2.3

Location	Result	Rank
Background	1.72	1
Site	2.67	2
Background	2.87	3
Background	3.19	4
Background	3.36	5
Site	3.61	6
Background	4.77	7
Background	5.14	8
Site	5.47	9
Background	5.61	10
Background	5.95	11
Background	5.99	12
Site	7.15	13
Background	7.46	14
Site	7.96	15
Site	8.34	16

M-4.2.3.5. $C_{n,m,\alpha} = C_{10,6,0.05} = 5$; because the fifth largest value is 5.99, there is no need to increase C.

M-4.2.3.6. Only three of the largest five values are from population 1 (site concentrations), therefore the null hypotheses cannot be rejected. The result is that there is no difference between the site concentrations and the background concentrations of nickel.

M-5. Multiple Population Tests. This Paragraph describes procedures to evaluate data from more than two populations. One could accomplish the same objectives by applying the tests described above multiple times. However, doing so would underestimate the true false rejection decision error rate. In other words, if multiple individual tests are done, H_0 is rejected more frequently than desired. The tests described in this Paragraph control the overall false rejection decision error rate by making multiple comparisons simultaneously.

M-5.1. *One-Factor Analysis of Variance (ANOVA).* The one-factor ANOVA is a statistical procedure to determine whether differences in mean concentrations among two or more populations are statistically significant. When a single variable is being measured for multiple populations (e.g., the concentration of chromium at multiple sites), the one-factor ANOVA allows the comparison of multiple population means in one test. Because the ANOVA test compares all the means to one another simultaneously, large false positives rates associated with multiple separate pairwise mean comparisons are avoided. Multi-factor ANOVA tests would be used when comparing several variables from multiple populations (e.g., the concentration of arsenic and chromium at multiple sites), but these are more complex than one-factor ANOVA tests and are beyond the scope of this document.

M-5.1.1. *Introduction*. There are two types of ANOVAs: parametric and nonparametric. The parametric ANOVA assumes that the errors, called residuals, are normally distributed with equal variance. The one-way parametric ANOVA model is the following:

$$x_{i,j} = \mu_i + \varepsilon_{i,j} \ .$$

The $x_{i,j}$ denotes the j^{th} measured value of the i^{th} group, where the i^{th} group contains n_i values and $i = 1, 2, \ldots K$ (the number of groups or populations). The residuals $\varepsilon_{i,j}$ are assumed to be values of a random variable ε that possess a normal distribution with mean of zero and standard deviation of σ. The parameters μ_i are the populations means for the groups; each possessing a common standard deviation σ. The equation is a model in the sense that it is of the form:

Measured value = *Function one or more parameters* + *Residual* (*random error*).

(Also refer to the linear regression model in Appendix Q.) As the population means μ_i are unknown, they are estimated by the sample group means:

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{i,j}}{n_i} \text{ for } i = 1, 2, \ldots K.$$

M-5.1.1.1. Thus, the "true" residuals $\varepsilon_{i,j}$ are estimated by the "sample" residuals as follows:

$$e_{i,j} = x_{i,j} - \bar{x}_i \ .$$

The sample residuals for each group (e.g., the n_i residuals for group i) must each be tested for normality and must be normally distributed.

M-5.1.1.2. The ANOVA is especially useful in situations where sample sizes are small. To apply a parametric one-way ANOVA, at least two groups must be present in the data and at least two samples must be available for each group. Although the ANOVA assumes equal variances, the test is not sensitive to unequal variances as long as the violation is not severe.

M-5.1.1.3. Directions for the ANOVA are given in Paragraph M-5.1.2, followed by an example in Paragraph M-5.1.3.

M-5.1.2. *Directions for the ANOVA Test.* Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the K sample populations to be compared to one another. Let the values from each population be represented by $x_{i,j}$ where $i = 1, 2, \ldots, K$ for the K groups and $j = 1, 2, \ldots, n_i$ for the observations in the i^{th} group.

M-5.1.2.1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ (no difference among the population means).

M-5.1.2.2. H_A : at least one mean, μ_i is different from one or more of the other means.

M-5.1.2.3. Verify that the residuals are normally distributed with equal variances (see Appendix F and Appendix N, respectively).

M-5.1.2.4. Let $(1-\alpha)100\%$ represent the chosen significance level for the test, so α is the false rejection rate for the test. Set up the ANOVA table as follows:

Source of Variation	**Degrees of Freedom (*v*)**	**Sum of Squares**	**Mean Square**	***F*-Value**
Groups	$v_G = K-1$	SSG	$\text{MSG} = \text{SSG}/(K-1)$	$F = \dfrac{\text{MSG}}{\text{MSE}}$
Error	$v_E = \left(\sum\limits_{i=1}^{K} n_i\right) - K$	SSE	$\text{MSE} = \text{SSE} \Big/ \left(\sum\limits_{i=1}^{K} n_i - K\right)$	
Total	$v_T = \left(\sum\limits_{i=1}^{K} n_i\right) - 1$	SST	$\text{MST} = \text{SST} \Big/ \left(\sum\limits_{i=1}^{K} n_i\right) - 1$	

$$\text{SST} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j}^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j} \Big/ \sum_{i=1}^{k} n_i$$

$$\text{SSG} = \sum_{i=1}^{K} n_i\left(\bar{x}_i - \bar{x}\right)^2 = \sum_{i=1}^{k}\left[\left(\sum_{j=1}^{n_i} x_{i,j}\right)^2 \Big/ n_i\right] - \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j} \Big/ \sum_{i=1}^{k} n_i$$

$$\text{SSE} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}_i\right)^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(e_{i,j}\right)^2 = \text{SST} - \text{SSG}\,.$$

Note that

$$\nu_T = \nu_G + \nu_E$$

$$\text{SST} = \text{SSG} + \text{SSE}\,.$$

M-5.1.2.5. It may be convenient to calculate MSE using the formula:

$$\text{MSE} = \frac{\sum_{i=1}^{K}(n_i - 1)\, s_i^2}{\sum_{i=1}^{K} n_i - K}\ .$$

In this form, MSE is often referred to as the "pooled" variance for the *K* groups, where s_i^2 is the sample variance for the *i*th group:

$$\text{s}_i^2 = \frac{\sum_{j=1}^{n_i}(x_{i,j} - \bar{x}_i)^2}{n_i - 1}\ .$$

M-5.1.2.6. Use Table B-7 of Appendix B to determine the critical value, $F_{1-\alpha,v_G,v_E}$, where $F_{\gamma,m,n}$ is the $\gamma 100^{\text{th}}$ percentile of the *F* distribution with *m* degrees of freedom for the numerator and *n* degrees of freedom for the denominator. Compare *F* to $F_{1-\alpha,v_G,v_E}$. If $F > F_{1-\alpha,v_G,v_E}$, then reject H_0 (the means of the sample populations are not all equal). Otherwise, conclude that there is no difference among the sample population means. If H_0 is rejected, perform multiple comparison tests to determine which populations are significantly different.

M-5.1.2.7. Statistical software sometimes outputs the coefficient of determination for the ANOVA:

$$r_{ANOVA}^2 = \text{SSG/SST}\,.$$

The square root of this quantity is similar in function to the regression coefficient for an ordinary least squares regression line (refer to Appendix Q) in that it accounts for the variation in the measured values accounted for by the model (often referred to as the explained variation). A large value for r_{ANOVA}^2 (which ranges from 0 to 1) indicates that most of the variation is ascribable to differences between the group means. It can be shown that

$$F = \frac{r_{ANOVA}^2}{1 - r_{ANOVA}^2} \times \left(\frac{\nu_E}{\nu_G}\right)$$

$$r_{ANOVA}^2 = \frac{\nu_G\, F}{\left(\nu_E + \nu_G F\right)}\ .$$

Therefore, when the calculated value of the statistic *F* is small (i.e., when the null hypothesis is not rejected), r_{ANOVA}^2 will be near zero.

M-5.1.3. *Example of ANOVA*. Suppose manganese (Mn) groundwater concentrations are going to be compared among the seven different wells at Site A using the following test with 95% level of confidence.

M-5.1.3.1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ (no difference among the sample means).

M-5.1.3.2. H_A: at least one mean, μ_i is different from one or more of the other means.

M-5.1.3.3. Table M-6 presents the data. All Mn concentrations were detected, so no proxy concentrations are needed to evaluate the data.

M-5.1.3.4. The data were tested for equal variances using Bartlett's test for equal variances (see Paragraph N-3). The data were also tested for normality using the Shapiro-Wilk test. Because the data were not normal, the data were transformed so that the residuals would follow a normal distribution.

M-5.1.3.5. Summary statistics for each well are presented in Table M-7.

M-5.1.3.6. Let $(1-\alpha)100\%$ represent the chosen significance level for the test, where $\alpha = 0.05$. Note that in this example K = 7 and n_i = 8 for i = 1, 2, … 7. Set up the ANOVA table as follows:

Source of Variation	**Degrees of Freedom (v)**	**Sum of Squares**	**Mean Square**	***F* Value**
Groups	6	137.29	22.88	346.09
Error	49	3.24	0.066	
Total	55	**140.53**		

Table M-6.
Manganese (Mn) Groundwater Concentrations to be Compared Among the Wells at Site A

Well Location	Result	Log Result	Well Location	Result	Log Result
69-2-02	0.432	–0.839	69-2-06A	0.294	–1.224
69-2-02	0.44	–0.821	69-2-06A	0.301	–1.201
69-2-02	0.513	–0.667	69-2-06A	0.379	–0.970
69-2-02	0.704	–0.351	69-2-06A	0.352	–1.044
69-2-02	0.327	–1.118	69-2-06A	0.346	–1.061
69-2-02	0.316	–1.152	69-2-06B	0.13	–2.040
69-2-02	0.454	–0.790	69-2-06B	0.184	–1.693
69-2-02	0.401	–0.914	69-2-06B	0.209	–1.565
69-2-04	0.0504	–2.988	69-2-06B	0.2	–1.609
69-2-04	0.0502	–2.992	69-2-06B	0.0739	–2.605
69-2-04	0.054	–2.919	69-2-06B	0.0876	–2.435
69-2-04	0.0523	–2.951	69-2-06B	0.126	–2.071
69-2-04	0.0923	–2.383	69-2-06B	0.129	–2.048
69-2-04	0.0556	–2.890	69-2-07	0.0137	–4.290
69-2-04	0.0534	–2.930	69-2-07	0.019	–3.963
69-2-04	0.0517	–2.962	69-2-07	0.0163	–4.117
69-2-05	0.00684	–4.985	69-2-07	0.0195	–3.937
69-2-05	0.00639	–5.053	69-2-07	0.0112	–4.492
69-2-05	0.00631	–5.066	69-2-07	0.0112	–4.492
69-2-05	0.00813	–4.812	69-2-07	0.0102	–4.585
69-2-05	0.00747	–4.897	69-2-07	0.00946	–4.661
69-2-05	0.00679	–4.992	69-2-08	0.563	–0.574
69-2-05	0.00731	–4.919	69-2-08	0.512	–0.669
69-2-05	0.00444	–5.417	69-2-08	0.475	–0.744
69-2-06A	0.3	–1.204	69-2-08	0.546	–0.605
69-2-06A	0.286	–1.252	69-2-08	0.276	–1.287
69-2-06A	0.303	–1.194	69-2-08	0.383	–0.960
			69-2-08	0.33	–1.109
			69-2-08	0.27	–1.309

Table M-7.
Summary Statistics

Well	Sample Size	Mean of Log Result	Standard Deviation of Log Result
69-2-02	8	–0.832	0.2539
69-2-04	8	–2.877	0.2026
69-2-05	8	–5.018	0.1818
69-2-06A	8	–1.144	0.1031
69-2-06B	8	–2.008	0.3779
69-2-07	8	–4.317	0.2832
69-2-08	8	–0.907	0.3011

M-5.1.3.7. The power of an ANOVA F-test can be estimated prior to a study. Table B-28 in Appendix B lists the power for $K = 3$ to 10 groups and significance levels of $\alpha = 0.2$, 0.1, and 0.05, where each group contains an equal number of samples n. To use the tables, the "effect size," Δ, must be also estimated as:

$\Delta =$ (largest group mean – smallest group mean)/(MSE)$^{1/2}$.

M-5.1.3.8. The tables list various values of Δ. For a specified value of K, n, α, and Δ, the tables list the minimum power (probability) corresponding to the alternative hypothesis that all group means, other than the two extremes, are equal to the "grand mean," which is equal to the median of the largest and smallest group means. When comparing K groups of equal size n, the tables are useful for determining approximately how large a sample size for each group is required to achieve a particular level of confidence $1 - \alpha$ and power $1 - \beta$. For example, for $K = 3$ groups and $\alpha = 0.05$, to detect a size effect $\Delta = 1.0$ (i.e., a difference between the largest and smallest mean equal to MSE$^{1/2}$) with power of at least $1 - \beta = 0.80$, the required sample size for each group $n \approx 20$.

M-5.2. *Kruskal-Wallis Test.* The Kruskal-Wallis test is the nonparametric version of the ANOVA. It is a statistical procedure to determine whether differences in median concentrations among a number of groups or multiple populations are statistically significant. The Kruskal-Wallis allows the comparison of multiple population means in one test. If the test shows statistically significant differences among the groups, multiple comparison procedures can be used to identify which group or groups are different.

M-5.2.1. *Introduction.* In terms of hypothesis tests, the null hypothesis is that all group medians are equal and the alternative hypothesis is that at least one group is different from one or more other groups. To test this hypothesis, no assumptions are required about the shape of the distributions; each group may have a different distribution. The Kruskal-Wallis test is used to evaluate whether the distributions are identical. Directions for the Kruskal-Wallis test are given below in Paragraph M-5.2.2, followed by an example in Paragraph M-5.2.3.

M-5.2.2. *Directions for the Kruskal-Wallis Test.* Let $(1-\alpha)100\%$ represent the chosen significance level for the test.

M-5.2.2.1. Rank all $x_{i,j}$ observations from lowest to highest. Let $R_{i,j}$ denote the rank of the $x_{i,j}$ observation.

M-5.2.2.1.1. *Ties.* If two or more observations are numerically equal, then use an average rank for each observation. The average rank is calculated as the average of the ranks that the tied observations would have received had the observations been different.

M-5.2.2.1.2. *Censored Data.* If any values are not-detected, it is appropriate to consider the ranks for these values equal to zero. (It is irrelevant what number is assigned to the non-detected values as long as all such values are assigned the same number, and it is smaller than any detected value.)

M-5.2.2.2. Add the ranks of the observations in each group. Call the sum of the ranks for the i^{th} group R_i. Also calculate the average rank for each group, $\overline{R_i} = R_i / n_i$. If there are at least 50% detected results and no tied values, then compute the Kruskal-Wallis statistic:

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{K} n_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2$$

$$= \left[\frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} \right] - 3(N+1)$$

where

$$N = \sum_{i=1}^{K} n_i \ .$$

M-5.2.2.3. If there are at least 50% detected results and there are tied values present in the data, then compute the adjusted Kruskal-Wallis statistic:

$$H' = \frac{\left[\frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} \right] - 3(N+1)}{1 - \left(\sum_{k=1}^{g} W_k / (N^3 - N) \right)}$$

where *g* is the number of groups of distinct tied observations and $W_k = (w_k^3 - w_k)$, where w_k is the number of tied data in the tied group *k*. Note that the unique observations can be considered groups of size 1, with the corresponding $W_k = (1^3 - 1) = 0$. If all the group medians are equal, then $H = 0$. As the differences between the group medians increase, *H* will also increase; so the larger the value of *H*, the less probable H_0 is true.

M-5.2.2.4. Compare the calculated value *H* (or H') to the tabulated critical value for the chi-square distribution, $\chi^2_{1-\alpha,K-1}$, with $K - 1$ degrees of freedom and $(1-\alpha)100\%$ level of confidence (found in Table B-2 of Appendix B).

M-5.2.2.5. Reject H_0 if $H >$ $\chi^2_{1-\alpha,K-1}$. If H_0 is rejected use multiple comparison tests to determine which populations are significantly different.

M-5.2.3. *Example of the Kruskal-Wallis Test.* Suppose lead groundwater concentrations are going to be compared among seven wells using the Kruskal-Wallis test with 95% level of confidence.

M-5.2.3.1. $H_0 : \mu_1 = \mu_2 = \ldots = \mu_7$ (i.e., no difference among the well means).

M-5.2.3.2. H_A : at least one mean is different from one or more of the other means.

M-5.2.3.3. Table M-8 presents the data. All lead concentrations were detected so no proxy concentrations were needed to evaluate the data.

M-5.2.3.4. The sum of the ranks for each of the seven groups is:

$$R_1 = 272\,,\ R_2 = 168\,,\ R_3 = 62.5\,,\ R_4 = 420\,,\ R_5 = 304\,,\ R_6 = 73.5\,,\ R_7 = 296$$

M-5.2.3.5. Because there are at least 50% detected results and there are tied values present in the data, compute the adjusted Kruskal-Wallis statistic:

$$H' = \frac{\left[\dfrac{12}{N(N+1)}\sum_{i=1}^{K}\dfrac{R_i^2}{n_i}\right] - 3(N+1)}{1 - \left(\sum_{k=1}^{g} W_k \big/ (N^3 - N)\right)}$$

The table below summarizes the $g = 4$ tied groups:

Tied Rank	**Number of Tied Observations** w_k	$W_K = w_k^3 - w_k$
4	3	24
12.5	2	6
19.5	2	6
21.5	2	6

$$H' = \frac{\left[\dfrac{12}{56(56+1)} \times (\dfrac{272^2}{8} + \dfrac{168^2}{8} + \dfrac{62.5^2}{8} + \dfrac{420^2}{8} + \dfrac{304^2}{8} + \dfrac{73.5^2}{8} + \dfrac{296^2}{8})\right] - 3(56+1)}{1 - \left[(24+6+6+6)/(56^3 - 56)\right]} = 48.91$$

$$\chi^2_{1-\alpha,K-1} = \chi^2_{1-0.05,7-1} = \chi^2_{0.95,6} = 12.59\,.$$

M-5.2.3.6. Now compare the calculated value to the critical value, 48.91 > 12.59. As the calculated value exceeds the critical value, reject H_0.

M-5.2.3.7. Because there is a difference in the average lead concentration among the seven wells, a multiple comparison test should be done to determine which wells are significantly different. A multiple comparison test based on ranks is discussed in Conover (1980).

Table M-8.
Lead Concentrations for Example M-5.2.3

Well	Result	Rank	Well	Result	Rank
6	0.978	1	5	3.100	29
6	1.037	2	7	3.118	30
3	1.061	4	5	3.144	31
3	1.061	4	7	3.178	32
3	1.061	4	1	3.215	33
6	1.095	6	1	3.219	34
6	1.109	7	1	3.235	35
3	1.144	8	5	3.346	36
3	1.227	9	1	3.395	37
3	1.241	10	5	3.421	38
3	1.270	11	5	3.434	39
3	1.426	12.5	1	3.478	40
6	1.426	12.5	1	3.586	41
6	1.513	14	5	3.605	42
6	1.530	15	5	3.627	43
6	1.601	16	7	3.671	44
2	2.588	17	7	3.689	45
2	2.595	18	5	3.694	46
2	2.610	19.5	7	3.922	47
2	2.610	19.5	7	3.932	48
2	2.625	21.5	4	4.057	49
2	2.625	21.5	4	4.101	50
2	2.639	23	4	4.103	51
7	2.918	24	4	4.119	52
1	3.011	25	4	4.159	53
7	3.035	26	4	4.177	54
1	3.068	27	4	4.214	55
2	3.073	28	4	4.228	56

M-6. Multiple Comparison Tests. Multiple comparisons occur whenever more than one statistical test is performed with the same data. These comparisons can arise, for example, as a result of the need to test multiple down-gradient wells against a pool of up-gradient background data or to regularly test several indicator parameters for contamination. The multiple comparison tests described in this section may not be needed if a significant difference is not obtained from the ANOVA *F*-test.

M-6.1. *Introduction.* Comparisons are usually written in terms of linear combinations of the population means, and are often referred to as "contrasts." For example, we may want to know if the mean for population 1, μ_1, differs from the mean for population 2, μ_2. This contrast can be written as $\mu_1 - \mu_2$. In general, a contrast is a linear combination

$$\theta = \sum a_i \mu_i$$

where

$$\sum a_i = 0 \,.$$

Beyond comparing pairs of means, a contrast to compare the mean of population 1 to the means of populations 2 and 3 can be written as $2\mu_1 - \mu_2 - \mu_3$.

M-6.1.1. The Type I error rate for multiple comparison tests can be viewed in two ways. Comparison-wise significance considers the probability of rejecting the hypothesis that only a single contrast equals zero ($H_0 : \theta_1 = 0$) when it is actually true. Experiment-wise significance considers the probability of rejecting any of a set of *m* hypotheses on contrasts ($H_0 : \theta_j = 0, j = 1,..., m$) when all of them are true.

Table M-9.
Summary of Multiple Comparison Tests

Test	Purpose
Dunnett's	Comparing treatment means to a control mean
Fisher's LSD	Comparing all pairs of means
Duncan's multiple range	Comparing all pairs of means
Tukey's	Comparing all pairs of means
Bonferroni's	Comparing any set of contrasts
Scheffé's	Comparing any set of contrasts

M-6.1.2. Table M-9 summarizes the multiple comparison tests that will be covered in this document. The Fisher's Least Significant Difference (LSD) test and Bonferroni's test are multiple comparison tests that are based on the Student's *t* distribution, whereas the Tukey's test and Duncan's multiple range test are based on the Studentized range statistic. Scheffé's multiple comparison test is used to achieve an experiment-wise false positive rate for all possible contrasts or linear combinations of means at the same time. All the multiple comparison tests presented rely on the assumption of normality. Assumptions of normality should have been verified during the ANOVA process, which is typically performed prior to these multiple comparison tests. More information on multiple comparison tests can be found in Mason et al. (1989) and Montgomery (1997).

M-6.1.3. There is no clear answer to the question of which multiple comparison technique should be used. For comparing all pairs of treatment means, Fisher's LSD is the least conservative (most powerful) test for identifying differences between means (i.e., it rejects H_0 most often) followed by Duncan's Multiple Range, Tukey, and finally Sheffé. The relative conservatism of the Bonferroni Test will depend on the number of groups. Montgomery (1997) recommends Fisher's LSD or Duncan's multiple range test for comparing all treatment means as long as the ANOVA *F*-test is significant, based on Monte Carlo studies conducted by Carmer and Swanson (1973). Mason et al. (1989) recommend Fisher's LSD to control the comparison-wise error rate and Tukey's test to control the experiment-wise error rate for comparing all treatment means. When many comparisons need to be made, multiple range tests such as Duncan's multiple range test and Tukey's test should be used as a compromise between the desired experiment-wise error rate and an unacceptable comparison-wise error rate (Mason et al., 1989). Obviously, if one's purpose is to compare treatment means to a control or if contrasts other than pairwise comparisons of treatments are of interest, Dunnett's, Bonferroni's, or Scheffé's test may be preferred.

M-6.2. *Fisher's Least-Significant Difference Test.* Fisher's LSD test is an extension of the *t*-test for comparing all pairs of treatment means. Each pairwise comparison will have a Type I error rate (probability of declaring the pair of means different when they are not) of α. Therefore, the *experiment-wise* error rate (the probability of declaring any pair of means different when they are not) will be *larger* than α. The disadvantage to the Fisher's LSD test is that its experiment-wise error rate is not satisfactory for testing all possible pairs of group means when there are a moderate to large number of groups to be compared (Mason et al., 1989). Directions for Fisher's LSD test (from Mason et al., 1989) are given in Paragraph M-6.2.1 and an example is presented in Paragraph M-6.2.2.

M-6.2.1. *Directions for Fisher's LSD Test.* Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let the values from each population be represented by $x_{i,j}$ where *i* = 1, 2,…, *K* for the *K* groups and $j = 1, 2, ..., n_i$ for the observations at the i^{th} group. Let $(1-\alpha)100\%$ represent the chosen confidence level for the test.

M-6.2.1.1. Verify the assumptions of normality.

M-6.2.1.2. The means of two groups, $\bar{x}_i$ and $\bar{x}_k$, in an ANOVA are declared to be significantly different if:

$$\bar{x}_i - \bar{x}_k > \text{LSD}$$

where

$$\text{LSD} = t_{1-\alpha/2, v_E}\left[\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_k}\right)\right]^{1/2}$$

and

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} \ .$$

t_{γ, v_E} is the $\gamma 100^{\text{th}}$ percentile for the Student's *t* distribution with v_E degrees of freedom (see Table B-23 in Appendix B). MSE and v_E come from the ANOVA procedures previously defined. Note that for *K* groups, $K(K-1)/2$ differences $\bar{x}_i - \bar{x}_k$ need to be calculated.

M-6.2.2. *Example of Fisher's LSD Test*. Mean manganese groundwater concentrations in seven wells were compared to one another using the ANOVA. The null hypothesis was rejected. The LSD test is subsequently applied below using the 95% level of confidence.

M-6.2.2.1. The table in Paragraph M-5.1.3 presents the data. All manganese concentrations were detected so no proxy concentrations are needed to evaluate the data.

M-6.2.2.2. Assumptions of normality were verified for the log result during the ANOVA process.

$$\begin{aligned}\text{LSD} &= t_{1-\alpha/2,v_E}\left[\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_k}\right)\right]^{1/2}\\ &= t_{0.975,49}\left[0.066\times\left(\frac{1}{8}+\frac{1}{8}\right)\right]^{1/2}\\ &= 2.01\times 0.128\\ &= 0.2584\ .\end{aligned}$$

M-6.2.2.3. Means that differ by more then 0.2584 would be considered statistically different with 95% confidence. Alternatively, confidence intervals for the difference in means can be calculated as $(\bar{x}_i - \bar{x}_k) \pm \text{LSD}$. If zero is not in the confidence interval, the two population means are declared significantly different at the α significance level. Table M-10 summarizes the results. Comparisons significant at the 0.05 level are indicated by ***.

M-6.2.2.4. Another way to visualize the conclusions is to list the means in order and identify those that are not significantly different. In Table M-11, means designated with the same "group" letter (A, B, C, etc.) are not significantly different at $\alpha = 0.05$.

M-6.2.2.5. As Wells 69-2-02 and 69-2-08 are in LSD grouping A, the means for these wells are not statistically different. The preceding table indicates that the difference between the two means is 0.0758, which is less than LSD = 0.2584.

M-6.3. *Bonferroni's Test*. The Bonferroni's test is designed to control the *experiment-wise* error rate (the probability of declaring any two means different when they are not). The test uses the overall significance level divided by the number of selected comparisons as the comparison-wise significance level. Mason et al. (1989) warn that Bonferroni's test should not be used when the number of comparisons becomes very large, because this results in an extremely conservative comparison-wise test. However, they do state that the experiment-wise error rate can be better controlled using Bonferroni's test rather than the Fisher's LSD test (where comparison-wise error is controlled). Also, note that Bonferroni's test can be used to test any contrast of interest (Mason et al., 1989). Directions for Bonferroni's Test (from Mason et al., 1989) are presented in Paragraph M-6.3.1 and an example is presented in Paragraph M-6.3.2.

Table M-10.
Results for Example M-6.2.2

Well Comparison	Difference Between Means $\bar{x}_i - \bar{x}_k$	95% Confidence Interval
02 – 08	0.0758	(–0.1825, 0.3342)
02 – 06A	0.3123	(0.0539, 0.5706)***
02 – 06B	1.1769	(0.9186, 1.4353)***
02 – 04	2.0452	(1.7868, 2.3036)***
02 – 07	3.4857	(3.2273, 3.7440)***
02 – 05	4.1861	(3.9277, 4.4444)***
08 – 06A	0.2365	(–0.0219, 0.4948)
08 – 06B	1.1011	(0.8427, 1.3595)***
08 – 04	1.9694	(1.7110, 2.2277) ***
08 – 07	3.4098	(3.1515, 3.6682)***
08 – 05	4.1103	(3.8519, 4.3686)***
06A – 06B	0.8646	(0.6063, 1.1230)***
06A – 04	1.7329	(1.4746, 1.9913)***
06A – 07	3.1734	(2.9150, 3.4317)***
06A – 05	3.8738	(3.6154, 4.1322)***
06B – 04	0.8683	(0.6099, 1.1266)***
06B – 07	2.3088	(2.0504, 2.5671)***
06B – 05	3.0092	(2.7508, 3.2675)***
04 – 07	1.4405	(1.1821, 1.6988)***
04 – 05	2.1409	(1.8825, 2.3992)***
07 – 05	0.7004	(0.4420, 0.9588)***

Table M-11.
List of the Means in Order for Example M-6.2.2

Well	Mean	*n*	LSD Groupings
69-2-02	–0.8315	8	A
69-2-08	–0.9073	8	B A
69-2-06A	–1.1438	8	B
69-2-06B	–2.0084	8	C
69-2-04	–2.8767	8	D
69-2-07	–4.3172	8	E
69-2-05	–5.0176	8	F

M-6.3.1. *Directions for Bonferroni's Test.* Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let the values from each population be represented by $x_{i,j}$ where *i* = 1, 2,…, *K* for the *K* groups and $j = 1, 2, ..., n_i$ for the observations in the *i*th group. Let $(1-\alpha)100\%$ represent the selected confidence level for the test.

M-6.3.1.1. Verify the assumptions of normality.

M-6.3.1.2. Let

$$\theta = \sum a_i \mu_i$$

represent one of *m* linear combinations of the means, μ_k, for which the hypothesis $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$ is being tested.

M-6.3.1.3. Reject H_0 if

$$|\theta| = \left| \sum a_i \bar{x}_i \right|$$

exceeds

$$\text{BSD} = t_{1-\alpha/2m, v_E} \left[\text{MSE} \sum a_i^2 / n_i \right]^{1/2}$$

where n_i is the number of observations used to calculate

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j} \quad .$$

t_{γ, v_E} is the $\gamma 100^{\text{th}}$ percentile for the Student's *t* distribution with v_E degrees of freedom (see Table B-23 in Appendix B), and *m* is the number of comparisons. For *K* means (groups), there are

$$m = \frac{K(K-1)}{2}$$

possible comparisons. MSE and v_E are determined from the ANOVA procedures previously defined.

M-6.3.2. *Example of Bonferroni's Test.* Suppose manganese groundwater concentrations are going to be compared among the seven different wells at Site A using Bonferroni's test with 95% level of confidence.

M-6.3.2.1. Table M-6 presents the data. All manganese concentrations were detected, so no proxy concentrations are needed to evaluate the data.

M-6.3.2.2. The assumptions of normality were verified during the ANOVA process. The contrasts to make pairwise comparisons of all 7 means are the 21 differences (where $a_i = \pm 1$):

$\mu_{69-2-02} - \mu_{69-2-04}$

$\mu_{69-2-02} - \mu_{69-2-05}$

$\mu_{69-2-02} - \mu_{69-2-06A}$

$\mu_{69-2-02} - \mu_{69-2-06B}$

$\mu_{69-2-02} - \mu_{69-2-07}$

$\mu_{69-2-02} - \mu_{69-2-08}$

$\mu_{69-2-04} - \mu_{69-2-05}$

$\mu_{69-2-04} - \mu_{69-2-06A}$

$\mu_{69-2-04} - \mu_{69-2-06B}$

$\mu_{69-2-04} - \mu_{69-2-07}$

$\mu_{69-2-04} - \mu_{69-2-08}$

$\mu_{69-2-05} - \mu_{69-2-06A}$

$\mu_{69-2-05} - \mu_{69-2-06B}$

$\mu_{69-2-05} - \mu_{69-2-07}$

$\mu_{69-2-05} - \mu_{69-2-08}$

$\mu_{69-2-06A} - \mu_{69-2-06B}$

$\mu_{69-2-06A} - \mu_{69-2-07}$

$\mu_{69-2-06A} - \mu_{69-2-08}$

$\mu_{69-2-06B} - \mu_{69-2-07}$

$\mu_{69-2-06B} - \mu_{69-2-08}$

$\mu_{69-2-07} - \mu_{69-2-08}$

$$\text{BSD} = t_{1-\alpha/2m,v_E}\left[\text{MSE}\sum a_i^2 / n_i\right]^{1/2} = t_{0.999,49}\left[0.066\left(\frac{1}{8}+\frac{1}{8}\right)\right]^{1/2} = 3.20 \times 0.128 = 0.412 \ .$$

Means that differ by more than 0.412 would be considered statistically different with 95% confidence. Alternatively, confidence intervals for the difference in means can be calculated as $\bar{x}_i - \bar{x}_k \pm \text{BSD}$. If zero is not covered by the confidence interval, the two population means are declared significantly different at the α significance level.

M-6.3.2.3. In Table M-12, means with the same letter are not significantly different at $\alpha = 0.05$. For example, the mean for 69-2-02 does not differ from the mean for 69-2-08 by more than 0.412, so we accept

$$H_0 : \mu_{69-2-02} - \mu_{69-2-08} = 0 .$$

Table M-12.
Means with the Same Letter are not Significantly Different at $\alpha = 0.05$ in Example M-6.3.2

Well	Mean	*n*	Bonferroni Grouping
69-2-02	–0.8315	8	A
69-2-08	–0.9073	8	A
69-2-06A	–1.1438	8	A
69-2-06B	–2.0084	8	B
69-2-04	–2.8767	8	C
69-2-07	–4.3172	8	D
69-2-05	–5.0176	8	E

On the other hand, we can reject

$$H_0 : \mu_{69-2-02} - \mu_{69-2-05} = 0$$

because the two observed means differ by more than 0.412. Notice that the more conservative Bonferroni test does not reject

$$H_0 : \mu_{69-2-02} - \mu_{69-2-06A} = 0$$

with 95% confidence while Fisher's LSD test did.

M-6.4. *Tukey's Test*. Tukey's test is designed to control the experiment-wise chance of a Type I error (declaring any two population means different when they are not) at α assuming equal sample sizes (Mason et al., 1989). Because of this, it is less powerful than Fisher's LSD or Duncan's multiple range test (Montgomery, 1997). Directions and an example for Tukey's Test (from Mason et al., 1989) are given in Paragraphs M-6.4.1 and M-6.4.2, respectively.

M-6.4.1. *Directions for Tukey's Test*. Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let the values from each population be represented by $x_{i,j}$, where *i* = 1, 2,…, *K* for the *K* groups and $j = 1, 2, \ldots, n_i$ for the observations at the i^{th} group. Let $(1-\alpha)100\%$ be the confidence level.

M-6.4.1.1. Verify the assumptions of normality. Two averages, $\bar{x}_i$ and $\bar{x}_r$, are based on n_i and n_r samples, respectively, where

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} \ .$$

Two means are significantly different if $|\,\bar{x}_i - \bar{x}_r\,| > \text{TSD}$ where:

$$TSD = q_{\alpha,k,v_E}\left[MSE\left(\frac{1/n_i + 1/n_r}{2}\right)\right]^{1/2} \quad .$$

M-6.4.1.2. The quantity q_{α,k,v_E} is the Studentized range statistic in Table B-22 of Appendix B, where *k* is the number of means being compared (typically equal to the number of groups *K*); MSE and v_E are from the ANOVA procedure previously defined, and α represents the desired significance level.

M-6.4.2. *Example of Tukey's Test*. Manganese groundwater concentrations are compared among the seven different wells at Site A using Tukey's Test with 95% level of confidence.

M-6.4.2.1. Table M-6 presents the data. All manganese concentrations were detected so no proxy concentrations are needed to evaluate the data. Assumptions of normality were verified during the ANOVA process.

$$\begin{aligned} TSD &= q_{\alpha,k,v_E}\left[MSE\left(\frac{1/n_i + 1/n_r}{2}\right)\right]^{1/2} \\ &= q_{0.05,7,49}\left[0.066\left(\frac{1/8+1/8}{2}\right)\right]^{1/2} = 4.35 \times 0.0908 = 0.3952 \ . \end{aligned}$$

M-6.4.2.2. Means that differ by more then 0.3952 would be considered statistically different with 95% confidence. Alternatively, confidence intervals for the difference in means can be calculated for the difference of any two means as $\bar{x}_i - \bar{x}_r \pm \mathrm{TSD}$. If zero is not in the confidence interval, the two population means are significantly different at the α significance level. Table M-13 summarizes the results. Comparisons significant at $\alpha = 0.05$ are indicated by ***.

M-6.4.2.3. In Table M-14, means with the same letter are not significantly different at $\alpha = 0.05$.

Table M-13.
Results from Example M-6.4.2

Well Comparison	Difference Between Means	Simultaneous 95% Confidence Intervals
69-2-02–69-2-08	0.0758	(–0.3194, 0.4710)
69-2-02–69-2-06A	0.3123	(–0.0829, 0.7075
69-2-02–69-2-06B	1.1769	(0.7817, 1.5721)***
69-2-02–69-2-04	2.0452	(1.6500, 2.4404) ***
69-2-02–69-2-07	3.4857	(3.0905, 3.8809)***
69-2-02–69-2-05	4.1861	(3.7909, 4.5813)***
69-2-08–69-2-06A	0.2365	(–0.1587, 0.6317)
69-2-08–69-2-06B	1.1011	(0.7059, 1.4963)***
69-2-08–69-2-04	1.9694	(1.5742, 2.3646)***
69-2-08–69-2-07	3.4098	(3.0146, 3.8051)***
69-2-08–69-2-05	4.1103	(3.7150, 4.5055)***
69-2-06A–69-2-06B	0.8646	(0.4694, 1.2598)***
69-2-06A–69-2-04	1.7329	(1.3377, 2.1281)***
69-2-06A–69-2-07	3.1734	(2.7782, 3.5686)***
69-2-06A–69-2-05	3.8738	(3.4786, 4.2690)***
69-2-6B–69-2-04	0.8683	(0.4731, 1.2635)***
69-2-06B–69-2-07	2.3088	(1.9135, 2.7040)***
69-2-06B–69-2-05	3.0092	(2.6139, 3.4044)***
69-2-04–69-2-07	1.4405	(1.0453, 1.8357)***
69-2-04–69-2-05	2.1409	(1.7457, 2.5361)***
69-2-07–69-2-05	0.7004	(0.3052, 1.0956)***

Table M-14.
Means with the Same Letter are not Significantly Different at α = 0.05

Tukey Grouping	Mean	*N*	Well
A	–0.8315	8	69-2-02
A	–0.9073	8	69-2-08
A	–1.1438	8	69-2-06A
B	–2.0084	8	69-2-06B
C	–2.8767	8	69-2-04
D	–4.3172	8	69-2-07
E	–5.0176	8	69-2-05

M-6.5. *Duncan's Multiple Range Test.* Duncan's multiple range test is used to test for differences in all pairs of means. Considering the ordered list of means, this procedure provides an experiment-wise error rate of

$$1-(1-\alpha)^{p-1}$$

when the pair of means are p steps apart in the ordered list (Montgomery, 1997). Thus, the experiment-wise probability of a Type I error depends on how far apart in the ordered list the two means lie (Mason et al., 1989). Duncan's multiple range test is similar to Tukey's test except that it has greater power to detect differences but does not control the experiment-wise error rate as well. Directions for Duncan's multiple range test (from Mason et al., 1989, and Montgomery, 1997) are presented in Paragraph M-6.5.1 followed by an example in Paragraph M-6.5.2.

M-6.5.1. *Directions for Duncan's Multiple Range Test.* Let K represent the total number of populations to be compared. Let n represent the sample sizes of each of the K sample populations. Let the values from each population be represented by $x_{i,j}$ where i = 1, 2,…, K and j = 1, 2,…, n for the observations in the i^{th} group (population).

M-6.5.1.1. Verify the assumptions of normality. The means

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n_i} x_{i,j}$$

are sorted from smallest to largest. The two extreme means are compared first. The largest and smallest of $p = K$ averages, $\bar{x}_a$ and $\bar{x}_b$ (each based on a sample size of n), are significantly different if $\left|\bar{x}_a - \bar{x}_b\right| > R_p$ where

$$R_p = q_{\alpha,p,v_E}\left(\frac{\text{MSE}}{n}\right)^{1/2}.$$

M-6.5.1.2. The quantity

$$q_{\alpha,p,v_E}$$

is the Studentized range critical value (see Table B-6 of Appendix B). MSE and v_E are from the ANOVA procedure previously defined, and α represents the comparison-wise error rate. The experiment-wise significance level for comparing the extremes of p means is

$$\alpha_p = 1-(1-\alpha)^{p-1}.$$

M-6.5.1.2.1. If the smallest and largest means are not significantly different, then no more comparisons are made and all other comparisons are declared not significantly different at the $(1-\alpha_p)100\%$ level of confidence.

M-6.5.1.2.2. If the smallest and largest averages are significantly different, then two comparisons are made where $p = k - 1$: one for the second smallest and the largest averages, and one for the smallest and the second largest averages.

M-6.5.1.2.3. For the two comparisons, if neither of these tests is significantly different, then no more comparisons are performed and only two extreme means ($\bar{x}_a$ and $\bar{x}_b$) are concluded to be significantly different.

M-6.5.1.2.4. If one or both of these tests are statistically significant, testing should continue with groups of averages lying within the two extremes that have been declared significantly different.

M-6.5.1.3. Testing continues until no further significant differences are obtained.

M-6.5.2. *Example of Duncan's Multiple Range Test.* Suppose manganese groundwater concentrations are going to be compared among the seven different wells at Site A using Duncan's multiple range test with 95% level of confidence.

M-6.5.2.1. Table M-6 presents the data. All manganese concentrations were detected so no proxy concentrations are needed to evaluate the data.

M-6.5.2.2. The assumptions of normality were verified during the ANOVA process.

M-6.5.2.3. There are seven groups to compare so we begin by comparing the one with the smallest mean to the one with the largest mean.

$$R_7 = q_{\alpha,7,v_E}(\mathrm{MSE}/n)^{1/2} = q_{0.05,7,49}(0.066/8)^{1/2} = 3.255 \times 0.0908 = 0.296\,.$$

Considering

$$\left|\bar{x}_{69-2-02} - \bar{x}_{69-2-05}\right| = \left|-0.8315 - (-5.0176)\right| = 4.186 > 0.296$$

we can conclude that the population means for these two wells differ at the

$$1-(1-\alpha)^{p-1} = 1-(1-0.05)^{7-1} = 0.26$$

significance level. As the two extreme means were significantly different, we now test means that are 6 levels apart.

$$R_6 = q_{\alpha,6,v_E}(\mathrm{MSE}/n)^{1/2} = q_{0.05,6,49}(0.066/8)^{1/2} = 3.212 \times 0.0908 = 0.292\,.$$

Considering

$$\left|\bar{x}_{69-2-02} - \bar{x}_{69-2-07}\right| = \left|-0.8315 - (-4.3172)\right| = 3.486 > 0.292$$

and

$$\left| \bar{x}_{69-2-08} - \bar{x}_{69-2-05} \right| = \left| -0.9073 - (-5.0176) \right| = 4.110 > 0.292$$

we can conclude that the population means for these two comparisons differ at the

$$1-(1-\alpha)^{p-1} = 1-(1-0.05)^{6-1} = 0.23$$

significance level.

M-6.5.2.4. Because means 6 levels apart are significantly different, continue the process with means 5 levels apart and so on. The final results are summarized in the Table M-15, where means with the same letter are not significantly different at an experiment-wise significance level of α = 0.05.

Table M-15.
Means with the same Letter are not Significantly Different at Significance of α = 0.05

Duncan Grouping	Mean	*N*	Well
69-2-02	–0.8315	8	A
69-2-08	–0.9073	8	B A
69-2-06A	–1.1438	8	B
69-2-06B	–2.0084	8	C
69-2-04	–2.8767	8	D
69-2-07	–4.3172	8	E
69-2-05	–5.0176	8	F

M-6.6. *Dunnett's Test for Simple Random and Systematic Samples.* Dunnett's test is used to test the difference between sample or "treatment" means from different populations against a control population. Dunnett's method is the same as the standard two-sample *t*-test (Paragraph M-2), except for the use of a larger pooled estimate of variance and the need for special *t* type tables (Table B-26 of Appendix B). The experiment-wise significance level for all comparisons will be α (Montgomery, 1997). Directions for the use of Dunnett's method for a simple random sample or a systematic random sample are presented in Paragraph M-6.6.1 and followed by an example in Paragraph M-6.6.2.

M-6.6.1. *Directions for Dunnett's Test for Simple Random and Systematic Samples*. Let *K* represent the total number of populations to be compared so there are (*K* – 1) sample populations and a single control population. Let $n_1, n_2, \ldots, n_{K-1}$ represent the sample sizes of each of the (*K* – 1) sample populations and let *m* represent the sample size of the control population.

M-6.6.1.1. H_0: $\mu_i - \mu_C \leq 0$ (no difference between the sample means and the control mean).

M-6.6.1.2. H_A: $\mu_i - \mu_C > 0$ for *i* = 1, 2,…, *K* – 1 where μ_i represents the mean of the i^{th} sample population and μ_C represents the mean of the control population.

M-6.6.1.3. Let α represent the chosen significance level for the test.

M-6.6.1.4. Verify the assumptions of normality. For each sample population, make sure that $0.5 < m / n_i < 2$. If not, Dunnett's Test should *not* be used.

M-6.6.1.5. Calculate the sample mean, $\bar{x}_i$, and the variance, s_i^2, for each of the K – 1 populations and the control ($i = 1,\ 2, \ldots,\ K - 1,\ C$).

M-6.6.1.6. Calculate the pooled standard deviation:

$$s_p = \sqrt{\frac{(m-1)s_C^2 + (n_1 - 1)s_1^2 + \ldots + (n_{K-1} - 1)s_{K-1}^2}{(m-1) + (n_1 - 1) + \ldots + (n_{K-1} - 1)}} \quad .$$

For each of the K – 1 sample populations, compute

$$t_i = \frac{\bar{x}_i - \bar{x}_C}{s_p \sqrt{1/n_i + 1/n_C}} \quad .$$

M-6.6.1.7. Use Table B-26 of Appendix B to determine the critical value, $t_{1-\alpha,v_E}$, where the degrees of freedom v_E = $(m-1) + (n_1 - 1) + \ldots + (n_{K-1} - 1)$. Compare t_i to $t_{1-\alpha,v_E}$ for each of the K – 1 sample populations.

M-6.6.1.7.1. If t_i > $t_{1-\alpha,v_E}$ for any sample population, then reject H_0 and conclude that the mean of the sample population exceeds the mean of the control population.

M-6.6.1.7.2. Otherwise, conclude that the mean of the sample population does not exceed the mean of the control population.

M-6.6.2. *Example of Dunnett's Test for Simple Random and Systematic Samples*. Suppose manganese (Mn) groundwater concentrations at six wells are going to be compared to a background well at Site A using the following test with 95% level of confidence.

M-6.6.2.1. H_0: $\mu_i - \mu_C \leq 0$ (no difference between the sample means and the control mean).

M-6.6.2.2. H_A: $\mu_i - \mu_C > 0$ for i = 1, 2,…, K – 1 where μ_i represents the mean of the i^{th} sample population and μ_C represents the mean of the control population.

M-6.6.2.3. All Mn concentrations were detected so no proxy concentrations are needed to evaluate the data.

M-6.6.2.4. The assumptions of normality were verified during the ANOVA process. Because the sample population for each well is equal to 8, we only have to calculate m/n_i once. As $m/n_i = 8/8 = 1$ is between 0.5 and 2, it is reasonable to apply Dunnett's test.

Well	69-2-02	69-2-04	69-2-08	69-2-05	69-2-06B	69-2-06A	Bkgd
Mean	–0.832	–2.877	–0.907	–5.018	–2.008	–1.144	–4.317
Variance	0.064	0.041	0.091	0.033	0.143	0.011	0.080

$$s_p = \sqrt{\frac{(m-1)s_c^2 + (n_1-1)s_1^2 + \ldots + (n_{K-1}-1)s_{K-1}^2}{(m-1)+(n_1-1)+\ldots+(n_{K-1}-1)}}$$

$$= \sqrt{\frac{7(0.080+0.064+0.041+0.091+0.033+0.143+0.011)}{7+7+7+7+7+7+7}} = \sqrt{\frac{3.240}{49}} = 0.2571$$

$$t_i = \frac{\bar{x}_i - \bar{x}_C}{s_p\sqrt{1/n_i + 1/n_C}} = \frac{\bar{x}_i - (-4.317)}{0.2571\sqrt{1/8+1/8}} = \frac{\bar{x}_i + 4.317}{0.1286}$$

so for each sample well

Sample Well, *i*	***t***
69-2-02	27.11
69-2-04	11.20
69-2-08	26.52
69-2-05	–5.45
69-2-06B	17.96
69-2-06A	24.68

M-6.6.2.5. The degrees of freedom are $(8-1)+(8-1)+\ldots+(8-1)=49$. So, using Table B-26 of Appendix B with 49 degrees of freedom, the critical value $t_{0.95,49}$ = 2.32.

M-6.6.2.5. For all wells except Well 69-2-05, $t_i > t_{0.95,\,49}$. We then reject H_0 and conclude that the means of the sample well populations exceed the mean of the control well population, except for Well 69-2-05.

Table M-16.
Data for Example M-6.6.2

Well Location	Result	Log Result	Well Location	Result	Log Result
69-2-02	0.432	–0.839	69-2-06A	0.294	–1.224
69-2-02	0.44	–0.821	69-2-06A	0.301	–1.201
69-2-02	0.513	–0.667	69-2-06A	0.379	–0.970
69-2-02	0.704	–0.351	69-2-06A	0.352	–1.044
69-2-02	0.327	–1.118	69-2-06A	0.346	–1.061
69-2-02	0.316	–1.152	69-2-06B	0.13	–2.040
69-2-02	0.454	–0.790	69-2-06B	0.184	–1.693
69-2-02	0.401	–0.914	69-2-06B	0.209	–1.565
69-2-04	0.0504	–2.988	69-2-06B	0.2	–1.609
69-2-04	0.0502	–2.992	69-2-06B	0.0739	–2.605
69-2-04	0.054	–2.919	69-2-06B	0.0876	–2.435
69-2-04	0.0523	–2.951	69-2-06B	0.126	–2.071
69-2-04	0.0923	–2.383	69-2-06B	0.129	–2.048
69-2-04	0.0556	–2.890	bkgd	0.0137	–4.290
69-2-04	0.0534	–2.930	bkgd	0.019	–3.963
69-2-04	0.0517	–2.962	bkgd	0.0163	–4.117
69-2-05	0.00684	–4.985	bkgd	0.0195	–3.937
69-2-05	0.00639	–5.053	bkgd	0.0112	–4.492
69-2-05	0.00631	–5.066	bkgd	0.0112	–4.492
69-2-05	0.00813	–4.812	bkgd	0.0102	–4.585
69-2-05	0.00747	–4.897	bkgd	0.00946	–4.661
69-2-05	0.00679	–4.992	69-2-08	0.563	–0.574
69-2-05	0.00731	–4.919	69-2-08	0.512	–0.669
69-2-05	0.00444	–5.417	69-2-08	0.475	–0.744
69-2-06A	0.3	–1.204	69-2-08	0.546	–0.605
69-2-06A	0.286	–1.252	69-2-08	0.276	–1.287
69-2-06A	0.303	–1.194	69-2-08	0.383	–0.960
			69-2-08	0.33	–1.109
			69-2-08	0.27	–1.309

M-6.7. *Scheffé's Test.* Scheffé's test is designed to allow the comparison of any set of contrasts while controlling the experiment-wise Type I error rate (the probability of declaring any contrast different from 0 when it is not) to be no more then α (Montgomery, 1997). When the experimenter is only interested in comparing pairs of treatment means, Scheffé's test is not the most sensitive. Directions for Scheffé's Test and an example are presented in Paragraphs M-6.7.1 and M-6.7.2, respectively.

M-6.7.1. *Directions for Scheffé's Test.* Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let

$$N = \sum_{i=1}^{K} n_i$$

be the overall sample size. Let the values from each population be represented by $x_{i,j}$ where i = 1, 2,…, K for the K groups and $j = 1, 2, ..., n_i$ for the observations in the i^{th} group. Let $(1-\alpha)100\%$ be the confidence level for the test.

M-6.7.1.1. Verify the assumptions of normality. Let

$$\theta = \sum a_i \mu_i$$

represent one of m linear combinations of the means u_i being tested for $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$.

M-6.7.1.2. Reject H_0 if $|\theta| = \left| \sum a_i \bar{x}_i \right|$ exceeds the critical value

$$S_\alpha = \sqrt{MSE \sum_{i=1}^{K} \left(a_i^2 / n_i \right)} \sqrt{(K-1) F_{1-\alpha, K-1, N-K}}$$

where n_i is the number of observations in the i^{th} group of

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}$$

and

$$F_{1-\alpha, K-1, N-K}$$

is the $(1-\alpha)100\%$ percentile for the F distribution with K – 1 numerator degrees of freedom and N – K denominator degrees of freedom (see Table B-7 in Appendix B).

M-6.7.2. *Example of Scheffé's Test.* Suppose manganese concentrations in groundwater are going to be compared in six different sampling wells and a background well using Scheffé's test with a 95% level of confidence.

M-6.7.2.1. Table M-16 presents the data. All manganese concentrations were detected, so no proxy concentrations are needed to evaluate the data. The assumptions of normality were verified during the ANOVA process.

M-6.7.2.2. Suppose two contrasts are of interest: comparing the background well to all of the other wells combined and comparing well 69-2-06A to well 69-2-06B. These two contrasts can be written:

$$\theta_1 = 6\mu_{bkgd} - \mu_{69-2-02} - \mu_{69-2-04} - \mu_{69-2-05} - \mu_{69-2-06A} - \mu_{69-2-06B} - \mu_{69-2-08}$$

$$\theta_2 = \mu_{69-2-06A} - \mu_{69-2-06B} \ .$$

The contrast estimates are:

$$\hat{\theta}_1 = 6\bar{x}_{bkgd} - \bar{x}_{69-2-02} - \bar{x}_{69-2-04} - \bar{x}_{69-2-05} - \bar{x}_{69-2-06A} - \bar{x}_{69-2-06B} - \bar{x}_{69-2-08}$$

$$\begin{aligned} &= 6(-4.317)-(-0.832)-(-2.877)-(-5.018)-(-1.144)-(-2.008)-(-0.907) \\ &= -13.1177 \end{aligned}$$

$$\hat{\theta}_2 = \bar{x}_{69-2-06A} - \bar{x}_{69-2-06B} = -1.144-(-2.008) = 0.8646 .$$

The critical values are:

$$\begin{aligned} S_{\alpha_1} &= \sqrt{\text{MSE}\sum_{i=1}^{K}\left(a_i^2/n_i\right)} \times \sqrt{(K-1)F_{1-\alpha,K-1,N-K}} \\ &= \sqrt{0.066 \times \left(\frac{36}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}\right)} \times \sqrt{(7-1)F_{0.95,6,49}} \\ &= 0.589 \times \sqrt{6 \times 2.29} = 2.1841 \end{aligned}$$

$$\begin{aligned} S_{\alpha_2} &= \sqrt{\text{MSE}\sum_{i=1}^{K}\left(a_i^2/n_i\right)} \times \sqrt{(K-1)F_{1-\alpha,K-1,N-K}} \\ &= \sqrt{0.066 \times \left(\frac{1}{8}+\frac{1}{8}\right)} \times \sqrt{(7-1)F_{0.95,6,49}} \\ &= 0.128 \times \sqrt{6 \times 2.29} = 0.4766 \ . \end{aligned}$$

M-6.7.2.3. Because the absolute value of each contrast exceeds the relevant critical value, we reject $H_0 : \theta_1 = 0$ and $H_0 : \theta_2 = 0$ with 95% confidence. In other words, the average measurement at the background well is significantly different from the average measurement at the other six wells, and the average measurement at well 69-2-06A differs significantly from the average at well 69-2-06B.