CHAPTER 3. METHODS FOR EVALUATING DATA

3.1. INTRODUCTION

Once data have been collected, it is necessary to statistically summarize and analyze the data. EPA recommends that the data analysis methods be selected before collecting the first sample. Many statistical methods have been computerized in easy-to-use software that is available for use on personal computers. Inclusion or exclusion in this section does not imply an endorsement or lack thereof by the U.S. Environmental Protection Agency. Commercial-off-the-shelf software that covers a wide range of statistical and graphical support includes SAS, Statistica, Statgraphics, Systat, Data Desk (Macintosh only), BMDP, and JMP. Numerous spreadsheets, database management packages, and other graphics software can also be used to perform many of the needed analyses. In addition, the following programs, written specifically for environmental analyses, are also available:

SCOUT: A Data Analysis Program, EPA, NTIS Order Number PB93-505303.

WQHYDRO (WATER QUALITY/HYDROLOGY GRAPHICS/ANALYSIS SYSTEM), Eric R. Aroner, Environmental Engineer, P.O. Box 18149, Portland, OR 97218.

WQSTAT, Jim C. Loftis, Department of Chemical and Bioresource Engineering, Colorado State University, Fort Collins, CO 80524.

Computing the proportion of sites implementing a certain BMP or the average number of acres that are under a certain BMP follows directly from the equations presented in Section 2.3 and is not repeated. The remainder of this section is focused on evaluating changes in BMP implementation. The methods provided in this section provide only a cursory overview of the type of analyses that might be of interest. For a more thorough discussion on these methods, the reader is referred to Gilbert (1987), Snedecor and Cochran (1980), and Helsel and Hirsch (1995). Typically the data collected for evaluating changes will typically come as two or more sets of random samples. In this case, the analyst will test for a shift or step change.

Depending on the objective, it is appropriate to select a one- or two-sided test. For example, if the analyst knows that BMP implementation will only go up as a result of an operator education program, a one-sided test could be formulated. Alternatively, if the analyst does not know whether implementation will go up or down, a two-sided test is necessary. To simply compare two random samples to decide whether they are significantly different, a two-sided test is used. Typical null hypotheses (H_o) and alternative hypotheses (H_a) for one- and two-sided tests are provided below:

<u>One-sided test</u>

H_o: BMP Implementation (Post education) $\leq$ BMP Implementation (Pre education)

H_a: BMP Implementation (Post education) > BMP Implementation (Pre education)

Two-sided test

H_o: BMP Implementation (Post education) = BMP Implementation (Pre education)

H_a: BMP Implementation (Post education) ≠ BMP Implementation (Pre education)

Selecting a one-sided test instead of a two-sided test results in an increased power for the same significance level (Winer, 1971). That is, if the conditions are appropriate, a corresponding one-sided test is more desirable than a two-sided test given the same " and sample size. The manager and analyst should take great care in choosing one- or two-sided tests.

3.2. COMPARING THE MEANS FROM TWO INDEPENDENT RANDOM SAMPLES

The Student's *t* test for two samples and the Mann-Whitney test are the most appropriate tests for these types of data. Assuming the data meet the assumptions of the *t* test, the two-sample *t* statistic with n_1+n_2-2 degrees of freedom is (Remington and Schork, 1970)

$$t = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \textbf{(3-1)}$$

where n_1 and n_2 are the sample sizes of the first and second data sets, respectively, and $\overline{x}_1$ and $\overline{x}_2$ are the estimated means from the first and second data sets, respectively. The pooled standard deviation, s_p, is defined by

Tests for Two Independent Random Samples

Test*	Key Assumptions
Two-sample *t*	• Both data sets must be normally distributed • Data sets should have equal variances†
Mann-Whitney	• None

* The standard forms of these tests require independent random samples.

† The variance homogeneity assumption can be relaxed.

$$s_p = \left[\frac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{n_1 + n_2 - 2} \right]^{0.5} \qquad \textbf{(3-2)}$$

where s_1^2 and s_2^2 correspond to the estimated variances of the first and second data sets, respectively. The difference quantity (Δ_o) can be any value, but here it is set to zero. Δ_o can be set to a non-zero value to test whether the difference between the two data sets is greater than a selected value. If the variances are not equal, Snedecor and Cochran (1980) can be used as a source for methods for computing the *t* statistic. In a two-sided test, the value from Equation 2-18 is compared to the *t* value from Table A2 with $\alpha/2$ and n_1+n_2-2 degrees of freedom.

The Mann-Whitney test can also be used to compare two independent random samples. This test is very flexible since there are no assumptions about the distribution of either sample or whether the distributions have to be the same (Helsel and Hirsch, 1995). Wilcoxon (1945) first introduced this test for equal-sized samples. Mann and Whitney (1947) modified the original Wilcoxon's test to apply it to

different sample sizes. Here, it is determined whether one data set tends to have larger observations than the other.

If the distributions of the two samples are similar except for location (i.e., similar spread and skew), H_a can be refined to imply that the median concentration from one sample is "greater than," "less than," or "not equal to" the median concentration from the second sample. To achieve this greater detail in H_a, transformations such as logs can be used.

Tables of Mann-Whitney test statistics (e.g., Conover, 1980) can be consulted to determine whether to reject H_o for small sample sizes. If n_1 and n_2 are greater than or equal to 10 observations, the test statistic can be computed from the following equation (Conover, 1980):

$$T_1 = \frac{T - n_1\frac{n+1}{2}}{\sqrt{\frac{n_1 n_2}{n(n-1)}\sum_{i=1}^{n} R_i^2 - \frac{n_1 n_2 (n+1)^2}{4(n-1)}}} \quad \textbf{(3-3)}$$

where

n_1 = number of observations in sample with fewer observations,
n_2 = number of observations in sample with more observations,
n = $n_1 + n_2$,
T = sum of ranks for sample with fewer observations, and
R_i = rank for the ith ordered observation used in both samples.

T_1 is normally distributed and Table A1 can be used to determine the appropriate quantile. Helsel and Hirsch (1995) and USEPA (1996) provide detailed examples for both of these tests.

3.3. COMPARING THE PROPORTIONS FROM TWO INDEPENDENT SAMPLES

Consider the example in which the proportion of waterbars that effectively divert water from the skid trail has been estimated during two time periods to be p_1 and p_2 using sample sizes of n_1 and n_2, respectively. Assuming a normal approximation is valid, the test statistic under a null hypothesis of equivalent proportions (no change) is

$$\frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad \textbf{(3-4)}$$

where p is a pooled estimate of proportion and is equal to $(x_1+x_2)/(n_1+n_2)$ and x_1 and x_2 are the number of successes during the two time periods. An estimator for the difference in proportions is simply p_1-p_2.

In an earlier example, it was determined that 129 observations in each sample were needed to detect a difference in proportions of 0.20 with a two-sided test, α equal to 0.05, $1-\beta$ equal to 0.90. Assuming that 130 samples were taken and p_1 and p_2 were estimated from the data as 0.6 and 0.4, the test statistic would be estimated as

$$\frac{0.6 - 0.4}{\sqrt{0.5(0.5)\left(\frac{1}{130}+\frac{1}{130}\right)}} = 3.22 \quad \textbf{(3-5)}$$

Comparing this value to the *t* value from Table A2 ($\alpha/2 = 0.025$, df=258) of 1.96, H_o is rejected.

3.4. COMPARING MORE THAN TWO INDEPENDENT RANDOM SAMPLES

The analysis of variance (ANOVA) and Kruskal-Wallis are extensions of the two-sample *t* and Mann-Whitney tests, respectively, and can be used for analyzing more than two independent random samples when the data are continuous (e.g., average SMA width). Unlike the *t* test described earlier, the ANOVA can have more than one factor or explanatory variable. The Kruskal-Wallis test accommodates only one factor, whereas the Friedman test can be used for two factors. In addition to applying one of the above tests to determine if one of the samples is significantly different from the others, it is also necessary to perform postevaluations to determine which of the samples is different. This section recommends Tukey's method to analyze the raw or rank-transformed data only if one of the previous tests (ANOVA, rank-transformed ANOVA, Kruskal-Wallis, Friedman) indicates a significant difference between groups. Tukey's method can be used for equal or unequal sample sizes (Helsel and Hirsch, 1995). The reader is cautioned, when performing an ANOVA using standard software, to be sure that the ANOVA test used matches the data. USEPA (1996) provides a more detailed discussion on comparing more than two independent random samples.

3.5. COMPARING CATEGORICAL DATA

In comparing categorical data it is important to distinguish between whether the categories are nominal (e.g., land ownership, county location, type of BMP) or ordinal (e.g., BMP implementation rankings, low-medium-high scales).

The starting point for all evaluations is the development of a contingency table. In Table 3-1, the preference of three BMPs is compared to harvest site type in a contingency table. In this case both categorical variables are nominal. In this example, 45 of the 102 observations on federal lands used BMP_1. There were a total of 174 observations.

To test for independence, the sum of the squared differences between the expected (E_{ij}) and observed (O_{ij}) count summed over all cells is computed as (Helsel and Hirsch, 1995)

$$\chi_{ct} = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \textbf{(3-6)}$$

where E_{ij} is equal to $A_i C_j / N$. χ_{ct} is compared to the $1-\alpha$ quantile of the χ^2 distribution with $(m-1)(k-1)$ degrees of freedom (see Table A3).

In the example presented in Table 3-1, the symbols listed in the parentheses correspond to the above equation. Note that *k* corresponds to the three types of BMPs and *m* corresponds to the three different types of harvest site. Table 3-2 shows computed values of E_{ij} and $(O_{ij}-E_{ij})^2/E_{ij}$ in parentheses for the example data. χ_{ct} is equal to 14.60.

Table 3-1. Contingency table of harvest site type and implemented BMP.

Harvest Site Type	BMP_1	BMP_2	BMP_3	Row Total, A_i
Private	10 (O_{11})	30 (O_{12})	17 (O_{13})	57 (A_1)
Federal	45 (O_{21})	32 (O_{22})	25 (O_{23})	102 (A_2)
State	8 (O_{31})	3 (O_{32})	4 (O_{33})	15 (A_3)
Column Total, C_j	63 (C_1)	65 (C_2)	46 (C_3)	174 (N)

Key to Symbols:
O_{ij} = number of observations for the *i*th harvest site and *j*th BMP type
A_i = row total for the *i*th harvest site type (total number of observations for a given harvest site type)
C_j = column total for the *j*th BMP type (total number of observations for a given BMP type)
N = total number of observations

From Table A3, the 0.95 quantile of the P^2 distribution with 4 degrees of freedom is 9.488. H_o is rejected; the selection of BMP is not random among the different harvest site types. The largest values in the parentheses in Table 3-2 give an idea as to which combinations of harvest site type and BMP are noteworthy. In this example, it appears that BMP_2 is preferred to BMP_1 in comparison to federal and state harvest sites.

Now consider that in addition to evaluating information regarding the harvest site and BMP type, we also recorded a value from 1 to 5 indicating how well the BMP was installed and maintained, with 5 indicating the best results. In this case, the BMP implementation rating is ordinal. Using the same notation as before, the average rank of observations in row *x*, R_x, is equal to (Helsel and Hirsch, 1995)

$$R_x = \sum_{i=1}^{x-1} A_i + (A_x+1)/2 \qquad \textbf{(3-7)}$$

where A_i corresponds to the row total. The average rank of observations in column *j*, D_j, is equal to

$$D_j = \frac{\sum_{i=1}^{m} O_{ij} R_i}{C_j} \qquad \textbf{(3-8)}$$

where C_j corresponds to the column total. The Kruskal-Wallis test statistic is then computed as

$$K = (N-1)\,\frac{\sum_{j=1}^{k} C_j D_j^2 - N\left[\frac{N+1}{N}\right]^2}{\sum_{i=1}^{m} A_i R_i^2 - N\left[\frac{N+1}{N}\right]^2} \qquad \textbf{(3-9)}$$

where K is compared to the P^2 distribution with *k-1* degrees of freedom. This is the most general form of the Kruskal-Wallis test since it is a comparison of distribution shifts rather than shifts in the median (Helsel and Hirsch, 1995).

Table 3-2. Contingency table of expected harvest site type and implemented BMP. (Values in parentheses correspond to $(O_{ij}-E_{ij})^2/E_{ij}$.)

Harvest Site Type	BMP_1	BMP_2	BMP_3	Row Total
Private	20.64 (5.48)	21.29 (3.56)	15.07 (0.25)	57
Federal	36.93 (1.76)	38.10 (0.98)	26.97 (0.14)	102
State	5.43 (1.22)	5.60 (1.21)	3.97 (0.00)	15
Column Total	63	65	46	174

Table 3-3 is a continuation of the previous example indicating the BMP implementation rating for each BMP type. For example, 29 of the 70 observations that were given a rating of 4 are associated with BMP_2. The terms inside the parentheses of Table 3-3 correspond to the terms used in Equations 3-7 to 3-9. Note that *k* corresponds to the three types of BMPs and *m* corresponds to the five different levels of BMP implementation. Using Equation 3-9 for the data in Table 3-3, *K* is equal to 14.86. Comparing this value to 5.991 obtained from Table A3, there is a significant difference in the quality of implementation between the three BMPs.

The last type of categorical data evaluation considered in this chapter is that in which both variables are ordinal. The Kendall J_b for tied data can be used for this analysis. The statistic J_b is calculated as (Helsel and Hirsch, 1995)

$$\tau_b = \frac{S}{\frac{1}{2}\sqrt{(N^2 - SS_a)(N^2 - SS_b)}} \quad \textbf{(3-10)}$$

where *S*, SS_a, and SS_c are computed as

$$S = \sum_{allxy}\left[\sum_{i>x}\sum_{j>y} O_{xy}O_{ij} - \sum_{i<x}\sum_{j<y} O_{xy}O_{ij}\right] \quad \textbf{(3-11)}$$

$$SS_a = \sum_{i=1}^{m} A_i^2 \quad \textbf{(3-12)}$$

$$SS_c = \sum_{j=1}^{k} C_j^2 \quad \textbf{(3-13)}$$

To determine whether J_b is significant, *S* is modified to a normal statistic, using

$$Z_S = \begin{cases} \dfrac{S-1}{\sigma_S} & if\ S>0 \\ \dfrac{S+1}{\sigma_S} & if\ S<0 \end{cases} \quad \textbf{(3-14)}$$

Table 3-3. Contingency table of implemented BMP and rating of installation and maintenance.

BMP Implementation Rating	BMP_1	BMP_2	BMP_3	Row Total, A_i
1	1 (O_{11})	2 (O_{12})	2 (O_{13})	5 (A_1)
2	7 (O_{21})	3 (O_{22})	5 (O_{23})	15 (A_2)
3	15 (O_{31})	16 (O_{32})	26 (O_{33})	57 (A_3)
4	32 (O_{41})	29 (O_{42})	9 (O_{43})	70 (A_4)
5	8 (O_{51})	15 (O_{52})	4 (O_{53})	27 (A_5)
Column Total, C_j	63 (C_1)	65 (C_2)	46 (C_3)	174 (N)

Key to Symbols:
O_{ij} = number of observations for the *i*th BMP implementation rating and *j*th BMP type
A_i = row total for the *i*th BMP implementation rating (total number of observations for a given BMP implementation rating)
C_j = column total for the *j*th BMP type (total number of observations for a given BMP type)
N = total number of observations

where

$$\sigma_s = \sqrt{\frac{N^3}{9}\left(1-\sum_{i=1}^{m} a_i^3\right)\left(1-\sum_{j=1}^{k} c_j^3\right)} \qquad \textbf{(3-15)}$$

where Z_S is zero if S is zero. The values of a_i and c_i are compute as A_i/N and C_i/N, respectively.

Table 3-4 presents the BMP implementation ratings that were taken in three separate years. For example, 15 of the 57 observations that were given a rating of 3 are associated with Year 2. Using Equations 3-11 and 3-15, S and F_s are equal to 2,509 and 679.75, respectively. Therefore, Z_s is equal to (2509-1)/679.75 or 3.69. Comparing this value to a value of 1.96, obtained from Table A1 ($\alpha/2$=0.025), indicates that BMP implementation is improving with time.

Table 3-4. Contingency table of implemented BMP and sample year.

BMP Implementation Rating	Year 1	Year 2	Year 3	Row Total, A_i	a_i
1	2 (O_{11})	1 (O_{12})	2 (O_{13})	5 (A_1)	0.029
2	5 (O_{21})	7 (O_{22})	3 (O_{23})	15 (A_2)	0.086
3	26 (O_{31})	15 (O_{32})	16 (O_{33})	57 (A_3)	0.328
4	9 (O_{41})	32 (O_{42})	29 (O_{43})	70 (A_4)	0.402
5	4 (O_{51})	8 (O_{52})	15 (O_{53})	27 (A_5)	0.155
Column Total, C_j	46 (C_1)	63 (C_2)	65 (C_3)	174 (N)	
c_i	0.264	0.362	0.374		

Key to Symbols:
O_{ij} = number of observations for the *i*th BMP implementation rating and *j*th year
A_i = row total for the *i*th BMP implementation rating (total number of observations for a given harvest type)
C_j = column total for the *j*th BMP type (total number of observations for a given year)
N = total number of observations
a_i = A_i /N
c_j = C_j /N