

Chapter 7

Comparing Several Independent Groups

Concentrations of volatile organic compounds are measured in shallow ground waters across a several county area. The wells sampled can be classified as being contained in one of seven land-use types: undeveloped, agricultural, wetlands, low-density residential, high-density residential, commercial, and industrial/transportation. Do the concentrations of volatiles differ between these types of surface land-use, and if so, how?

Alkalinity, pH, iron concentrations, and biological diversity are measured at low flow for small streams draining areas mined for coal. Each stream drains either unmined land, land strip-mined and then abandoned, or land strip-mined and then reclaimed. The streams also drain one of two rock units, a sandstone or a limestone formation. Do drainages from mined and unmined lands differ in quality? What affect has reclamation had? Are there differences in chemical or biological quality due to rock type separate and distinct from the effects due to mining history?

Three methods for field extraction and concentration of an organic chemical are to be compared at numerous wells. Are there differences among concentrations produced by the extraction processes? These must be discerned above the well-to-well differences in concentration which contribute considerable noise to the data.

The methods of this chapter can be used to answer questions such as those above. These methods are extensions of the ones introduced in Chapters 5 and 6, where now more than two groups of data are to be compared. The classic technique in this situation is analysis of variance. More robust nonparametric techniques are also presented for the frequent situations where data do not meet the assumptions of analysis of variance.

Suppose a set of continuous data, such as concentration or water levels, is collected. It is suspected that one or more influences on the magnitude of these data comes from grouped variables, variables whose values are simply "from group X". Examples include season of the year ("from summer", "winter", etc.), aquifer type, land-use type, and similar groups. Each observation will be classified into one of these groups.

First consider the effect of only one grouped variable, calling it an **explanatory variable** because it is believed to explain some of the variation in magnitude of the data at hand. This variable is also called a **factor**. It consists of a set of k groups, with each data point belonging in one of the k groups. For example, the data could be calcium concentrations from wells in one of k aquifers, and the objective is to determine whether the calcium concentrations differ among the k aquifers. Within each group (aquifer) there are n_j observations (the sample size of each group is not necessarily the same). Observation y_{ij} is the i th of n_j observations in group j , so that $i=1, \dots, n_j$ for the j th of k groups $j=1, \dots, k$. The total number of observations N is thus

$$N = \sum_{j=1}^k n_j, \quad \text{which simplifies to} \quad N = k \cdot n$$

when the sample size $n_j = n$ for all k groups (equal sample sizes).

The tests in this chapter determine if all k groups have the same central value (median or mean, depending on the test), or whether at least one of the groups differs from the others. When data within each of the groups are normally distributed and possess identical variances, an analysis of variance (ANOVA) can be used. Analysis of variance is a parametric test, determining whether each group's mean is identical. When there are only two groups, the ANOVA becomes identical to a t -test. Thus ANOVA is like a t -test between three or more groups of data, and is restricted by the same types of assumptions as was the t -test. When every group of data cannot be assumed to be normally distributed or have identical variance, a nonparametric test should be used instead. The Kruskal-Wallis test is much like a rank-sum test extended to more than two groups. It compares the medians of groups differentiated by one explanatory variable (one factor).

When the effect of more than one factor is to be evaluated simultaneously, such as both rock type and mining history in one of the examples which began this chapter, the one-way tests can no longer be used. For data which can be assumed normal, several factors can be tested simultaneously using multi-factor analysis of variance. However, the requirements of normality and equal variance now apply to data grouped by each unique combination of factors. This becomes quite restrictive and is rarely met in practice. Therefore nonparametric alternatives are also presented.

The following sections begin with tests for differences due to one factor. Subsequent sections discuss tests for effects due to more than one factor. All of these have as their null hypothesis

that each group median (or mean) is identical, with the alternative that at least one is different. However, when the null hypothesis is rejected, these tests do not tell which group or groups are different! Therefore sections also follow on multiple comparison tests -- tests performed after the ANOVA or Kruskal-Wallis null hypothesis has been rejected, for determining which groups differ from others. A final section on graphical display of results finishes the chapter.

7.1 Tests for Differences Due to One Factor

7.1.1 The Kruskal-Wallis Test

The Kruskal-Wallis test, like other nonparametric tests, may be computed by an exact method used for small sample sizes, by a large-sample approximation (a chi-square approximation) available on statistical packages, and by ranking the data and performing a parametric test on the ranks. Tables for the exact method give p-values which are exactly correct. The other two methods produce approximate p-values that are only valid when sample sizes are large, but do not require special tables. Tables of exact p-values for all sample sizes would be huge, as there are many possible combinations of numbers of groups and sample sizes per group. Fortunately, large sample approximations for all but the smallest sample sizes are very close to their true (exact) values. Thus exact computations are rarely required. All three versions have the same objective, as stated by their null and alternate hypotheses.

7.1.1.1 Null and alternate hypotheses

In its most general form, the Kruskal-Wallis test has the following null and alternate hypotheses:

- H_0 : All of the k groups of data have identical distributions, versus
 H_1 : At least one group differs in its distribution.

No assumptions are required about the shape(s) of the distributions. They may be normal, lognormal, or anything else. If the alternate hypothesis is true, they may have different distributional shapes. In this form, the only interest in the data is to determine whether all groups are identical, or whether some tend to produce observations different in value than the others. This difference is not attributed solely to a difference in median, though that is one possibility. Thus the Kruskal-Wallis test, like the rank-sum test, may be used to determine the general equivalence of groups of data.

In practice, the test is usually performed for a more specific purpose -- to determine whether all groups have the same median, or whether at least one median is different. This form requires that all other characteristics of the data distributions, such as spread or skewness, are identical -- though not necessarily in the original units. Any data for which a monotonic transformation, such as in the ladder of powers, produces similar spreads and skewness are also valid. This parallels the rank-sum test (see Chapter 5). As a test for difference in medians, the Kruskal-Wallis null and alternate hypotheses are:

- H_0 : The medians of the k groups are identical,
 H_1 : At least one median differs from the others. (a 2-sided test).

As with the rank-sum test, the Kruskal-Wallis test statistic and p -value computed for data that are transformed using any monotonic transformation are identical to the test statistic and p -value using the original units. Thus there is little incentive to search for transformations (to normality or otherwise) -- the test is applicable in many situations.

7.1.1.2 Computation of the exact test

The exact form of the Kruskal-Wallis test is required when comparing 3 groups with sample sizes of 5 or less per group, or with 4 or more groups of size 4 or less per group (Lehmann, 1975). For larger sample sizes the large-sample approximation is sufficiently accurate. As there are few instances where sample sizes are small enough to warrant using the exact test, exact tables for the Kruskal-Wallis test are not included in this book. Refer to either Conover (1980) or Lehmann (1975) for those tables.

Should the exact test be required, compute the exact test statistic K as shown in the large sample approximation of the following section. K is computed identically for both the exact form or large sample approximation. When ties occur, the large sample approximation must be used.

7.1.1.3 The large-sample approximation

To compute the test, the data are ranked from smallest to largest, from 1 to N . At this point the original values are no longer used; their ranks are used to compute the test statistic. If the null hypothesis is true, the average rank for each group should be similar, and also be close to the overall average rank for all N data. When the alternative hypothesis is true, the average rank for some of the groups will differ from others, reflecting the difference in magnitude of its observations. Some of the average group ranks will then be significantly higher than the overall average rank for all N data, and some will be lower. The test statistic K uses the squares of the differences between the average group ranks and the overall average rank, to determine if groups differ in magnitude. K will equal 0 if all groups have identical average ranks, and will be positive if group ranks are different. The distribution of K when the null hypothesis is true can be approximated quite well by a chi-square distribution with $k-1$ degrees of freedom.

The degrees of freedom is a measure of the number of independent pieces of information used to construct the test statistic. If all data are divided by the overall group mean to standardize the data set, then when any $k-1$ average group ranks are known, the final (k th) average rank can be computed from the others as

$$\bar{R}_k = \frac{N}{n_k} \cdot \left(1 - \sum_{j=1}^{k-1} \frac{n_j}{N} \bar{R}_j \right)$$

Therefore there are actually only $k-1$ independent pieces of information represented by the k average group ranks. From these the k th average rank is fixed.

Large Sample Approximation for the Kruskal-Wallis test	
Situation	Several groups of data are to be compared, to determine if their medians are significantly different. For a total sample size of N , the overall average rank will equal $(N+1)/2$. If the average rank within a group (average group rank) differs considerably from this overall average, not all groups can be considered similar.
Computation	All N observations are jointly ranked from 1 to N , smallest to largest. These ranks R_{ij} are then used for computation of the test statistic. Within each group, the average group rank \bar{R}_j is computed: $\bar{R}_j = \frac{\sum_{i=1}^{n_j} R_{ij}}{n_j} .$
Tied data	When observations are tied, assign the average of their ranks to each.
Test Statistic	The average group rank \bar{R}_j is compared to the overall average rank $\bar{R} = (N+1)/2$, squaring and weighting by sample size, to form the test statistic K : $K = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left[\bar{R}_j - \frac{N+1}{2} \right]^2 .$
Decision Rule	To reject H_0 : all groups have identical distributions, versus H_1 : at least one distribution differs Reject H_0 if $K \geq \chi^2_{1-\alpha, (k-1)}$ the $1-\alpha$ quantile of a chi-square distribution with $(k-1)$ degrees of freedom; otherwise do not reject H_0 .

Example 1.

Fecal coliforms, in organisms per 100 ml, were measured in the Illinois River from 1971 to 1976 (Lin and Evans, 1980). A small subset of those data are presented here. Do all four seasons exhibit similar values, or do one or more seasons differ? Boxplots for the four seasons are shown in figure 7.1.

	<u>Summer</u>	<u>Fall</u>	<u>Winter</u>	<u>Spring</u>
	100	65	28	22
	220	120	58	53
	300	210	120	110
	430	280	230	140
	640	500	310	320
	1600	1100	500	1300
PPCC p-value	0.05	0.06	0.50	0.005

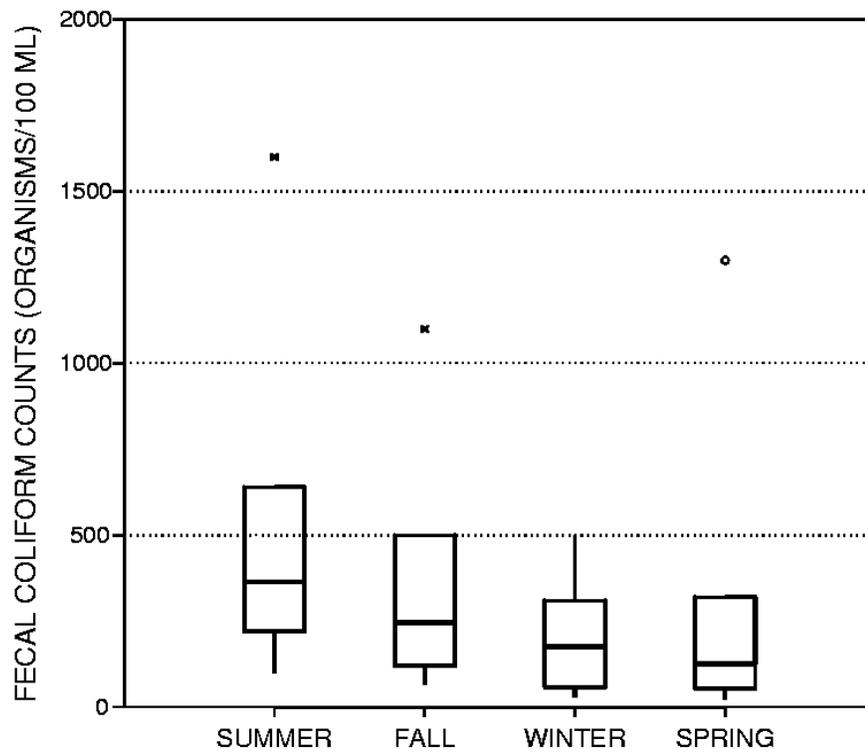


Figure 7.1 Boxplots of Fecal Coliform Data from the Illinois River

Should a parametric or nonparametric test be performed on these data? If even one of the four groups exhibits non-normality, the assumptions of parametric analysis of variance are violated. The consequences of this violation is an inability to detect differences which are truly present -- a lack of power. The PPCC test for normality rejects normality at $\alpha = 0.05$ for two of the seasons, summer and spring (table 7.1). Outliers and skewness for the fall samples also argue for non-normality. Based solely on the skewness and outliers evident in the boxplot, a nonparametric test should be used on these data.

Computation of the Kruskal-Wallis test is shown in table 7.2. This is compared to a table of the chi-square distribution available in many statistics texts, such as Iman and Conover (1983). We conclude that there is not enough evidence in these data to reject the assumption that fecal coliform counts are distributed similarly in all four seasons.

	Summer	Fall	Winter	Spring	
Ranks R_{ij}	6	5	2	1	
	12	8.5	4	3	
	15	11	8.5	7	
	18	14	13	10	
	21	19.5	16	17	
	<u>24</u>	<u>22</u>	<u>19.5</u>	<u>23</u>	
\bar{R}_j	16	13.3	10.5	10.2	$\bar{R}_j = 12.5$
$K=2.69$	$\chi^2_{0.95,(3)} = 7.815$	$p=0.44$	so, do not reject equality of distributions.		

7.1.1.4 The rank transform approximation

The rank transform approximation to the Kruskal-Wallis test is computed by performing a one-factor analysis of variance on the ranks R_{ij} . This approximation compares the mean rank within each group to the overall mean rank, using an F-distribution for the approximation of the distribution of K. The F and chi-square approximations will result in very similar p-values. The rank transform method should properly be called an "analysis of variance on the ranks".

For the example 1 data, the rank transform approximation results in a p-value of 0.47, essentially identical to that for the large sample approximation. Detailed computations are shown following the discussion of ANOVA in the next section.

7.1.2 Analysis of Variance (One Factor)

Analysis of variance is the parametric equivalent to the Kruskal-Wallis test. It compares the mean values of each group with the overall mean for the entire data set. If the group means are dissimilar, some of them will differ from the overall mean, as in figure 7.2. If the group means are similar, they will also be similar to the overall mean, as in figure 7.3.

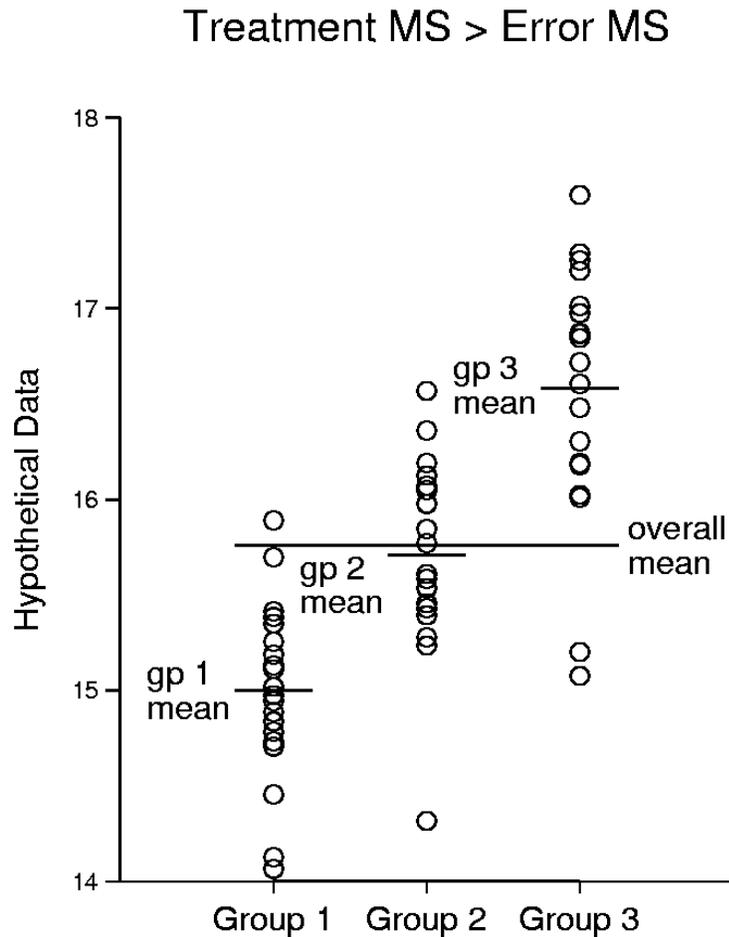


Figure 7.2 Hypothetical data for three groups.
Treatment mean square > Error mean square.

Why should a test of differences between means be named an analysis of variance? In order to determine if the differences between group means (the signal) can be seen above the variation within groups (the noise), the total noise in the data as measured by the total sum of squares is split into two parts:

$$\begin{aligned}
 \text{Total sum of squares} &= \text{Treatment sum of squares} + \text{Error sum of squares} \\
 (\text{overall variation}) &= (\text{group means} - \text{overall mean}) + (\text{variation within groups}) \\
 \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 &= \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2
 \end{aligned}$$

If the total sum of squares is divided by $N-1$, where N is the total number of observations, it equals the variance of the y_{ij} 's. Thus ANOVA partitions the variance of the data into two parts, one measuring the signal and the other the noise. These parts are then compared to determine if the means are significantly different.

7.1.2.1 Null and alternate hypotheses

The null and alternate hypotheses for the analysis of variance are:

- H₀: the k group means are identical $\mu_1 = \mu_2 = \dots = \mu_k$.
- H₁: at least one mean is different.

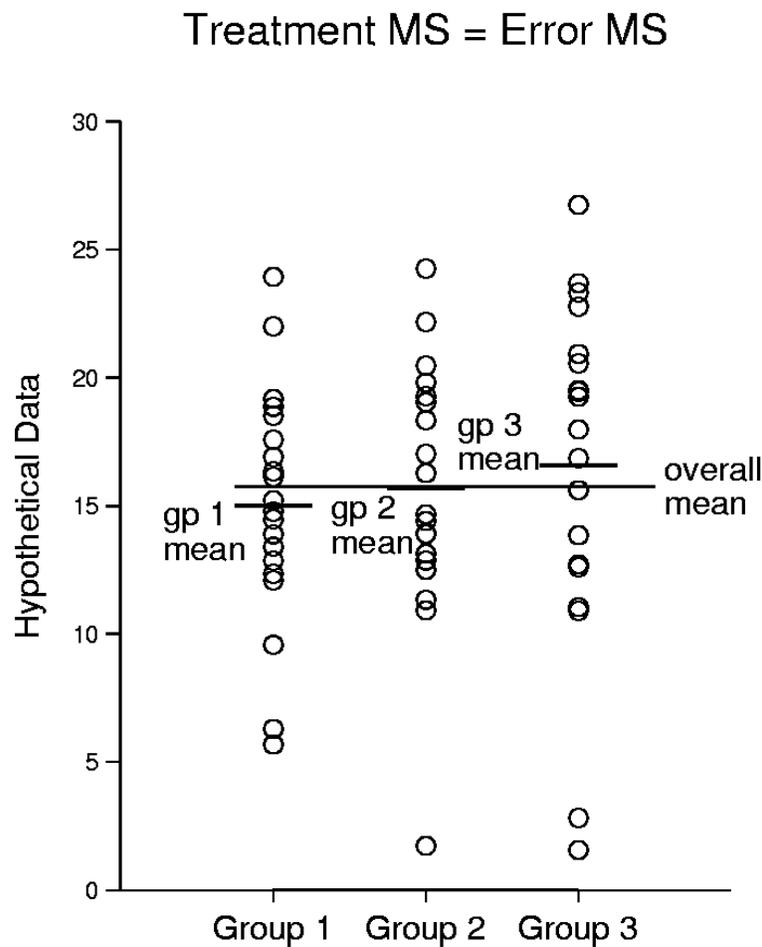


Figure 7.3 Hypothetical data for three groups. Treatment mean square \cong Error mean square.

7.1.2.2 Assumptions of the test

If ANOVA is performed on two groups, the F statistic which results will equal the square of the two-sample t-test statistic $F=t^2$, and will have the same p-value. It is not surprising, then, that the same assumptions apply to both tests:

1. All samples are random samples from their respective populations.
2. All samples are independent of one another.
3. Departures from the group mean ($y_{ij} - \bar{y}_j$) are normally distributed for all j groups.
4. All groups have equal population variance σ^2 estimated for each group by s_j^2

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n_j - 1}$$

Violation of either the normality or constant variance assumption results in a loss of ability to see differences between means (a loss of power). The analysis of variance suffers from the same five problems as did the t-test: 1) lack of power when applied to non-normal data, 2) dependence on an additive model, 3) lack of applicability for censored data, 4) assumption that the mean is a good measure of central tendency for skewed data, and 5) difficulty in assessing whether the normality and equality of variance assumptions are valid for small sample sizes. See Chapter 5 for a detailed discussion of these problems.

Difficulties arise when using prior tests of normality to "prove" non-normality before allowing use of the nonparametric Kruskal-Wallis test. Small samples sizes may inhibit detecting non-normality, as mentioned above. Second, transformations must be done on more than two groups of data. It is usually quite difficult to find a single transformation which when applied to all groups will result in each becoming normal with constant variance. Even the best transformation based on sample data may not alleviate the power loss inherent when the assumptions of ANOVA are violated. Finally, if all groups are actually from a normal distribution, one or more may be "proven" non-normal simply by chance (there is an $\alpha\%$ chance for each group). Thus the results of testing for normality can be quite inconclusive prior to performing ANOVA. The value of nonparametric approaches here is that they are relatively powerful for a wide range of situations.

7.1.2.3 Computation

Each observation y_{ij} can be written as a sum of the overall true mean μ , plus the difference α_j between μ and the true mean of the jth group μ_j , plus the difference ϵ_{ij} between the individual observation y_{ij} and the jth group mean μ_j :

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij},$$

where: y_{ij} is the i th individual observation in group j , $j=1,\dots,k$;
 μ is the overall mean (over all groups);
 α_j is the "group effect", or $(\mu_j - \mu)$, and
 ϵ_{ij} are the residuals or "error" within groups.

If H_0 is true, all j groups have the same mean equal to the overall mean μ , and thus $\alpha_j = 0$ for all j . If group means differ, $\alpha_j \neq 0$ for some j . In order to detect a difference between means, the variation within a group around its mean due to the ϵ_{ij} 's must be sufficiently small in comparison to the difference between group means so that the group means may be seen as different (see figure 7.2). The variation within a group is estimated by the within-group or error mean square (MSE), computed from the data. The variation between group means is estimated by the treatment mean square (MST). Their computation is shown below.

Sum of Squares

The error or within-group sum of squares

$$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

estimates the total within-group noise using departures from the sample group mean \bar{y}_j . Error in this context refers not to a mistake, but to the inherent variability within a group. The treatment (between-group) sum of squares

$$SST = \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2$$

estimates the treatment effect using differences between group means and the overall mean of the sample, weighted by sample size.

Degrees of freedom

Each of the sums of squares has an associated degrees of freedom, the number of independent pieces of information used to calculate the statistic. For the treatment sum of squares this equals $k-1$, as when $k-1$ of the group means are known, the k th group mean can be calculated. The total sum of squares has $N-1$ degrees of freedom, the denominator of the formula for the variance of y_{ij} . The error sum of squares has degrees of freedom equal to the difference between the above two, or $N-k$.

Mean Squares and the F-test

Dividing the sums of squares by their degrees of freedom produces the total variance, and the mean squares for treatment (MST) and error (MSE). These mean squares are also measures of the variance of the data.

<u>Mean Square</u>		<u>Formula</u>	<u>Estimates:</u>
Variance of y_{ij}	=	Total SS / N-1	Total variance of the data
MST	=	SST / k-1	Variance within groups + variance between groups.
MSE	=	SSE / N-k	Variance within groups.

If H_0 is true, there is no variance between group means (no difference between means), and the MST will on average equal the MSE (figure 7.3). As $\alpha_j = 0$, all variation is simply around the overall mean μ , and the MST and MSE both estimate the total variance. However when H_1 is true, the MST is larger on average than the MSE (figure 7.2), as most of the noise is that between groups. Therefore a test is constructed to compare these two estimates of variance, MST and MSE. The F-ratio

$$F = \text{MST} / \text{MSE}$$

is computed and compared to quantiles of an F distribution. If MST is sufficiently larger than MSE, F is large and H_0 is rejected. When H_0 is true and there is no evidence for differences in group means, F is expected to equal 1 ($\mu_F = 1$ when H_0 is true). In other words, an $F = 1$ has a p-value near 0.50, varying with the degrees of freedom. If F were below 1, which could happen due to random variation in the data, generally $p > 0.50$ and no evidence exists for differences between group means.

The computations and results of an ANOVA are usually organized into an ANOVA table. For a one-way ANOVA, the table looks like:

<u>Source</u>	<u>df</u>	<u>SS</u>	<u>MS</u>	<u>F</u>	<u>p-value</u>
Treatment	(k-1)	SST	MST	MST/MSE	p
<u>Error</u>	<u>(N-k)</u>	<u>SSE</u>	MSE		
Total	N-1	Total SS			

Example 1, cont.

For the fecal coliform data from the Illinois River, the ANOVA table is given below. The F statistic is quite small, indeed below 1. At $\alpha=0.05$ or any reasonable α -level, the mean counts would therefore not be considered different between seasons.

<u>Source</u>	<u>df</u>	<u>SS</u>	<u>MS</u>	<u>F</u>	<u>p-value</u>
Season	3	361397	120466	0.67	0.58
<u>Error</u>	<u>20</u>	<u>3593088</u>	179654		
Total	23	3954485			

However, this ANOVA has been conducted on non-normal data. Without knowing the results of the Kruskal-Wallis test, concern should be expressed that the result of "no difference" may be an artifact of the lack of power of the ANOVA, and not of a true equivalence of means. Some

statisticians have recommended performing both tests. This may be unnecessary if the data exhibit sufficient non-normality to suspect an inability of ANOVA to reject. Also assumed by performing ANOVA is that group means are an appropriate data summary. For the obviously skewed distributions found for all but the winter season, means will make little sense as estimates of the values which might be expected to occur. Means would be useful when estimating the mass of bacteria transported per season, but not in the hypothesis testing realm.

One factor analysis of variance	
Situation	Several groups of data are to be compared, to determine if their means are significantly different. Each group is assumed to have a normal distribution around its mean. All groups have the same variance.
Computation	<p>The treatment mean square and error mean square are computed as their sum of squares divided by their degrees of freedom (df). When the treatment mean square is larger than the error mean square as measured by an F-test, the group means are significantly different.</p> $MST = \frac{\sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2}{k - 1}$ <p style="text-align: right;">where $k-1$ = treatment degrees of freedom</p> $MSE = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{N - k}$ <p style="text-align: right;">where $N-k$ = error degrees of freedom</p>
Tied data	No alterations necessary.
Test Statistic	<p>The test statistic F:</p> $F = MST / MSE$
Decision Rule	<p>To reject H_0: the mean of every group is identical, versus H_1: at least one mean differs .</p> <p>Reject H_0 if $F \geq F_{1-\alpha, k-1, N-k}$ the $1-\alpha$ quantile of an F distribution with $k-1$ and $N-k$ degrees of freedom; otherwise do not reject H_0.</p>

7.2 Tests For The Effects of More Than One Factor

It is quite common that more than one factor is suspected to be influencing the magnitudes of observations. In these situations it is desirable to measure the influence of all factors simultaneously. Sequential one-factor tests are an inadequate alternative to a single multi-factor

test. Even when only one factor is actually influencing the data and a one-way ANOVA for that factor soundly rejects H_0 , a second one-way test for a related factor may erroneously reject H_0 simply due to the association between the two factors. The test for the second factor should remove the effect of the first before establishing that the second has any influence. By evaluating all factors simultaneously, the influence of one can be measured while compensating for the others. This is the objective of a multi-factor analysis of variance, and of the nonparametric analogue.

7.2.1 Nonparametric Multi-Factor Tests

For two-factor and more complex ANOVA's where the data within one or more treatment groups are not normally distributed and may not have equal variances, there are two possible approaches for analysis. The first is a class of tests which include the Kruskal-Wallis and Friedman tests as simpler cases. These tests, described by Groggel and Skillings (1986), do not allow for interactions between factors. The tests reformat multiple factors into two factors, one the factor being tested, and the other the collection of all other treatment groups for all remaining factors. The data are then ranked within treatment groups for analysis, much as in a Friedman test. The reader is referred to their paper for more detail.

The second procedure is a rank transformation test (Conover and Iman, 1981). All data are ranked from 1 to N, and an ANOVA computed on the ranks. This procedure is far more robust to departures from the assumptions of normality and constant variance than is an ANOVA on the original data. The rank transformation produces values which are much closer to meeting the two critical assumptions than are the original values themselves. The tests determine whether the mean rank differs between treatment groups, rather than the mean. The mean rank is interpreted as an estimate of the median. Multiple comparison procedures on the ranks can then differentiate which groups differ from others.

Examples of the computation and performance of these rank transformation tests will be delayed until after discussion of parametric factorial ANOVA.

7.2.2 Multi-Factor Analysis of Variance -- Factorial ANOVA

The effects of two or more factors may be simultaneously evaluated using a factorial ANOVA design. A factorial ANOVA occurs when none of the factors is a subset of the others. If subsetted factors do occur, the design includes "nested" factors and the equations for computing the F test statistics will differ from those here (nested ANOVA is briefly introduced in a later section). A two-factor ANOVA will be fully described -- more than two factors can be incorporated, but are beyond the scope of this book. See Neter, Wasserman and Kutner (1985) for more detail on higher-way and nested analysis of variance.

For a two-factor ANOVA, the influences of two explanatory variables are simultaneously tested. The first page of this chapter presented a two-factor ANOVA, the determination of chemical concentrations among basins at low flow. The objective was to determine whether concentrations differed as a function of mining history (whether or not each basin was mined, and if so whether it was reclaimed) and of rock type.

7.2.2.1 Null and alternate hypotheses

Call the two factors A and B. There are $i=1, \dots, a \geq 2$ categories of factor A, and $j=1, \dots, b \geq 2$ categories of factor B. Treatment groups are defined as all the possible combinations of factors A and B, so there are $a \cdot b$ treatment groups. Within each treatment group there are n_{ij} observations. The test determines whether mean concentrations are identical among all the $a \cdot b$ treatment groups, or whether at least one differs.

$$H_0 : \text{all } a \cdot b \text{ treatment group means } \mu_{ij} \text{ are equal} \quad \mu_{11} = \mu_{12} = \dots = \mu_{ab}$$

$$H_1 : \text{at least one } \mu_{ij} \text{ differs from the rest.}$$

The magnitude of any observation y_{ijk} can be affected by several possible influences:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}, \text{ where}$$

- α_i = influence of the i th category of A
- β_j = influence of the j th category of B
- $\alpha\beta_{ij}$ = interaction effect between A and B beyond those of α_i and β_j separately for the ij th treatment group, and
- ε_{ijk} = residual error, the difference between the k th observation ($k=1, \dots, n_{ij}$) and the treatment group mean μ_{ij} .

The null hypothesis states that treatment group means μ_{ij} all equal the overall mean μ .

Therefore α_i , β_j , and $\alpha\beta_{ij}$ all equal 0 -- there are no effects due to any of the factors or to their interaction. If any one of α_i , β_j , or $\alpha\beta_{ij}$ are nonzero, the null hypothesis is rejected, and at least one treatment group evidences a difference in its mean.

7.2.2.2 Interaction between factors

If $\alpha\beta_{ij} = 0$ in the equation above, there is no interaction present. Without interaction, the effect of factor B is identical for all groups of factor A, and the effect of factor A is identical for all groups of factor B. Suppose there are 3 groups of factor A (a_1 , a_2 , and a_3) and 2 groups of factor B (b_1 and b_2), resulting in six treatment groups overall. Lack of interaction can be visualized by plotting the means for all treatment groups as in figure 7.4. The parallelism of the lines shows that no interaction is present. The effect of A going from a_1 to a_2 to a_3 is identical regardless of which B group is involved. The increase going from b_1 to b_2 for factor B is identical for every group of factor A.

When interaction is present ($\alpha\beta_{ij} \neq 0$) the treatment group means are not determined solely by the additive effects of factors A and B alone. Some of the groups will have mean values larger

or smaller than those expected just from the results of the individual factors. The effect of factor A can no longer be discussed without reference to which group of factor B is of interest, and the effect of factor B can likewise not be stated apart from a knowledge of the group of factor A. In a plot of the treatment group means, the lines are no longer parallel (figure 7.5). The pattern of differences going from a1 to a2 to a3 depends on which group of factor B is of interest, and likewise for the differences between b1 and b2 -- the pattern differs for the three A groups.

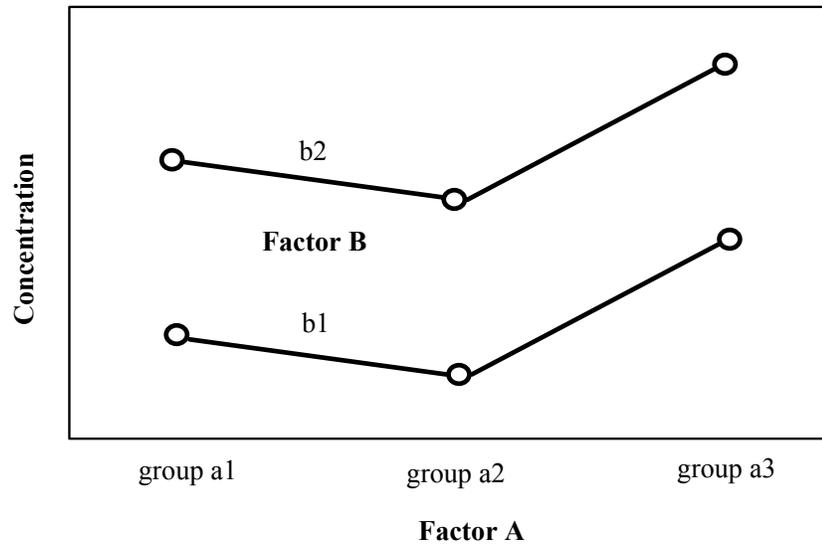


Figure 7.4 Six treatment group means with no interaction present

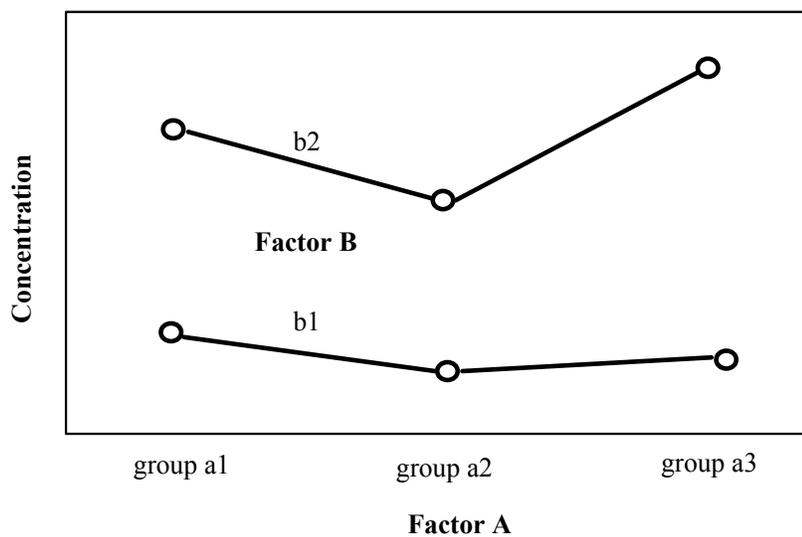


Figure 7.5 Six treatment group means with interaction present

Interaction can result from a synergistic or antagonistic effect. As an example, fish may not die instream due only to higher water temperatures, or to slightly higher copper concentrations, but combine the two and the result could be deadly. This type of interrelation between factors results in a significant interaction effect. For k factors there are $(k-1)$ possible interaction terms between the factors. Unless it is known ahead of time that interactions are not possible, interaction terms should always be included and tested for in multi-factor ANOVA models.

7.2.2.3 Assumptions for factorial ANOVA

Assumptions are the same as for a one-way ANOVA. Departures from each treatment group mean μ_{ij} (every combination of factors A and B) are assumed normally distributed with identical variance. This is a consequence of the ϵ_{ij} , which are normally distributed and of variance σ^2 , being randomly distributed among the treatment groups. The normality and constant variance assumptions can be checked by inspecting boxplots of the data for each treatment group.

7.2.2.4 Computation

The influences of factors A, B, and their interaction are evaluated separately by again partitioning the total sums of squares into component parts due to each factor. After dividing by their respective degrees of freedom, the mean squares for factors A, B, and interaction are produced. As with a one-way ANOVA, these are compared to the error mean square (MSE) using F-tests to determine their significance.

Sum of Squares

The equations for the sums of squares for factor A (SSA), factor B (SSB), interaction (SSI), and error, assuming constant sample size n per treatment group, are:

$SSA = \sum^a \frac{(\sum^b \sum^n y)^2}{bn} - \frac{(\sum^a \sum^b \sum^n y)^2}{abn}$	<p><u>due to</u></p> <p>$\mu_i - \mu$</p>
$SSB = \sum^b \frac{(\sum^a \sum^n y)^2}{an} - \frac{(\sum^a \sum^b \sum^n y)^2}{abn}$	<p>$\mu_j - \mu$</p>
$SSI = \text{Total SS} - SSA - SSB - SSE$	<p>$\mu_{ij} - (\mu_i + \mu_j) + \mu$</p>
$SSE = \sum^a \sum^b \sum^n (y)^2 - \sum^a \sum^b \frac{(\sum^n y)^2}{n}$	<p>$y_{ijk} - \mu_{ij}$</p>
$\text{Total SS} = \sum^a \sum^b \sum^n (y)^2 - \frac{(\sum^a \sum^b \sum^n y)^2}{abn}$	<p>$y_{ijk} - \mu$</p>

Mean Squares and the F-test

Dividing the sums of squares by their degrees of freedom produces the mean squares for factors A, B, interaction, and error as in the ANOVA table below. If H_0 is true and α_i , β_j , and $\alpha\beta_{ij}$ all equal 0, all variation is simply around the overall mean μ . The MSA, MSB, and MSI will then all be measures of the error variance, as is the MSE, and all three F-tests will have ratios not far from 1. However when H_1 is true, at least one of the mean squares in the numerators should be larger than the MSE, and the resulting F-ratio will be larger than the appropriate quantile of the F distribution. When F is large, H_0 can be rejected, and that influence be considered to significantly affect the magnitudes of the data at a level of risk equal to α .

The two-factor ANOVA table is as follows when there is an equal number of observations for each treatment (all $n_{ij} = n$).

Source	df	SS	MS	F	p-value
Factor A	(a-1)	SSA	SSA/(a-1)	MSA/MSE	
Factor B	(b-1)	SSB	SSB/(b-1)	MSB/MSE	
Interaction	(a-1)(b-1)	SSI	SSI/(a-1)(b-1)	MSI/MSE	
Error	<u>ab(n-1)</u>	<u>SSE</u>	SSE/[ab(n-1)]		
Total	abn-1	Total SS			

Multi-factor analysis of variance													
Situation	Two or more influences are to be simultaneously tested, to determine if either cause significant differences between treatment group means. Each group is assumed to have a normal distribution around its mean. All groups have the same variance.												
Computation	Compute the sums of squares and mean squares as above.												
Tied data	No alterations necessary.												
Test Statistic	<table style="width: 100%; border: none;"> <tr> <td style="width: 33%;">To test factor A:</td> <td style="width: 33%;">To test factor B:</td> <td style="width: 33%;">To test for interaction:</td> </tr> <tr> <td>$F_A = MSA / MSE$</td> <td>$F_B = MSB / MSE$</td> <td>$F_I = MSI / MSE$</td> </tr> <tr> <td colspan="3" style="text-align: center;">with degrees of freedom for the numerator of:</td> </tr> <tr> <td>dfn = (a-1)</td> <td>dfn = (b-1)</td> <td>dfn = (a-1)(b-1)</td> </tr> </table>	To test factor A:	To test factor B:	To test for interaction:	$F_A = MSA / MSE$	$F_B = MSB / MSE$	$F_I = MSI / MSE$	with degrees of freedom for the numerator of:			dfn = (a-1)	dfn = (b-1)	dfn = (a-1)(b-1)
To test factor A:	To test factor B:	To test for interaction:											
$F_A = MSA / MSE$	$F_B = MSB / MSE$	$F_I = MSI / MSE$											
with degrees of freedom for the numerator of:													
dfn = (a-1)	dfn = (b-1)	dfn = (a-1)(b-1)											
Decision Rule	<p>To reject H_0: the mean of every group is identical (no treatment effects for either factor or interaction), versus</p> <p>H_1: at least one mean differs.</p> <p>Reject H_0 if $F \geq F_{1-\alpha, \text{dfn}, ab(n-1)}$ the $1-\alpha$ quantile of an F distribution with dfn and $ab(n-1)$ degrees of freedom; otherwise do not reject H_0.</p>												

Example 2

Iron concentrations were measured at low flow in numerous small streams in the coal-producing areas of eastern Ohio (Helsel, 1983). Each stream drains either an unmined area, a reclaimed coal mine, or an abandoned coal mine. Each site is also underlain by either a sandstone or limestone formation. Are iron concentrations influenced by upstream mining history, by the underlying rock type, or by both?

There are several scenarios which would cause H_0 to be rejected. Factor A (say mining history) could be significant ($\alpha_i \neq 0$), but factor B insignificant. Or factor B (rock type) could be significant ($\beta_j \neq 0$), but not A. Both factors could be significant ($\alpha_i, \beta_j \neq 0$). Both factors could be significant, plus an additional interaction effect because one or more treatment groups (say unreclaimed sandstone basins) exhibited much different iron concentrations than those expected from either influence alone ($\alpha_i, \beta_j, \alpha\beta_{ij} \neq 0$). Finally, both factor A and B could be not significant ($\alpha_i, \beta_j = 0$) but concentrations be elevated for one specific treatment group ($\alpha\beta_{ij} \neq 0$). This would be interpreted as no overall mining or rock type effect, but one combination of mining history and rock type would have differing mean concentrations.

Boxplots for a subset of the iron concentration data from Helsel (1983) are presented in figure 7.6. Note the skewness, as well as the differences in variance as depicted by differing box heights. A random subset was taken in order to produce equal sample sizes per treatment group, yet preserving the essential data characteristics. The subset data are listed in Appendix C5. In the section 7.2.2.5, analysis of unequal sample sizes per treatment group will be presented and the entire iron data set analyzed.

There are six treatment groups, combining the three possible mining histories (unmined, abandoned mine, and reclaimed mine) and the two possible rock types (sandstone and limestone). An analysis of variance conducted on this subset which has $n=13$ observations per treatment group produced the following ANOVA table. Tested was the effect of mining history alone, rock type alone, and their interaction (Mine*Rock). A*B is a common abbreviation for the interaction between A and B.

ANOVA table for the subset of iron data

Source	df	SS	MS	F	p-value
Rock	1	15411	15411	2.38	0.127
Mine	2	32282	16141	2.49	0.090
Rock*Mine	2	25869	12934	2.00	0.143
<u>Error</u>	<u>72</u>	<u>466238</u>	6476		
Total	77	539801			

None of the three possible influences is significant at the $\alpha = 0.05$ level, as their p-values are all larger than 0.05. However, the gross violation of the test's assumptions of normality and equal

variance shown in the boxplots must be considered. Perhaps the failure to reject H_0 is due not to a lack of an influence evidenced in the data, but of the parametric test's lack of power to detect these influences because of the violation of test assumptions. To determine whether this is so, the equivalent rank transformation test is performed.

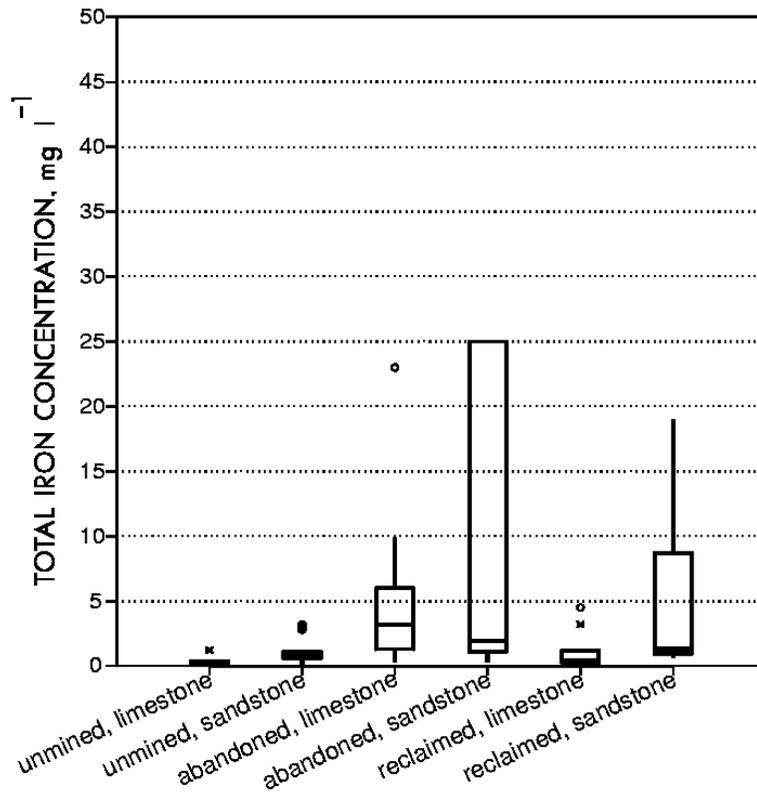


Figure 7.6 A subset of the iron concentrations at low flow from Helsel (1983)

To compute the rank transformation test, the data are ranked from smallest to largest, 1 to $n=78$. An analysis of variance is then performed on the ranks of the data. The ANOVA table is below, while a boxplot of data ranks is shown in figure 7.7.

ANOVA table for the ranks of the subset of iron data

Source	df	SS	MS	F	p-value
Rock	1	4121.7	4121.7	13.38	0.000
Mine	2	10933.9	5467.0	17.74	0.000
Rock*Mine	2	2286.2	1143.1	3.71	0.029
<u>Error</u>	<u>72</u>	<u>22187.2</u>	308.2		
Total	77	39529.0			

Results for the rank transformation tests are startlingly different than those for the parametric ANOVA. All three influences, mining history, rock type, and their interaction, are significant at $\alpha = 0.05$. Gross violations of the assumptions of ANOVA by these data have clearly inhibited the parametric test from detecting the influences of these factors. The rejection of H_0 for the rank test indicates that the median iron concentrations differ between treatment groups. Mean concentrations will be distorted by the skewness and outliers present in most of the treatment groups.

Analysis of variance on data ranks is an "asymptotically distribution-free" technique. That is, for sufficiently large sample sizes it tests hypotheses which do not require the assumption of data normality. For the cases where equivalent, truly nonparametric techniques exist such as the Kruskal-Wallis and Friedman tests, the rank transformation procedures have been shown to be large-sample approximations to the test statistics for those techniques. Where no equivalent nonparametric methods have yet been developed such as for the two-way design, rank transformation results in tests which are more robust to non-normality, and resistant to outliers and non-constant variance, than is ANOVA without the transformation.

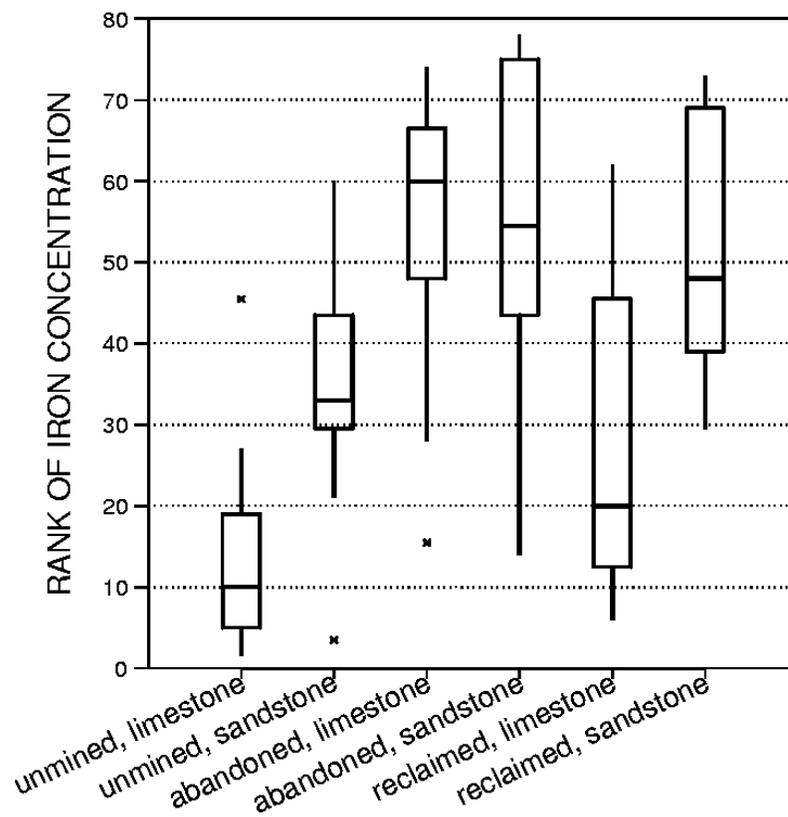


Figure 7.7 Boxplots of the ranks of the iron data shown in Figure 7.6

A third option for analysis of the two-way design is ANOVA on data transformed by a power transformation. The purpose of the power transformation is to produce a more nearly-normal

and constant variance data set. As water resources data are usually positively skewed, the log transformation is often employed. Using logarithms for ANOVA implies that the influences of each factor are multiplicative in the original units, as the influences of the logarithms are additive. The primary difficulty in using a power transformation is in producing a normally distributed error structure for every treatment group. Groups which are skewed may be greatly aided by a transformation, but be side-by-side with a group which was symmetric in the original units, and is now asymmetric after transformation! Boxplots for each treatment group should be inspected prior to performing the ANOVA to determine if each group is at least symmetric. When only some of the treatment groups exhibit symmetry, much less normality, concerns over the power of the procedure remain. F tests which appear to be not significant are always suspect.

In figure 7.8, boxplots of the base 10 logarithms of the low-flow iron concentrations are presented. Most of the treatment groups still remain distinctly right-skewed even after the transformation, while the unmined limestone group appears less symmetric following transformation! There is nothing magic in the log transformation -- any other transformation going down the ladder of powers might also remedy positive skewness. It may also alter a symmetric group into one that is left-skewed. The search for a transformation which results in all groups being symmetric is often fruitless. In

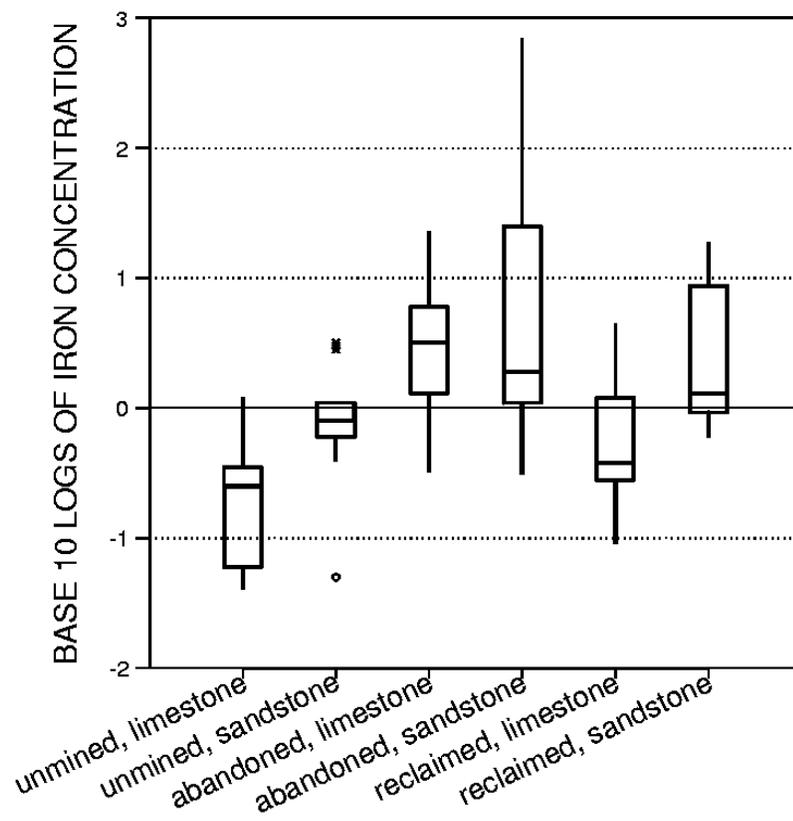


Figure 7.8 Boxplots of the base 10 logarithms of the iron data shown in Figure 7.6

addition, the "best" power transformation will likely change going from one data set to another, one location to another, and one time period to another. In comparison, the rank transformation has simplicity, comparability among locations and time periods, and general validity as being asymptotically distribution-free. When the assumptions of normality and constant variance are questionable, the rank transformation is the most generally appropriate alternative.

7.2.2.5 Unequal sample sizes

Equations presented in the previous section are appropriate only when the number of data per treatment group is identical for each group. This is also called a "balanced" design. Computations for unequal sample sizes ("unbalanced" designs) are more complex. Smaller statistics software packages often encode tests valid only for balanced designs, though that is not always obvious from their output. Yet water resources data rarely involve situations when all sample sizes are equal. Sample bottles are broken, floods disrupt the schedule, etc. When data are unbalanced, the sums of squares for the above equations no longer test

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

but test instead an hypothesis involving weighted group means, where the weights are a function of treatment group sample sizes. This is of little use to the practitioner. Some software will output the (useless and incorrect) results valid only for equal sample sizes even when unbalanced data are provided as input, with no warnings of their invalidity. Be sure that when unequal sample sizes occur, tests which can incorporate them are performed.

To perform ANOVA on unbalanced data, a regression approach is necessary. This is done on larger statistical packages such as Minitab or SAS. SAS's "type I" sums of squares (called "sequential sums of squares" by Minitab) are valid only for balanced cases, but SAS's "type III" sums of squares (Minitab's "adjusted sums of squares") are valid for unbalanced cases as well. Unbalanced ANOVAs are computed in the same fashion as nested F-tests for comparing regression models in analysis of covariance, discussed in Chapter 11. Because the equations for the sums of squares are "adjusted" for unequal sample sizes, they do not sum to the total sum of squares as for balanced ANOVA. See Neter, Wasserman and Kutner (1985) for more detail on the use of regression models for performing unbalanced ANOVA.

Example 2, continued

The complete 241 observations (Appendix C6) from Helsel (1983) are analyzed with an unbalanced ANOVA. Boxplots for the six treatment groups are shown in figure 7.9. They are quite similar to those in figure 7.6, showing that the subsets adequately represented all the data. An ANOVA table for the complete iron data set is as follows. Note that the sums of squares do not add together to equal the total sum of squares for this unbalanced ANOVA. Results for these data would be incorrect if performed by software capable only of balanced ANOVA. Conclusions reached (do not reject for all tests) agree with those previously given for ANOVA on the data subset.

ANOVA table for the complete (unbalanced) iron data

Source	df	SS	MS	F	p-value
Rock	1	71409	71409	0.51	0.476
Mine	2	262321	131160	0.93	0.394
Rock*Mine	2	178520	89260	0.64	0.530
<u>Error</u>	<u>235</u>	32978056	140332		
Total	240	34062640			

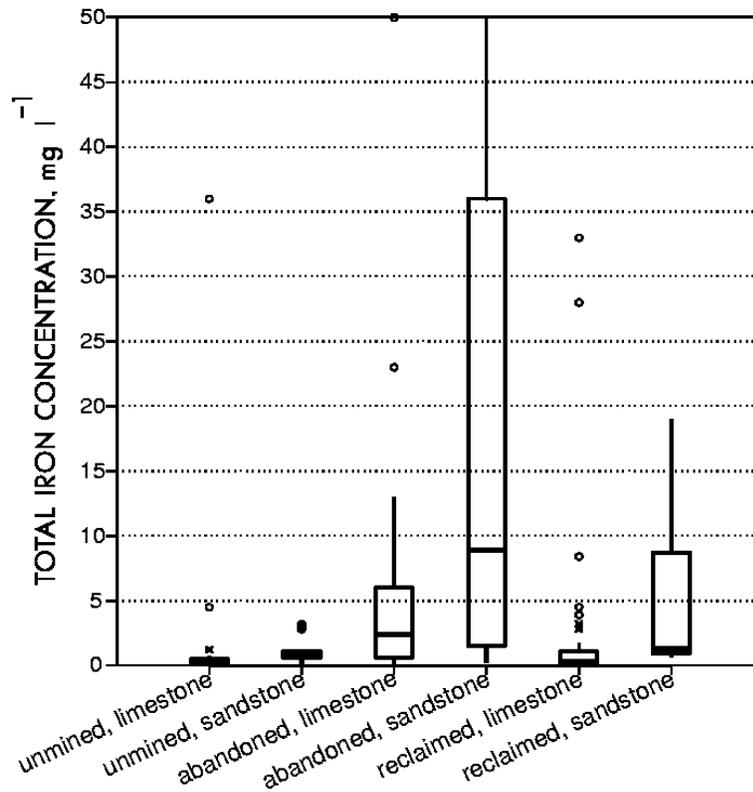


Figure 7.9 Iron concentrations at low flow from Helsel (1983)

7.2.2.6 Fixed and random factors

An additional requirement for the F tests previously given is that both factors are fixed. With a fixed factor, the inferences to be made from the results extend only to the treatment groups under study. For example, the influences of unmined, abandoned, and reclaimed mining histories were previously compared. Differences in resulting chemical concentrations between these three specific mining histories are of interest, and hence this is a fixed factor. A random factor would result from a random selection of several groups out of a larger possible set to represent the overall factor. Inferences from the test results would be extended beyond the specific groups being tested to the generic factor itself. Thus there is little or no interest in attributing test results to a specific individual group, but only in ascertaining a generic effect due to that factor.

As an example, suppose soil concentrations of a trace metal are to be compared between three particle size fractions all across the state, to determine which of the three fractions is most appropriate as a reconnaissance medium. Particle size is a fixed effect -- there is interest in those specific sizes. However, there is only enough funding to sample sparsely if done all across the state, so instead a random factor is incorporated to determine whether spatial differences occur. Several counties are selected at random, and intensive sampling occurs within those counties. No sampling is done outside of those counties. The investigator will determine not only which size fraction is best, but whether this is consistent among the counties (the random effect), which by inference is extended to the entire state. There is no specific interest in the counties selected, but only as they represent spatial variability.

If every factor were random, F tests would use the mean squares for interaction as denominators rather than the mean square for error. If a mix of random and fixed factors occurs (called a "mixed effects" design) as in the example above, there would be a mixture of mean squares used as denominators. In general the fixed factors in the design use the interaction mean squares as denominators, and the random factors the error mean square, the reverse of what one might intuitively expect! However, the structure of mixed effects F tests can get much more complicated, especially for more than two factors, and texts such as Neter, Wasserman and Kutner (1985) or Sokal and Rohlf (1981) should be consulted for the correct setup of F tests when random factors are present. Note that computer software uses the MSE in the denominator unless otherwise specified, and thus assumes that all factors are fixed. Therefore F tests automatically produced will not be correct when random factors are present, and the correct F ratio must be specifically requested and computed.

7.3 Blocking -- The Extension of Matched-Pair Tests

In Chapter 6, tests for differences between matched-pairs of observations were discussed. Each pair of observations had one value in each of two groups, such as "before" versus "after". The advantage of this type of design is that it "blocks out" the differences from one matched-pair to another that is contributing unwanted noise. Such noise may mask the differences between the two groups (the treatment effect being tested) unless matched-pairs are used.

Similar matching schemes can be extended to test more than two treatment groups. Background noise is eliminated by applying the treatment to blocks (rather than pairs) of similar or identical individuals. Only one observation is usually available for each combination of treatment and block. This is called a "randomized complete block design", and is a common design in the statistical literature.

The third example at the beginning of this chapter, detecting differences between three extraction methods used at numerous wells, is an example of this design. The treatment effect is

the extraction method, of which there are three types (three groups). The blocking effect is the well location; the well-to-well differences are to be "blocked out". One sample is analyzed for each extraction method at each well.

Four methods for analysis of a randomized complete block design will be presented. Each of them attempts to measure the same influences. To do this, each observation y_{ij} is broken down into the effects of four influences:

$$y_{ij} = \mu + \alpha_j + \beta_i + \epsilon_{ij},$$

where

- y_{ij} is the individual observation in block i and group j ;
- μ is the overall mean or median (over all groups),
- α_j is the " j th group effect", $j=1,k$
- β_i is the " i th block effect", $i=1,n$
- ϵ_{ij} is the residual or "error" between the individual observation and the combined group and block effects.

Median polish provides resistant estimates of the overall median, of group effects and block effects. It is an exploratory technique, not an hypothesis test procedure. Related graphical tools determine whether the two effects are additive or not, and whether the ϵ_{ij} are normal, as assumed by an ANOVA. If not, a transformation should be employed to achieve additivity and normality before an ANOVA is performed. The Friedman and median aligned ranks tests are nonparametric alternatives for testing whether the treatment effect is significant in the presence of blocking.

7.3.1 Median Polish

Median polish (Hoaglin et al., 1983) is an iterative process which provides a resistant estimate m of the overall median μ , as well as estimates a_j of the group effects α_j and b_i of the block effects β_i . Its usefulness lies in its resistance to effects of outliers. The polishing is begun by subtracting the medians of each row from the data table, leaving the residuals. The median of these row medians is then computed as the first estimate of the overall median, and subtracted from the row medians. The row medians are now the first estimates of the row effects. Then the median of each column is subtracted from the residual data table and set aside. The median of the column medians is subtracted from the column medians, and added to the overall median. The column medians now become the first estimates of the column effects. The entire process is repeated a second time, producing an estimated overall median m , row and column departures from the overall median (estimates a_j and b_i), and a data table of residuals e_{ij} estimating the ϵ_{ij} .

Example 3

Mercury concentrations were measured in periphyton at six stations along the South River, Virginia, above and below a large mercury contamination site (Walpole and Myers, 1985). Measurements were made on six different dates. Of interest is whether the six stations differ in mercury concentration. Is this a one-way ANOVA setup? No, because there may be

differences among the six dates -- the periphyton may not take up mercury as quickly during some seasons as others, etc. Differences caused by sampling on six different dates are unwanted noise which should be blocked out, hence date is a blocking effect. The data are presented in table 7.3, and boxplots by station in figure 7.10. There appears to be a strong increase in mercury concentration going downstream from station 1 to station 6, reflecting an input of mercury along the way.

Station:	1	2	3	4	5	6
<u>Date</u>						
1	0.45	3.24	1.33	2.04	3.93	5.93
2	0.10	0.10	0.99	4.31	9.92	6.49
3	0.25	0.25	1.65	3.13	7.39	4.43
4	0.09	0.06	0.92	3.66	7.88	6.24
5	0.15	0.16	2.17	3.50	8.82	5.39
6	0.17	0.39	4.30	2.91	5.50	4.29

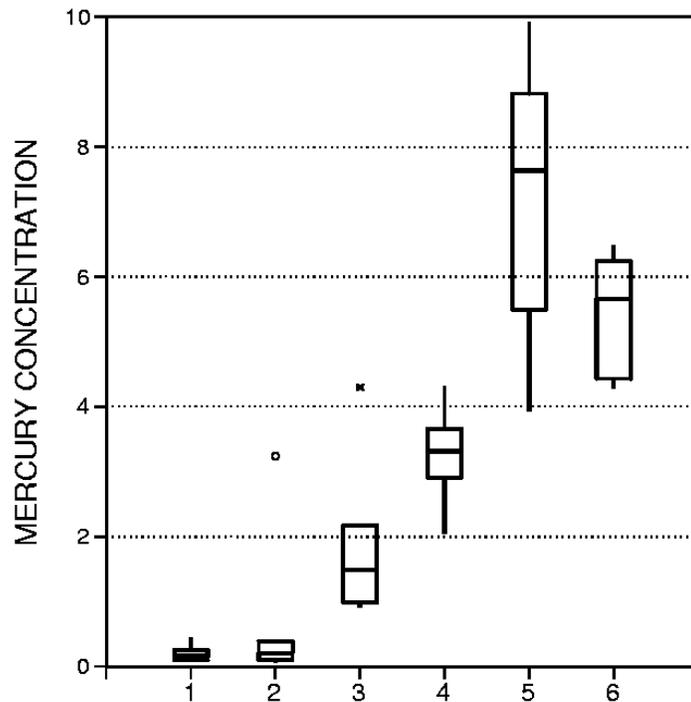


Figure 7.10 Periphyton Mercury Upstream (1) to Downstream (6) of Input to River

The first step in median polish is to compute the median of each row (date), and subtract it from that row's data. The residuals remain in the table.

Station:	1	2	3	4	5	6	row med
<u>Date</u>							(b_i)
1	-2.190	0.600	-1.310	-0.600	1.290	3.290	2.64
2	-2.550	-2.550	-1.660	1.660	7.270	3.840	2.65
3	-2.140	-2.140	-0.740	0.740	5.000	2.040	2.39
4	-2.200	-2.230	-1.370	1.370	5.590	3.950	2.29
5	-2.685	-2.675	-0.665	0.665	5.985	2.555	2.84
6	-3.430	-3.210	0.700	-0.690	1.900	0.690	3.60

Next the median of the row medians (2.64) is computed as the first estimate of the overall median m . This is subtracted from each of the row medians:

Station:	1	2	3	4	5	6	row med
<u>Date</u>							(b_i)
1	-2.19	0.60	-1.31	-0.60	1.29	3.29	0.00
2	-2.55	-2.55	-1.66	1.66	7.27	3.84	0.01
3	-2.14	-2.14	-0.74	0.74	5.00	2.04	-0.25
4	-2.20	-2.23	-1.37	1.37	5.59	3.95	-0.35
5	-2.69	-2.68	-0.67	0.67	5.99	2.56	0.20
6	-3.43	-3.21	0.70	-0.69	1.90	0.69	0.96
							m=2.64

The median of each column (station) is then computed and subtracted from that column's data. The residuals from the subtractions remain in the table.

Station:	1	2	3	4	5	6	row med
<u>Date</u>							(b_i)
1	0.19	2.99	-0.29	-1.31	-4.01	0.37	0.00
2	-0.17	-0.16	-0.64	0.95	1.97	0.92	0.01
3	0.24	0.25	0.28	0.03	-0.30	-0.88	-0.25
4	0.18	0.16	-0.35	0.66	0.29	1.03	-0.35
5	-0.31	-0.29	0.35	-0.04	0.69	-0.36	0.20
6	-1.05	-0.82	1.72	-1.40	-3.40	-2.23	0.96
a _j col med:	-2.38	-2.39	-1.02	0.71	5.30	2.92	m=2.64

Then the median of the column medians (-0.16) is subtracted from each of the column medians, and added to the overall median:

Station:	1	2	3	4	5	6	row med
<u>Date</u>							(b_i)
1	0.19	2.99	-0.29	-1.31	-4.01	0.37	0.00
2	-0.17	-0.16	-0.64	0.95	1.97	0.92	0.01
3	0.24	0.25	0.28	0.03	-0.30	-0.88	-0.25
4	0.18	0.16	-0.35	0.66	0.29	1.03	-0.35
5	-0.31	-0.29	0.35	-0.04	0.69	-0.36	0.20
6	-1.05	-0.82	1.72	-1.40	-3.40	-2.23	0.96
a_j col med:	-2.22	-2.23	-0.86	0.87	5.46	3.08	$m=2.48$

This table now exhibits the first "polish" of the data. Usually two complete polishes are performed in order to produce more stable estimates of the overall median and row and column effects. For the second polish, the above process is repeated on the table of residuals from the first polish. After a second complete polish, little change in the estimates is expected from further polishing. The table then looks like:

Station:	1	2	3	4	5	6	row med
<u>Date</u>							(b_i)
1	0.22	3.02	-0.19	-1.26	-3.77	0.31	0.03
2	-0.57	-0.56	-0.97	0.57	1.78	0.43	0.47
3	0.08	0.09	0.19	-0.11	-0.24	-1.12	-0.03
4	-0.08	-0.09	-0.54	0.42	0.24	0.69	-0.03
5	-0.17	-0.14	0.56	0.11	1.04	-0.31	0.12
6	0.15	0.38	2.99	-0.18	-1.98	-1.11	-0.18
a_j col med:	-2.18	-2.19	-0.89	0.89	5.29	3.20	$m=2.38$

The above table shows that

- 1) The station effects are large in comparison to the date effects (the a_j are much larger in absolute magnitude than the b_i).
- 2) There is a clear progression from smaller to larger values going downstream (a_j generally increases from stations 1 to 6), with the maximum at station 5.
- 3) A large residual occurs for station 5 at date 1 (smaller concentration than expected).

7.3.1.1 Plots related to median polish for checking assumptions

Median polish can be used to check the assumptions behind an analysis of variance. The first assumption is that the residuals ϵ_{ij} are normally distributed. Boxplots of the residuals e_{ij} in the table provide a look at the distribution of errors after the treatment and block effects have been removed. Figure 7.11 shows that for the periphyton mercury data the residuals are probably not normal due to the large proportion of outliers, but at least are relatively symmetric:

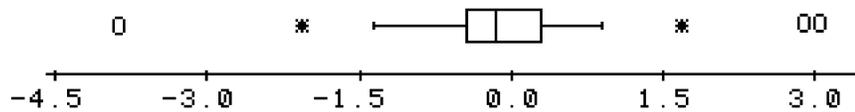


Figure 7.11 Residuals from the median smooth of periphyton mercury data

In addition, the additivity of the table can be checked. An ANOVA assumes that the treatment and block effects are additive. In other words, if being in group 1 adds -2.18 units of concentration to the overall mean or median, and if being at time 1 adds 0.03 units, these add together for treatment group 1 at time 1. If this is not the case, a transformation of the data prior to ANOVA must be performed to produce additivity. To check additivity, the "comparison value" c_{ij} (Hoaglin et al., 1983) is computed for each combination ij of block and treatment group, where

$$c_{ij} = a_i \cdot b_j / m.$$

A residuals plot of the tabled residuals e_{ij} versus c_{ij} will appear to have a random scatter around 0 if the data are additive. If not, the pattern of residuals will lead to an appropriate transformation to additivity -- for a nonzero slope s , the data should be raised to the $(1-s)$ power in the ladder of powers. In figure 7.12, a residuals plot for the mercury median polish indicate no clear nonzero slope (most of the data are clustered in a central cloud), and therefore no transformation is necessary.

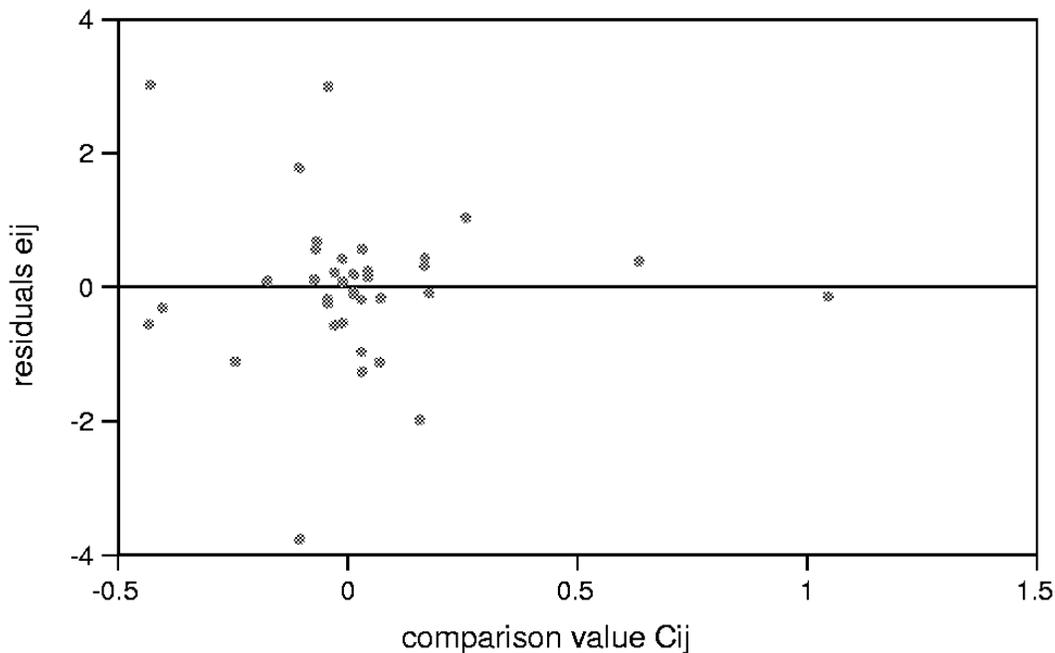


Figure 7.12 Median polish residuals plot showing random scatter around $e_{ij}=0$

7.3.2 The Friedman Test

The Friedman test is the most common nonparametric test used for the randomized complete block design. It computes the ranks of the data only within each block, not making cross-comparisons between blocks. Treatment effects are determined from the within-block ranks each treatment has received. The Friedman test is an extension of the sign test, and reduces to the sign test when comparing only two treatment groups. Its advantages and disadvantages in comparison to the analysis of variance are the same as that of the sign test to the t-test. When the errors ϵ_{ij} can be considered normal, the ANOVA should be preferred. For the many situations where the errors are not normal, the Friedman test will generally have equal or greater power to detect differences between treatment groups, and should be performed. The Friedman test is especially useful when the data can be ranked but differences between observations cannot be computed, such as when comparing a <1 to a 5.

7.3.2.1 Null and alternate hypotheses

The Friedman test is used to determine whether

H_0 : the median values for k groups of data are identical, or

H_1 : at least one median is significantly different.

As with the Kruskal-Wallis test, the test does not provide information on which medians are significantly different from others. That information must come from a multiple comparison test.

7.3.2.2 Computation of the exact test

Rank the data within each block from 1 to k , from smallest to largest. If the null hypothesis is true, the ranks within each block will vary randomly with no consistent pattern. Summing across blocks, the average rank for each treatment group will be similar for all groups, and also be close to the overall average rank. When the alternative hypothesis is true, the ranks in most of the blocks for one or more of the groups will be consistently higher or lower than others. The average group rank for those groups will then differ from the overall average rank. A test statistic X_f is constructed which uses the square of the differences between the average group ranks and the overall rank, to determine if groups differ in magnitude.

The exact test statistic for the Friedman test is a function of both the number of blocks and treatments. Iman and Davenport (1980) state that the exact test should be used for all cases where the number of treatment groups plus the number of blocks ($k + n$) is ≤ 9 . For larger sample sizes a large sample approximation is sufficiently accurate for use. When the number of blocks n is small, the F approximation should be preferred over the chi-square approximation (see the next section).

Should the exact test be required, compute the exact test statistic X_f as shown for the large sample approximation of the following section. X_f is computed identically for both the exact form and large sample approximation. When ties occur, either a corrected large sample

approximation must be used, or the rank transform (F approximation) calculated. The rank transform may be easier to compute.

7.3.2.3 Large sample approximation

For years the Friedman test statistic was approximated using a chi-square distribution with $k-1$ degrees of freedom. This is the approximation used by statistics packages, and is presented here because of its common use. However, it does not take into account the number of blocks in the data set, and can be in serious error for small n and small α ($\alpha < 0.1$) (Iman and Davenport, 1980). An F approximation which is more accurate for small n is also available. It can be computed from the chi-square approximation, or directly from the data as a rank transform method (an analysis of variance on the within-block ranks R_{ij}).

The box on the next page outlines the computation process for the large sample approximation to the Friedman test statistic.

Example 3, continued.

The Friedman test is used to determine if the median concentration of periphyton mercury differs for the 6 stations along the South River of Virginia. The boxplots of this data were shown in figure 7.10, and the data given in table 7.3. The within-block ranks are given below. For 6 blocks (date) and 6 stations, sample sizes are large enough to employ an approximation, so the preferred F approximation is computed.

Station:	1	2	3	4	5	6	
<u>Date</u>							
1	1	4	2	3	5	6	
2	1.5	1.5	3	4	6	5	
3	1.5	1.5	3	4	6	5	
4	2	1	3	4	6	5	
5	1	2	3	4	6	5	
6	1	2	5	3	6	4	
\bar{R}_j	1.33	2.0	3.17	3.67	5.83	5.0	

$$\text{overall median} = (k+1)/2 = 3.5$$

The Friedman test

Situation Measurements of k treatment groups are performed on the same or related sets of subjects, called blocks. There are n blocks. One observation is made on each group-block combination ($N = k \cdot n$).

Computation Within each block, observations are ranked from 1 to k , smallest to largest. These within-block ranks R_{ij} are then used to compute the average group rank \bar{R}_j for each of the $j=1, k$ treatment groups:

$$\bar{R}_j = \frac{\sum_{i=1}^n R_{ij}}{n} .$$

Test Statistic The average group rank \bar{R}_j is compared to the overall average rank $\bar{R} = (k+1)/2$ in the test statistic X_f :

$$X_f = \frac{12 n}{k(k+1)} \sum_{j=1}^k \left[\bar{R}_j - \frac{k+1}{2} \right]^2 .$$

X_f is compared either to an exact table or approximated by a chi-square distribution with $(k-1)$ degrees of freedom. However, a better approximation is available which is compared to an F distribution (Iman and Davenport, 1980). This form is more accurate for small n .

$$f = \frac{(n-1) X_f}{n(k-1) - X_f} .$$

Tied data When observations are tied within a block, assign the average of their ranks to each. X_f must be corrected when more than a few ties occur.

$$X_f = \frac{12 n}{k(k+1) - \frac{1}{n(k-1)} \sum_{i=1}^n \sum_{j=1}^k (t_{ij} (j^3 - j))} \sum_{j=1}^k \left[\bar{R}_j - \frac{k+1}{2} \right]^2 .$$

where t_{ij} equals the number of ties of extent j in row i . The test statistic f is then computed from this corrected X_f as above. An alternative to computing X_f and then f is the rank transform ANOVA (next section).

Decision Rule To reject H_0 : the median of every group is identical, versus H_1 : at least one median differs

Exact test: Reject H_0 if $X_f > x_{\alpha}$, the $(1-\alpha)$ th quantile of the Friedman test statistic distribution from table B7 of the Appendix; otherwise do not reject H_0 .

F-approximation: Reject H_0 if $f \geq F_{1-\alpha, k-1, (n-1)(k-1)}$ the $1-\alpha$ quantile of an F distribution with $k-1$ and $(n-1)(k-1)$ degrees of freedom; otherwise do not reject H_0 .

There are only two ties, so ignoring the formula for the tie correction to the variance,

$$\begin{aligned} Xf &= \frac{12(6)}{6(7)} \sum_{j=1}^6 \left[\bar{R}_j - \frac{7}{2} \right]^2 = \frac{12}{7} \sum (-2.17)^2 + (-1.5)^2 + (-0.33)^2 + (0.17)^2 + (2.33)^2 + (1.5)^2 \\ &= \frac{12}{7} \cdot 14.78 \\ &= 25.33. \end{aligned}$$

This can be compared to a chi-square distribution having $k-1 = 5$ df.

To be more exact, the tie correction will be computed. For rows $i=1,4,5,6$ there are no ties. So for $j=1$, $t_{ij} = 6$ (there are 6 "ties" of extent 1), and for $j=2$ to 6, $t_{ij} = 0$ (no true ties). For these four rows

$$\sum_{j=1}^k (t_{ij} (j^3-j)) = 6(1-1)+0(8-2)+0(27-3)+0(64-4)+0(125-5)+0(216-6) = 0.$$

Rows without ties will always add to zero. Also note that "ties" of extent 1 will always contribute 0 to the sum, as $1^3-1 = 0$. For rows $i=2$ and 3 there is one pair of tied values per row. Thus for $j=1$, $t_{ij} = 4$ (4 single values); for $j=2$, $t_{ij} = 1$ (1 tie of extent 2), and for $j=3$ to 6, $t_{ij} = 0$ (no triplicates, etc.). For each of these two rows

$$\sum_{j=1}^k (t_{ij} (j^3-j)) = 4(1-1)+1(8-2)+0(27-3)+0(64-4)+0(125-5)+0(216-6) = 6.$$

Therefore $\sum_{i=1}^n \sum_{j=1}^k (t_{ij} (j^3-j)) = 0+6+6+0+0+0 = 12$, and

$$Xf = \frac{12 \cdot 6}{6(7) - \frac{1}{6(5)} \cdot 12} \cdot 14.78 = 25.58$$

which can be compared to a chi-square distribution with 5 degrees of freedom.

The better approximation is the F approximation, or

$$f = \frac{(5) 25.58}{6(5) - 25.58} = 28.94, \text{ which is compared to } F_{0.95, 5, 25} = 4.5$$

Therefore reject H_0 that the medians are the same with a p-value of <0.0001 .

7.3.2.4 Rank transform approximation: analysis of variance: on within-block ranks

Again an approximation to the exact test statistic may be computed by performing the parametric two-factor ANOVA on the ranks. For the Friedman test, the appropriate ranks are the within-block ranks of table 7.5. Ties are automatically corrected for by assigning the average rank to all ties within a block. A two-factor ANOVA on the within-block ranks has an ANOVA table as in section 7.3.4. The resulting F statistic, the ratio of the MST for the treatment group over the MSE, is the same as the statistic f derived from the chi-square approximation above. Thus the ANOVA on within-block ranks gives a better approximation

than does X_f for the cases ($\alpha < 0.1$ and small n) where the chi-square approximation is inaccurate (Groggel, 1987).

Example 3, continued.

The ANOVA table for the within-block ranks of table 7.5 is:

Source	df	SS	MS	F	p-value
Date (block)	5	0.000	0.000		
Station	5	88.667	17.733	28.93	<0.0001
<u>Error</u>	<u>25</u>	<u>15.333</u>	0.613		
Total	35	104.000			

Note that all differences between blocks have been nullified by transforming the data to the identical within-block ranks, 1 to k . As the blocks all have the same values within them, the block sum of squares equals 0. Also note that the F statistic is identical to that previously calculated from the large-sample approximation after tie correction. Therefore the ANOVA on within-block ranks provides a convenient way to avoid the complicated tie correction to the Friedman statistic.

7.3.3 Median Aligned-Ranks ANOVA

The Friedman test is the multi-treatment equivalent of the sign test. In Chapter 6 the signed-rank test was presented in addition to the sign test, and was favored over the sign test when the differences between the two treatments were symmetric. In this section a multi-treatment equivalent to the signed-rank test is presented, called the Median Aligned-Ranks ANOVA (MARA). MARA is one of several possible extensions of the signed-rank test; others include Quade's test (Conover, 1980). Groggel (1987) and Fawcett and Salter (1984) have shown that an aligned-rank method has substantial advantages in power over other possible signed-rank extensions.

Friedman's test avoids any comparisons across blocks, just as the sign test avoids comparisons of the magnitudes of paired differences across blocks. This avoids the confusion produced by block-to-block differences, but does not take advantage of the information contained in such comparisons. MARA allows comparisons between blocks by first subtracting the within-block median from all of the data within that block. This "aligns" the data across blocks to a common center. It is equivalent to the ranking of block-to-block differences done in the signed-ranks test. To derive the benefits of cross-block comparisons, a cost is incurred. This is an assumption that the residuals ϵ_{ij} are symmetric. Symmetry can be evaluated by estimating the residuals using median polish, and producing a boxplot as in figure 7.11.

Note that just as for the Friedman's test and two-way ANOVA without replication there are $(k-1)(n-1)$ error degrees of freedom, $(n-1)$ less than a one-way ANOVA. MARA is a two-

factor analysis, with alignment contributing the block effect. However, MARA is computed using a one-way ANOVA on the aligned ranks, so the correct F-test will differ from that performed automatically by a computerized analysis. The error degrees of freedom must be $(k-1)(n-1)$, not $k(n-1)$ as for a one-way ANOVA. MARA is identical to the aligned ranks procedure of Fawcett and Salter (1984), except that the block median is used for alignment rather than the block mean.

The Median Aligned-Ranks ANOVA test																					
Situation	Measurements of k treatment groups are performed on the same or related sets of subjects, called blocks. There are n blocks. One observation is made on each group-block combination ($N = k \cdot n$).																				
Computation	<p>Within each of the n blocks, the observations are aligned by subtracting the block median, forming the aligned o_{ij}.</p> $o_{ij} = (y_{ij} - b_i), \quad \text{where block median } b_i = [\text{median}(y_{ij}), j=1, \dots, k]$ <p>The o_{ij} are then ranked from 1 to N, forming aligned ranks AR_{ij}:</p> $AR_{ij} = \text{rank}(o_{ij}) .$																				
Test Statistic	<p>One-way analysis of variance is computed on the AR_{ij}. However, the F statistic is $F = \text{MST}/\text{MSE}$, where the error degrees of freedom are $(n-1)$ less than in a one-way ANOVA because of the alignment procedure. The ANOVA table is:</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">Source</th> <th style="text-align: center;">df</th> <th style="text-align: center;">SS</th> <th style="text-align: center;">MS</th> <th style="text-align: center;">F</th> </tr> </thead> <tbody> <tr> <td>Treatment</td> <td style="text-align: center;">$(k-1)$</td> <td style="text-align: center;">SST</td> <td style="text-align: center;">$SST/(k-1)$</td> <td style="text-align: center;">MST/MSE</td> </tr> <tr> <td><u>Error</u></td> <td style="text-align: center;"><u>$(k-1)(n-1)$</u></td> <td style="text-align: center;"><u>SSE</u></td> <td style="text-align: center;">$SSE/[(k-1)(n-1)]$</td> <td></td> </tr> <tr> <td>Total</td> <td style="text-align: center;">$n(k-1)$</td> <td style="text-align: center;">Total SS</td> <td></td> <td></td> </tr> </tbody> </table>	Source	df	SS	MS	F	Treatment	$(k-1)$	SST	$SST/(k-1)$	MST/MSE	<u>Error</u>	<u>$(k-1)(n-1)$</u>	<u>SSE</u>	$SSE/[(k-1)(n-1)]$		Total	$n(k-1)$	Total SS		
Source	df	SS	MS	F																	
Treatment	$(k-1)$	SST	$SST/(k-1)$	MST/MSE																	
<u>Error</u>	<u>$(k-1)(n-1)$</u>	<u>SSE</u>	$SSE/[(k-1)(n-1)]$																		
Total	$n(k-1)$	Total SS																			
Tied data	Average ranks are assigned to all tied o_{ij} .																				
Decision Rule	<p>To reject H_0: the median of every group is identical, versus H_1: at least one median differs</p> <p>Reject H_0 if $F \geq F_{1-\alpha, k-1, (n-1)(k-1)}$ the $1-\alpha$ quantile of an F distribution with $k-1$ and $(n-1)(k-1)$ degrees of freedom; otherwise do not reject H_0.</p>																				

7.3.3.1 Null and alternate hypotheses

The null and alternate hypotheses are identical to those of the Friedman test

H_0 : the median values for k groups of data are identical, or

H_1 : at least one median is significantly different.

Here, however, it is assumed that the residuals ϵ_{ij} are symmetric. MARA does not provide information on which medians are significantly different from others. That must come from a multiple comparison test.

7.3.3.2 Computation

MARA is a rank transform approximation test; p-values for an exact test have not been computed.

Example 3, continued

The aligned o_{ij} for the periphyton mercury data were computed during the first step of the median polish, and listed in table 7.4. These o_{ij} are then ranked from 1 to $N=36$ to form aligned ranks, which are presented in table 7.6:

Station:	1	2	3	4	5	6	
<u>Date</u>							
1	9	19	14	18	24	30	
2	5.5	5.5	12	26	36	31	
3	10.5	10.5	15	23	33	28	
4	8	7	13	25	34	32	
5	3	4	17	20	35	29	
6	1	2	22	16	27	21	

A one-way analysis of variance is conducted on these aligned ranks. However, the computerized F-test is ignored, as the error degrees of freedom used were $n(k-1)=30$, and do not reflect the alignment process. The appropriate ANOVA table and F-test are below, and the p-value shows that H_0 is to be rejected. Significant differences are found between treatment group medians:

Source	df	SS	MS	F	p-value
Station	5	3290.3	658.1	27.71	<0.0001
<u>Error</u>	<u>25</u>	<u>593.7</u>	23.8		
Total	30	3884.0			

7.3.4 Parametric Two-Factor ANOVA Without Replication

The traditional parametric test for the randomized complete block design is again an analysis of variance -- a two-factor ANOVA without replication. One factor is the contrast between treatment groups while the second is the block effect. There is one observation (no replicates) per treatment-block combination. The block effect is of no interest except to remove its masking of the treatment effect, so no test for its presence is required.

7.3.4.1 Null and alternate hypotheses

The hypotheses are similar to those of the Friedman and MARA tests, except that treatment group means, rather than medians, are being tested.

H₀: the k treatment group means are identical, $\mu_1 = \mu_2 = \dots = \mu_k$, versus

H₁: at least one mean is significantly different.

The ANOVA model is identical to that for all of the tests of this section:

$$y_{ij} = \mu + \alpha_j + \beta_i + \epsilon_{ij},$$

where

y_{ij} is the individual observation in block i and group j;

μ is the overall mean,

α_j is the "jth group effect", $j=1,k$

β_i is the "ith block effect", $i=1,n$

ϵ_{ij} is the residual or "error" between the individual observation and the combined group and block effects.

Here, however, it is assumed that the residuals ϵ_{ij} follow a normal distribution. ANOVA does not provide information on which means differ from others. That must come from a multiple comparison test.

7.3.4.2 Computation

As with other analysis of variance procedures, the treatment and error mean squares are computed, and their ratio forms the F statistic to be compared to a table of the F distribution for evaluation of its significance. Again there are k treatment groups and n blocks.

In comparison to a one-way ANOVA without blocking, the error sum of squares SSE is split into two parts, the SSE and the sum of squares for the block effect SSB. The variation due to differences between blocks is thereby removed from the background noise (MSE). If there is an appreciable block effect, removal of the SSB lowers the SSE and MSE in comparison to their values for a one-way ANOVA. This produces a higher F statistic, allowing the treatment effect to be more easily discerned.

Example 3, continued

An analysis of variance is calculated directly on the periphyton mercury data. The ANOVA table is:

Source	df	SS	MS	F	p-value
Date	5	3.26	0.65		
Station	5	230.13	46.03	26.15	<0.0001
Error	<u>25</u>	<u>44.02</u>	1.76		
Total	35	277.40			

The null hypothesis is again soundly rejected. The treatment group means are declared different at any reasonable alpha level. As in all of the tests applied to this data set, the block effect (Date) is minimal.

Two-factor ANOVA without replication																											
Situation	Measurements of k treatment groups are performed on the same or related sets of subjects, called blocks. There are n blocks. One observation is made on each group-block combination (N = k•n).																										
Computation	Sums of squares for treatment, block and error are computed using the following formula. These are divided by their appropriate degrees of freedom to form mean squares.																										
	$SST = \frac{\sum^k \left[\sum^n y \right]^2}{n} - \frac{\left[\sum^k \sum^n y \right]^2}{kn}$ $SSB = \frac{\sum^n \left[\sum^k y \right]^2}{k} - \frac{\left[\sum^k \sum^n y \right]^2}{kn}$ $SSE = \text{Total SS} - SST - SSB$ $\text{Total SS} = \sum^k \sum^n y^2 - \frac{\left[\sum^k \sum^n y \right]^2}{kn}$																										
Test Statistic	The F statistic is computed as $F = MST/MSE$. The ANOVA table is:																										
	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">Source</th> <th style="text-align: center;">df</th> <th style="text-align: center;">SS</th> <th style="text-align: center;">MS</th> <th style="text-align: center;">F</th> <th style="text-align: center;">p-value</th> </tr> </thead> <tbody> <tr> <td>Treatment</td> <td style="text-align: center;">k-1</td> <td style="text-align: center;">SST</td> <td style="text-align: center;">SST/(k-1)</td> <td rowspan="3" style="text-align: center; vertical-align: middle;">MST/MSE</td> <td></td> </tr> <tr> <td>Block</td> <td style="text-align: center;">n-1</td> <td style="text-align: center;">SSB</td> <td style="text-align: center;">SSB/(n-1)</td> </tr> <tr> <td>Error</td> <td style="text-align: center;">$\frac{(k-1)(n-1)}{}$</td> <td style="text-align: center;">SSE</td> <td style="text-align: center;">SSE/[(k-1)(n-1)]</td> </tr> <tr> <td>Total</td> <td style="text-align: center;">N-1</td> <td style="text-align: center;">Total SS</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>	Source	df	SS	MS	F	p-value	Treatment	k-1	SST	SST/(k-1)	MST/MSE		Block	n-1	SSB	SSB/(n-1)	Error	$\frac{(k-1)(n-1)}{}$	SSE	SSE/[(k-1)(n-1)]	Total	N-1	Total SS			
Source	df	SS	MS	F	p-value																						
Treatment	k-1	SST	SST/(k-1)	MST/MSE																							
Block	n-1	SSB	SSB/(n-1)																								
Error	$\frac{(k-1)(n-1)}{}$	SSE	SSE/[(k-1)(n-1)]																								
Total	N-1	Total SS																									
Tied data	No corrections necessary.																										
Decision Rule	To reject H_0 : the mean of every group is identical, versus H_1 : at least one mean differs Reject H_0 if $F \geq F_{1-\alpha, k-1, (n-1)(k-1)}$ the $1-\alpha$ quantile of an F distribution with k-1 and (n-1)(k-1) degrees of freedom; otherwise do not reject H_0 .																										

7.4 Multiple Comparison Tests

In most cases the analyst is interested not only in whether group medians or means differ, but which differ from others. This is information not supplied by the tests presented in the previous sections, but by methods called multiple comparison tests (MCTs). MCTs compare all possible pairs of treatment group medians or means, and are performed only after the null hypothesis of

"all medians or means identical" has been rejected. Of interest is the "pattern" of group medians or means:

$$\text{group A} \cong \text{group B} < < \text{group C},$$

etc. MCT's are not efficient methods for contrasting specific sets of groups known to be of interest before an ANOVA or Kruskal-Wallis test is done, such as a treatment versus a control. Other tests are available for making specific contrasts. Instead, MCT's compare all possible combinations of treatment group centers, ranking the centers in order and indicating which are similar or different from others.

Stoline (1981) reviews the many types of parametric multiple comparison tests. Campbell and Skillings (1985) discuss nonparametric multiple comparisons.

7.4.1 Parametric Multiple Comparisons

Parametric MCT's compare treatment group means. They often calculate a "least significant range" or LSR, the distance between any two means which must be exceeded in order for the two groups to be considered significantly different at a significance level α .

$$\text{If } |\bar{y}_1 - \bar{y}_2| > LSR = R\sqrt{s^2/n}, \quad \bar{y}_1 \text{ and } \bar{y}_2 \text{ are significantly different.}$$

The statistic R is analogous to the t-statistic in a t-test. R depends on the test used (is some function of either a t- or studentized range statistic q), the error degrees of freedom from the ANOVA, and on α . The variance s^2 is just the MSE from the ANOVA. Parametric MCT's can be classified into four types, based on their method of computation and on whether a pairwise or overall α level is used (figure 7.13).

	α pairwise	α overall
MST (equal n only)	Duncans Multiple Range test	REGWQ *
	SNK	REGWF *
SIM (equal or unequal n)	Fisher's t-tests (LSD)	Tukey * Scheffe Bonferroni

Figure 7.13 Types of Parametric Multiple Comparison Tests

Methods with an asterisk * in figure 7.13 have the most power to detect differences between group means of those methods using the overall error rate. The REGW methods are the most powerful (have the smallest LSR) for equal sample sizes, though Tukey's test is close in power. For unequal sample sizes, Tukey's method is the most powerful of those listed. Therefore

Tukey's method is a generally applicable and powerful multiple comparison test for a variety of situations.

Multiple-stage tests, MST, are valid only when group sample sizes are equal. Examples are the Duncan's, Student-Newman-Keuls (SNK), and REGW tests. Their R statistic varies for each pairwise comparison as a function of the number of group means in between the two being compared. A new least significant range ($\bar{y}_1 - \bar{y}_2$) must then be computed for each pairwise comparison of means. If sample sizes were unequal, test results could be non-intuitive, as in: $A > B$, $B > C$, but $A = C$ where " $A > B$ " means that A is larger and significantly different from B, and " $A = C$ " means A is not significantly different from C. This could arise if B had a large sample size so that comparisons involving it had a lower LSR than those not involving B. Thus MSTs are valid only for equal sample sizes within all groups.

Simultaneous inference methods, SIM, are valid for both equal and unequal group sample sizes. Examples are Tukey's, Sheffe's, and Fisher's t-tests. These tests use one R value to calculate a single least significant range for all pairwise comparisons. The harmonic mean

$$\text{harmonic mean of } n_1 \text{ and } n_2 = \frac{2 n_1 n_2}{n_1 + n_2}$$

is substituted for n in the case of unequal group sample sizes. So for unequal sample sizes a SIM should be used.

The second classification criteria for MCTs is based on the type of error rate α used for comparisons (figure 7.13). One class of tests uses the stated α level for each pairwise comparison (α_p = pairwise error rate). When there are multiple comparisons each having a pairwise error rate of α , the overall probability of declaring at least one false difference (the overall error rate α_o) is much greater than α_p . This overall error rate is the error rate for the "pattern" of group means, and is more often of interest than a pairwise error rate in water resources applications. For example, when comparing six group means, there are $(6 \cdot 5)/2 = 15$ pairwise comparisons. If $\alpha_p = 0.05$ is used for each test, then there will be an overall error rate $\alpha_o = 1 - (1 - \alpha_p)^{15} = 0.54$ of making at least one error in the overall comparisons of the six group means.

Unfortunately, the distinction between the overall and pairwise error rates is often not understood, and pairwise rates are presented as if they were overall rates. The pairwise rate is much like the probability of being robbed today, while the overall rate is like the probability of ever being robbed in your lifetime. To claim that the (very small) probability of being robbed today is actually the probability of ever being robbed leads to a false sense of security. Similarly, citing that according to a Duncan's multiple range test, $A > B = C = D > E = F$ with an error rate of $\alpha = 0.05$ when in fact 0.05 was used for each test, also presents a false sense of security in the results.

Duncan's test is often used in this incorrect fashion. Individual paired differences found at the $\alpha = 0.05$ level results in the overall rate of at least one error in the pattern of group means at something higher, such as the 0.54 chance for the 6 groups above. When the primary interest is in the overall pattern and its accuracy, methods which set the error rate equal to the overall α , such as Tukey's test, should be performed.

Some authors report only results of a MCT, usually Duncan's multiple range test, skipping the required prior ANOVA F-tests. **NEVER DO THIS!** The likely reason that this has been done is that ANOVA did not find significant differences at a true (overall) significance level of 0.05, but the Duncan's test did find differences. Why does this occur? Duncan's test was performed at a pairwise significance level of 0.05, but at an overall level of something much higher (0.54 for the six means above). An overall error level of 0.54 states there a 54 percent chance that two means will be declared significantly different when in fact they are not. An ANOVA at $\alpha = 0.54$ would also be "significant" (the p-value is somewhere below 0.54), but a test having this large an error rate is essentially useless! ANOVA should always be performed first as the appropriate test for determining whether any differences occur between group means. If they do not, stop there. By performing only a MCT, an $\alpha=0.54$ test is conducted while declaring it to be an $\alpha = 0.05$ test of whether differences occur. This is quite misleading.

7.4.1.1 Assumptions

All MCTs discussed thus far have the same assumptions as does ANOVA -- data within each treatment group are normally distributed, and each treatment group has equal variance. Violations of these assumptions will result in a loss of power to detect differences which are actually present.

7.4.1.2 Computation of Tukey's test

Two group means \bar{y}_i and \bar{y}_j can be considered different if

$$\left| \bar{y}_i - \bar{y}_j \right| > q(1-\alpha), k, N-k \cdot \sqrt{\text{MSE} / n}$$

where

- q is the studentized range statistic from Neter, Wasserman and Kutner (1985),
- α is the overall significance level,
- k is the number of treatment group means compared,
- $N-k$ are the degrees of freedom for the MSE, and
- n is the sample size per group.

For unequal sample sizes

$$\left| \bar{y}_i - \bar{y}_j \right| > q(1-\alpha), k, N-k \cdot \sqrt{\text{MSE} \cdot \frac{n_i + n_j}{2 n_i n_j}}$$

where n has been replaced with the harmonic mean of the unequal sample sizes for the two groups being compared, n_i and n_j . For only two groups, q becomes the student's t statistic, and

Tukey's test is identical to Fisher's all-possible t-tests. Formulas for other MCT's can be found in SAS Institute (1985).

Example 4

Knopman (1990) tested wells located in the Appalachian mountains of Pennsylvania to see if their specific capacities differed among four rock types -- dolomites, limestones, siliciclastics (sandstones, shales, etc.), and metamorphic plus igneous rocks. To make the data more nearly normal, the natural log of specific capacity was used. A subset of 200 observations across the four rock types were randomly selected from the over 4000 original observations. This subset is presented in Appendix C7. Boxplots are shown in figure 7.14. The ANOVA table below indicates that the log specific capacities differed significantly between the four rock types.

Source	df	SS	MS	F	p-value
Rock type	3	54.03	18.010	4.19	0.007
Error	196	842.15	4.297		
Total	199	896.18			

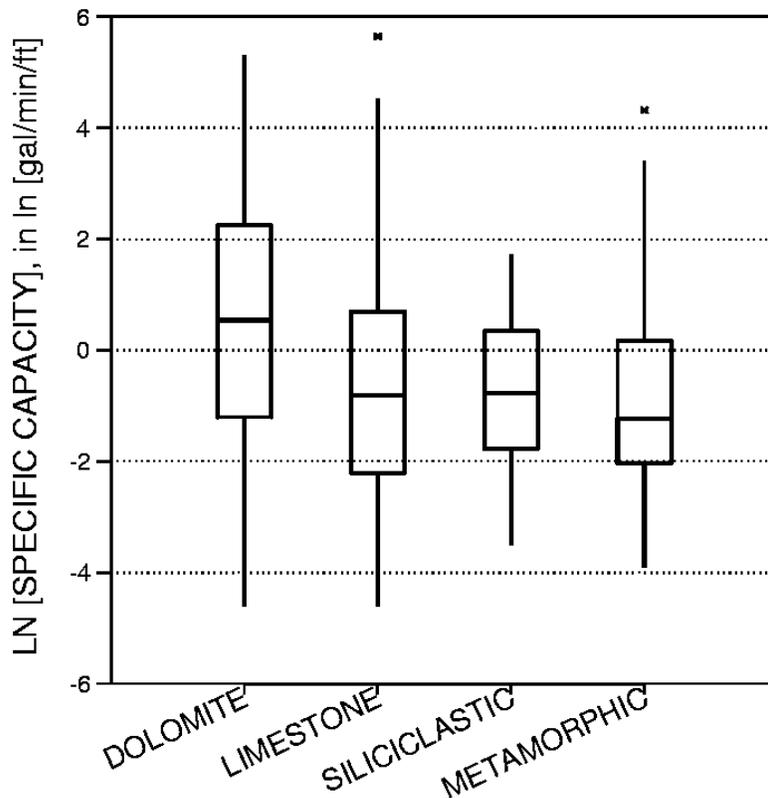


Figure 7.14 Natural logs of specific capacity of wells in four rock types, Pennsylvania

Since the null hypothesis is rejected, Tukey's test can be computed. The four group means are :

$$\begin{array}{ll}
 \bar{y} \text{ [dolomite]} &= \bar{y}_d = 0.408 & \bar{y} \text{ [limestone]} &= \bar{y}_l = -0.688 \\
 \bar{y} \text{ [siliciclastic]} &= \bar{y}_s = -0.758 & \bar{y} \text{ [metamorphic]} &= \bar{y}_m = -0.894
 \end{array}$$

The least significant range LSR is computed as:

$$\begin{aligned} \text{LSR} &= q_{(0.95, 4, 196)} \cdot \sqrt{4.297/50} \cong q_{(0.95, 4, \infty)} \cdot \sqrt{4.297 / 50} = 3.63 \cdot 0.293 \\ &= 1.06 \end{aligned}$$

Therefore, any group means of log specific capacity which differ by more than 1.06 are significantly different by the Tukey's multiple comparison test. \bar{y}_d is then seen to be significantly different and larger than the other three groups, which are not significantly different from each other, or:

$$\bar{y}_d > \bar{y}_l = \bar{y}_s = \bar{y}_m$$

REGWQ could also be computed because sample sizes in each subset group are equal. The choice of REGWQ versus Tukey's would largely depend on which were available. First the k group means are ordered by magnitude ($\bar{y}_d, \bar{y}_l, \bar{y}_s, \bar{y}_m$). The first comparison is made between the extremes, \bar{y}_d versus \bar{y}_m . The studentized range is again used, accounting for the number of means between and including the two being compared; $k=4$ in this first case. If this test proves to be significant, the two possible comparisons with $p=k-1$ intervening group means are made -- \bar{y}_d versus \bar{y}_s and \bar{y}_l versus \bar{y}_m . Continue working inward until an insignificant difference is found. No comparisons of group means contained between means already found to be insignificant need be made.

For REGWQ, two group means differ at an overall significance level α if:

$$\begin{aligned} \bar{y}_i - \bar{y}_j > q_{\alpha_p, p, N-p} \cdot \sqrt{\text{MSE} / n} \\ \text{where } \alpha_p &= 1 - (1-\alpha)^{p/k} \quad \text{for } p < (k-1) \\ &= \alpha \quad \text{for } p \geq (k-1). \end{aligned}$$

Using the log specific capacity data, comparing \bar{y}_d versus \bar{y}_m using $\alpha_p = \alpha = 0.05$:

the least significant range = $q_{0.05, 4, 196} \cdot \sqrt{4.297 / 50} = 1.06$, identical to Tukey's LSR.

Therefore $\bar{y}_d > \bar{y}_m$. Next, compare \bar{y}_d versus \bar{y}_s and \bar{y}_l versus \bar{y}_m . Both of these have

$p=3$ and an LSR of $q_{0.05, 3, 197} \cdot \sqrt{4.297 / 50} = 3.31 \cdot 0.293 = 0.97$. Therefore

$\bar{y}_d > \bar{y}_s$ and $\bar{y}_l = \bar{y}_m$. Since the limestone and metamorphic group means are not

significantly different there is no reason to test the siliciclastic versus the metamorphic group means. For the final comparison, \bar{y}_d is compared to \bar{y}_l . The LSR is based on $p=2$ and

$\alpha_p = 1 - (0.95)^{2/4} = 0.025$. Therefore LSR = $q_{0.025, 2, 198} \cdot \sqrt{4.297 / 50} = 3.31 \cdot 0.293 = 0.97$. So $\bar{y}_d > \bar{y}_l$ and the overall pattern is again:

$$\bar{y}_d > \bar{y}_l = \bar{y}_s = \bar{y}_m$$

7.4.2 Nonparametric Multiple Comparisons

Statisticians are actively working in this area (see Campbell and Skillings, 1985). The simplest procedures for performing nonparametric multiple comparisons are rank transformation tests. Ranks are substituted for the original data, and a multiple comparison test such as Tukey's is

performed on the ranks. These are logical follow-ups to the rank transform approximation approaches to the Kruskal-Wallis, Friedman, and two-way ANOVA tests previously presented.

For the one-way situation, Campbell and Skillings (1985) recommend a multiple-stage test using the Kruskal-Wallis (KW) statistic. The process resembles the REGWQ test above. After a significant KW test occurs for k groups, place the groups in order of ascending average rank. Perform new KW tests for the two possible comparisons between $p = (k-1)$ groups, noting that this involves re-ranking the observations each time. If significant results occur for one or both of these tests, continue attempting to find differences in smaller subsets of $p < (k-1)$. In order to control the overall error rate, follow the pattern of REGWQ for the critical alpha values:

$$\alpha_p = 1 - (1-\alpha)^{p/k} \quad \text{for } p < (k-1)$$

$$= \alpha \quad \text{for } p \geq (k-1)$$

Example 4 continued

First, Tukey's test will be performed on the ranks of the Pennsylvania log specific capacity data. Then a second nonparametric MCT, the multiple-stage Kruskal-Wallis (MSKW) test using REGWQ alpha levels, is performed.

The ANOVA table for testing data ranks shows a strong rejection of H_0 :

Source	df	SS	MS	F	p-value
Rock type	3	38665	12888	4.02	0.008
<u>Error</u>	<u>196</u>	<u>627851</u>	3203		
Total	199	666515			

The four group mean ranks are :

$$\begin{aligned} \bar{R} \text{ [dolomite]} &= \bar{R}_d = 124.11 & \bar{R} \text{ [limestone]} &= \bar{R}_l = 94.67 \\ \bar{R} \text{ [siliciclastic]} &= \bar{R}_s = 95.06 & \bar{R} \text{ [metamorphic]} &= \bar{R}_m = 88.16 \end{aligned}$$

The least significant range LSR for a Tukey's test on data ranks is computed as:

$$\begin{aligned} \text{LSR} &= q_{(0.95, 4, 196)} \cdot \sqrt{3203/50} \cong q_{(0.95, 4, \infty)} \cdot \sqrt{3203/50} = 3.63 \cdot 8.00 \\ &= 29.06 \end{aligned}$$

Pairs of group mean ranks which are at least 29.06 units apart are significantly different.

Therefore (within 0.01) $\bar{R}_d > \bar{R}_s = \bar{R}_l = \bar{R}_m$.

To compute the MSKW test, the first step is merely the Kruskal-Wallis test on the four groups.

The overall mean rank \bar{R} equals 100.5. Then

$$K=11.54 \quad \chi^2_{0.95,(3)} = 7.815 \quad p=0.009 \quad \text{so, reject equality of group medians.}$$

Proceeding, new Kruskal-Wallis tests are performed between the two sets of three contiguous treatment groups: \bar{R}_d vs. \bar{R}_l vs. \bar{R}_s and \bar{R}_l vs. \bar{R}_s vs. \bar{R}_m . This requires that the data all be re-ranked each time. Their respective test statistics are denoted K_{dls} and K_{lsm} . The significance level is as in REGWQ, so for $(k-1) = 3$ groups, $\alpha_p = \alpha = 0.05$.

$$\begin{array}{llll} K_{dls} = 8.95 & \chi^2_{0.95,(2)} = 5.99 & p=0.012 & \text{so, reject equality of group medians.} \\ K_{lsm} = 0.61 & & p=0.74 & \text{group medians not significantly different.} \end{array}$$

Finally, the $k-2 = 2$ group comparisons are performed. There is no need to do these for the limestone versus siliciclastic and siliciclastic versus metamorphic comparisons, as the 3-group Kruskal-Wallis test found no differences among those group medians. Therefore the only remaining 2-group comparison is for dolomite versus limestone. The 2-group Kruskal-Wallis test is performed at a significance level of $\alpha_p = 1 - (0.95)^{2/4} = 0.025$.

$$K_{dl} = 5.30 \quad \chi^2_{0.975,(1)} = 5.02 \quad p=0.021 \quad \text{so, reject equality of group medians.}$$

The pattern is the same as for the other MCT's,

$$\text{median}_d > \text{median}_l = \text{median}_s = \text{median}_m.$$

7.5 Presentation of Results

Following the execution of the tests in this chapter, results should be portrayed in an easily-understandable manner. This is best done with figures. A good figure provides a visual confirmation of the outcome of the hypothesis test. Differences between groups are clearly portrayed. A poor figure gives the impression that the analyst has something to hide, and is hiding it effectively! The following sections provide a quick survey of good and bad figures for illustrating differences between three or more treatment groups.

7.5.1 Graphical Comparisons of Several Independent Groups

Perhaps the most common method used to report comparisons between groups is a table, and not a graph. Table 7.7a is the most common type of table in water resources, one which presents only the mean and standard deviations. As has been shown several times, the mean and standard deviation alone do not capture much of the important information necessary to compare groups, especially when the data are skewed. Table 7.7b provides much more information -- important percentiles such as the quartiles are listed as well.

Table 7.7a A simplistic table comparing the four groups of log specific capacity data

	<u>Mean</u>	<u>Std.Dev.</u>
Dolomite	0.408	2.557
Limestone	-0.688	2.360

Siliciclastics	-0.758	1.407
Metamorphic	-0.894	1.761

Table 7.7b A more complete table for the log specific capacity data

	N	Mean	Median	Std.Dev.	Min	Max	P25	P75
Dolomite	50	0.408	0.542	2.557	-4.605	5.298	-1.332	2.264
Limestone	50	-0.688	-0.805	2.360	-4.605	5.649	-2.231	0.728
Siliciclastics	50	-0.758	-0.777	1.407	-3.507	1.723	-1.787	0.381
Metamorphic	50	-0.894	-1.222	1.761	-3.912	4.317	-2.060	0.178

However, neither table provides quick intuitive insight into the data structure. Neither sufficiently illustrates the differences between groups found by the hypothesis tests in example 4, or how they differ.

Histograms are commonly used to display the distribution of one or more data sets, and have been employed to attempt to illustrate differences between three or more groups of data. They are not usually successful. The many crossing lines, coupled with an artificial division of the data into categories, results in a cluttered and confusing graph. Figure 7.15 displays four overlapping histograms, one for each of the data groups. It is impossible to discern anything about the relative characteristics of any of the data groups from this figure. Overlapping histograms should be avoided unless one is purposefully trying to confuse the audience! In figure 7.16, side-by-side bar charts display the same information. This too is confusing and difficult to interpret. From the graph one could not easily say which group had the highest mean or median, much less anything about the groups' variability or skewness. Many business software packages allow speedy production of such useless graphs as these.

Figure 7.17 shows a quantile plot of the same four data groups. The quantile plot far exceeds the histogram and bar chart in clarity and information content. The dolomite group stands apart from the other three throughout most of its distribution, illustrating both the ANOVA and multiple comparison test results. An experienced analyst can look for differences in variability and skewness by looking at the slope and shapes of each group's line. A probability plot of the four groups would have much the same content, with the additional ability to look for departures from a straight line as a visual clue for non-normality.

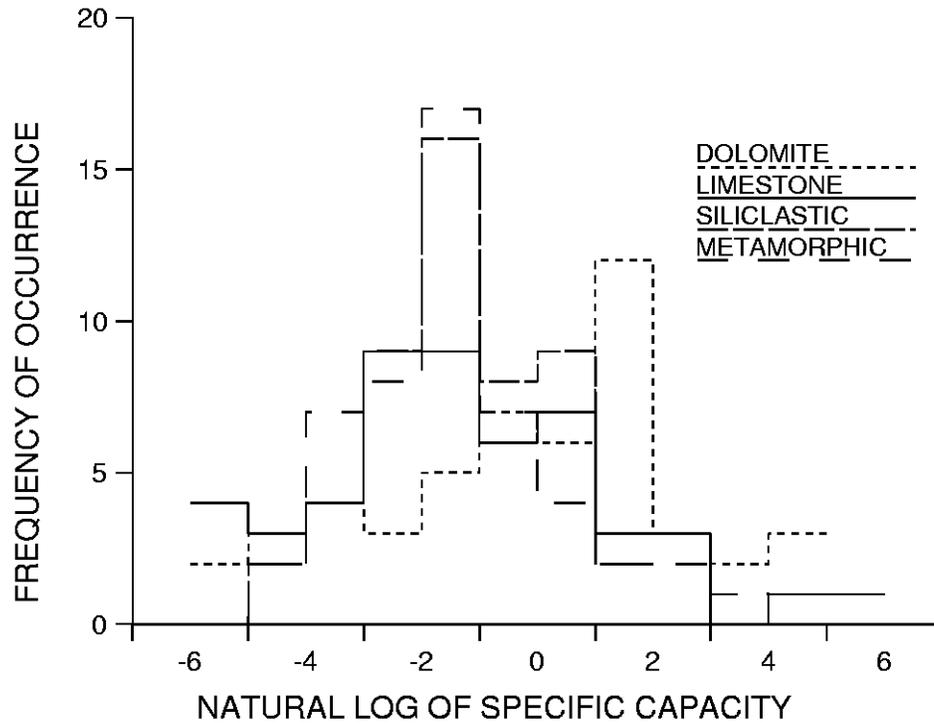


Figure 7.15 Overlapping histograms fail to differentiate between four groups of data

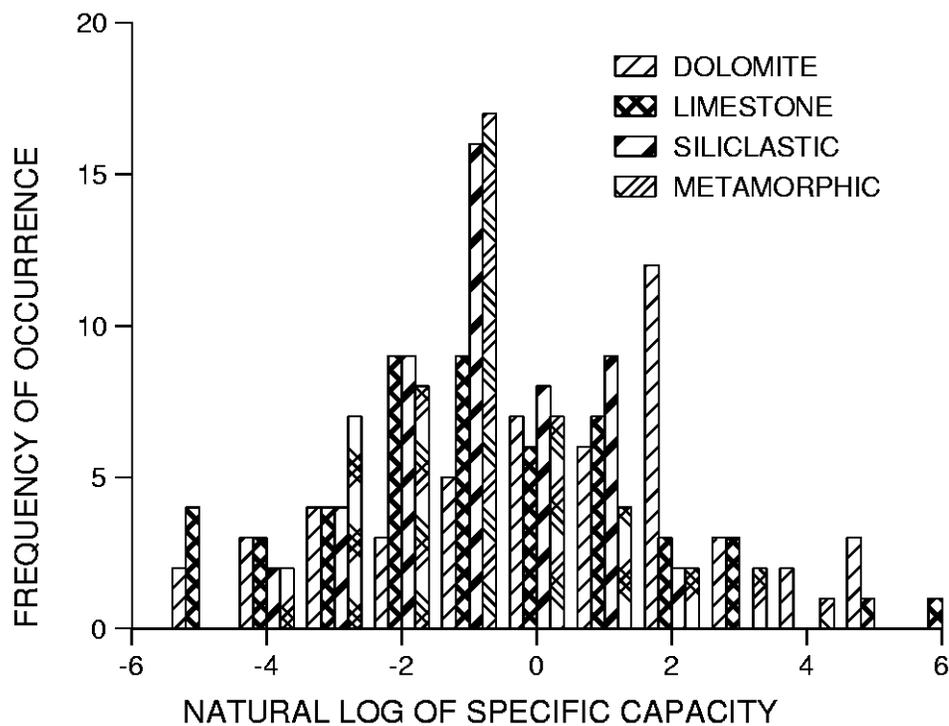


Figure 7.16 Side-by-side bars fail to clearly differentiate between four groups of data

Compare figures 7.15 to 7.17 with boxplots of the log specific capacity data shown previously in figure 7.14. Boxplots clearly demonstrate the difference between the dolomite and other group medians. Variability is also documented by the box height, and skewness by the heights of the top and bottom box halves. See Chapter 2 for more detail on boxplots. Boxplots illustrate the results of the tests of this chapter more clearly than commonly-used alternate methods such as histograms.

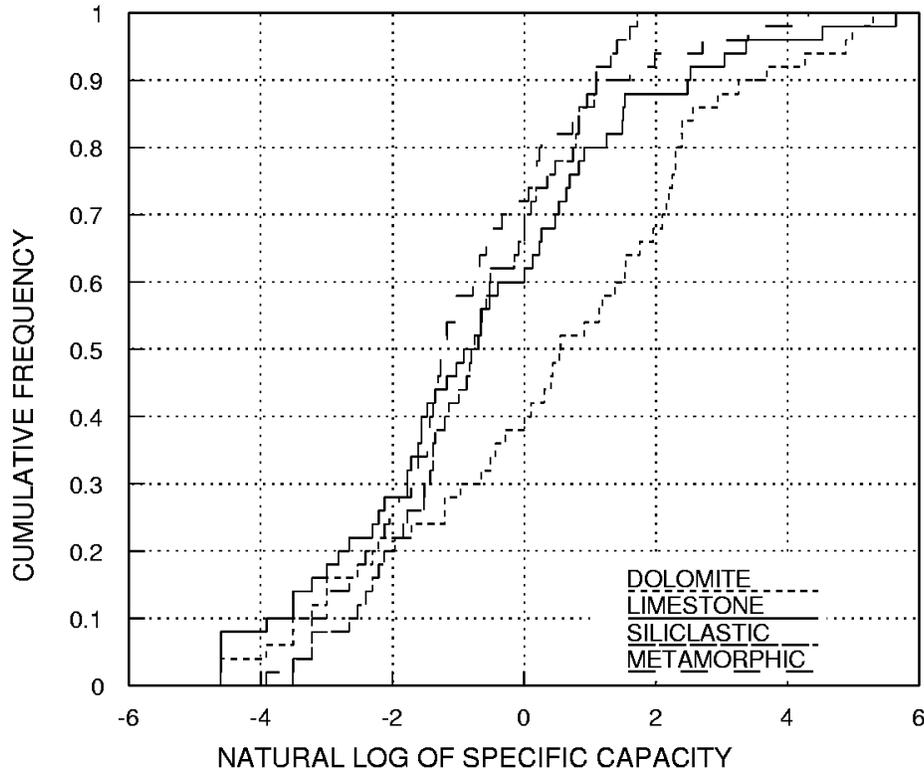


Figure 7.17 Quantile plots differentiate between four groups of data

7.5.2 Presentation of Multiple Comparison Tests

Suppose a multiple comparison test resulted in the following:

$$\begin{array}{llll}
 \bar{y}_1 = \bar{y}_2 & \bar{y}_1 \neq \bar{y}_3 & \bar{y}_1 \neq \bar{y}_4 & (= : \text{not significantly different}) \\
 \bar{y}_2 = \bar{y}_3 & \bar{y}_2 \neq \bar{y}_4 & & (\neq : \text{significantly different}) \\
 \bar{y}_3 = \bar{y}_4 & & &
 \end{array}$$

for four treatment groups having $\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$.

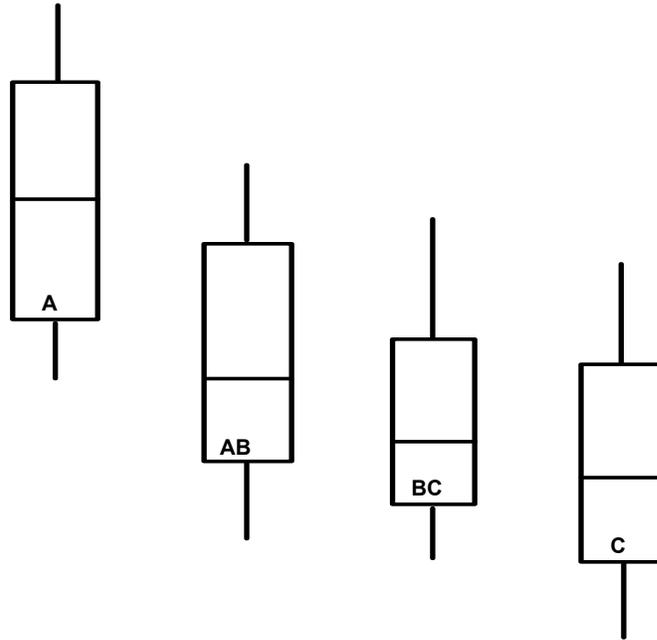
The results are often presented in one of the two following formats:

- Letters

$$\begin{array}{cccc}
 \bar{y}_1 & \bar{y}_2 & \bar{y}_3 & \bar{y}_4 \\
 A & AB & BC & C
 \end{array}$$

Treatment group means are ordered, and those having the same letter underneath them are not significantly different. The convenience of this presentation format is that letters can easily be

positioned somewhere within side-by-side boxplots, illustrating the results of a MCT as well as the overall test for equality of all means or medians (see figure 7.18).



MCT results: Boxes with same letter are not significantly different.

Figure 7.18 Boxplots with letters showing the results of a MCT.

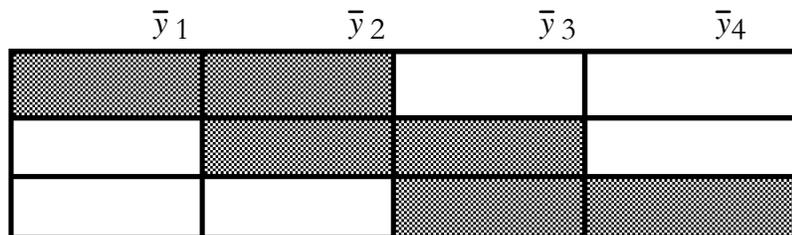
2. Lines



In this presentation format, group means connected by a single unbroken line are not significantly different. This format is suited for inclusion in a table listing group means or medians.

A third method is somewhat more visual:

3. Shaded Boxes



These shaded boxes can be thought of as thick versions of the lines presented above. Group means with boxes shaded along the same row are not significantly different. Shaded boxes allow group means to be ordered by something other than mean or median value. For example, the order of stations going upstream to downstream might be 3,1,2,4. Boxes put in that order show a significant increase in concentration between 3 and 1 and a significant drop off again between 2 and 4. So in addition to displaying multiple comparison test results, the shaded boxes below also illustrate the pattern of concentration levels of the data.

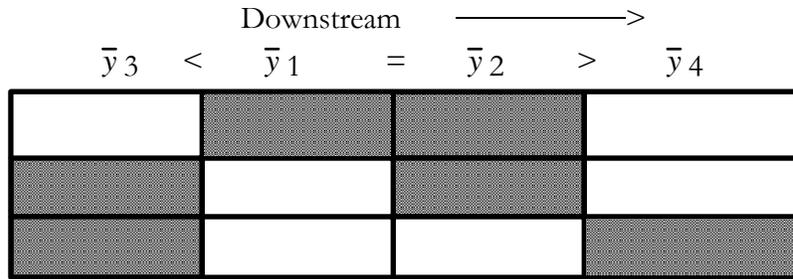


Figure 7.19 Shaded boxes for illustration of a multiple comparison test. Station means not significantly different have boxes shaded within the same row.

Exercises

- 7.1 Discharge from pulp liquor waste may have contaminated shallow groundwater with caustic, high pH effluent (Robertson, et al., 1984). Determine whether the pH of samples taken from three sets of piezometers are all identical -- one piezometer group is known to be uncontaminated. If not, which groups are different from others? Which are contaminated?

	<u>pH of samples taken from piezometer groups</u>					
BP-1	7.0	7.2	7.5	7.7	8.7	7.8
BP-2	6.3	6.9	7.0	6.4	6.8	6.7
BP-9	8.4	7.6	7.5	7.4	9.3	9.0

- 7.2 In addition to the waters from granitic terrain given in Exercise 2.3, Feth et al. (1964) measured chloride concentrations of ephemeral springs. These additional data are listed below (use the zero value as is). Test whether concentrations in the three groups are all identical. If not, which differ from others?

	<u>Chloride concentration, in mg/L</u>					
<u>Ephemeral Springs</u>	0.0	0.9	0.1	0.1	0.5	0.2
	0.3	0.2	0.1	2.0	1.8	0.1
	0.6	0.2	0.4			

- 7.3 The number of Corbicula (bottom fauna) per square meter for a site on the Tennessee River was presented by Jensen (1973). The data are found in Appendix C8. Perform a median polish for the data of strata 1. Graph the polished estimates of year and seasonal effects. Is any transformation suggested by the residuals?
- 7.4 Test the Corbicula data of strata 1 to determine whether season and year are significant determinants of the number of organisms.
- 7.5 Test for significant differences in the density of Corbicula between seasons and strata for the 1969 data.