Chapter 11

Tests for the Effects of Two Factors

Chapter Objectives

- State and interpret the two-factor ANOVA models with and without the interaction effect.
- Interpret sums of squares, degrees of freedom, and mean squares.
- Carry out two-factor ANOVA F tests for the main effects of two factors and for their interaction effect.
- Obtain and interpret fitted values and residuals associated with the ANOVA model.
- Carry out a Friedman test for main effects in a two-factor study.
- Decide which test (the two-factor ANOVA F test or the Friedman test) is more appropriate for a given set of data.
- Carry out a Bonferroni multiple comparison procedure to identify which of levels of a given factor differ from each other.

Key Takeaways

- The two-factor ANOVA F tests are parametric tests for main effects of two factors and for their interaction effect. They require either that the samples are from normal populations or the sample sizes are all large. A log transformation can make right-skewed data more normal prior to conducting the ANOVA F tests.
- A two-factor ANOVA model describes several sources variation in a response variable: the non-random main effects of two factors, their non-random interaction effect, and within-groups random error.
- Sums of squares in two-factor ANOVA are statistics that measure between-rows, between-columns, interaction, and within-groups variation in the observed values of a response variable.
- Mean squares are another way to measure between-rows, between-columns, interaction, and within-groups variation. They're obtained by dividing sums of squares by their degrees of freedom. The degrees of freedom associated with a sum of squares is determined by how many of its squared deviations are "free to vary." The values of two mean squares are directly comparable, but the values of two sums of squares aren't necessarily comparable.
- The two-factor ANOVA F test statistics are ratios of two mean squares. Their numerator measures either between-rows, between-columns, or interaction variation and their denominator within-groups variation.
- Blocking can be used to control for variation in a response variable due to extraneous factors that aren't necessarily of interest in a study.
- The Friedman test is a nonparametric test for the effects of two factors that doesn't require a normality assumption or large sample sizes.
- A multiple comparison procedure, such as the Bonferroni procedure, is used to identify *which* factor

levels differ from each other after a two-factor ANOVA F test (or Friedman test) has indicated that such differences exist.

11.1 Introduction

Environmental studies often involve *simultaneously* investigating the effects of *two* factors (categorical explanatory variables) on a response variable. These studies can arise in the context of taking samples from populations as well as in the context of conducting experiments. Here are a few examples.

Example 11.1: Two-Factor Studies

Enrichment of soils by nutrients such as phosphorus can lead to invasion by exotic weeds that spread quickly (because they have few predators, competitors, parasites, and diseases) and then replace native species that have important ecological functions.

To investigate the role that topography and soil type play in soil phosphorus levels, a study was carried out in three national parks outside of Sydney, Australia [3]. Two soil types, shale-derived and sandstone-derived, were examined in each of four topographies, valleys, north-facing slopes, south-facing slopes, and hilltops. In each of the eight combinations of soil type and topography, three 250 m^2 quadrats were selected. From each quadrat, five soil cores were collected from randomly selected locations, then *composited* (combined), and the phosphorus (ppm) was determined for each of the 24 composited specimens.

The phosphorus concentrations are shown in the two-way layout below, where the the two *factors*, soil type, which has two *levels*, and topography, which has four *levels*, are represented by the margins of the table.

Factor B: Topography

		Valley (j=1)	North-Facing (j=2)	South-Facing (j=3)	Hilltop (j=4)	
Factor A: Soil Type	Shale (i=1)	98	78	117	83	
		172	77	54	12	$\bar{Y}_{1\cdot} = 90.5$
		185	100	96	14	
	Sandstone (i=2)	19	27	28	55	
		39	49	53	21	$\bar{Y}_{2\cdot} = 35.9$
		25	24	72	19	
		$\bar{Y}_{\cdot 1} = 89.7$	$\bar{Y}_{\cdot 2} = 59.2$	$\bar{Y}_{\cdot 3} = 70.0$	$\bar{Y}_{\cdot 4} = 34.0$	$\bar{Y} = 63.2$

There were three research questions:

1. Is there a significant difference between phosphorus concentrations for the two soils types?
2. Are there significant differences in phosphorus concentrations among the four topographies?
3. Does the effect of soil type (if any) on phosphorus concentrations differ depending on the topography?

These three research questions refer to a ***soil type main effect***, a ***topography main effect***, and the effect of an ***interaction*** between soil type and topography, respectively.

A bar plot, side-by-side boxplots, and a three-dimensional individual value plot of the data are shown below.

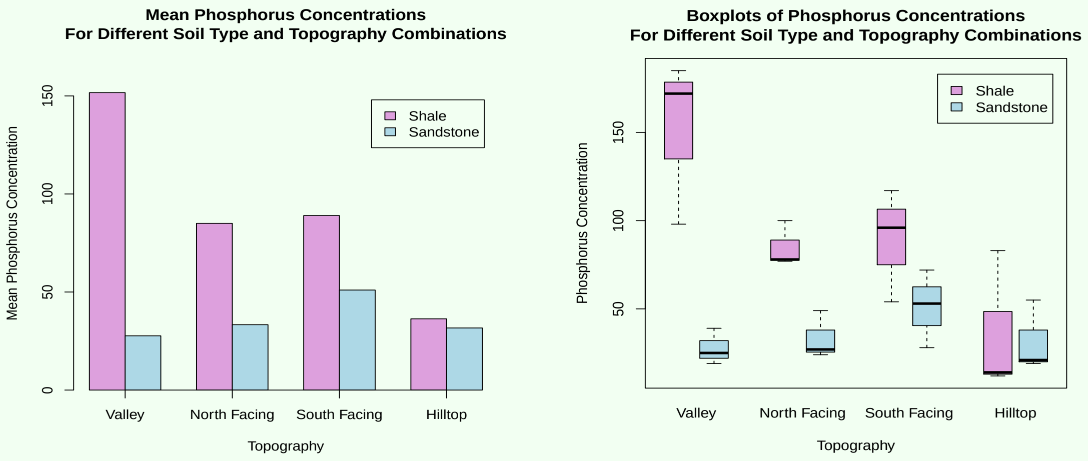


Figure 11.1: Bar plot (left) and boxplots (right) of phosphorus concentrations for two soil types and four topographies.

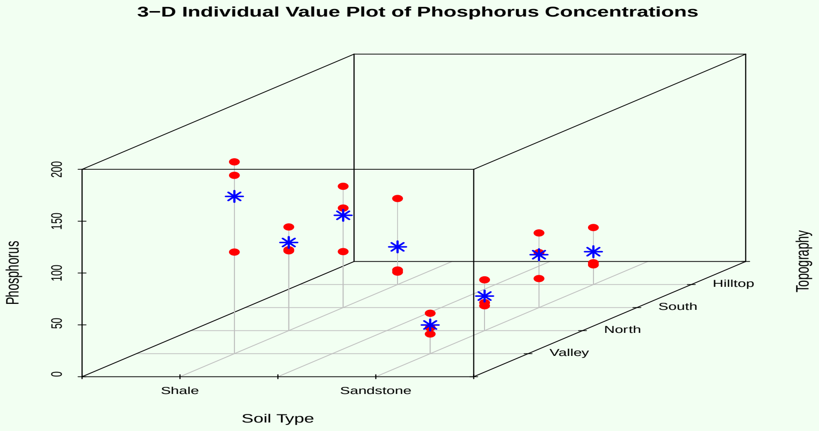


Figure 11.2: Three dimensional individual value plot of phosphorus concentrations for two soil types and four topographies. Sample means are depicted as blue asterisks.

Example 11.2: Two-Factor Studies

In quality assurance studies at wastewater treatment laboratories, a quality assurance engineer will often ”spike” water specimens by adding a known concentration of an analyte before the specimen is processed. The goal is to determine whether any inaccuracy, or bias, exists in the analysis methods. The response variable is the percent recovery, defined as

$$\text{Percent Recovery} \;=\; \frac{\text{Measured} - \text{Background}}{\text{Spike Amount}} \cdot 100\%\,,$$

where

Measured = The measured concentration in the spiked specimen (mg/L)

Background = The measured concentration in the unspiked specimen (mg/L)
Spike Amount = The concentration of spike added to the specimen (mg/L)

In one study, lab technicians analyzed spiked samples for two types of water, wastewater effluent and tap water, and three different pH levels, 6.0, 7.0 and 8.0 [1]. For each pH level, three specimens of each type of water were spiked with ammonia (as NH_3–N), and the percent recovery was determined for each specimen. The table below shows the data.

Factor B: pH

		pH=6.0	pH=7.0	pH=8.0	
		(j=1)	(j=2)	(j=3)	
	Effluent	100	98	102	
Factor A:	(i=1)	88	99	101	$\bar{Y}_{1\cdot} = 98.6$
Water		101	99	99	
Type	Tap	98	95	95	
	Water	96	95	98	$\bar{Y}_{2\cdot} = 96.0$
	(i=2)	96	97	94	
		$\bar{Y}_{\cdot 1} = 96.5$	$\bar{Y}_{\cdot 2} = 97.2$	$\bar{Y}_{\cdot 3} = 98.2$	$\bar{Y} = 97.3$

The questions to be addressed by the quality assurance were:

1. Is there a significant difference between the percent recoveries for the two types of water?
2. Are there significant differences in the percent recoveries among the three pH levels?
3. Does the effect of pH (if any) on percent recovery differ depending on the type of water being analyzed?

These three questions refer to a ***water type main effect***, a ***pH main effect***, and the effect of an ***interaction*** between water type and pH, respectively.

In this chapter we'll look at two hypothesis test procedures for data involving two factors:

1. The two-factor analysis of variance (ANOVA) F tests.
2. The Friedman test.

Both tests are applicable to observational studies (as in Example 11.1) and experiments (as in Example 11.2). The first is parametric, assuming normality of the response variable, and is used to test for the so-called *main effects* of the two factors and their *interaction effect*. The second test is nonparametric, making no normality assumption about the response variable, but its use is restricted to data with only one observation per combination of the levels of the two factors, and when applied to experiments is only applicable when a so-called *randomized block design* was used.

11.2 Two-Factor Analysis of Variance

11.2.1 Introduction

We'll call the factors in a two-factor study ***factor A*** and ***factor B***, and we'll denote the number of ***levels*** of factor A by $\boldsymbol{a}$ and the number of levels of factor B by $\boldsymbol{b}$. We can think of the factors as *two* explanatory variables in the study, both of which are categorical.

In the context of sampling from populations, the two factors are used to define the populations, and each *combination* of a factor A level with a factor B level defines a population from which a sample is drawn.

In a randomized experiment, *both* factors are manipulated by the experimenter, and each *combination* of a factor A level with a factor B level is one set of experimental conditions to which individuals may be randomly assigned.

When the data are organized in a two-way table as in Examples 11.1 and 11.2, with rows representing levels of factor A and columns levels of factor B, each of the ab row-column intersections is called ***cell*** of the table. Each cell corresponds to a ***population***, and the individual observations within the cell are the sample, which we'll refer to as a ***group***. In a two-factor *experiment*, each cell corresponds to a set of experimental conditions, or ***treatment***, and the individuals within that cell form the ***treatment group***.

Two-factor analysis of variance (or ***ANOVA***) is a procedure for deciding if either of the factors affects the response variable. We refer to the effect of factor A as the ***factor A main effect*** and the effect of factor B as the ***factor B main effect***. Later, starting with Section 11.2.7, we'll see that two-factor ANOVA can also be used to decide whether the effect of one factor is different depending on the level of the other one. If it is, we say there's an ***interaction*** between the effects of the two factors.

11.2.2 Notation

Suppose that we have independent random samples from ab populations defined by the levels of two factors in a in a two-way table like the ones in Examples 11.1 and 11.2, where rows represent levels of factor A and columns levels of factor B. The group sample sizes sizes don't necessarily all have to be the same.

Note: Although two-factor ANOVA *can* be carried out when the sample sizes or group sizes are unequal, *we'll only examine details of two-factor ANOVA for the case in which they're equal.* Thus, for the remainder of this chapter, we'll let

$$\boldsymbol{n} = \text{The common sample size for the } ab \text{ groups,}$$

with the understanding that in practice they *don't* necessarily all have to have the same size.

As was done in Examples 11.1 and 11.2, it's useful to organize the data in a two-way table having the form shown below.

		Factor B				
		Level $j = 1$	Level $j = 2$	$\cdots$	Level $j = b$	
	Level $i = 1$	$\left.\begin{matrix}Y_{111}\\Y_{112}\\\vdots\\Y_{11n}\end{matrix}\right\}\bar{Y}_{11}$	$\left.\begin{matrix}Y_{121}\\Y_{122}\\\vdots\\Y_{12n}\end{matrix}\right\}\bar{Y}_{12}$	$\cdots$	$\left.\begin{matrix}Y_{1b1}\\Y_{1b2}\\\vdots\\Y_{1bn}\end{matrix}\right\}\bar{Y}_{1b}$	$\bar{Y}_{1\cdot}$
Factor A	Level $i = 2$	$\left.\begin{matrix}Y_{211}\\Y_{212}\\\vdots\\Y_{21n}\end{matrix}\right\}\bar{Y}_{21}$	$\left.\begin{matrix}Y_{221}\\Y_{222}\\\vdots\\Y_{22n}\end{matrix}\right\}\bar{Y}_{22}$	$\cdots$	$\left.\begin{matrix}Y_{2b1}\\Y_{2b2}\\\vdots\\Y_{2bn}\end{matrix}\right\}\bar{Y}_{2b}$	$\bar{Y}_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	Level $i = a$	$\left.\begin{matrix}Y_{a11}\\Y_{a12}\\\vdots\\Y_{a1n}\end{matrix}\right\}\bar{Y}_{a1}$	$\left.\begin{matrix}Y_{a21}\\Y_{a22}\\\vdots\\Y_{a2n}\end{matrix}\right\}\bar{Y}_{a2}$	$\cdots$	$\left.\begin{matrix}Y_{ab1}\\Y_{ab2}\\\vdots\\Y_{abn}\end{matrix}\right\}\bar{Y}_{ab}$	$\bar{Y}_{a\cdot}$
		$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$	$\cdots$	$\bar{Y}_{\cdot b}$	$\bar{Y}$

As seen in the table, we use the notation

$\boldsymbol{Y_{ijk}}$ = The kth observation at the ith level of factor A and jth level of factor B.

The first subscript, $\boldsymbol{i}$, indicates the row (factor A level) and takes the values $1, 2, \ldots, a$. The second, $\boldsymbol{j}$, indicates the column (factor B level) and takes the values $1, 2, \ldots, b$. The third subscript, $\boldsymbol{k}$, distinguishes among individuals within a group, and takes the values $1, 2, \ldots, n$.

In the margins of the table are the following means.

$\boldsymbol{\bar{Y}_{i\cdot}}$ = The sample mean of the observations made at the ith level of factor A, called the ***ith row mean*** or ***ith factor A level mean***.

$\boldsymbol{\bar{Y}_{\cdot j}}$ = The sample mean of the observations made at the jth level of factor B, called the ***jth column mean*** or ***jth factor B level mean***.

The ”dot” in a subscript indicates that we're averaging over the levels of that factor. For example, for the ith factor A level mean $\bar{Y}_{i\cdot}$, the ”dot” indicates that we're averaging over the levels of factor B, or columns in the table. Likewise, for the jth factor B level mean $\bar{Y}_{\cdot j}$, the ”dot” indicates that we're averaging over the levels of factor A, or rows in the table.

Another set of means, shown in the *cells* of the table above, are defined by the following.

$\boldsymbol{\bar{Y}_{ij}}$ = The sample mean of the n observations at the ith level of factor A and jth level of factor B, called the sample ***group mean***.

We'll also let

$\boldsymbol{N}$ = The ***overall sample size***, or total number of observations in all ab groups combined.

Since there are ab groups, each of which has n observations,

$$N = abn.$$

Lastly, we'll let

$\boldsymbol{\bar{Y}}$ = The ***overall sample mean***, that is, the mean of all N observations in the ab groups combined.

The next fact says that the overall mean can be obtained by averaging row means, column means, group means.

Fact 11.1 When the group sample sizes are all the same, the overall mean $\bar{Y}$ is equal to all of the following:

1. The average of the a row means $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{a\cdot}$.
2. The average of the b column means $\bar{Y}_{\cdot 1}, \bar{Y}_{\cdot 2}, \ldots, \bar{Y}_{\cdot b}$.
3. The average of all ab group means $\bar{Y}_{ij}$.

Example 11.3: Two-Factor ANOVA Means

For the study of phosphorus in two soil types and four topographies (Example 11.1), the common sample size per group and overall sample size are

$$n = 3 \qquad \text{and} \qquad N = 24,$$

and the overall mean is obtained by averaging either the 24 phosphorus observations, the eight group means (given in Example 11.6), the two row means (from Example 11.1), as below,

$$\bar{Y} = \frac{1}{2}(\bar{Y}_{1\cdot} + \bar{Y}_{2\cdot}) = \frac{1}{2}(90.5 + 35.9) = 63.2,$$

or the four column means (also from Example 11.1), as below.

$$\bar{Y} = \frac{1}{4}(\bar{Y}_{\cdot 1} + \bar{Y}_{\cdot 2} + \bar{Y}_{\cdot 3} + \bar{Y}_{\cdot 4}) = \frac{1}{4}(89.7 + 59.2 + 70.0 + 34.0) = 63.2\,.$$

11.2.3 Variation Between Rows, Between Columns, and Within Groups

The two row means given in Example 11.1 aren't equal, so there's evidence that soil type affects phosphorus concentrations. But we wouldn't expect them to be equal, even if soil type had no effect, because naturally occurring spatial variation in phosphorus leads to random sampling error. Likewise, there's evidence for a topography effect because the four column means aren't equal, but we wouldn't expect them to be equal either, even in the absence of a topography effect, for the same reason. So in order to decide if either factor has an effect, we'll need to be able to tell if the differences among row or column means are larger than can be explained by chance variation (sampling error).

We can visually inspect for factor A or factor B main effects by graphing the row or column means in a so-called ***main effects plot*** (or ***level means plot***), as in the next example.

Example 11.4: Main Effects Plots

For the study of phosphorus in two soil types and four topographies, the main effects plots, using the row means $\bar{Y}_{1\cdot} = 90.5$ and $\bar{Y}_{2\cdot} = 35.9$ and the column means $\bar{Y}_{\cdot 1} = 89.7$, $\bar{Y}_{\cdot 2} = 59.2$, $\bar{Y}_{\cdot 3} = 70.0$, and $\bar{Y}_{\cdot 4} = 34.0$ from Example 11.1, are shown below.

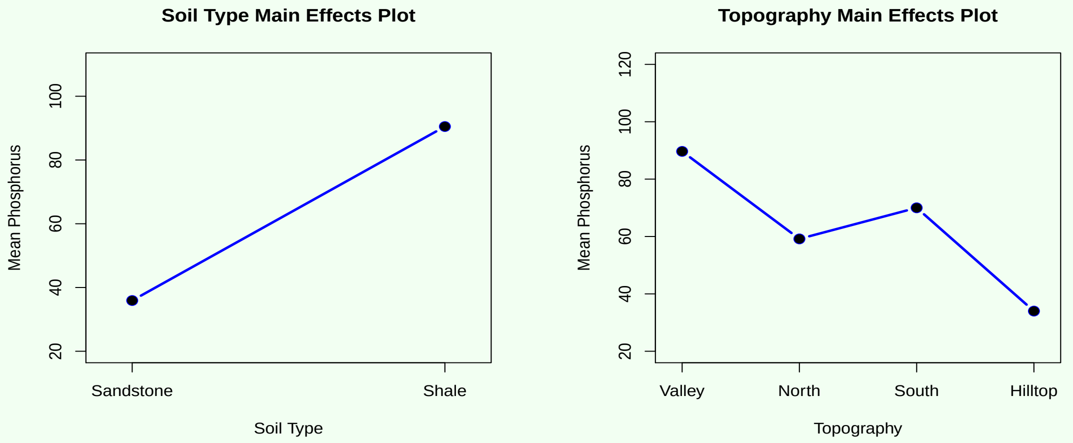


Figure 11.3: Plots of the main effects of soil type (left) and topography (right) on phosphorus concentrations.

The left plot shows that shale-based soil has a higher mean phosphorus concentration than sandstone-based soil. The right one shows that the phosphorus is highest in the valley and lowest on the hilltop.

The variation among row means is called ***between-rows variation*** (or ***factor A variation***), and the variation among column means is called ***between-columns variation*** (or ***factor B variation***). To decide if the between-rows or between-columns variation is more than can be explained by chance alone, measures of these types of variation will be compared to a measure of the variation that's due purely to random fluctuations in the response variable, called ***within-groups variation***.

11.2.4 Two-Factor ANOVA Model (Group Means Version)

We'll describe data from a two-factor study using a statistical model with *nonrandom* parts representing the between-rows and between columns variation in the data, that is, variation due to factor A and B main effects, and *random* part representing variation of individual observations within groups. The model will be consistent with the assumption that the observations in each group are a random sample from a normal population. The table below is a schematic of the normality assumption and depicts some of the notation we'll use for the model.

Factor B

$$\begin{array}{cc|c|c|c|c|c}
 & & \text{Level } j=1 & \text{Level } j=2 & \cdots & \text{Level } j=b & \\
\hline
 & \begin{array}{c}\text{Level}\\ i=1\end{array} & \left.\begin{array}{l} Y_{111}\\ Y_{112}\\ \vdots \\ Y_{11n}\end{array}\right\} \sim N(\mu_{11},\sigma) & \left.\begin{array}{l} Y_{121}\\ Y_{122}\\ \vdots \\ Y_{12n}\end{array}\right\} \sim N(\mu_{12},\sigma) & \cdots & \left.\begin{array}{l} Y_{1b1}\\ Y_{1b2}\\ \vdots \\ Y_{1bn}\end{array}\right\} \sim N(\mu_{1b},\sigma) & \mu_{1\cdot} \\
\hline
\begin{array}{c}\textbf{Factor}\\ \textbf{A}\end{array} & \begin{array}{c}\text{Level}\\ i=2\end{array} & \left.\begin{array}{l} Y_{211}\\ Y_{212}\\ \vdots \\ Y_{21n}\end{array}\right\} \sim N(\mu_{21},\sigma) & \left.\begin{array}{l} Y_{221}\\ Y_{222}\\ \vdots \\ Y_{22n}\end{array}\right\} \sim N(\mu_{22},\sigma) & \cdots & \left.\begin{array}{l} Y_{2b1}\\ Y_{2b2}\\ \vdots \\ Y_{2bn}\end{array}\right\} \sim N(\mu_{2b},\sigma) & \mu_{2\cdot} \\
\hline
 & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
\hline
 & \begin{array}{c}\text{Level}\\ i=a\end{array} & \left.\begin{array}{l} Y_{a11}\\ Y_{a12}\\ \vdots \\ Y_{a1n}\end{array}\right\} \sim N(\mu_{a1},\sigma) & \left.\begin{array}{l} Y_{a21}\\ Y_{a22}\\ \vdots \\ Y_{a2n}\end{array}\right\} \sim N(\mu_{a2},\sigma) & \cdots & \left.\begin{array}{l} Y_{ab1}\\ Y_{ab2}\\ \vdots \\ Y_{abn}\end{array}\right\} \sim N(\mu_{ab},\sigma) & \mu_{a\cdot} \\
\hline
 & & \mu_{\cdot 1} & \mu_{\cdot 2} & \cdots & \mu_{\cdot b} & \mu
\end{array}$$

Note that the population means $\boldsymbol{\mu_{ij}}$ depend on the level i of factor A and the level j of factor B, but population standard deviations are the same, $\boldsymbol{\sigma}$, regardless of the levels of the two factors. We call μ_{ij} the ***group true mean***.

The assumption that the observations in each group are a random sample from a normal population can be stated much more succinctly by saying that for each i and j, within the i,jth group,

$$Y_{ijk} \sim \mathrm{N}(\mu_{ij},\, \sigma),$$

with k taking the values $1, 2, \ldots, n$, and the observations Y_{ijk} are collected independently of each other.

Another way to state the assumption that the observations in each group are a random sample from a normal population is via the ***group means version*** of the ***two-factor ANOVA model***.

Two-Factor ANOVA Model (Group Means Version): A statistical model for describing data in random samples in a two-factor study (or randomized groups in a two-factor experiment) is:

$$Y_{ijk} \;=\; \mu_{ij} + \epsilon_{ijk}\,, \tag{11.1}$$

where

Y_{ijk} is the kth observation ($k = 1, 2, \ldots, n$) at the ith level of factor A ($i = 1, 2, \ldots, a$) and jth level of factor B ($j = 1, 2, \ldots, b$).

μ_{ij} is the mean of the population corresponding to the ith level of factor A and jth level of factor B (or the true mean response to the i,jth treatment) and is called the ***i, jth true group mean***.

ϵ_{ijk} is a random error term following a $\mathrm{N}(0, \sigma)$ distribution, and the ϵ_{ijk}'s are independent of each other.

The unknown ***model parameters*** are the group means $\mu_{11}, \mu_{12}, \ldots, \mu_{ab}$, and the standard deviation σ. In practice, these will need to be *estimated* from the data. We'll see how to estimate them in Section 11.2.8.

11.2.5 Additive Effects Two-Factor ANOVA Model

In the group means version of the ANOVA model, the group means μ_{ij} may differ with the levels of the two factors i and j, so *effects* of the factors are modeled via differences among the group means. If *neither* factor had an effect, the group means would all be equal.

Thus we *could* consider the ab groups as levels of a *single* factor and carry out a *one-factor* ANOVA F test of

$$\begin{aligned} H_0 &: \quad \mu_{11} = \mu_{12} = \cdots = \mu_{ab} \\ H_a &: \quad \text{Not all } \mu_{ij}\text{'s are equal,} \end{aligned}$$

But by doing it this way, even if we rejected the null hypothesis, we couldn't tell from that result alone *which* factor has an effect, only that at least one of them does.

Thus the group means version of the two-factor ANOVA model turns out to not be very useful. Instead, we'll use a version of the model that allows us to test separately for effects of the two factors. The model will have the form

$$Y \;=\; \text{Overall Mean} + \text{Factor } A \text{ Effect} + \text{Factor } B \text{ Effect} + \text{Error}$$

The model is called the ***additive effects two-factor ANOVA model***, and is defined more formally below. Later, in Section 11.2.7, we'll look at another version of the model for which the effects *aren't* additive.

Additive Effects Two-Factor ANOVA Model: Another statistical model for describing data in random samples in a two-factor study (or randomized groups in a two-factor experiment) is:

$$Y_{ijk} \;=\; \underbrace{\mu + \alpha_i + \beta_j}_{\text{This is } \mu_{ij}} + \epsilon_{ijk}\,, \tag{11.2}$$

where

Y_{ijk} is the kth observation ($k = 1, 2, \ldots, n$) at the ith level of factor A ($i = 1$, $2, \ldots, a$) and jth level of factor B ($j = 1, 2, \ldots, b$).
$\boldsymbol{\mu}$ is a constant called the ***overall true mean***.
$\boldsymbol{\alpha_i}$ is the ***effect*** of the ith level of factor A.
$\boldsymbol{\beta_j}$ is the ***effect*** of the jth level of factor B.
ϵ_{ijk} is a random error term following a N(0, σ) distribution, and the ϵ_{ijk}'s are independent of each other.

In this model, the (unknown) ***model parameters*** are μ, $\alpha_1, \alpha_2, \ldots, \alpha_a$, $\beta_1, \beta_2, \ldots, \beta_a$, and σ. In practice, their values will have to be estimated from the data. Their formal definitions will be given along with their estimators in Section 11.2.8.

The additive effects model re-expresses each group mean μ_{ij} in the group means version as an *overall mean plus an effect of factor A plus an effect of factor B*:

$$\mu_{ij} \;=\; \mu + \alpha_i + \beta_j\,.$$

This is the *nonrandom* part of the model, the *random* part being the error term. In practice, we're usually most interested the effects of factors A and B and generally not too interested in the overall mean μ. Later, we'll see how to carry out separate hypothesis tests for the effects of the two factors.

Example 11.5: Additive Two-Factor ANOVA Model

For the soil phosphorus study (Example 11.1), the *additive effects two-factor ANOVA model* for describing a phosphorus concentration Y is of the form

$$Y \ = \ \text{Overall Mean} + \text{Soil Type Effect} + \text{Topography Effect} + \text{Error}$$

In order for the additive effects model to accurately describe the variation a set of two-factor data, the effects of the two factors should truly be *additive*. We'll see what this means in the next section, and we'll also see how to check whether it's the case. In Section 11.2.7, we'll look at a different model that doesn't require the additivity assumption and *which should be used if we have any doubt as to whether or not the effects of the two factors are additive.*

11.2.6 Checking for Additivity of Effects

In order for the additive effects two-factor ANOVA model to accurately describe a set of two-factor data, the effects of the two factors should be *additive*. We say that the effects of factors A and B are ***additive*** if the effect of factor A is the same *regardless of the level of factor B*, and the effect of factor B is the same *regardless of the level of factor A*. For the soil phosphorus study, this would mean that the *difference* in phosphorus concentrations for the two soil types is the same regardless of whether it's on a hilltop, in a valley, or on a north- or south-facing slope, and the *differences* among concentrations for the four topographies are the same regardless of whether they're in shale- or sandstone-based soil.

When the effects of two factors *aren't* additive, we say that there's an *interaction effect* between them. An ***interaction effect*** is present when *the effect of each factor is different depending on the level of the other factor.*

To check for an interaction effect, we use a graph called an ***interaction plot***, which shows the group means (y axis) versus the levels of one of the factors (x axis), and connecting lines distinguishing levels of the other factor. As it turns out, there *is* an interaction effect between soil type and topography on phosphorus, as the interaction plot in the next example illustrates, and the hypothesis test in Example 11.11 will confirm.

Example 11.6: Interaction Plots

The eight group means for the soil phosphorus study (Examples 11.1 and 11.4) are shown in the main body of the table below.

Factor B: Topography

		Valley (j=1)	North-Facing (j=2)	South-Facing (j=3)	Hilltop (j=4)	
Factor A: Soil Type	Shale (i=1)	$\bar{Y}_{11} = 151.7$	$\bar{Y}_{12} = 85.0$	$\bar{Y}_{13} = 89.0$	$\bar{Y}_{14} = 36.3$	$\bar{Y}_{1\cdot} = 90.5$
	Sandstone (i=2)	$\bar{Y}_{21} = 27.7$	$\bar{Y}_{22} = 33.3$	$\bar{Y}_{23} = 51.0$	$\bar{Y}_{24} = 31.7$	$\bar{Y}_{2\cdot} = 35.9$
		$\bar{Y}_{\cdot 1} = 89.7$	$\bar{Y}_{\cdot 2} = 59.2$	$\bar{Y}_{\cdot 3} = 70.0$	$\bar{Y}_{\cdot 4} = 34.0$	$\bar{Y} = 63.2$

These means are graphed in interaction plots below.

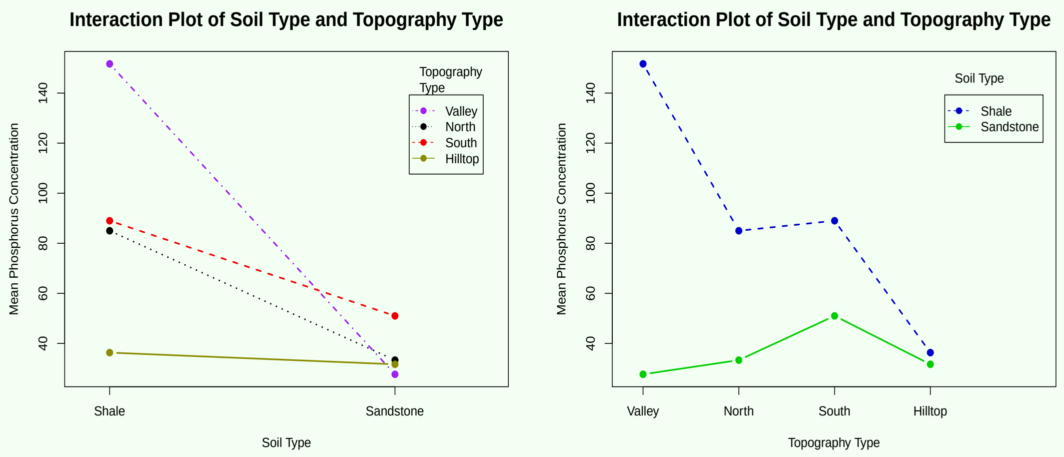


Figure 11.4: Interaction plots showing the effects of soil type on phosphorus for different topographies (left) and the effects of topography on phosphorus for different soil types (right).

In the left plot, each line connects the group means for the two soil types in a given topography. *Because the lines aren't parallel*, the effect of soil type appears to be different depending on the topography, that is, the effects of the two factors appear to be *non-additive*.

In the right plot, each line connects the group means for the four topographies in a given soil type. *The lines aren't parallel*, and it appears that the effect of topography is more pronounced in shale-based soil than in sandstone-based soil. This indicates (again) that the effects of the two factors are *aren't additive*.

The last example showed how to use interaction plots to check whether the effects of two factors are additive:

- If the lines in an interaction plot are (approximately) parallel, it suggests that the effects of the two factors are *additive*.
- If the lines are very non-parallel, it suggests that the effects are *not additive*, in other words, that there's an *interaction effect* between the two factors.

Here's an example in which the effects are additive.

Example 11.7: Interaction Plots

Consider the following (hypothetical) true group means.

		Factor B		
		Level $j = 1$	Level $j = 2$	Level $j = 3$
Factor A	Level $i = 1$	$\mu_{11} = 7.2$	$\mu_{12} = 7.8$	$\mu_{13} = 9.6$
	Level $i = 2$	$\mu_{21} = 9.6$	$\mu_{22} = 10.2$	$\mu_{23} = 12.0$

These group means μ_{ij} were generated from

$$\mu_{ij} \;=\; \mu + \alpha_i + \beta_j,$$

with $\mu = 9.4$, $\alpha_1 = -1.2$, $\alpha_2 = 1.2$, $\beta_1 = -1.0$, $\beta_2 = -0.4$, and $\beta_3 = 1.4$, so they're consistent with an *additive* effects model. Their interaction plots are below.

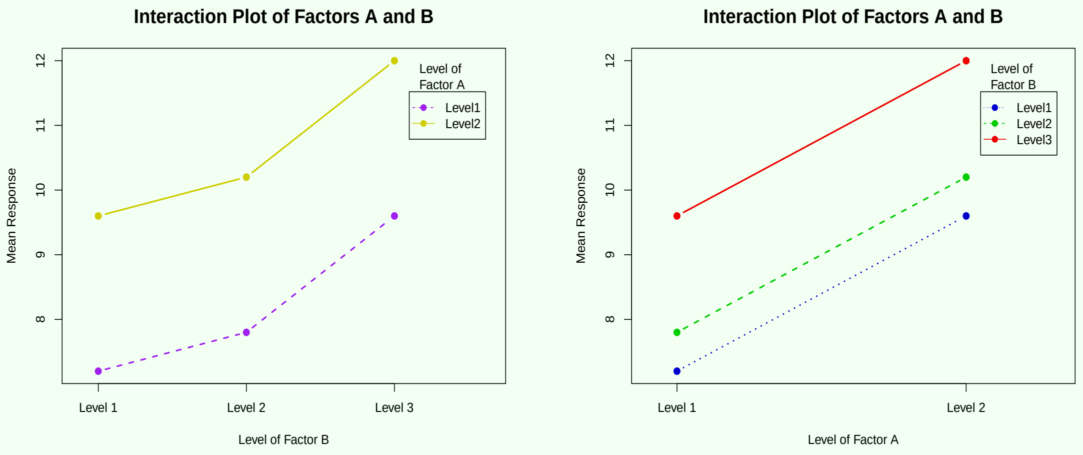


Figure 11.5: Interaction plots when the effects of factors A and B are additive.

In the left plot the ”trend” is the same for both lines, showing that the effect of factor B is the same regardless of the level of factor A. In the right plot the slopes of the lines are the same, which shows that factor B has the same effect regardless of the level of factor A.

Before turning to a model for non-additive factor effects, a few comments about the practical use of interaction plots are worth mentioning. First, for a given set of data, we'd only need to look at one of the interaction plots, not both, because it turns out that they'll always be in agreement regarding whether they suggest the effects are additive or not. Second, for real data the lines, in the plot will almost never be exactly parallel, even when the effects are additive, because of sampling error in the values of the group means. Later we'll see how to perform an F test to decide if an observed interaction effect is more than can be explained by chance variation.

11.2.7 Two-Factor ANOVA Model With Interaction Effect

When we suspect that there may be an interaction effect between two factors, the additive model won't adequately describe the data. Instead, we'll need to include in the model a term that represents the interaction effect that may be present.

Example 11.8: Two-Factor ANOVA Model With Interaction

Because the interaction plots of Example 11.5 suggest that the effects of soil type and topography aren't additive, we'll model a phosphorus concentration Y using a model having the form

$$Y \;=\; \text{Overall Mean} + \text{Soil Type Effect} + \text{Topography Effect} + \text{Interaction Effect} + \text{Error}$$

The model we'll use when the effects of two factors aren't additive is the ***two-factor ANOVA model with interaction effect***.

Two-Factor ANOVA Model With Interaction Effect: Another statistical model for describing data in random samples in a two-factor study (or randomized groups in a two-factor experiment) is:

$$Y_{ijk} \;=\; \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\text{This is } \mu_{ij}} + \epsilon_{ijk}\,, \tag{11.3}$$

where

Y_{ijk} is the kth observation ($k = 1, 2, \ldots, n$) at the ith level of factor A ($i = 1, 2, \ldots, a$) and jth level of factor B ($j = 1, 2, \ldots, b$).
μ is a constant called the *overall true mean.*
α_i is the *effect* of the ith level of factor A.
β_j is the *effect* of the jth level of factor B.
$\boldsymbol{(\alpha\beta)_{ij}}$ is called the ***interaction effect*** and represents the *combined* effect of the ith level of factor A and jth level of factor B *above and beyond* their additive effects.
ϵ_{ijk} is a random error term following a N(0, σ) distribution, and the ϵ_{ijk}'s are independent of each other.

In this model, the (unknown) ***model parameters*** are μ, $\alpha_1, \alpha_2, \ldots, \alpha_a$, $\beta_1, \beta_2, \ldots, \beta_a$, and now the ab interaction terms $(\alpha\beta)_{11}, (\alpha\beta)_{12}, \ldots, (\alpha\beta)_{ab}$, and σ. In practice, their values will have to be estimated from the data. Their formal definitions and their estimators will be given in Section 11.2.8.

The model with interaction effect modifies the additive effects version of the model by adding on a so-called ***interaction effect*** $\boldsymbol{(\alpha\beta)_{ij}}$ when re-expressing each group mean μ_{ij} in terms of the effects of the two factors,

$$\mu_{ij} \;=\; \mu \,+\, \alpha_i \,+\, \beta_j \,+\, (\alpha\beta)_{ij}\,.$$

This is the *nonrandom* part of the model. The *random* part is the error term. Later, we'll see how to carry out separate hypothesis tests for the main effects *and* the interaction effect of the two factors.

The interaction effect can take a different for each of the ab groups. Writing it as

$$(\alpha\beta)_{ij} \;=\; \mu_{ij} \,-\, (\mu + \alpha_i + \beta_j)$$

shows that it's the amount by which the true mean response in the i, jth group is heightened *above and beyond the additive effects* of the ith level of factor A and jth of factor B.

If the ab interaction effects $(\alpha\beta)_{11}, (\alpha\beta)_{12}, \ldots, (\alpha\beta)_{ab}$ are all equal to zero, the interaction model reduces to the additive effects version. In other word, the additive effects version is a special case of the interaction effects model (for which the interaction effects are all zero). This means that the interaction effects model is more widely applicable – it can be used regardless of whether or not the effects are additive – so it's generally preferred over the additive model for describing data from two-factor studies.

For this reason, the remainder of this chapter is focused mainly on the interaction effect model. We'll see later how to test the hypothesis that the interaction effects are all zero.

11.2.8 Model Parameter Estimates, Fitted Values, and Residuals

Because the two-factor ANOVA model *with* the interaction term can be used to describe data regardless of whether or not the effects of the factors are *additive*, we'll focus on that model. Thus from this point on, unless otherwise stated, all computational formulas and hypothesis test procedures given in this section are pertinent to the interaction effects version of the model, but not necessarily to the additive model. Computations and test procedures for the additive model, though, are similar.

Model Parameters and Their Estimators

As mentioned in Chapter 10, when a statistical model accurately reflects the true underlying process that generated a set of data, then *estimates* of the (unknown) model parameters will inform us about that process. For the two-factor ANOVA model, in order to be able to estimate the parameters, we'll need precise definitions of what they are.

We define

$\boldsymbol{\mu_{i\cdot}}$ = The average of the group means μ_{ij} at the ith level of factor A, called the ***ith true row mean*** or ***ith true factor A level mean***.

$\boldsymbol{\mu_{\cdot j}}$ = The average of the group means μ_{ij} at the jth level of factor B, called the ***jth true column mean*** or ***jth true factor B level mean***.

We also define

$\boldsymbol{\mu}$ = The average of all ab group means μ_{ij}, called the ***overall true mean***.

These true means are shown in the margins of the table in Subsection 11.2.4.

The definitions of the parameters α_i, β_j, and $(\alpha\beta)_{ij}$ are given below, along with their estimators based on the data and an alternative notational form for the estimators that will be useful later.

Model Parameter Estimators

Model Parameter	Estimator	Alternate Notation for the Estimator
μ_{ij}	$\bar{Y}_{ij}$	$\hat{\mu}_{ij}$
μ	$\bar{Y}$	$\hat{\mu}$
$\alpha_i = \mu_{i\cdot} - \mu$	$\bar{Y}_{i\cdot} - \bar{Y}$	$\hat{\alpha}_i$
$\beta_j = \mu_{\cdot j} - \mu$	$\bar{Y}_{\cdot j} - \bar{Y}$	$\hat{\beta}_j$
$(\alpha\beta)_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)$	$\bar{Y}_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}$	$\widehat{(\alpha\beta)}_{ij} = \hat{\mu}_{ij} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)$

Estimation of the parameter σ, the standard deviation of the $\mathrm{N}(0, \sigma)$ error distribution, will be covered in Subsection 11.2.13.

Comment: Formally, the true row, column, and overall means given above are defined by

$$\mu_{i\cdot} = \frac{\sum_{j=1}^{b} \mu_{ij}}{b}, \quad \mu_{\cdot j} = \frac{\sum_{i=1}^{a} \mu_{ij}}{a}, \quad \text{and} \quad \mu = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} \mu_{ij}}{ab}.$$

Fitted Values

Once the parameters of the two-factor ANOVA model have been estimated, we say that the model has been ***fitted*** to the data.

For each of the n individuals in a given group, we define the individual's ***fitted value*** (or ***predicted value***) to be the estimate of $\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

Fitted Values (Two-Factor ANOVA Model With Interaction Effect):

$$\begin{aligned}\text{Fitted Values for}\quad&\\ \text{Individuals in } i,j\text{th Group} \quad &= \hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+(\hat{\alpha\beta})_{ij}\\ &= \bar{Y}+(\bar{Y}_{i\cdot}-\bar{Y})+(\bar{Y}_{\cdot j}-\bar{Y})+(\bar{Y}_{ij}-\bar{Y}_{i\cdot}-\bar{Y}_{\cdot j}+\bar{Y})\\ &= \bar{Y}_{ij}\end{aligned}$$

Note that *the fitted values are just the group means* $\bar{Y}_{11}, \bar{Y}_{12}, \ldots, \bar{Y}_{ab}$. For the study of phosphorus in soil, the fitted values are the blue asterisks in the three-dimensional individual value plot in Fig. 11.2.

Residuals

The fitted values provide estimates of the nonrandom ”overall pattern” part of the two-factor ANOVA model. We'll also be interested in evaluating the random ”deviations” away from that pattern corresponding to the error term ϵ in the model.

The ***residual*** associated with the kth individual in the i, jth group, denoted $\boldsymbol{e_{ijk}}$, is the deviation of that individual's observed response away from its fitted value.

Residuals (Two-Factor ANOVA Model With Interaction Effect):

$$\begin{aligned} e_{ijk} &= Y_{ijk}-(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+(\hat{\alpha\beta})_{ij})\\ &= Y_{ijk}-\bar{Y}_{ij}\,. \end{aligned} \tag{11.4}$$

Thus *a residual is just a deviation of an observation* Y_{ijk} *away from its corresponding group mean* $\bar{Y}_{ij}$. In Fig. 11.2, the residuals are the vertical gaps between the points and the asterisks, and they correspond to random experimental error in the phosphorus measurements.

Note: Rearranging (11.4), we can write an observation Y_{ijk} as

$$Y_{ijk} \;=\; \hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+(\hat{\alpha\beta})_{ij}+e_{ijk}\,,$$

which has the form

Observed Value = Fitted Value + Residual.

Comparing this to the interaction two-factor ANOVA model (11.3) makes it clear that *the residual* e_{ijk} *approximates the random error term* ϵ_{ijk}. In Section 11.2.13, we'll use the residuals to estimate the standard deviation σ of the $\mathrm{N}(0,\sigma)$ error distribution, and in Section 11.2.20 we'll use them to check the normality assumption.

Because residuals are deviations away from group means, they sum to zero within each group.

Fact 11.2 In a two-factor study, for each $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b$, the residuals within the i, jth group sum to zero, that is,

$$\sum_{k=1}^{n} e_{ijk} \;=\; 0$$

for each fixed i and j.

11.2.9 The Triple Summation Notation

The *triple summation notation* is a convenient way to express a sum that's carried out over all rows (the leftmost $\sum$), columns (the middle $\sum$), and individuals within groups (rightmost $\sum$). As shown below, the operation can be carried out by first summing within each of the ab groups, calling the results T_{ij}, and then summing those group totals over all rows and columns.

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk} \;=\; \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\sum_{k=1}^{n} Y_{ijk}\right) \;=\; \sum_{i=1}^{a}\sum_{j=1}^{b} T_{ij}$$

11.2.10 Sums of Squares

Introduction

To decide if there's a statistically significant main effect of factor A, we'll compare the *between-rows variation* among the row means to the *within-groups variation* among individual observations within groups, and to decide if there's a statistically significant factor B main effect, we'll compare the *between-columns variation* to the *within-groups variation.* In this section, we'll look at sums of squares that measure these types of variation. We'll also look at a sum of squares that will be used to test for an interaction effect.

Between-Rows Variation

Between-rows variation refers to variation in the row means $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{a\cdot}$. In two-factor ANOVA, we measure this variation using the ***factor A sum of squares***, denoted **SSA** and defined as follows (where we invoke the alternate notation for the model parameter estimators given in Section 11.2.8).

Factor *A* Sum of Squares:

$$\text{SSA} \;=\; nb\sum_{i=1}^{a}(\bar{Y}_{i\cdot} - \bar{Y})^2 \;=\; nb\sum_{i=1}^{a}\hat{\alpha}_i^2\,.$$

The nb in front serves a purpose similar to that of the n in front of the treatment sum of squares in one-factor ANOVA – it helps make the between-rows variation comparable to the within-groups variation (see Section 10.2.11 of Chapter 10). The larger the differences are among the row means, the larger SSA will be, and it's in this sense that measures *between-rows* variation.

Between-Columns Variation

Between-columns variation refers to variation in the column means $\bar{Y}_{\cdot 1}, \bar{Y}_{\cdot 2}, \ldots, \bar{Y}_{\cdot b}$. We measure of this variation using the ***factor B sum of squares***, denoted **SSB** and defined as follows.

Factor *B* Sum of Squares:

$$\text{SSB} \;=\; na\sum_{j=1}^{b}(\bar{Y}_{\cdot j} - \bar{Y})^2 \;=\; na\sum_{j=1}^{b}\hat{\beta}_j^2\,.$$

The larger the differences are among the column means, the larger SSB will be, and it's in this sense that measures *between-columns* variation.

Nonadditive Between-Groups Variation

Nonadditive between-groups variation refers to variation among the group means $\bar{Y}_{11}, \bar{Y}_{12}, \ldots, \bar{Y}_{ab}$ *above and beyond* the variation that can be explained by the *addititve* effects of the two factors. We measure this variations by the ***AB interaction sum of squares***, denoted **SSAB** and defined by the following.

***AB* Interaction Sum of Squares**:

$$\text{SSAB} \;=\; n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y})^2 \;=\; n\sum_{i=1}^{a}\sum_{j=1}^{b}(\hat{\alpha\beta})_{ij}^{2}\,.$$

This sum of squares will be large when the group means aren't consistent with an additive model, which will be the case when the there's an interaction effect. Thus a large SSAB is an indication of an interaction effect.

Within-Groups Variation

Finally, the *within-groups variation* refers to variation of individual observations away from their corresponding group means or, put another way, variation in the residuals. This variation is measured by the ***error sum of squares***, also called the ***residual sum of squares***, denoted **SSE** and given by the following.

Error Sum of Squares:

$$\text{SSE} \;=\; \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{ij})^2 \;=\; \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} e_{ijk}^{2}\,.$$

The triple summation is needed because we're summing squared residuals for all N individuals, that is, for all n individuals $(k = 1, 2, \ldots, n)$ in every combination of the a rows $(i = 1, 2, \ldots, a)$ and b columns $(j = 1, 2, \ldots, b)$. The larger the residuals are, the larger SSE will be, and so SSE measures variation among individual observations *within-groups* due purely to random *experimental error.*

11.2.11 The Two-Factor ANOVA Partition of the Variation in the Data

Total Variation

In a two-factor study, we can think of variation in the data as arising either from one of the nonrandom effects of the two factors or from random experimental error due to heterogeneity among individuals in the samples and measurement error. The factor A effect (if any) contributes to between-rows variation, the factor B effect to between-columns variation, the interaction effect to nonadditive between-groups variation, and the random experimental error to within-groups variation (but it contributes to the other types of variation too).

We'll see shortly that if we were to combine the observations from all ab samples into one big sample, *all* of the variation in the data can be attributed either to the effects of the factors or to the random error.

The *total variation* in the data is measured by squaring the deviations of individual observations Y_{ijk} away from the overall mean $\bar{Y}$ and summing them, giving the ***total sum of squares***, denoted **SSTo**, as follows.

Total Sum of Squares:

$$\text{SSTo} \;=\; \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y})^2\,.$$

Because SSTo measures total variation in the data, it will be large if either the row, column, or group means differ substantially from each other or the observations within each group vary substantially. Therefore, SSTo reflects between-rows, between columns, nonadditive between-groups, and within-groups variation.

Partition of the Total Variation

The dissecting of the total variation in the data into parts attributable to effects of the factors and random error, described at the beginning of this section, is called the ***two-factor ANOVA partition*** of the total variation, and is stated formally in the following fact.

Fact 11.3 The sums of squares defined above satisfy the following relation.

$$\text{SSTo} \;=\; \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}. \tag{11.5}$$

This decomposes the variation in the data as:

$$\begin{aligned}\text{Total Variation} \;=\;& \text{Between-Rows Variation} + \text{Between-Columns Variation}\\ &+ \text{Nonadditive Between-Groups Variation} + \text{Within-Groups Variation}\end{aligned}$$

A mathematical verification of the two-factor ANOVA partition is given in Subsection 11.2.21.

Example 11.9: Sums of Squares and the ANOVA Partition

For the data from the soil phosphorus study (Example 11.1), statistical software reports the following sums of squares.

$$\begin{aligned}\text{SSTo} &= 51406.0 && \text{(Total variation)}\\ \text{SSA} &= 17876.0 && \text{(Variation due to soil type)}\\ \text{SSB} &= 9693.8 && \text{(Variation due to topography)}\\ \text{SSAB} &= 11390.8 && \text{(Variation due to soil type, topography interaction)}\\ \text{SSE} &= 12445.3 && \text{(Variation due to random error)}\end{aligned}$$

We see that the two-factor ANOVA partition holds since

$$51406.0 \;=\; 17876.0 + 9693.8 + 11390.8 + 12445.3.$$

11.2.12 Degrees of Freedom

As for one-factor ANOVA, the ***degrees of freedom*** (or ***df***) associated with a sum of squares is the number of deviations, among those used to compute the sum of squares, that are ”free to vary.” The reason why degrees of freedom are important is because later, when we carry out hypothesis tests for factor effects, they'll determine the F distributions from which p-values are obtained.

Here are the degrees of freedom associated with the sums of squares in two-factor ANOVA.

Degrees of Freedom: For two-factor ANOVA, the degrees of freedom are:

$$\begin{aligned} df \text{ for SSTo} &= N-1 \\ df \text{ for SSA} &= a-1 \\ df \text{ for SSB} &= b-1 \\ df \text{ for SSAB} &= (a-1)(b-1) \\ df \text{ for SSE} &= ab(n-1) \;=\; N-ab \end{aligned}$$

An explanation for why these degrees of freedom correctly reflect the number of deviations that are ”free to vary” will be given in Subsection 11.2.22.

As was the case for one-factor ANOVA, the degrees of freedom, like their corresponding sums of squares, are additive in the following sense.

Fact 11.4 The degrees of freedom given above satisfy the following relation.

$$df \text{ for SSTo} \;=\; df \text{ for SSA} \;+\; df \text{ for SSB} \;+\; df \text{ for SSAB} \;+\; df \text{ for SSE}. \tag{11.6}$$

Example 11.10: Degrees of Freedom

Continuing from the previous soil phosphorus example, we have $a = 2$ soil types, $b = 4$ topographies, and $n = 3$ phosphorus observations per group. Thus the total number of phosphorus observations is $N = 24$, and we have

$$\begin{aligned} df \text{ for SSTo} &= 23 \\ df \text{ for SSA} &= 1 \\ df \text{ for SSB} &= 3 \\ df \text{ for SSAB} &= 3 \\ df \text{ for SSE} &= 16. \end{aligned}$$

As expected, (11.6) holds because $23 = 1 + 3 + 3 + 16$.

11.2.13 Mean Squares

Introduction

As for one-factor ANOVA, a ***mean square*** for two-factor ANOVA is defined as a sum of squares divided by its degrees of freedom. The ***mean square for factor A***, denoted **MSA**, ***mean square for factor B***, denoted **MSB**, ***mean square for an AB interaction***, denoted **MSAB**, and ***mean squared error***, denoted **MSE**, are defined below.

Mean Squares: For two-factor ANOVA, the mean squares are

$$\text{MSA} = \tfrac{\text{SSA}}{a-1} \qquad \text{MSB} = \tfrac{\text{SSB}}{b-1}$$

$$\text{MSAB} = \tfrac{\text{SSAB}}{(a-1)(b-1)} \qquad \text{MSE} = \tfrac{\text{SSE}}{ab(n-1)}$$

MSA, MSB, MSAB, and MSE Under H_0 and H_a

The factor A mean square MSA measures variation among the row means $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{a\cdot}$, which will vary when there's a factor A effect, but also to a lesser extent just due to random sampling error. Therefore, to test for a factor A effect, we'll need to distinguish between variation in $\bar{Y}_{i\cdot}$'s that's due purely to sampling error and variation that's due a factor A effect (*in addition to* sampling error). We'll do this by comparing the MSA to the MSE, which measures variation that's due purely to random error. Likewise, to test for a factor B effect and an AB interaction effect, we'll compare the MSB and MSAB, respectively, to the MSE.

More formally, consider the following three sets of hypotheses, for which the null hypothesis in each case says there's *no effect* and the alternative says there's an *effect*:

1. Concerning a ***factor A main effect***:

$$\begin{array}{lcl} H_0^A : \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{a\cdot} & \quad\text{or}\quad & H_0^A : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \\ H_a^A : \text{Not all } \mu_{i\cdot}\text{'s are equal} & & H_a^A : \text{Not all } \alpha_i\text{'s equal } 0 \end{array}$$

2. Concerning a ***factor B main effect***:

$$\begin{array}{lcl} H_0^B : \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot b} & \quad\text{or}\quad & H_0^B : \beta_1 = \beta_2 = \cdots = \beta_b = 0 \\ H_a^B : \text{Not all } \mu_{\cdot j}\text{'s are equal} & & H_a^B : \text{Not all } \beta_i\text{'s equal } 0 \end{array}$$

3. Concerning an ***AB interaction effect***:

$$\begin{array}{l} H_0^{AB} : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab} = 0 \\ H_a^{AB} : \text{Not all } (\alpha\beta)_{ij}\text{'s equal } 0 \end{array}$$

The following facts will be used later to test these hypotheses.

Fact 11.5 Consider random samples from ab populations in a two-factor study. Suppose that the two-factor ANOVA model (11.3) is appropriate and that the ϵ_{ijk}'s are independent and $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma)$.

Then the MSE is *always* an *unbiased* estimator of σ^2 (regardless of which of the above hypotheses are true). On the other hand,

1. - MSA is an *unbiased* estimator of σ^2 when H_0^A is true.
   - But MSA will tend to *overestimate* σ^2 when H_a^A is true.
2. - MSB is an *unbiased* estimator of σ^2 when H_0^B is true.
   - But MSB will tend to *overestimate* σ^2 when H_a^B is true.
3. - MSAB is an *unbiased* estimator of σ^2 when H_0^{AB} is true.
   - MSAB will tend to *overestimate* σ^2 when H_a^{AB} is true.

Estimating σ

Because the MSE estimates σ^2, the common population variance, or equivalently, the variance of the N(0, σ) distribution of the error term ϵ in the ANOVA model, regardless of whether or not either of the factors has an effect, its square root estimates σ.

Estimator of $\boldsymbol{\sigma}$: For a two-factor study, the estimator of σ, denoted $\hat{\boldsymbol{\sigma}}$, is

$$\hat{\sigma} \;=\; \sqrt{\text{MSE}}.$$

11.2.14 Two-Factor ANOVA *F* Tests

Hypotheses and Test Statistics

For the three sets of hypotheses given in the previous section, here are the corresponding ***two-factor ANOVA F test statistics***.

Two-Factor ANOVA *F* Test Statistics:

1. For a ***factor A main effect***:

   Two-Factor ANOVA *F* Test Statistic for Factor *A*:

$$F_A \;=\; \frac{\text{MSA}}{\text{MSE}}$$

2. For a ***factor B main effect***:

   Two-Factor ANOVA *F* Test Statistic for Factor *B*:

$$F_B \;=\; \frac{\text{MSB}}{\text{MSE}}$$

3. For an ***AB interaction effect***: **Two-Factor ANOVA *F* Test Statistic for the *AB* Interaction**:

$$F_{AB} \;=\; \frac{\text{MSAB}}{\text{MSE}}$$

Properties of the *F* Test Statistics: The following will help us interpret the observed values of the F test statistics.

1. - We can think of F_A as

$$F_A \;=\; \frac{\text{Between-Rows Variation}}{\text{Within-Groups Variation}}$$

   - If H_0^A was true, we'd expect $F_A \approx 1$ because MSA and MSE would both estimate σ^2.
   - But if H_a^A was true, we'd expect $F_A > 1$ because MSE would estimate σ^2 (still) but MSA would likely overestimate σ^2.

2. - We can think of F_B as

$$F_B \;=\; \frac{\text{Between-Columns Variation}}{\text{Within-Groups Variation}}$$

   - If H_0^B was true, we'd expect $F_B \approx 1$ because MSB and MSE would both estimate σ^2.
   - But if H_a^B was true, we'd expect $F_B > 1$ because MSE would estimate σ^2 (still) but MSB would likely overestimate σ^2.

3. - We can think of F_{AB} as

$$F_{AB} \;=\; \frac{\text{Nonadditive Between-Groups Variation}}{\text{Within-Groups Variation}}$$

- If H_0^{AB} was true, we'd expect $F_{AB} \approx 1$ because MSAB and MSE would both estimate σ^2.
- But if H_a^{AB} was true, we'd expect $F_{AB} > 1$ because MSE would estimate σ^2 (still) but MSAB would likely overestimate σ^2.

Therefore,

Large values of F_A, F_B, and F_{AB} (larger than about 1) provide evidence in favor of H_a^A, H_a^B, or H_a^{AB}, respectively.

P-Values

To decide whether an observed value of F_A, F_B, or F_{AB} provides statistically significant evidence in support of the alternative hypothesis, we'll need to know their sampling distributions under the null hypotheses.

Sampling Distributions of $\boldsymbol{F_A}$, $\boldsymbol{F_B}$, and $\boldsymbol{F_{AB}}$ under $\boldsymbol{H_0^A}$, $\boldsymbol{H_0^B}$, and $\boldsymbol{H_0^{AB}}$: Consider random samples from ab populations in a two-factor study. Suppose that the two-factor ANOVA model (11.3) is appropriate and that the ϵ_{ijk}'s are independent and $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma)$. Then

1. When H_0^A is true (so there's no factor A effect),
$$F_A \;\sim\; F(a-1,\; ab(n-1)).$$
2. When H_0^B is true (so there's no factor B effect),
$$F_B \;\sim\; F(b-1,\; ab(n-1)).$$
3. When H_0^{AB} is true (so there's no AB interaction effect),
$$F_{AB} \;\sim\; F((a-1)(b-1),\; ab(n-1)).$$

Because *large* values of the F statistics provide evidence against the null hypotheses, the rejection regions for the ANOVA F tests are comprised of F values in the *rightmost* $100\alpha\%$ of the F distributions, and the p-values are the tail probabilities to the *right* of the observed F values.

The Two-Factor ANOVA F Test Procedures

The two-factor ANOVA F test procedures are summarized in the table below.

Two-Factor ANOVA *F* Tests

Assumptions: Data are random samples from ab populations corresponding to combinations of the levels of two factors A and B (or they're responses to treatments in a randomized two-factor experiment), the two-factor ANOVA model (11.3) is appropriate, the ϵ_{ijk}'s are independent and either they follow a N(0, σ) distribution or the ab sample sizes n are all large.*

Null hypotheses:

1. $H_0^A : \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0.$
2. $H_0^B : \beta_1 = \beta_2 = \ldots = \beta_b = 0.$
3. $H_0^{AB} : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{ab} = 0.$

Test statistic values: $F_A = \dfrac{\text{MSA}}{\text{MSE}}, \quad F_B = \dfrac{\text{MSB}}{\text{MSE}}, \quad F_{AB} = \dfrac{\text{MSAB}}{\text{MSE}}.$

Decision rule: Reject any H_0 if corresponding p-value $< \alpha$ or F is in rejection region.

Alternative hypotheses	P-value = area	Rejection region = **
1. H_a : Not all α_i's equal 0.	to the right of F_A under $F(a-1, N-ab)$ distribution	F_A values such that $F_A \geq F_{\alpha,a-1,N-ab}$
2. H_a : Not all β_j's equal 0.	to the right of F_B under $F(b-1, N-ab)$ distribution	F_B values such that $F_B \geq F_{\alpha,b-1,N-ab}$
3. H_a : Not all $(\alpha\beta)_{ij}$'s equal 0.	to the right of F_{AB} under $F((a-1)(b-1), N-ab)$ distribution	F_{AB} values such that $F_{AB} \geq F_{\alpha,(a-1)(b-1),N-ab}$

* Two-factor ANOVA F tests can also be carried out (using statistical software) when the ab sample sizes $n_{11}, n_{12}, \ldots, n_{ab}$ aren't all the same. The sample sizes are considered to be large if they're all at least 15, unless the samples exhibit strong skewness, in which case they should all be at least 40.

** $F_{\alpha,m,n}$ is the $100(1-\alpha)$th percentile of the F distribution with m and n d.f.

11.2.15 The ANOVA Table

The results of a two-factor analysis of variance (degrees of freedom, sums of squares, mean squares, observed F test statistic values, and p-values) are usually summarized in a ***two-factor ANOVA table*** having the form shown below.

Source	DF	SS	MS	F	P-value
Factor A	$a-1$	SSA	MSA = SSA/$(a-1)$	F =MSA/MSE	p
Factor B	$b-1$	SSB	MSB = SSB/$(b-1)$	F =MSB/MSE	p
Interaction	$(a-1)(b-1)$	SSAB	MSAB = SSAB/$((a-1)(b-1))$	F =MSAB/MSE	p
Error	$N-ab$	SSE	MSE = SSE/$(N-ab)$		
Total	$N-1$	SSTo			

The table is arranged so that the first row (labeled ”Factor A”) pertains to *between-rows* variation, the second row (”Factor B”) to *between-columns* variation, the third (”Interaction”) to *nonadditive between-groups* variation, the fourth (”Error”) to *within-groups* variation, and the last (”Total”) to *total* variation.

11.2.16 Procedure for Reading the ANOVA Table

It’s important to use the following procedure when assessing significance of effects in a two-factor ANOVA. ***First look at the test results for the AB interaction effect***.

1. If the interaction is statistically significant, there’s *no need proceed* to the tests for main effects because we *already know*, *regardless of what those test results are*, that both factors have effects (but each factor’s effect differs according to the level of the other factor). Instead, examine the effects of factor A *separately* for each level of factor B, or vice versa, for example by performing multiple comparison tests on the *group means* as described in Section 11.5.

2. If the interaction *isn’t* statistically significant, proceed to the tests for main effects. If either main effect is significant, perform multiple comparison tests on the *level means* for that factor, as described in Section 11.5, to decide *which* levels of the factor differ from each other.

We’ll see later that an interaction effect can ”mask” the effects of the two factors, rendering the main effect test results *not* statistically significant even though both factors have effects. Before looking at the ”masking” problem, though, here are a few examples illustrating the procedure for reading the ANOVA table.

Example 11.11: Two-Factor ANOVA Table

For the soil phosphorus study (Example 11.1), the ANOVA table (obtained using software) is below.

Source	DF	SS	MS	F	P-value
Soil Type	1	17876.0	17876.0	22.98	0.000
Topography Type	3	9693.8	3231.3	4.15	0.024
Interaction	3	11390.8	3796.9	4.88	0.013
Error	16	12445.3	777.8		
Total	23	51406.0			

We first look at the results of the F test for an interaction effect between soil type by topography. The null hypothesis is that there’s *no interaction effect*, that is, that the effects of soil type and topography and are *additive*. The observed test statistic value is $F = 4.88$ and the p-value, from the right tail of the $F(3, 16)$ distribution, is 0.013. Thus, using level of significance $\alpha = 0.05$, the interaction effect is statistically significant. In other words, the effects of topography and soil type aren’t additive, meaning that the effect of soil type on phosphorus is different depending on the topography (and the effect of topography is different depending on the soil type).

At this point *there’s no need to proceed* with the tests for main effects because we already know, no matter what the results of those tests say, that both soil type and topography have effects, but their effects are different depending on the level of the other factor. Instead we should investigate the effect of soil type separately for each topography (or the effect of topography separately for each soil type).

Example 11.12: Two-Factor ANOVA Table

For the study of the effects of water type and pH on percent recovery of ammonia (Example 11.2), the ANOVA table (obtained using software) is below.

Source	DF	SS	MS	F	P-value
Water Type	1	29.4	29.4	2.84	0.118
PH Level	2	8.4	4.2	0.41	0.673
Interaction	2	21.8	10.9	1.05	0.379
Error	12	124.0	10.3		
Total	17	183.6			

The first step is to examine the results of the F test for an interaction effect between water type and pH level. The null hypothesis is that there's *no interaction effect*, that is, that the effects of water type and pH are *additive.* The observed test statistic value is $F = 1.05$ and the p-value, from the $F(2, 12)$ distribution, is 0.379. Thus, using $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that there's no interaction effect between water type and pH. In other word, their effects are additive.

Because there's no interaction effect, it makes sense to proceed to the tests for main effects. For the pH main effect, the null hypothesis says pH has *no effect.* The observed test statistic value is $F = 0.41$ and the p-value, from the $F(2, 12)$ distribution, is 0.673. Thus, using $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the pH level has no effect on the ammonia percent recovery.

Finally, for the water type main effect, the null hypothesis says water type has *no effect.* The test statistic value is $F = 2.84$ and the p-value, from the $F(1, 12)$ distribution, is 0.118. Using $\alpha = 0.05$, we again fail to reject the null hypothesis and conclude that water type has no effect on the ammonia percent recovery.

11.2.17 Main Effects Masked by an Interaction Effect

The next example shows how main effects can be ”masked” by an interaction effect when the effects of a factor are discordant depending on the level of the other factor.

Example 11.13: Main Effects Masked by Interaction

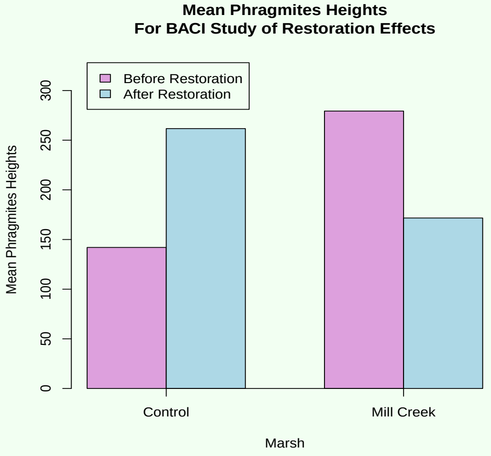
On June 8, 2000 the oil tanker T/V Posavina was rammed and punctured by its own tug, spilling 59,000 gallons of oil into the Chelsea River, Chelsea, Massachusetts, oiling a substantial stretch of shoreline. The response to the oil spill included restoration in fall, 2005 of a degraded 1.5-acre salt marsh along Mill Creek, located on the upper reach of the Chelsea River. The marsh restoration involved removal of roots and rhizomes of the invasive reed *Phragmites australis.*

A *before-after-control-impact* study was carried out to examine the effectiveness of the restoration project in decreasing Phragmites cover and increasing the cover of native plants. Heights of Phragmites' were measured in the Mill Creek marsh and an adjacent, unrestored control marsh on September 1, 2005, just before the restoration of the Mill Creek marsh, and again in 2007, two years after the restoration [2].

The table below shows, for three 1 m^2 quadrats selected from each marsh before the restoration and three selected from each marsh after the restoration, the mean height (cm) of the three tallest Phragmites plants in the quadrat.

		Period		
		Before	After	
Site	Control	64	179	$\bar{Y}_{1\cdot} = 201.8$
		80	300	
		282	306	
	Mill Creek	254	210	$\bar{Y}_{2\cdot} = 225.5$
		300	154	
		284	154	
		$\bar{Y}_{\cdot 1} = 210.7$	$\bar{Y}_{\cdot 2} = 216.7$	

The goal is to decide if the heights of Phragmites decreased more at the restored Mill Creek site than at the unrestored control site. A bar plot of the group means is below.



The ANOVA table (obtained using software) is below.

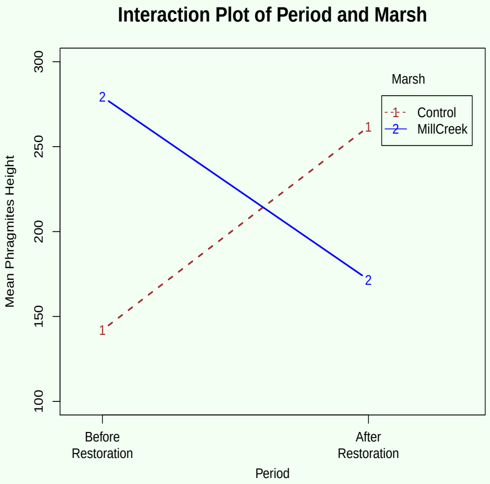
Source	DF	SS	MS	F	P-value
Period	1	108	108	0.020	0.8909
Site	1	1680	1680	0.312	0.5918
Interaction	1	38760	38760	7.195	0.0278
Error	8	43096	5387		
Total	11	83644			

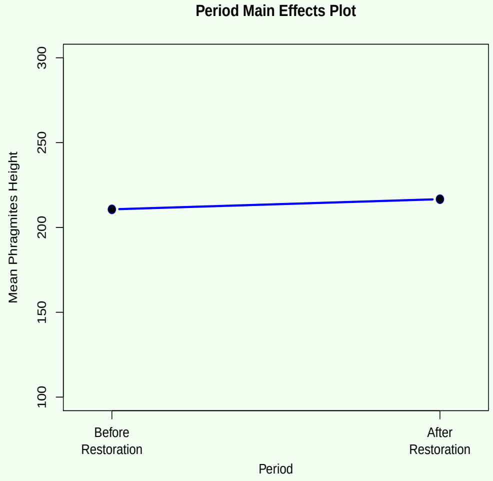
The interaction effect is statistically significant, indicating that Phragmites heights decreased more at the Mill Creek site than at the control site.

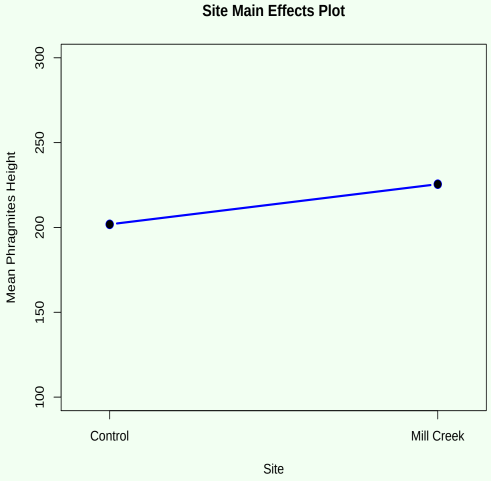
Notice, though, that neither the period nor the site main effect is significant. It's tempting to conclude that neither of these factors has any effect, but because the interaction is significant, we know that they both have effects, but the effect of each is different depending on the level of the other.

To see why neither main effect was statistically significant, we can look at the group means, row means, and column means, shown in the table below, and at an interaction plot and the level means plots, also below.

		Period		
		Before	After	
Site	Control	$\bar{Y}_{11} =$ **142.0**	$\bar{Y}_{12} =$ **261.7**	$\bar{Y}_{1\cdot} =$ 201.8
	Mill Creek	$\bar{Y}_{21} =$ **279.3**	$\bar{Y}_{22} =$ **171.7**	$\bar{Y}_{2\cdot} =$ 225.5
		$\bar{Y}_{\cdot 1} =$ 210.7	$\bar{Y}_{\cdot 2} =$ 216.7	







During a given period, the means for the two sites are very different, but when averaged over the two periods, the discordance in the effects of period for the two sites lead to site means that are so close that the site main effect isn't significant. We say that the effect of site is ”masked” by the interaction. A similar phenomenon explains why the period effect is ”masked.”

11.2.18 One Observation Per Group

Introduction

Occasionally, we're only able to make one observation at each combination of the levels of the factors in a two-factor study, so the common sample size is $n = 1$. This might be the case, for example, when the data collection process is prohibitively expensive or time consuming.

Example 11.14: One Observation Per Group

Carbonyls are a group of chemical compounds that play an important role in atmospheric chemical processes, including the production of smog and ozone.

Atmospheric carbonyls are measured by taking air samples through cartridges that trap the carbonyls on adsorbent material coated with dinitrophenylhydrazine, where they're converted to hydrazone derivatives for later analysis using high performance liquid chromatography.

An experiment was carried out to investigate the potential bias associated with the cartridge-based sampling method for measuring acetaldehyde, a member of the carbonyls group [7]. Two factors suspected of influencing measurement bias were sampling volume and concentration level.

A supply of standard gas with known acetaldehyde concentration was purchased for use in the experiment. Specimens of the standard were then prepared to four concentrations, using dilution factors of 1000 (S1), 200 (S2), 100 (S3), and 50 (S4), and sampled at four total sampling volumes, 1, 2, 4, and 20 L. One gas specimen was analyzed under each of the 16 experimental conditions, giving one observation per group. The response variable is the acetaldehyde recovery rate, defined as

$$\text{Recovery Rate} = \frac{\text{Measured Mass } (\mu\text{g})}{\text{Known True Mass } (\mu\text{g})} \cdot 100\%$$

The results of the experiment are below.

Sampling Volume

		1 L	2 L	4 L	20 L	
	S1	78	67	78	63	$\bar{Y}_{1\cdot} = 71$
Concentration Level	S2	78	73	77	76	$\bar{Y}_{2\cdot} = 76$
	S3	80	81	80	82	$\bar{Y}_{3\cdot} = 81$
	S4	107	106	106	106	$\bar{Y}_{4\cdot} = 106$

$\bar{Y}_{\cdot 1} = 86$ $\quad \bar{Y}_{\cdot 2} = 82$ $\quad \bar{Y}_{\cdot 3} = 85$ $\quad \bar{Y}_{\cdot 4} = 82$

Two-Factor ANOVA with One Observation Per Group

When there's only one observation per group in a two-factor study, it turns out that *if the interaction term is included in the model, we can't carry out F tests for effects.* Thus, if we're willing to assume that the effects of the two factors are additive, we use the *additive model.*

To see why we can't test for main effects if we include the interaction in the model, note that each group mean would equal the one observation in the group,

$$\bar{Y}_{ij} \;=\; Y_{ijk},$$

so *the residuals would all equal zero*,

$$\begin{aligned} e_{ijk} &= Y_{ijk} - \bar{Y}_{ij} \\ &= Y_{ijk} - Y_{ijk} \\ &= 0 \end{aligned}$$

This would lead to both an error sum of squares SSE and a mean squared error MSE of zero. In other words, *all* of the variation in the data would be explained by the nonrandom part of the model, which would be said to ***saturate*** the variation. Because MSE is the denominator of the F test statistics, we'd be *unable to perform F tests* for main effects or an interaction effect.

But if instead we fitted the *additive* model, the fitted values would be

$$\begin{aligned}\text{Fitted Value} &= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j \\ &= \bar{Y} + (\bar{Y}_{i\cdot} - \bar{Y}) + (\bar{Y}_{\cdot j} - \bar{Y}) \\ &= \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}\end{aligned}$$

and thus the residuals would be

$$\begin{aligned}e_{ijk} &= Y_{ijk} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) \\ &= Y_{ijk} - (\bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y})\,.\end{aligned}$$

These no longer would all equal zero, and so the F tests (for main effects) could be carried out.

Example 11.15: One Observation Per Group

For the experiment to investigate the bias associated with measuring acetaldehyde, the additive model was fitted to the data using software, giving the ANOVA table below.

Source	DF	SS	MS	F	P-value
Concentration Level	3	2916.9	972.3	63.01	0.000
Sampling Volume	3	50.9	17.0	1.10	0.398
Error	9	33.431	8.358		
Total	15	3106.7			

Because we fitted the *additive* model, there's no interaction term in the ANOVA table. Based on the results, we conclude, using $\alpha = 0.05$, that sampling volume has no effect on the acetaldehyde recovery rate, but concentration does have an effect.

11.2.19 Blocking: An Extension of Matched Pairs Study Designs

In Chapter 9, *matched pairs study designs* were used to "filter out" from a data set unwanted variation due to heterogeneity among individuals, such as sites or days, when testing for a difference between two groups, such as "before" versus "after" or "upstream" versus "downstream." In such studies, each pair is an example of what's called a ***block***, or set of individuals that are *homogeneous* with respect to some characteristic thought to be related to the response variable. The characteristic by which individuals are matched is an example of a so-called ***blocking factor***.

A ***block design*** is an extension of the matched pairs design in which *more than two* groups are compared. In a *block design*, instead of each *block* consisting of a pair of individuals, it consists of one individual for each of the groups being compared. In such studies, there's a ***factor of interest*** that defines the groups and whose effect we want to ascertain, but the blocking factor is of no interest except to remove its masking of the effect of the factor of interest.

Example 11.16: Blocking

Mercury concentrations were measured in periphyton (freshwater organisms that cling to surfaces) at six stations along the South River, Virginia, in the vicinity of a large mercury contamination site

[5]. Measurements were made on six different dates. The data are below.

Mercury

Date	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6
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

Of interest is whether the six stations differ in mercury concentration. Although we *could* run this as a one-factor ANOVA, there may be differences among the six dates (for example, the periphyton may not take up mercury as quickly during some seasons as others). Differences caused by sampling on six different dates are unwanted noise that contributes to within-groups variation, potentially making detection of a station effect difficult.

Instead of running the one-factor ANOVA, we'll consider date to be a blocking factor, and in order to filter out the unwanted variation due to date, we'll run a two-factor ANOVA (using the additive model since there's only one observation per group), with date and station as the factors.

The resulting ANOVA table is below.

Source	DF	SS	MS	F	P-value
Date	5	3.26	0.65	0.37	0.864
Station	5	230.13	46.03	26.14	0.000
Error	25	44.02	1.76		
Total	35	277.41			

To see the benefit of using date in the model as a blocking factor, in comparison, if we had run a one-factor ANOVA without blocking, the error sum of squares SSE would have been 47.38. With date included in the model, as above, this SSE is split into two parts, the SSE (for the two-factor model), 44.02, and the sum of squares for the blocking factor SSA, 3.26. The variation due to heterogeneity between blocks (dates) is thereby removed from the background noise (MSE). If there is an appreciable block effect, removal of the SSA lowers the SSE and MSE in comparison to their values for a one-factor ANOVA. This produces a higher F statistic, allowing the effect of the factor of interest to be more easily discerned.

11.2.20 Using Residuals to Check the ANOVA F Test Assumptions

The ANOVA F tests rely on three assumptions:

1. The observations Y_{ijk} in the ab samples are independent of each other, or equivalently, the errors ϵ_{ijk} in the ANOVA model are independent.
2. The observations in the samples (or responses to the treatments) follow normal distributions, or equivalently, the errors follow a normal distribution.
3. The ab populations (or responses to the ab treatments) have a common standard deviation σ, or equivalently, the standard deviation σ of the error distribution is the same from one group to the next.

The first assumption (independence) is usually addressed in the study design by separating observations sufficiently in space and time. The other assumptions (normality and common σ) are checked via plots of the residuals.

Checking the Normality Assumption

Instead of checking the normality assumption separately for each group, it's usually preferable check it by plotting the N *residuals* together in a histogram or normal probability plot. The normality assumption is tenable as long as the plot doesn't show strong signs of non-normality.

Checking the Common σ Assumption

There are a few ways to check the common standard deviation assumption.

1. **Plot the residuals versus the fitted values**: We can look at an individual value plot of the residuals versus the group means (fitted values), with a horizontal line at $y = 0$. The amount of vertical spread above and below the line should be roughly the same from one group to the next, and in particular, it shouldn't increase with the mean (fitted value). See Fig. 10.10 in Chapter 10 for an example.
2. **Compare sample standard deviations**: Another way to check the common σ assumption is to compare the sample standard deviations. As stated in Subsection 10.2.13 of Chapter 10, if the largest of the ab sample standard deviations is less than twice as large as the smallest, then it's reasonable to assume that the population standard deviations are equal. This is meant as a rough guideline only.

Example 11.17: Checking Assumptions

For the soil phosphorus study, a histogram and normal probability plot of the residuals are below.

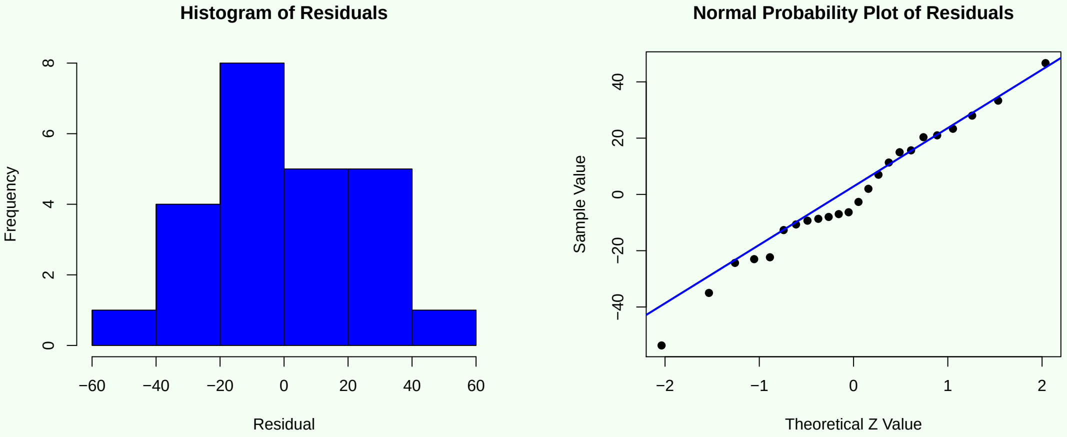


Figure 11.6: Histogram (left) and normal probability plot (right) of the residuals after fitting the two-factor ANOVA model to the soil phosphorus data.

The plots show that the normality assumption appears to be met. A plot the residuals versus the fitted values (group means) is below.

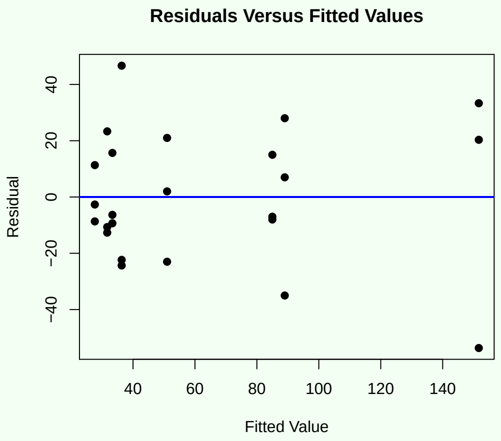


Figure 11.7: Plot of residuals versus fitted values after fitting the two-factor ANOVA model to the soil phosphorus data.

The amount of (vertical) spread of the points appears to be roughly constant from left to right, so it seems reasonable to assume that the true (unknown) standard deviation σ is the same from one group to the next.

Because the normality and constant standard deviation assumptions appear to be met, the results of the F tests performed in Example 11.11 are valid.

11.2.21 Comments on the ANOVA Partition of the Total Variation

Before turning to methods of analyzing two-factor data that don't satisfy the assumptions required for ANOVA, we'll look at why the ANOVA partition

$$\text{SSTo} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}. \tag{11.7}$$

holds is true. We can write a deviation $Y_{ijk} - \bar{Y}$ away from the overall mean as

$$\underbrace{Y_{ijk} - \bar{Y}}_{\substack{\text{Total deviation}\\ \text{of individual}\\ \text{observation away}\\ \text{from overall}\\ \text{mean}}} = \underbrace{\bar{Y}_{i\cdot} - \bar{Y}}_{\substack{\text{Deviation of row}\\ \text{mean away from}\\ \text{overall mean}}} + \underbrace{\bar{Y}_{\cdot j} - \bar{Y}}_{\substack{\text{Deviation of}\\ \text{column mean}\\ \text{away from}\\ \text{overall mean}}} + \underbrace{\bar{Y}_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}}_{\substack{\text{Deviation of}\\ \text{group mean}\\ \text{above and}\\ \text{beyond additive}\\ \text{row and column}\\ \text{effects}}} + \underbrace{Y_{ij} - \bar{Y}_{i}}_{\substack{\text{Deviation of}\\ \text{individual}\\ \text{observation away}\\ \text{from group mean}}} \tag{11.8}$$

If we square both sides of (11.8) and then sum over all $N = abn$ individuals, it can be shown that all of the the cross product terms on the right hand side all sum to zero, and we end up with (11.7).

11.2.22 Comments on the Degrees of Freedom

We'll now see why the degrees of freedom are as given in (11.2.12).

- SSTo has $N-1$ degrees of freedom because only $N-1$ of the N deviations $Y_{ijk}-\bar{Y}$ are "free to vary" since they sum to zero.

- SSA has $a-1$ degrees of freedom because only $a-1$ of the deviations $\bar{Y}_{i.}-\bar{Y}$ are "free to vary" since they sum to zero.

- SSB has $b-1$ degrees of freedom associated with it because only $b-1$ of the deviations $\bar{Y}_{.j}-\bar{Y}$ are "free to vary" since they sum to zero.

- SSE has $ab(n-1)=N-ab$ degrees of freedom associated with it because within each of the ab groups, only $n-1$ of the deviations $Y_{ijk}-\bar{Y}_{ij}$ are "free to vary" since they sum to zero.

- For SSAB, it can be shown that only $(a-1)(b-1)$ of the deviations $\bar{Y}_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}+\bar{Y}$ are "free to vary." Alternatively, we can use the additive property (11.6) of the degrees of freedom to get

$$\textit{df} \text{ for SSAB } = \textit{df} \text{ for SSA} + \textit{df} \text{ for SSB} + \textit{df} \text{ for SSE} - \textit{df} \text{ for SSTo.}$$

  Plugging the in the appropriate values on the right side gives $(a-1)(b-1)$.

11.3 Dealing With Unequal Standard Deviations: Transformations

As was the case for data from one-factor studies, when the standard deviation in a two-factor study increases with the group mean, it's sometimes possible to stabilize it across the groups by transforming the data, most commonly by using the log transformation. Because an increasing standard deviation is often associated with right skewed data, by taking logs we're able to both stabilize the standard deviation *and* transform the data to normality.

11.4 Dealing With Non-Normal Data: Transformations

The two-factor ANOVA F tests are *parametric* tests because they rely on the normality assumption. When this assumption isn't met (and the sample sizes aren't large), as usual there are two main courses of action:

1. **Transform the data to normality**: We could transform all ab samples (same transformation on every sample), for example by taking their logs or using one of the transformations in the Ladder of Powers, so that the transformed samples are each more normally distributed, and then carry out the two-factor ANOVA F tests on the transformed data.
2. **Carry out a nonparametric test**: There are two *nonparametric* procedures for two-factor data. The first is to carry out two-factor ANOVA on the ***rank transformed*** data. The steps are:

   - Combine the ab samples into one big sample, keeping track of which group each observation originally belonged to.
   - Sort and *rank* the observations in the combined sample from smallest (rank = 1) to largest (rank = N).
   - Carry out a two-factor ANOVA, as described in this chapter, using the *ranks* of the observations rather than the observations themselves.

   The second *nonparametric* procedure that's *sometimes* applicable is the ***Friedman test*** described in Section 11.6. This test can be used when a so-called *randomized block design* was used in the study and there's only *one observation per group.*

11.5 Multiple Comparisons After Two-Factor ANOVA

11.5.1 Introduction

When two-factor ANOVA F tests detect statistically significant effects, we usually want to know more about the nature of those effects. The way to proceed will depend on whether or not the interaction effect is significant.

1. **Interaction not significant**: In this case, if a main effect is significant, we can test for differences among the row means $\mu_{1\cdot}, \mu_{2\cdot}, \ldots, \mu_{a\cdot}$ (if factor A is significant) or among the column means $\mu_{\cdot 1}, \mu_{\cdot 2}, \ldots, \mu_{\cdot b}$ (if factor B is significant) using the multiple comparison procedure described below.
2. **Interaction significant**: In this case, it usually doesn't make sense to investigate the main effects because the effect of each factor will be different depending on the level of the other factor. Instead, we investigate the effect of a given factor separately for each level of the other factor by performing multiple comparison tests for differences among individual group means μ_{ij} as described below.

We'll look at one multiple comparison procedure for use after two-factor ANOVA, the ***Bonferroni procedure***, which is almost identical to Bonferroni procedure described in Section 10.7 of Chapter 10, but with slight modifications. In practice, the other multiple comparison procedures listed in Section 10.7 of Chapter 10 could also be used.

11.5.2 Multiple Comparisons When the Interaction is Not Significant

Multiple Comparison Tests for Factor A

When the interaction effect isn't significant, but the factor A main effect is (or both main effects are), to test for differences among the levels of factor A, that is, among the true row means $\mu_{1\cdot}, \mu_{2\cdot}, \ldots, \mu_{a\cdot}$, the number of pairwise comparisons we'll need to make is

$$\text{Number of pairs } \mu_{i\cdot} \text{ and } \mu_{i'\cdot} \text{ to compare} \quad = \quad \frac{a(a-1)}{2},$$

where each comparison will be a test of hypotheses of the form

$$\begin{aligned} H_0 : \mu_{i\cdot} - \mu_{i'\cdot} &= 0 \\ H_a : \mu_{i\cdot} - \mu_{i'\cdot} &\neq 0 \end{aligned}$$

To control the overall familywise Type I error rate at some level α_f using the Bonferroni method, we carry out each pairwise test using the *Bonferroni-corrected level of significance*

$$\alpha_p \ = \ \frac{\alpha_f}{a(a-1)/2}.$$

The appropriate test statistic for each pairwise test is the ***Bonferroni pairwise t test statistic*** defined below.

Bonferroni Pairwise t Test Statistic for Factor A:

$$\begin{aligned} t &= \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} - 0}{\sqrt{\frac{\text{MSE}}{bn} + \frac{\text{MSE}}{bn}}} \\ &= \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sqrt{\frac{2\cdot\text{MSE}}{bn}}}, \end{aligned}$$

where MSE is the mean squared error from the two-factor ANOVA.

Each t statistic is compared to the $t(ab(n-1))$ distribution to obtain the p-value for the pairwise test, which is then compared to α_p to reach a decision for that test.

Multiple Comparison Tests for Factor B

When the interaction isn't significant, but the factor B main effect is (or both main effects are), the procedure for testing for differences among the levels of factor B, that is, among the true column means $\mu_{\cdot 1}, \mu_{\cdot 2}, \ldots, \mu_{\cdot b}$, is similar to the procedure for testing for differences among the levels of factor A. Now, though, the number of pairwise comparisons is

$$\text{Number of pairs } \mu_{\cdot j} \text{ and } \mu_{\cdot j'} \text{ to compare} \;=\; \frac{b(b-1)}{2}$$

and the tests are of hypotheses of the form

$$\begin{aligned} H_0 : \mu_{\cdot j} - \mu_{\cdot j'} &= 0 \\ H_a : \mu_{\cdot j} - \mu_{\cdot j'} &\neq 0 \end{aligned}$$

To control the overall familywise Type I error rate at some level α_f, now we carry out each pairwise test using the *Bonferroni-corrected level of significance*

$$\alpha_p \;=\; \frac{\alpha_f}{b(b-1)/2}.$$

The test statistic for each pairwise test in this case is the ***Bonferroni pairwise t test statistic*** defined below.

Bonferroni Pairwise t Test Statistic for Factor B:

$$\begin{aligned} t &= \frac{\bar{Y}_{\cdot j} - \bar{Y}_{\cdot j'} - 0}{\sqrt{\frac{\text{MSE}}{an} + \frac{\text{MSE}}{an}}} \\ &= \frac{\bar{Y}_{\cdot j} - \bar{Y}_{\cdot j'}}{\sqrt{\frac{2\cdot\text{MSE}}{an}}}, \end{aligned}$$

where MSE is (again) the mean squared error from the two-factor ANOVA.

Each t statistic is again compared to the $t(ab(n-1))$ distribution to obtain the p-value for the pairwise test, which again is compared to α_p.

11.5.3 Multiple Comparisons When the Interaction is Significant

When the interaction is significant, we carry out multiple pairwise comparison tests for differences among pairs of group means μ_{ij}. Here, the total number of pairwise comparisons is

$$\text{Number of pairs } \mu_{ij} \text{ and } \mu_{i'j'} \text{ to compare} \;=\; \frac{ab(ab-1)}{2}$$

and each pairwise comparison test is of hypotheses of the form

$$\begin{aligned} H_0 : \mu_{ij} - \mu_{i'j'} &= 0 \\ H_a : \mu_{ij} - \mu_{i'j'} &\neq 0 \end{aligned}$$

To control the overall familywise Type I error rate at some level α_f, now we carry out each pairwise test using the *Bonferroni-corrected level of significance*

$$\alpha_p \;=\; \frac{\alpha_f}{ab(ab-1)/2}.$$

The test statistic for each pairwise test in this case is the ***Bonferroni pairwise t test statistic***

Bonferroni Pairwise *t* Test Statistic for Group Means:

$$\begin{aligned} t &= \frac{\bar{Y}_{ij} - \bar{Y}_{i'j'} - 0}{\sqrt{\frac{\text{MSE}}{n} + \frac{\text{MSE}}{n}}} \\ &= \frac{\bar{Y}_{ij} - \bar{Y}_{i'j'}}{\sqrt{\frac{2 \cdot \text{MSE}}{n}}}. \end{aligned}$$

Each t statistic is again compared to the $t(ab(n-1))$ distribution to obtain the p-value for the pairwise test, which again is compared to α_p.

Comment: In practice, we may be only interested in making comparisons of *specific pairs* of group means, for example to test for differences among group means within fixed levels of one of the factors. If that's the case, we proceed exactly as described in this section, but using the *Bonferroni-corrected level of significance* $\alpha_p = \alpha_f / c$, where c is the number of pairwise comparisons we're making.

11.6 The Friedman Test

The ***Friedman test*** is a non-parametric alternative to the two-factor ANOVA F test that's appropriate when we're interested in the effect of one factor, but want to *control* for the effect of the other factor, which is called the ***blocking factor***. In an experiment, we do this by using a ***randomized blocks*** design, that is, a study design in which individuals are first split into groups, called ***blocks***, according to the levels of the blocking factor (such as age class or gender if the individuals are animals), and then within each group, assigned to treatments corresponding to levels of the factor of interest.

If factor A is the blocking factor and factor B is of interest, the null hypothesis is that factor B has *no effect* and the alternative is that it has an *effect*. We can write these in terms of the factor B true level means as

$$\begin{aligned} H_0 : &\quad \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot b} \\ H_a : &\quad \text{Not all } \mu_{\cdot j}\text{'s are equal} \end{aligned}$$

The Friedman test is similar to the Kruskal-Wallis test, but ranking is done separately within each of the blocks. Here's how to compute the ***Friedman test statistic***.

Friedman Test Statistic: Suppose factor A is the blocking factor and factor B is of interest. Suppose also that we have one observation per group.

1. Within each of the a blocks, combine the observations for the b levels of factor B, keeping track of which level each observation was made at.
2. Sort the observations, and *rank* them from smallest (rank = 1) to largest (rank = b). If two or more observations are tied, assign to each of them the average of the ranks they would've been assigned if they hadn't been tied.
3. Compute the mean rank $\bar{R}_j$ for each level of the factor ($j = 1, 2, \ldots, b$) and the overall mean rank $\bar{R} = (b+1)/2$.
4. The Friedman test statistic, denoted Q, is

$$Q = \frac{12a}{b(b+1)} \sum_{j=1}^{b} (\bar{R}_j - \bar{R})^2. \qquad (11.9)$$

When the null hypothesis is true, mean ranks $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_b$ for different levels of factor B won't differ from each other much, except due to chance variation, and therefore won't differ much from the overall mean $\bar{R}$. As a result, when H_0 is true, Q should be pretty close to zero. On the other hand, when H_a is true, $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_b$ will differ substantially from each other and Q will be large. It follows that

Large values of Q provide evidence in favor of H_a: Not all $\mu_{\cdot j}$'s are equal.

To decide if an observed value of Q provides statistically significant evidence in support of the alternative hypothesis, we'll need to know its sampling distribution under the null hypothesis.

Sampling Distribution of Q Under H_0: Suppose we have data from a two-factor study in which factor A is a blocking factor and factor B is of interest. Then if the null hypothesis

$$H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot b}$$

is true, the Friedman test statistic Q given by (11.9) follows (approximately) a chi-square distribution with $b - 1$ degrees of freedom, which we write as

$$Q \sim \chi^2(b-1).$$

Because *large* values of Q provide evidence against the null hypothesis, the rejection region for the Friedman test is comprised of Q values in the *rightmost* $100\alpha\%$ of the $\chi^2(b-1)$ distribution, and the p-value is the tail probability to the *right* of the observed Q value.

The Friedman test procedure is summarized in the following table.

Friedman Test for Randomized Blocks Designs

Assumptions: Data are ab independent samples of size $n = 1$ from populations representing combinations of the levels of a blocking variable A and a factor B.*

Null hypothesis: $H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot b}$.

Test statistic value: $Q = \frac{12a}{b(b+1)} \sum_{j=1}^{b} \left(\bar{R}_j - \bar{R}\right)^2$.

Decision rule: Reject H_0 if p-value $< \alpha$ or Q is in rejection region.

Alternative hypothesis	P-value = area under χ^2 distribution with $b-1$ d.f.:	Rejection region = Q values such that:**
$H_a : \mu_{\cdot i} \neq \mu_{\cdot j}$ for some i, j	to the right of Q	$Q \geq \chi^2_{\alpha,b-1}$

* The Friedman test can also be carried on data from experiments in which either a completely randomized design or a randomized blocks design was used.
** $\chi^2_{\alpha,b-1}$ is the $100(1-\alpha)$th percentile of the χ^2 distribution with $b-1$ d.f.

11.7 Problems

11.1 A two-factor study was carried out with two levels of factor A, four levels of factor B, and five observations per group.

a) Give the degrees of freedom for the factor A sum of squares SSA.

b) Give the degrees of freedom for the factor B sum of squares SSB.

c) Give the degrees of freedom for the interaction sum of squares SSAB.

d) Give the degrees of freedom for the error sum of squares SSE.

e) Give the degrees of freedom for the total sum of squares SSTo.

f) Give the numerator and denominator degrees of freedom for the F distribution used to obtain the p-value in the test for a factor A main effect.

g) Give the numerator and denominator degrees of freedom for the F distribution used to obtain the p-value in the test for a factor B main effect.

h) Give the numerator and denominator degrees of freedom for the F distribution used to obtain the p-value in the test for an interaction effect.

11.2 A two-factor ANOVA was carried out using software. The resulting ANOVA table is shown below.

Source	DF	SS	MS	F	P-value
Factor A	3	233.79	77.93	32.81	0.000
Factor B	1	5.04	5.04	2.12	0.164
Interaction	3	5.79	1.93	0.81	0.505
Error	16	38.00	2.38		
Total	23	282.63			

Use the given information in the ANOVA table to answer the following questions.

a) How many levels of factor A are there?

b) How many levels of factor B are there?

c) How many total observations are there in the data set?

d) If the group sample sizes were all the same (same number of observations n per group), what is the value of the common sample size n?

11.3 Experiments have been used in forestry to assess the effects of various factors on the growth behavior of trees. In one study, researchers suspected that healthy spruce seedlings would bud sooner than diseased ones and that acidity might impact the buds by affecting virus uptake into the root system. They exposed healthy and diseased seeds to three levels of acidity (pH) before planting. The response variable is the average rating of the buds produced by a seedling, where a rating is one of

0	=	Bud not broken
1	=	Bud partially expanded
3	=	Bud fully expanded

For each combination of health status and pH, four seedlings' buds were rated. The data are below.

		pH		
		3	5.5	7
Health	Diseased	1.2	0.8	1.0
		1.4	0.6	1.0
		1.0	0.8	1.2
		1.2	1.0	1.4
		1.4	0.8	1.2
	Healthy	1.4	1.0	1.2
		1.6	1.2	1.4
		1.6	1.2	1.2
		1.6	1.4	1.2
		1.4	1.4	1.4

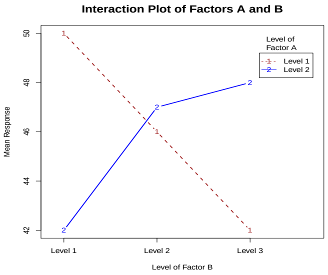
A two-factor ANOVA was carried out. The resulting ANOVA table is shown below, but with some values missing. Without carrying out the ANOVA, fill in the missing values.

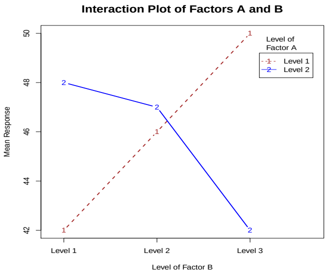
Source	DF	SS	MS	F	P-value
Health	?	0.588	?	?	0.000
pH	?	0.651	?	?	0.000
Interaction	?	0.128	?	?	0.069
Error	?	0.512	?		
Total	?	?			

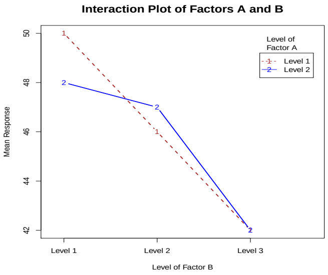
11.4 A two-factor experiment was carried out with two levels of factor A, three levels of factor B, and two observations per group. The data are shown below.

		Factor B		
		Level $j = 1$	Level $j = 2$	Level $j = 3$
Factor A	Level $i = 1$	41.0, 43.0	45.0, 47.0	49.0, 51.0
	Level $i = 2$	47.0, 49.0	46.0, 48.0	41.0, 43.0

a) An interaction plot of the data was made. Which of the following plots is the correct interaction plot?







b) The ANOVA table is shown below.

Source	DF	SS	MS	F	P-value
Factor A	1	0.333	0.3333	0.17	0.697
Factor B	2	4.667	2.3333	1.17	0.373
Interaction	2	100.667	50.3333	25.17	0.001
Error	6	12.000	2.0000		
Total	11	117.667			

Based on the results of the ANOVA, is the interaction effect statistically significant? Use a level of significance $\alpha = 0.05$.

c) Based on your answer to part *b*, would it make sense to use the results of the F test for a factor B main effect to decide if factor B has any effect on the response? Explain your answer.

d) Does factor B have *any* effect on the response? Explain.

11.5 On June 8, 2000 the oil tanker T/V Posavina was rammed and punctured by its own tug, spilling 59,000 gallons of oil into the Chelsea River, Chelsea, Massachusetts, oiling a substantial stretch of shoreline. The response to the oil spill included restoration in fall, 2005 of a degraded 1.5-acre salt marsh along Mill Creek, located on the upper reach of the Chelsea River. The marsh restoration involved removal of roots and rhizomes of the invasive common reed *Phragmites australis* as well as the associated sediments.

To examine the effectiveness of the restoration project towards meeting the objective of decreasing Phragmites cover and increasing the cover of native plants, heights of Phragmites' were measured in the Mill Creek marsh and an adjacent, unrestored control marsh on September 1, 2005, just before the restoration of the Mill Creek marsh, and again in 2007, two years after the restoration [2]. Thus a before-after-control-impact design was used.

The table below shows, for three 1 m^2 quadrats selected from each marsh before the restoration and three selected from each marsh after the restoration, the mean height (cm) of the three tallest Phragmites plants in the quadrat.

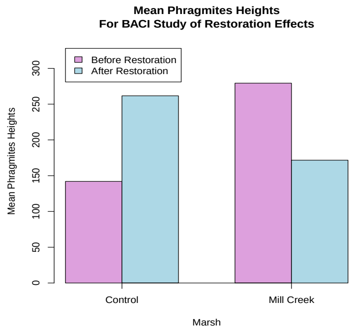
Plant Heights

Period	Marsh	Height
Before	Mill Creek	254
Before	Mill Creek	300
Before	Mill Creek	284
Before	Control	64
Before	Control	80
Before	Control	282
After	Mill Creek	210
After	Mill Creek	151
After	Mill Creek	154
After	Control	179
After	Control	300
After	Control	306

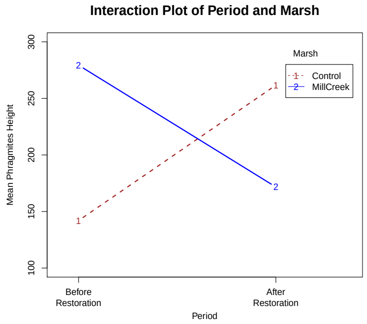
The goal is to decide if the heights of Phragmites decreased more at the restored Mill Creek site than at the control site. A bar plot of the group means is below.

a) Write out the two-factor ANOVA model, with interaction effect, including any assumptions about the random error term ϵ in the model.

b) Carry out a two-factor ANOVA, with interaction effect, and write out the resulting ANOVA table.

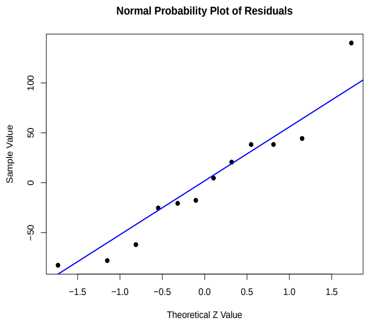


c) State the conclusion of the ANOVA F test for an interaction effect between period and marsh using a level of significance $\alpha = 0.05$. Based on the F test, did the heights of Phragmites decrease more at the restored Mill Creek site than at the control site? You may want to refer to the interaction plot of the group means below.



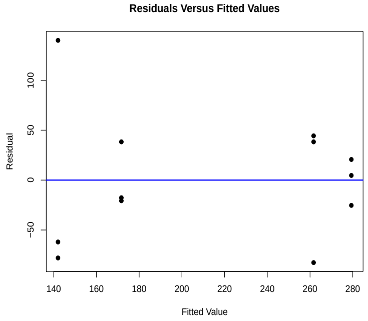
d) Based on the answer to part c, does it make sense to proceed with tests for main effects of marsh and period? Explain your answer.

e) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the F tests, that the error term ϵ is normally distributed appear to be met?

f) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the F tests, that the standard deviation σ of the error distribution is the same for the four marsh by period combinations appear to be met?

g) Assuming that σ is the same for the four marsh by period combinations, what's the estimated value of σ?

11.6 In wastewater effluent, synthetic dyes used to color textiles, paper, and plastics can be highly toxic to aquatic life. One inexpensive method for removing dyes from wastewater in developing countries where agriculture is abundant is adsorption by biomass remnants from agriculture.

A study was carried out in India to investigate the use of rice husk as a biomass adsorbent [9]. As part of the study, an experiment was carried out to determine how two factors, pH and temperature, affect the dye removal efficiency by rice husk. Dye was diluted in water at two pH levels (Low = 2.0 and High = 7.0) and two temperatures (Warm = 40° and Hot = 70° Celsius) and then subjected to the rice husk adsorption process. Two replications of the experiment were performed, giving two observations per group.

The response variable was dye removal efficiency (%), defined as

$$\text{Removal Efficiency} \;=\; \left(\frac{C_i - C_f}{C_i}\right) \cdot 100\%,$$

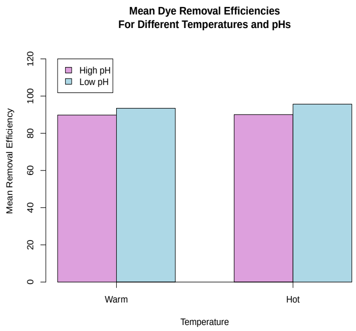
where C_i is the initial dye concentration and C_f is the final dye concentration after the rice husk adsorption process. The table below shows the data.

Dye Removal Efficiency

pH	Temperature	Removal Efficiency
Low	Warm	93.19
Low	Warm	93.69
Low	Hot	94.77
Low	Hot	96.53
High	Warm	88.17
High	Warm	91.41
High	Hot	89.32
High	Hot	90.75

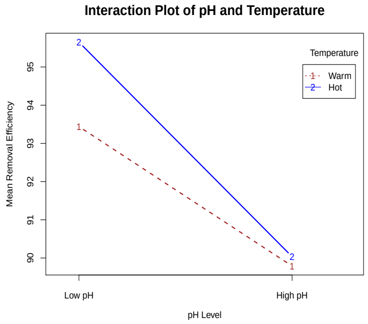
The goal is to decide if either pH or temperature has an effect on the dye removal efficiency. A bar plot of the group means is below.

a) Write out the two-factor ANOVA model, with interaction effect, including any assumptions about the random error term ϵ in the model.



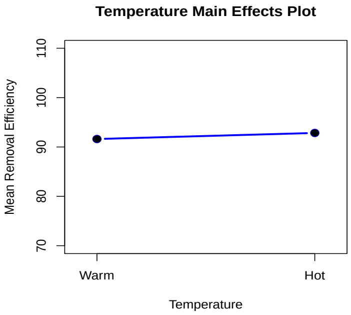
b) Carry out a two-factor ANOVA, with interaction effect, and write out the resulting ANOVA table.

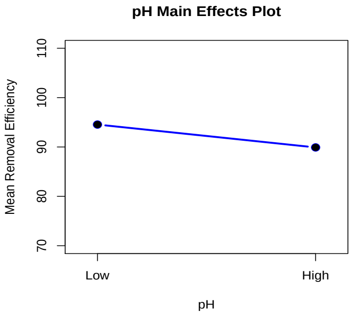
c) State the conclusion of the ANOVA F test for an interaction effect between pH and temperature using a level of significance $\alpha = 0.05$. Based on the F test, is the effect of pH, if any, on removal efficiency different depending on the temperature? You may want to refer to the interaction plot of the group means below.



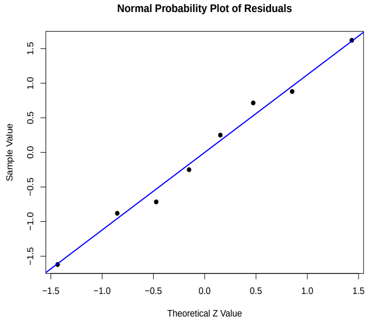
d) Based on the answer to part c, does it make sense to proceed with tests for main effects of pH and temperature? Explain your answer.

e) In part c, you should have found that the interaction effect isn't statistically significant, so it makes sense to proceed with tests for main effects. State the conclusion of the ANOVA F test for a temperature main effect using a level of significance $\alpha = 0.05$. Based on the F test, does temperature have any effect on removal efficiency? You may want to refer to the level means plot of temperature below.





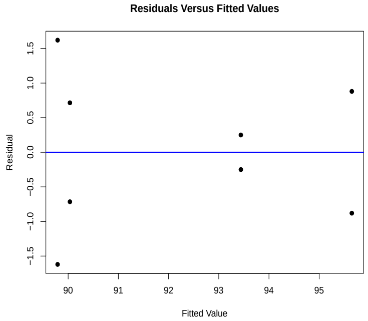
f) State the conclusion of the ANOVA F test for a pH main effect using a level of significance $\alpha = 0.05$. Based on the F test, does pH have any effect on removal efficiency? You may want to refer to the level means plot of pH above.



g) A normal probability plot of the residuals is below.

Based on the plot, does the assumption, required by the F tests, that the error term ϵ is normally distributed appear to be met?

h) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the F tests, that the standard deviation σ of the error distribution is the same for the four pH by temperature combinations appear to be met?

i) Assuming that σ is the same for the four pH by temperature combinations, what's the estimated value of σ?

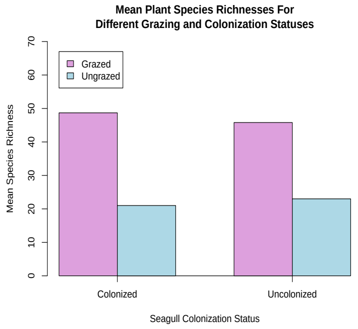
11.7 In a study of plant species diversity on small islands, the species richness (number of species present) was measured on several islands in the Aegean archipelago, Greece. In addition, each island was classified according to whether or not it had undergone animal grazing (by goats or sheep) and whether or not nesting seagulls had colonized it [8].

The table below shows the species richness measurements for 40 islands, 10 in each combination of grazing status (yes/no) and seagull colonization status (yes/no).

Plant Species Diversity on Islands

			Seagull
	Species	Grazing	Colonization
Island Name	Richness	Status	Status
Antidragonera	89	Grazed	Colonized
Aspronisi (north)	45	Grazed	Colonized
Diabates (east)	50	Grazed	Colonized
Faradonisi megalo	60	Grazed	Colonized
Imia (east)	17	Grazed	Colonized
Imia (west)	20	Grazed	Colonized
Minaronisi	45	Grazed	Colonized
Nisida Manoli	55	Grazed	Colonized
Psathi	67	Grazed	Colonized
Spartonisi	39	Grazed	Colonized
Fragkonisi	103	Grazed	Uncolonized
Kounelonisi	59	Grazed	Uncolonized
Lidia	15	Grazed	Uncolonized
Makronisi 1	76	Grazed	Uncolonized
Megalo Trachili	11	Grazed	Uncolonized
Mikro Trachili	6	Grazed	Uncolonized
Neronisi	27	Grazed	Uncolonized
Tigani	12	Grazed	Uncolonized
Velona	63	Grazed	Uncolonized
Zouka (Megali)	86	Grazed	Uncolonized
Aspronisi (east)	11	Ungrazed	Colonized
Aspronisi (east 1)	34	Ungrazed	Colonized
Aspronisi (northwest)	46	Ungrazed	Colonized
Aspronisi (west)	7	Ungrazed	Colonized
Faradonisi (northwest)	33	Ungrazed	Colonized
Faradonisi (south)	22	Ungrazed	Colonized
Kalapodi mikro	12	Ungrazed	Colonized
Plakousa	17	Ungrazed	Colonized
Saraki	16	Ungrazed	Colonized
(Unnamed 1)	12	Ungrazed	Colonized
East Gourna	33	Ungrazed	Uncolonized
Kapelo	1	Ungrazed	Uncolonized
Kommeno nisi	34	Ungrazed	Uncolonized
Kouloura 2	45	Ungrazed	Uncolonized
Paplomata	26	Ungrazed	Uncolonized
Prassonisi	14	Ungrazed	Uncolonized
Prassonisi 3	15	Ungrazed	Uncolonized
Vatopoula	54	Ungrazed	Uncolonized
West Gourna	7	Ungrazed	Uncolonized
(Unnamed 2)	1	Ungrazed	Uncolonized

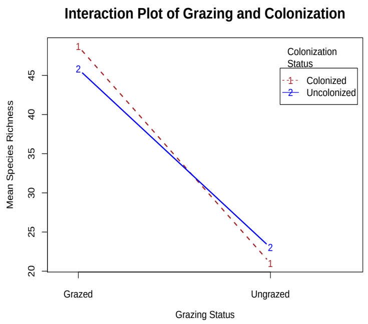
A goal of the study was to determine if either factor, grazing or seagull colonization, affects the plant species diversity. A bar plot of the group means is below.



a) Write out the two-factor ANOVA model, with interaction effect, including any assumptions about the random error term ϵ in the model.

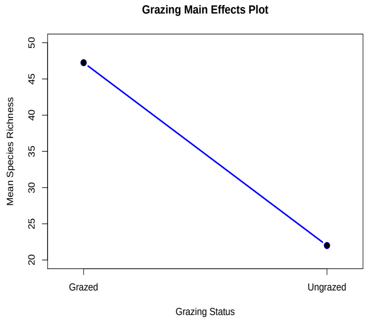
b) Carry out a two-factor ANOVA, with interaction effect, and write out the resulting ANOVA table.

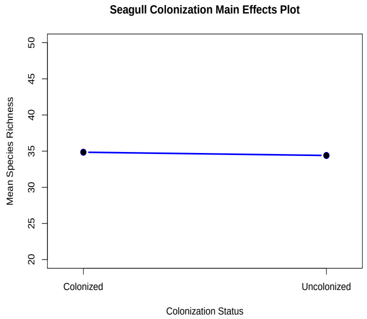
c) State the conclusion of the ANOVA F test for an interaction effect between grazing and seagull colonization using a level of significance $\alpha = 0.05$. Based on the F test, is the effect of grazing, if any, on plant species diversity different depending on the seagull colonization status of an island? You may want to refer to the interaction plot of the group means below.



d) Based on the answer to part c, does it make sense to proceed with tests for main effects of grazing and seagull colonization? Explain your answer.

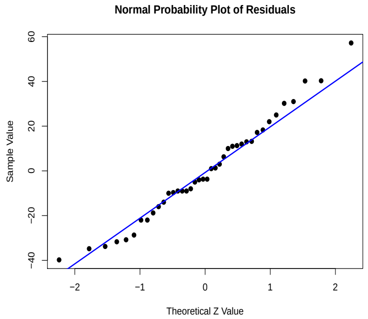
e) In part c, you should have found that the interaction effect isn't statistically significant, so it makes sense to proceed with tests for main effects. State the conclusion of the ANOVA F test for a grazing main effect using a level of significance $\alpha = 0.05$. Based on the F test, does grazing have any effect on plant species richness? You may want to refer to the level means plot of grazing below.





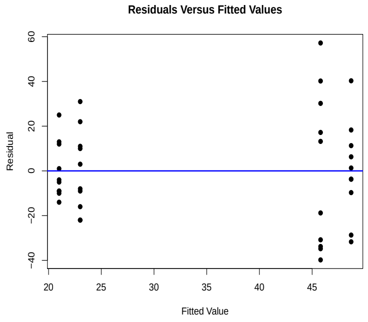
f) State the conclusion of the ANOVA F test for a seagull colonization main effect using a level of significance $\alpha = 0.05$. Based on the F test, does seagull colonization have any effect on plant species richness? You may want to refer to the level means plot of seagull colonization above.

g) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the F tests, that the error term ϵ is normally distributed appear to be met?

h) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the F tests, that the standard deviation σ of the error distribution is the same for the four grazing by colonization status combinations appear to be met?

i) Assuming that σ is the same for the four grazing by colonization status combinations, what's the estimated value of σ?

11.8 Hydroelectric power plants occasionally release sudden, short term increases in streamflow, called "pulsed flows", for recreational purposes mandated by license agreements. But pulsed flows can kill or displace benthic macroinvertebrates, an important food source for fish, birds, and mammals.

A before-after, control-impact study was carried out to investigate the effect on macroinvertebrates of pulsed flows for whitewater boating from a hydroelectric plant on the North Fork Feather River, Plumas County, California [4]. Macroinvertebrate data were collected before and after pulsed flows on two reaches of the river, one downstream of the dam (the impact site) and the other upstream (the control site).

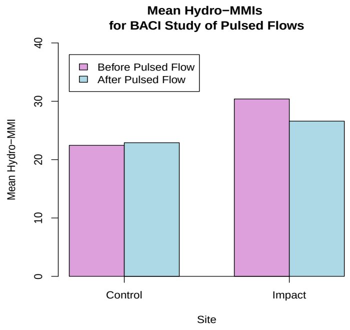
The response variable was an index of macroinvertebrate community health called the hydropower multimetric index (Hydro-MMI), which is based on abundances of five classes of macroinvertebrates and is scaled to lie between 0 and 100, with lower scores indicating less healthy macroinvertebrate communities.

The table below shows the Hydro-MMI scores two days before and two days after each of two pulsed flows in fall, 2004, at the control and impact reaches of the river.

Macroinvertebrate Community Health

Site	Period	Hydro-MMI
Control	Before	29.6
Control	After	24.8
Control	Before	31.2
Control	After	28.4
Impact	Before	22.2
Impact	After	23.0
Impact	Before	22.7
Impact	After	22.8

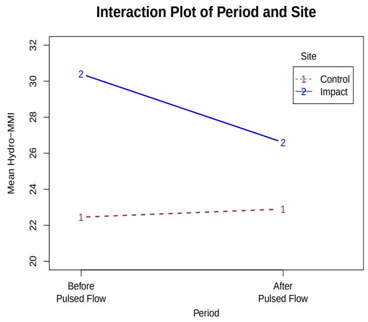
We want to decide if the Hyro-MMI decreased statistically significantly more at the impact site than at the control site. A bar plot of the group means is below.



a) Write out the two-factor ANOVA model, with interaction effect, including any assumptions about the random error term ϵ in the model.

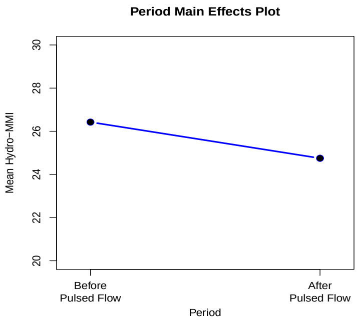
b) Carry out a two-factor ANOVA, with interaction effect, and write out the resulting ANOVA table.

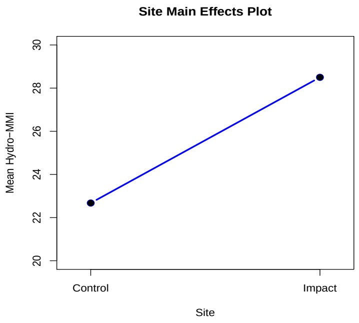
c) State the conclusion of the ANOVA F test for an interaction effect between period and site using a level of significance $\alpha = 0.05$. Based on the F test, did the Hydro-MMI decrease statistically significantly more at the impact site than at the control site? You may want to refer to the interaction plot of the group means below.



d) Based on the answer to part c, does it make sense to proceed with tests for main effects of period and site? Explain your answer.

e) In part c, you should have found that the interaction effect isn't statistically significant, so it makes sense to proceed with tests for main effects. State the conclusion of the ANOVA F test for a period main effect using a level of significance $\alpha = 0.05$. Based on the F test, was there a change in the mean Hydro-MMI from the period before the pulse flows to the period after? You may want to refer to the level means plot of period below.

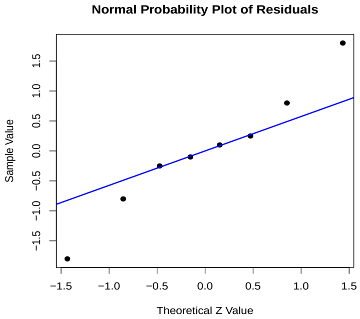




f) State the conclusion of the ANOVA F test for a site main effect using a level of significance $\alpha = 0.05$.

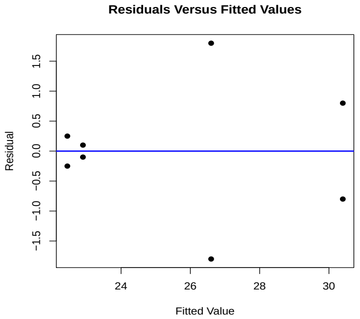
Based on the F test, is there a difference in the Hydro-MMI means for the two sites? You may want to refer to the level means plot of site above.

g) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the F tests, that the error term ϵ is normally distributed appear to be met?

h) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the F tests, that the standard deviation σ of the error distribution is the same for the four site by period combinations appear to be met?

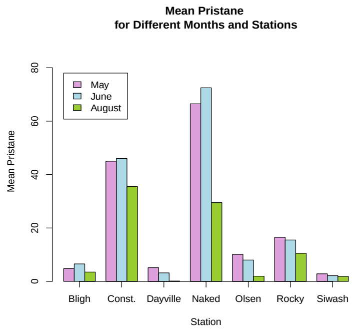
i) Assuming that σ is the same for the four site by period combinations, what's the estimated value of σ?

11.9 In order to establish baseline levels prior to the start of oil tanker movement through the Prince William Sound, Alaska, various hydrocarbons were measured in sediment at seven stations in the sound in May, June, and August, 1978 [6] (see also Problem 10.11 in Chapter 10). At each station, two observations of each hydrocarbon were made per month. The table below shows the pristane and phytane concentrations (ng/g).

Hydrocarbons in Sediment

Station	Month	Pristane	Phytane
Bligh Island	May	4.4	2.3
Bligh Island	May	5.2	1.7
Bligh Island	June	7.4	2.3
Bligh Island	June	5.7	1.9
Bligh Island	August	6.4	1.7
Bligh Island	August	0.6	2.4
Constantine Harbor	May	46.0	11.0
Constantine Harbor	May	44.0	10.0
Constantine Harbor	June	44.0	9.9
Constantine Harbor	June	48.0	11.0
Constantine Harbor	August	27.0	7.6
Constantine Harbor	August	44.0	10.0
Dayville Flats	May	6.2	1.2
Dayville Flats	May	4.1	0.7
Dayville Flats	June	3.4	1.2
Dayville Flats	June	3.0	0.9
Dayville Flats	August	0.2	0.1
Dayville Flats	August	0.1	0.1
Naked Island	May	67.0	0.9
Naked Island	May	66.0	1.8
Naked Island	June	55.0	1.4
Naked Island	June	90.0	0.9
Naked Island	August	32.0	0.8
Naked Island	August	27.0	0.7
Olsen Bay	May	9.2	1.7
Olsen Bay	May	11.0	3.7
Olsen Bay	June	5.0	1.4
Olsen Bay	June	11.0	9.2
Olsen Bay	August	0.6	2.4
Olsen Bay	August	3.3	1.4
Rocky Bay	May	15.0	1.6
Rocky Bay	May	18.0	2.6
Rocky Bay	June	18.0	1.0
Rocky Bay	June	13.0	1.4
Rocky Bay	August	12.0	0.8
Rocky Bay	August	9.0	0.5
Siwash Bay	May	2.6	2.7
Siwash Bay	May	3.1	4.0
Siwash Bay	June	1.6	1.2
Siwash Bay	June	2.7	3.2
Siwash Bay	August	0.7	3.3
Siwash Bay	August	3.0	3.2

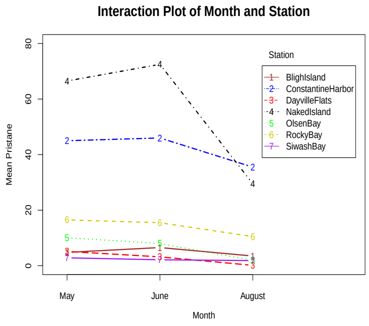
In this problem we'll analyze the pristane data. We want to decide if the pristane concentrations changed over the months or differed among the stations, and if it changed over the months, whether the change was different depending on the station. A bar plot of the group means is below.



a) Write out the two-factor ANOVA model, with interaction effect, including any assumptions about the random error term ϵ in the model.

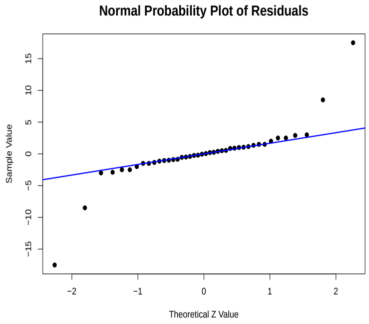
b) Carry out a two-factor ANOVA, with interaction effect, and write out the resulting ANOVA table.

c) State the conclusion of the ANOVA F test for an interaction effect between month and station using a level of significance $\alpha = 0.05$. Based on the F test, did the change in pristane over the months, if there was one, differ depending on the station? You may want to refer to the interaction plot of the group means below.



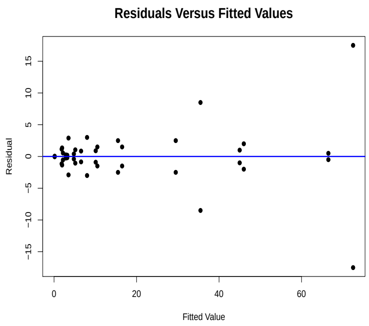
d) Based on the answer to part c, does it make sense to proceed with tests for main effects of month and station? Explain your answer.

e) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the F tests, that the error term ϵ is normally distributed appear to be met?

f) A plot of the residuals versus the fitted values is below.

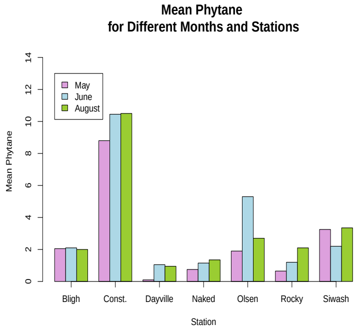


Based on the plot, does the assumption, required by the F tests, that the standard deviation σ of the error distribution is the same for the 21 month by station combinations appear to be met?

g) Assuming that σ is the same for the 21 month by station combinations, what's the estimated value of σ?

11.10 Refer to the study of hydrocarbons in Prince William Sound, Alaska, as described in Problem 11.9. In this problem, we'll analyze the phytane data.

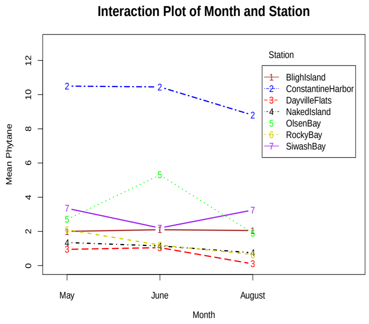
We want to decide if the phytane concentrations changed over the months or differed among the stations, and if it changed over the months, whether the change was different depending on the station. A bar plot of the group means is below.



a) Write out the two-factor ANOVA model, with interaction effect, including any assumptions about the random error term ϵ in the model.

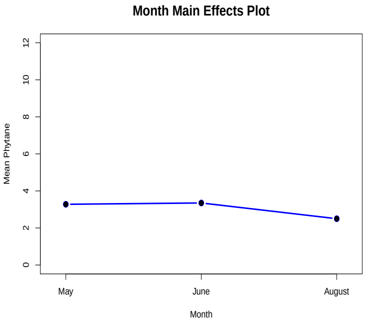
b) Carry out a two-factor ANOVA, with interaction effect, and write out the resulting ANOVA table.

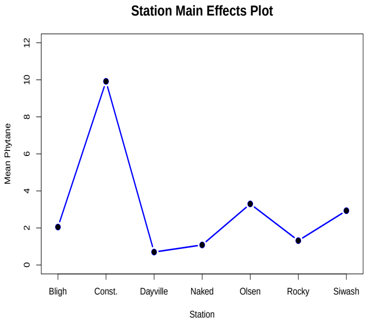
c) State the conclusion of the ANOVA F test for an interaction effect between month and station using a level of significance $\alpha = 0.05$. Based on the F test, did the change in phytane over the months, if there was one, differ depending on the station? You may want to refer to the interaction plot of the group means below.



d) Based on the answer to part c, does it make sense to proceed with tests for main effects of month and station? Explain your answer.

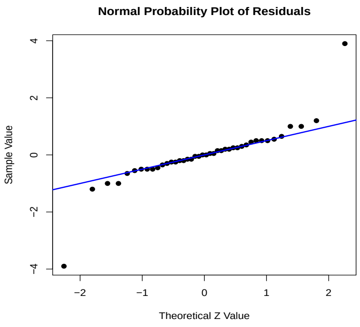
e) In part c, you should have found that the interaction effect isn't statistically significant, so it makes sense to proceed with tests for main effects. State the conclusion of the ANOVA F test for a month main effect using a level of significance $\alpha = 0.05$. Based on the F test, did the phytane concentrations change over the months? You may want to refer to the level means plot of month below.





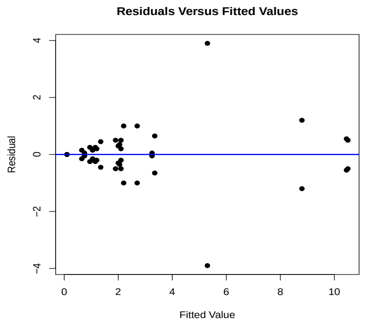
f) State the conclusion of the ANOVA F test for a station main effect using a level of significance $\alpha = 0.05$. Based on the F test, are there differences among the mean phytane concentrations for the seven stations? You may want to refer to the level means plot of station above.

g) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the F tests, that the error term ϵ is normally distributed appear to be met?

h) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the F tests, that the standard deviation σ of the error distribution is the same for the 21 month by station combinations appear to be met?

i) Assuming that σ is the same for the 21 month by station combinations, what's the estimated value of σ?

Bibliography

[1] Paul Berthouex and Linfield Brown. *Statistics for Environmental Engineers.* CRC Press LLC, second edition, 2002.

[2] David M. Burdick, Gregg Moore, and Chris R. Peter. Evaluation of Mill Creek Salt Marsh Restoration Project, Chelsea, Massachusetts, final report. Technical report, Submitted to: NOAA Restoration Center, Gloucester, Massachusetts, 2008.

[3] A. Clements. Suburban development and resultant changes in the vegetation of the bushland of the northern Sydney region. *Australian Journal of Ecology*, 8(3):307 – 320, 1983.

[4] Garcia and Associates (GANDA). Evaluating the impacts of manufactured recreation streamflows on the macroinvertebrate community of a regulated river. Technical Report CEC-500-2006-078, California Energy Commission, PIER Energy-Related Environmental Research Program, 2006.

[5] D.R. Helsel. *Nondetects and Data Analysis, Statistics for Censored Environmental Data.* John Wiley and Sons, Inc., 2005.

[6] J. F. Karinen, M. M. Babcock, D. W. Brown, W. D. Jr. MacLeod, L. S. Ramos, and J. W. Short. Hydrocarbons in intertidal sediments and mussels from Prince William Sound, Alaska, 1977-1980: Characterization and probable sources. Technical Report U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-9, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, 1993 (revised December 1994).

[7] Raktim Pal and Ki-Hyun Kim. Influences of sampling volume and sample concentration on the analysis of atmospheric carbonyls by 2,4-dinitrophenylhydrazine cartridge. *Analytica Chimica Acta*, 610:289 – 296, 2008.

[8] M. Panitsa, D. Tzanoudakis, K. A. Triantis, and S. Sfenthourakis. Patterns of species richness on very small islands: the plants of the Aegean archipelago. *Journal of Biogeography*, 33:1223 – 1234, 2006.

[9] V. Ponnusami, V. Krithika, R. Madhuram, and S.N. Srivastava. Biosorption of reactive dye using acid-treated rice husk: Factorial design analysis. *Journal of Hazardous Materials*, 142:397 – 403, 2007.