Chapter 10

Tests for Comparing k Populations

Chapter Objectives

- State and interpret the two versions of the one-factor ANOVA model.
- Interpret sums of squares, degrees of freedom, and mean squares.
- Carry out a one-factor ANOVA F test for differences among k population means.
- Obtain and interpret fitted values and residuals associated with a fitted ANOVA model.
- Carry out a Kruskal-Wallis test for differences among k population means.
- Decide which test (the ANOVA F test or Kruskal-Wallis test) is more appropriate for a given set of data.
- Carry out a Bonferroni multiple comparison procedure to identify which of k population means differ from each other.

Key Takeaways

- The ANOVA F test is a parametric test for differences among k population means that requires either that the samples are from normal populations or the sample sizes are all large. We can assess normality by graphing the data. A log transformation can make right-skewed data more normal prior to conducting an ANOVA F test.
- The one-factor ANOVA model describes two sources of variation in a response variable: between-groups non-random differences, and within-groups random error.
- Sums of squares in ANOVA are statistics that measure between-groups and within-groups variation in the observed values of a response variable.
- Mean squares are another way to measure between-groups 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 ANOVA F test statistic is a ratio of two mean squares. Its numerator measures between-groups variation and its denominator within-groups variation.
- The Kruskal-Wallis test is a nonparametric test for differences among k population means 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* population means differ from each other after an ANOVA F test (or Kruskal-Wallis test) has indicated that such differences exist.

10.1 Introduction

Environmental studies often involve taking samples from *several* populations to decide if their means differ. For example, the populations might correspond to different land-use types, different hazardous waste sites, or different industrial regions.

Example 10.1: Comparing *k* Populations

Phosphate concentrations were measured in water sampled from three watersheds near Toronto, Canada that differed by land-use type: highly urbanized, moderately urbanized, and rural [17]. One question of interest was whether the mean phosphate concentration differs from one land-use type to the next.

The hypotheses would be

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3$$
$$H_a\text{: the } \mu_i\text{'s aren't all equal.}$$

where μ_1, μ_2, and μ_3 are the population mean phosphate concentrations for the three land-use types.

We'll denote the number of populations being compared by $\boldsymbol{k}$, and we'll look at two hypothesis test procedures for deciding whether differences exist among their means:

1. The one-factor analysis of variance (ANOVA) F test
2. The Kruskal-Wallis test

The first one is a parametric test, requiring either that the populations are normal or that the sample sizes are large. The second is a nonparametric test, so it doesn't have the normality requirement. When only two populations are being compared ($k = 2$), ANOVA becomes identical to the two-sample t test (pooled version) of Chapter 8, so ANOVA is like a t test for three or more samples. Likewise, the Kruskal-Wallis test is like a rank sum test for three or more samples (and is identical to that test when only two populations are being compared).

We'll sometimes refer to the k samples as ***groups***, and we'll think of the categorical variable that defines the groups as the ***explanatory variable***, also referred to as a ***factor*** whose ***levels*** are the grouping categories. In the phosphate study (Example 10.1), the *factor* is land-use type, with three *levels*: highly urbanized, moderately urbanized, and rural.

Both hypothesis test procedures can also be used to decide if the mean responses to *several* treatments in a randomized *experiment* differ. In this case, the groups, or factor levels, correspond to treatments.

Example 10.2: Comparing *k* Experimental Treatments

To investigate the bioaccumulation of selenium (Se) in snakes via maternal transfer, 33 female house snakes were randomly assigned to three treatment groups, 11 per group [8]. The first group was fed mice having Se levels typical of those found in nature, about 1 μg/g dry mass of Se. The second the third groups were fed mice injected with solutions containing 10 and 20 μg/g of Se, respectively. After six months the snakes were allowed to mate, and their eggs and hatchlings analyzed for Se.

The hypotheses would be

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3$$
$$H_a\text{: the } \mu_i\text{'s aren't all equal.}$$

where μ_1, μ_2, and μ_3 are the population mean Se concentrations in offspring of snakes fed the three diets.

10.2 One-Factor Analysis of Variance

10.2.1 Introduction

One-factor analysis of variance (or ***ANOVA***) is a procedure for deciding if there are any differences among the means of k populations from which samples have been drawn, or among the mean responses to k treatments.

Example 10.3: One Factor ANOVA

A quality assurance study was carried out to compare lead measurement results for water sent to $k = 5$ laboratories [1]. If the labs are systematically producing different results, it may signify improperly calibrated equipment or poorly trained technicians.

A large quantity of wastewater was split into 50 specimens, 10 of which were assigned randomly each lab for analysis. The table below shows the lead concentration measurements (μg/L) and their summary statistics for each lab:

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
3.4	4.5	5.3	3.2	3.3
3.0	3.7	4.7	3.4	2.4
3.4	3.8	3.6	3.1	2.7
5.0	3.9	5.0	3.0	3.2
5.1	4.3	3.6	3.9	3.3
5.5	3.9	4.5	2.0	2.9
5.4	4.1	4.6	1.9	4.4
4.2	4.0	5.3	2.7	3.4
3.8	3.0	3.9	3.8	4.8
4.2	4.5	4.1	4.2	3.0
$\bar{Y}_1 = 4.30$	$\bar{Y}_2 = 3.97$	$\bar{Y}_3 = 4.46$	$\bar{Y}_4 = 3.12$	$\bar{y}_5 = 3.34$
$S_1 = 0.904$	$S_2 = 0.440$	$S_3 = 0.642$	$S_4 = 0.764$	$S_5 = 0.737$

Four plots for visually comparing these five samples are below.

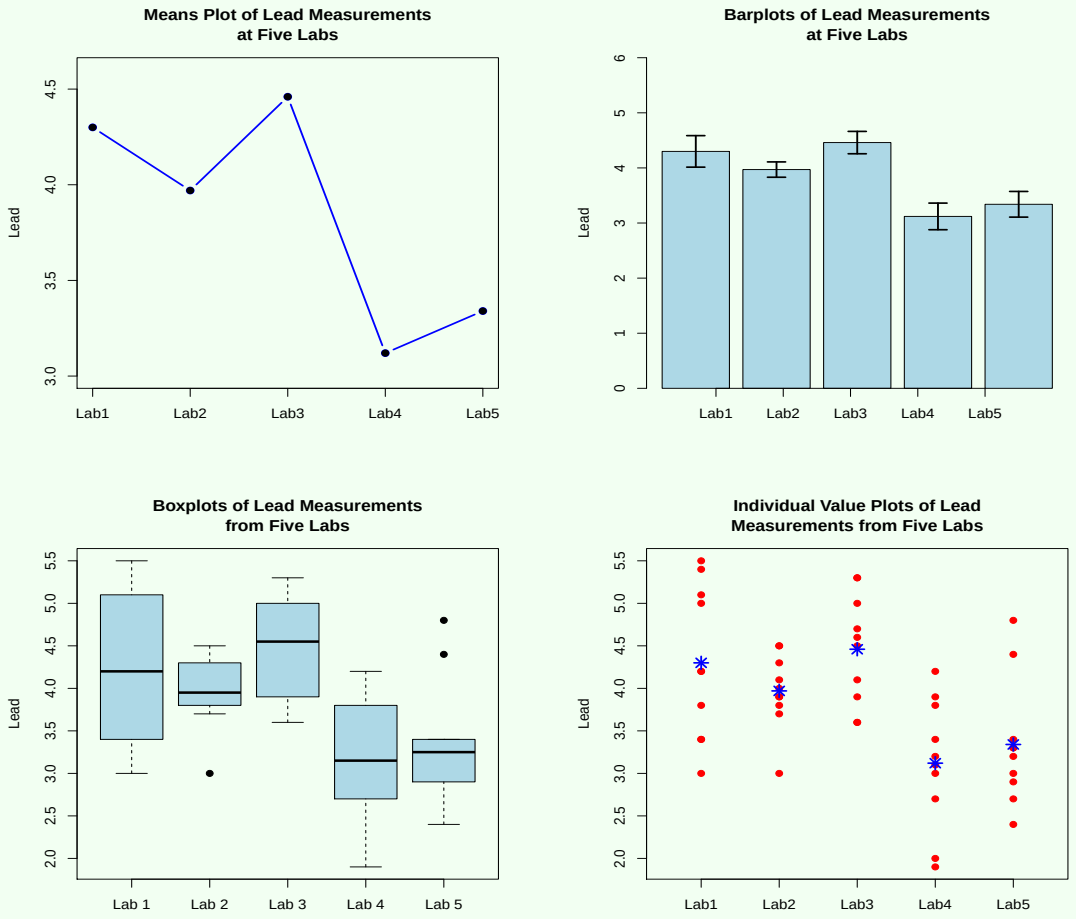


Figure 10.1: Means plot (top left), bar plot of the means (top right), boxplots (bottom left), and individual value plot (bottom right) of lead measurements from five labs. Error bars in the bar plot are ± one standard error. Asterisks in the individual value plot are the sample means.

The plots suggest that the labs aren't all producing the same results, and the sample means shown at the bottom of the data table seem to support this conclusion. But are the observed differences among the sample means statistically significant, or can they be explained by chance variation? A *one-factor ANOVA F test* will answer this question.

10.2.2 Notation

Suppose either that we have independent random samples of sizes $\boldsymbol{n_1, n_2, \ldots n_k}$ (not necessarily all the same) from k populations or responses for k treatment groups of sizes $\boldsymbol{n_1, n_2, \ldots n_k}$ in a randomized experiment. *We'll focus only on the case in which the sample sizes are equal*, even though ANOVA *can* be carried out when they're not equal. Focusing on the equal-sample size case simplifies much of the notation. Thus, for the remainder of this chapter, we'll let

$\boldsymbol{n}$ = The common sample size for the k samples.

As in Example 10.3, the data and summary statistics may be presented in a table of the form below.

Group 1	Group 2	$\cdots$	Group k
Y_{11}	Y_{21}	$\cdots$	Y_{k1}
Y_{12}	Y_{22}	$\cdots$	Y_{k2}
$\vdots$	$\vdots$	$\cdots$	$\vdots$
Y_{1n}	Y_{2n}	$\cdots$	Y_{kn}
$\bar{Y}_1$	$\bar{Y}_2$	$\cdots$	$\bar{Y}_k$
S_1	S_2	$\cdots$	S_k

where

$\boldsymbol{Y_{ij}}$ = The jth observation in the ith group.

In this *double-subscript* notation, the first subscript, $\boldsymbol{i}$, indicates the group to which an individual belongs, and it takes the values $1, 2, \ldots, k$. The second subscript, $\boldsymbol{j}$, distinguishes individuals within a group. It takes the values $1, 2, \ldots, n$.

Also, we'll let

$\bar{\boldsymbol{Y}}_{\boldsymbol{i}}$ = The sample mean for the ith group, and is called the ***i*th group mean** or ***i*th factor level mean**.

$\boldsymbol{S_i}$ = The sample standard deviation for the ith group.

We'll let

$\boldsymbol{N}$ = The ***overall sample size***, that is, the total number of observations in all k groups combined.

Because each of the k samples has n observations,

$$N = kn.$$

Finally, we let

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

The next fact says that the overall mean can obtain by averaging the group means.

Fact 10.1 When the group sample sizes are equal, the overall sample mean $\bar{Y}$ is equal to the average of the k group means $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$, that is,

$$\bar{Y} = \frac{1}{k}\sum_{i=1}^{k} \bar{Y}_i \, . \tag{10.1}$$

Example 10.4: One Factor ANOVA

For the lead measurements made at five labs (Example 10.3),

$$n = 10 \qquad \text{and} \qquad N = 50,$$

and the overall mean is obtained either by averaging the 50 observations or as below by averaging

the five group means shown at the bottom of the table in Example 10.3.

$$\begin{aligned}\bar{Y} &= \frac{1}{k}\sum_{i=1}^{k}\bar{Y}_i \\ &= \frac{1}{5}(4.30 + 3.97 + 4.46 + 3.12 + 3.34) \\ &= 3.838.\end{aligned}$$

When the sample sizes *aren't* equal, the overall mean is a *weighted average* of the group means, with weights proportional to the sample sizes.

10.2.3 Variation Between and Within Groups

The five group means given in Example 10.3 aren't all equal, but we wouldn't expect them to be, even if the technicians at the labs were well trained and the equipment was properly calibrated, because slight variations in lead concentrations from one water specimen to the next as well as measurement error would lead to chance variation (sampling error) among the group means. We want to know if the differences among the five means are larger than can be explained by chance.

Differences among group means ($\bar{Y}_i$'s) are referred to as ***between-groups variation***. Variation of individual observations (Y_{ij}'s) within a group is called ***within-groups variation***. To decide if the variation *between* groups is more than can be explained by chance, we'll compare a measure of that variation to a measure of the variation *within* groups. If the between-groups variation is large relative to the within-groups variation, we'll deem the differences between group means to be statistically significant, that is, larger than can be explained by chance.

The plots below illustrate this concept. In both plots, the group means are $\bar{Y}_1 = 112$, $\bar{Y}_2 = 120$, and $\bar{Y}_3 = 105$. Thus the between-groups variation is the same for the two sets of data. But in the left plot, the within-groups variation is much smaller, as indicated by the shorter boxes. An ANOVA F test would find statistically significant differences among the three groups in the left plot, but not among those in the right plot.

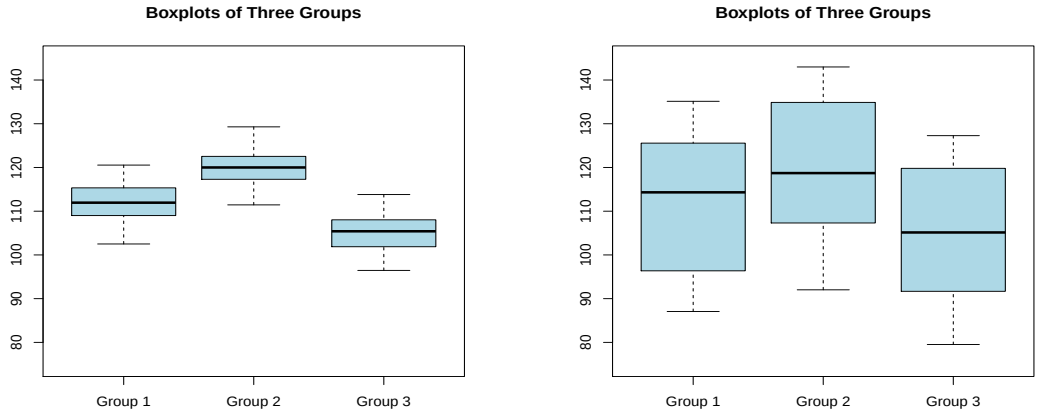


Figure 10.2: Two sets of boxplots showing three groups. The between-groups variation is the same for the two plots, but the within-groups variation is much smaller for the plot on the left.

10.2.4 The One-Factor ANOVA Model

Statistical Models

A common approach to detecting patterns in noisy data is to think of the variation in the data as arising from two types of sources: nonrandom and random. *Nonrandom sources* of variation are what drive the overall pattern in the data. They're usually few in number and either known or identifiable as being the source of the pattern. Often, they correspond to some underlying physical, chemical, or biological process. *Random sources* of variation result from the random selection of the individuals (e.g. specimens) upon which measurements are made and the differing characteristics that those individuals have. When those differing characteristics affect the measured variable, it produces the "noise" in noisy data, that is, the deviations of individual observations away from the overall pattern. There are usually many such differing characteristics, but they're often unknown and unidentifiable, and each one's contribution to the variation might be small or tiny.

For the study of phosphate in water (Example 10.1), land-use type (the *factor*) would be considered a nonrandom source of phosphate variation, and the multitude of characteristics that differ from one water specimen to the next (e.g. sample site location, sampling time of day, weather conditions, etc.) would be considered random sources of variation. For the study of selenium in snakes (Example 10.2), the amount of Se fed to the adult snakes (the *factor*) would be considered a nonrandom source of variation in their offsprings' Se levels, and individual characteristics of the adult snakes (e.g. their prior histories, genetic backgrounds, physical attributes, etc.) would be considered random sources of variation.

By thinking about data in terms of nonrandom and random sources of variation, we can formulate a ***statistical model*** to convey that thinking. Usually, statistical models have the form

$$Y \;=\; \text{Nonrandom Overall Pattern} \;+\; \text{Random Deviation From the Pattern}$$

This expresses an observation Y as the sum of a term representing one or more nonrandom sources of variation and another representing random sources. The nonrandom piece is written in terms of unknown constants called ***model parameters***. It might also involve one or more explanatory variables. The random piece is a random variable, usually assumed to follow a normal distribution whose mean is zero.

The simplest example of a statistical model is the one used in Chapter 4 to describe a measurement that includes measurement error,

$$Y \;=\; \mu + \epsilon\,.$$

Here, Y is an observed measurement, μ is the true (unknown) concentration, and ϵ is a $\mathrm{N}(0,\,\sigma)$ measurement error. According to this model, the "overall pattern" in a set of measurements is just the true concentration μ, which would be called the (unknown) *model parameter*. Recall that this model is equivalent to saying

$$Y \;\sim\; \mathrm{N}(\mu,\,\sigma)\,.$$

Once a statistical model has been specified for a given set of data, we use the data to *estimate* the values of the unknown parameters in the model and to *test hypotheses* about those values. When the null hypothesis is rejected, a pattern is considered to have been detected in the data.

The One-Factor ANOVA Model: Two Versions

In an ANOVA setting, the statistical model will contain a *nonrandom* piece representing *between-groups* variation and a *random* piece representing *within-groups* variation. There are two (equivalent) ways to specify the nonrandom piece: the *group means* version and the *treatment effects* version.

We'll assume that our k samples are drawn from *normal* populations whose means might differ but whose standard deviations are equal. In the context of the lead measurements from the five labs (Example 10.3), this can be written as

Lab 1: $Y_{11}, Y_{12}, \ldots, Y_{1n}$ are a sample from a N(μ_1, σ) population.
Lab 2: $Y_{21}, Y_{22}, \ldots, Y_{2n}$ are a sample from a N(μ_2, σ) population.
$\vdots$
Lab 5: $Y_{51}, Y_{52}, \ldots, Y_{5n}$ are a sample from a N(μ_5, σ) population.

where $\boldsymbol{\mu_1, \mu_2, \ldots, \mu_5}$ are referred to as the ***group population means***. Note that the population means might differ from one lab to the next, but the population standard deviations are supposed to be the same.

One-Factor ANOVA Model (Group Means Version)

The assumptions about the data can be written more succinctly as the *statistical model*

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where Y_{ij} is the jth lead measurement made at Lab i and

$$\epsilon_{ij} \sim \text{N}(0, \sigma).$$

In this model, the nonrandom "overall pattern" is determined by the population means $\mu_1, \mu_2, \ldots, \mu_5$, which, if they differ, contribute to *between-groups* variation in the data (see Fig. 10.3). Larger differences among these means correspond to more *between-groups* variation. The random "deviation" of a lab's lead measurement away from that lab's population mean,

$$\epsilon_{ij} = Y_{ij} - \mu_i,$$

is called ***random error***. It results from natural variation from one water specimen to the next, and contributes to *within-groups* variation. The error standard deviation $\boldsymbol{\sigma}$ (which is also each lab's population standard deviation) represents the size of a typical random error. A larger σ corresponds to more *within-groups* variation. A graphical depiction of the model is below.

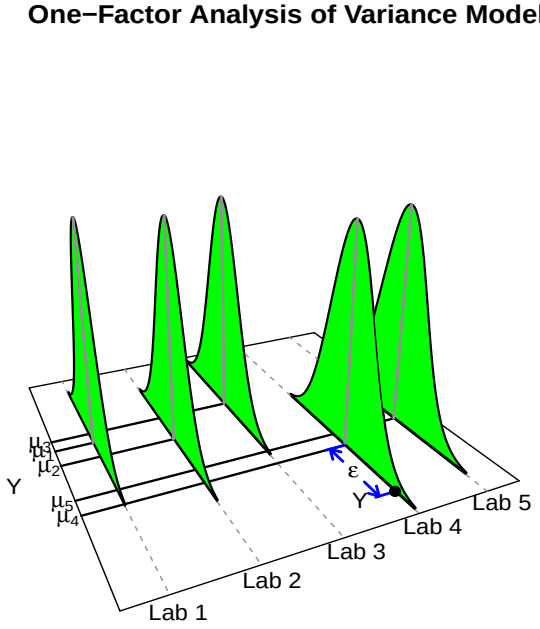


Figure 10.3: Graphical depiction of the group means version of the one-factor ANOVA model for Example 10.3.

Generalizing from this lead measurements example gives the so-called ***group means version*** of the ***one-factor ANOVA model***.

One-Factor ANOVA Model (Group Means Version): A statistical model for describing data in random samples from k populations is:

$$Y_{ij} \;=\; \mu_i \,+\, \epsilon_{ij}\,, \tag{10.2}$$

where

Y_{ij} is the jth observation ($j = 1, 2, \ldots, n$) in the ith group ($i = 1, 2, \ldots, k$).
$\boldsymbol{\mu_i}$ is the ***ith group's population mean***.
ϵ_{ij} is a random error term following a N(0, σ) distribution, and the ϵ_{ij}'s are independent of each other.

The unknown ***model parameters*** are the population means $\mu_1, \mu_2, \ldots, \mu_k$ and the error standard deviation σ. In practice, these will be *estimated* from the data. We'll see how they're estimated in Subsection 10.2.6, and we'll see how to test hypothesis about the μ_i's in Section 10.2.12.

One-Factor ANOVA Model (Treatment Effects Version)

Sometimes it's preferable to write the statistical model in terms of the *effects* of the treatments in an experiment. We define

$\boldsymbol{\mu}$ = The ***overall population mean***, defined to be the average of the group population means,

that is,

$$\mu \;=\; \frac{1}{k}\sum_{i=1}^{k}\mu_i\,.$$

We also define

$\boldsymbol{\alpha_i}$ = The ***ith treatment effect***, defined to be the discrepancy between the ith group's population mean μ_i and the overall mean μ,

that is,

$$\alpha_i \;=\; \mu_i - \mu\,. \tag{10.3}$$

With these definitions, we can write the ith group mean, μ_i, as the *overall mean plus a treatment effect*:

$$\begin{aligned}\mu_i \;&=\; \mu \,+\, (\mu_i - \mu)\\ &=\; \mu \,+\, \alpha_i\,.\end{aligned} \tag{10.4}$$

Plugging the right side of (10.4) in for μ_i in (10.2), we get the ***treatment effects version*** of the ***one-factor ANOVA model***.

One-Factor ANOVA Model (Treatment Effects Version): The statistical model for describing data in random samples from k populations can also be written as:

$$Y_{ij} \;=\; \underbrace{\mu + \alpha_i}_{\text{This is } \mu_i} \,+\, \epsilon_{ij}\,, \tag{10.5}$$

where

Y_{ij} is the jth observation ($j = 1, 2, \ldots, n$) in the ith sample ($i = 1, 2, \ldots, k$).

μ is a constant representing an overall population mean.
α_i is the effect of the ith treatment.
ϵ_{ij} is a random error term following a N(0, σ) distribution, and the ϵ_{ij}'s are independent of each other.

In this version of the model, the unknown ***model parameters*** are the overall population mean μ, the treatment effects $\alpha_1, \alpha_2, \ldots, \alpha_k$, and the error standard deviation σ. In practice, these will be *estimated* from the data. We'll see how they're estimated in Subsection 10.2.6, and we'll see how to test hypothesis about the α_i's in Section 10.2.12.

For the study of lead measurements at the five labs, μ is the overall population mean measurement across all five labs, and α_i is the *effect* of a measurement being made at the ith lab. Note that α_i will be positive if the ith lab's measurement results are systematically high (relative to the overall mean) and negative if its results are systematically low. Fig. 10.4 below illustrates.

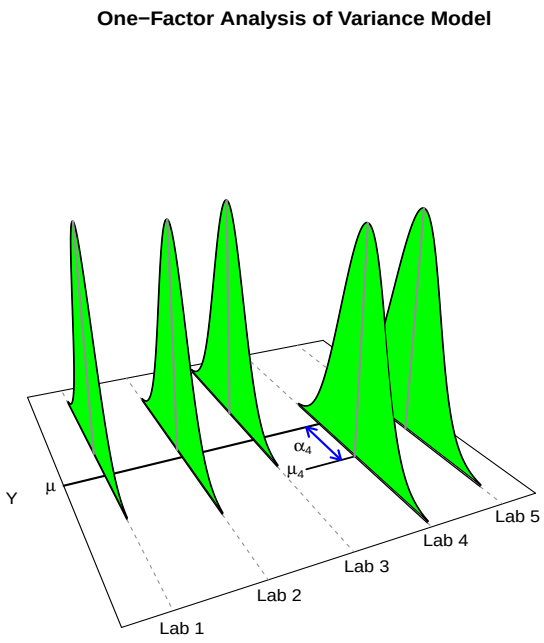


Figure 10.4: Graphical depiction of the treatment effects version of the one-factor ANOVA model for Example 10.3.

10.2.5 Null and Alternative Hypotheses

The hypotheses that we'll be testing are stated differently depending on which of the two versions of the ANOVA model is being used to describe the data. If the group means version is being used, the hypotheses are stated in terms of the group population means $\mu_1, \mu_2, \ldots, \mu_k$. If the treatment effects version is being used, they're stated in terms of the treatment effects $\alpha_1, \alpha_2, \ldots, \alpha_k$. But the two sets of hypotheses are equivalent, just two different ways of saying the same thing. Here they are.

	Hypothesis About the $\boldsymbol{\mu_i}$'s	**Equivalent Hypothesis About the $\boldsymbol{\alpha_i}$'s**
Null	$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$	$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$
Alternative	H_a : The μ_i's aren't all equal	H_a : The α_i's don't all equal 0

To see that the two sets of hypotheses say the same thing, note that by (10.3), the α_i's will all equal zero only if the μ_i's are all the same (and equal to their average μ).

Comment: The alternative hypothesis *doesn't* necessarily say the μ_i's are *all different* from each other (or that the α_i's are *all* non-zero). It just says that for *at least one* pair, the μ_i's don't equal each other (and that *at least one* of the α_i's doesn't equal zero).

10.2.6 Model Parameter Estimates, Fitted Values, and Residuals

Model Parameter Estimates

If the nonrandom part of a statistical model accurately reflects the true underlying physical, biological, chemical, etc. process that generated a pattern in a set of data, then *estimates* of the (unknown) model parameters will provide insight into that underlying process.

Recall that for the group means version of the one-factor ANOVA model, the parameters are $\mu_1, \mu_2, \ldots, \mu_k$ and σ, and for the treatment effects version, the parameters are μ, $\alpha_1, \alpha_2, \ldots, \alpha_k$, and σ. Estimation of σ will be covered in Subsection 10.2.10. The other parameters are listed in the table below along with their estimators based on the data. Also shown is alternative notation for each estimator that will come in handy later.

Model Parameter Estimators

Model Parameter	Estimator	Alternate Notation for the Estimator
μ_i	$\bar{Y}_i$	$\hat{\mu}_i$
μ	$\bar{Y}$	$\hat{\mu}$
$\alpha_i = \mu_i - \mu$	$\bar{Y}_i - \bar{Y}$	$\hat{\alpha}_i$

The figures below illustrate these parameter estimates for the data on lead measurements from the five labs (Example 10.3). These figures should be compared with the graphs depicting the ANOVA models in Figs. 10.3 and 10.4.

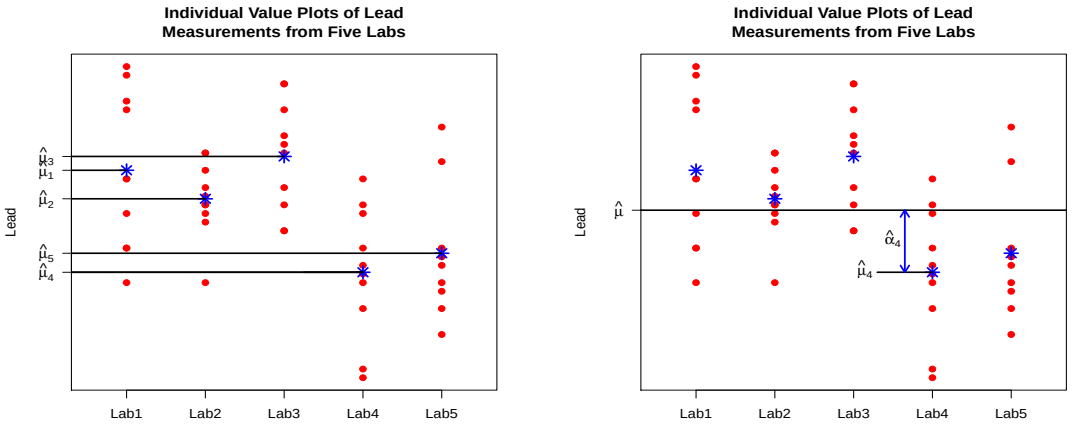


Figure 10.5: Graphical depiction of the parameter estimates for the group means (left) and treatment effects (right) versions of the one-factor ANOVA model for Example 10.3. Asterisks are the sample means.

Fitted Values

Once the parameters in a statistical model have been estimated, we say that the model has been ***fitted*** to the data. Fitting a model to a set of data provides a summary of the nonrandom "overall pattern" in the data.

For each individual in a given group, we define the individual's ***fitted value*** (also called ***predicted value***) to be the estimate of that group's population mean μ_i (or $\mu + \alpha_i$).

Fitted Values (Group Means Version):

$$\text{Fitted Values for Individuals in } i\text{th Group} \;=\; \hat{\mu}_i \;=\; \bar{Y}_i$$

Fitted Values (Treatment Effects Version):

$$\text{Fitted Values for Individuals in } i\text{th Group} \;=\; \hat{\mu} + \hat{\alpha}_i \;=\; \bar{Y} + (\bar{Y}_i - \bar{Y}) \;=\; \bar{Y}_i$$

Note that regardless of which version of the model is being used, *the fitted values are just the group means* $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$. Note also that all n individuals in a given group share the same fitted value. For the study of lead measurements at the five labs, the fitted values are the blue asterisks in Fig. 10.5, and all ten water specimens sent to a given lab share that same fitted value. The alternative name *predicted value* is a reference to a fitted value's interpretation as the lead measurement value we'd "predict" for a water specimen sent to the given lab.

Residuals

The fitted values provide estimates of the nonrandom "overall pattern" piece of the statistical model. *Residuals* are estimates of the random "deviations" ϵ away from the overall pattern, that is, of the random error in the data. A ***residual*** is the difference between an individual's observed Y value and the fitted value for that individual.

$$\text{Residual} \;=\; \text{Observed } Y \text{ Value} \;-\; \text{Fitted Value}$$

We'll denote the residual for the jth individual in the ith group by $\boldsymbol{e_{ij}}$. In symbols, residuals are defined as below for the two versions of the ANOVA model.

Residuals (Group Means Version):

$$\begin{aligned} e_{ij} &= Y_{ij} - \hat{\mu}_i \\ &= Y_{ij} - \bar{Y}_i \end{aligned} \tag{10.6}$$

Residuals (Treatment Effects Version):

$$\begin{aligned} e_{ij} &= Y_{ij} - (\hat{\mu} + \hat{\alpha}_i) \\ &= Y_{ij} - \bar{Y}_i \end{aligned} \tag{10.7}$$

Note that regardless of which version of the model is being used, *the residuals are just the deviations of the* Y_{ij}*'s away from their corresponding group means* $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$. There'll be one residual for each observed value Y_{ij} in the data set (N residuals total). For the study of lead measurements at the five labs, the residuals are the gaps in Fig. 10.5 between the points and the blue asterisks, and correspond to random

error in the lead measurements.

Note: By rearranging (10.6) or (10.7), we can write an observation Y_{ij} as

$$Y_{ij} \;=\; \hat{\mu}_i + e_{ij}$$

or

$$Y_{ij} \;=\; \hat{\mu} + \hat{\alpha}_i + e_{ij}\,,$$

both of which are of the form

$$\text{Observed Value} \;=\; \text{Fitted Value} \;+\; \text{Residual}.$$

Comparing these expressions to the ANOVA models (10.2) and (10.5), it's clear that *the residual e_{ij} approximates the random error term ϵ_{ij}* (which is unobservable because the true model parameter values are unknown). In Subsection 10.2.11, we'll use the residuals to estimate the standard deviation σ of the $\mathrm{N}(0, \sigma)$ error distribution, and in Subsection 10.2.13 we'll use them to check the normality assumption.

10.2.7 The Double Summation Notation

We'll use the residuals to illustrate the *double summation* notation. From Chapter 3 we know that the deviations of observations away from a sample mean alway sum to zero. But residuals *are* deviations away from sample (group) means, so they sum to zero within each group. This is stated formally in the next fact.

Fact 10.2 For data in random samples from k populations, for each $i = 1, 2, \ldots, k$, the residuals within the ith group sum to zero, that is,

$$\sum_{j=1}^{n} e_{ij} \;=\; 0$$

for each fixed i.

It follows that the sum of *all* N residuals will be zero too because this sum can be obtained by first summing within each group and then summing those results:

$$\sum_{j=1}^{n} e_{1j} + \sum_{j=1}^{n} e_{2j} + \cdots + \sum_{j=1}^{n} e_{kj} \;=\; 0 + 0 + \cdots + 0 \;=\; 0. \tag{10.8}$$

The left hand side above is more concisely written using the ***double summation*** notation, whereby (10.8) becomes

$$\sum_{i=1}^{k} \sum_{j=1}^{n} e_{ij} \;=\; 0.$$

This notation says first, for each fixed value of i, sum over the n individuals ($j = 1, 2, \ldots, n$) within the ith group, and then sum those results over the k groups ($i = 1, 2, \ldots, k$). The double summation notation will be used extensively in the remainder of this chapter and the next.

10.2.8 Sums of Squares

Introduction

As mentioned in Subsection 10.2.3, to decide if the observed differences among the group means in a one-factor study are larger than can be explained by chance variation, we'll compare a measure of the *between-groups* variation in those means to a measure of the *within-groups* variation in individual observations. In this section we'll look at a way to measure these two types of variation.

Recall (Chapter 3) that one measure of variation in a data set $X_1, X_2, \ldots, X_n$ is the sample variance (squared standard deviation),

$$S^2 \;=\; \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\,, \tag{10.9}$$

which we think of as an "average" squared deviation of an observation X_i away from the sample mean $\bar{X}$. The numerator on the right side above is an example of what's called a ***sum of squares***. The measures of variation used in ANOVA will be based on certain *sums of squares*.

Between-Groups Variation

Between-groups variation refers to variation among the group means $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$. In ANOVA, we measure this variation by the ***treatment sum of squares***, denoted **SSTr**, defined as follows (where we also invoke the definition, from Subsection 10.2.6, of the estimators $\hat{\alpha}_i$ of the treatment effects α_i).

Treatment Sum of Squares (Between-Groups Variation):

$$\text{SSTr} \;=\; n\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2 \;=\; n\sum_{i=1}^{k}\hat{\alpha}_i^2. \tag{10.10}$$

The reason for the n in front will be explained later. For now, just know that it'll make things easier when we compare the between-groups variation to the within-groups variation. The treatment sum of squares reflects the sizes of the deviations of the group means away from the overall mean $\bar{Y}$. Those deviations will be large when there are substantial differences among the group means, and in this case SSTr will be large. It's in this sense that SSTr measures *between-groups* variation.

Within-Groups Variation

Within-groups variation refers to variation of the observations Y_{ij} within the groups. One way of measuring this variation is to compute the sample variances $S_1^2, S_2^2, \ldots, S_k^2$ one group at at time and then combine them. In ANOVA, they're combined by multiplying each one by $n-1$ and then summing the results. This gives the so-called ***error sum of squares***, denoted **SSE**.

Error Sum of Squares (Within Groups Variation):

$$\text{SSE} \;=\; (n-1)S_1^2 + (n-1)S_2^2 + \cdots + (n-1)S_k^2\,. \tag{10.11}$$

Notice from the definition (10.9) of a sample variance that when we multiply by $n-1$ we get the sum of squares within that group, that is,

$$(n-1)S_i^2 \;=\; \sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2\,.$$

Thus another way to obtain the SSE is to add together the within-group sums of squares:

$$\text{SSE} \;=\; \sum_{j=1}^{n}(Y_{1j} - \bar{Y}_1)^2 + \sum_{j=1}^{n}(Y_{2j} - \bar{Y}_2)^2 + \cdots + \sum_{j=1}^{n}(Y_{kj} - \bar{Y}_k)^2\,.$$

We can write this more concisely using the double summation notation, as below, where we also use the fact that each deviation $Y_{ij} - \bar{Y}_i$ is just a *residual*.

Error Sum of Squares (Alternative Formula):

$$\text{SSE} \;=\; \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_i)^2 \;=\; \sum_{i=1}^{k}\sum_{j=1}^{n}e_{ij}^2\,. \tag{10.12}$$

This says that the error sum of squares is simply the ***sum of squared residuals***. The double summation notation indicates that we're summing the squared residuals for all N individuals, that is, for the n individuals ($j = 1, 2, \ldots, n$) in every one of the k groups ($i = 1, 2, \ldots, k$). The error sum of squares reflects the sizes of the deviations of individual observations away from their corresponding group means, that is, the sizes of the residuals. Those deviations (the residuals) will be large if there's substantial variation within each group, and in this case SSE will be large. It's in this sense that SSE measures *within-groups* variation. Furthermore, because the residuals are approximations of the *random errors* in the ANOVA model, a large value of SSE indicates a large degree of variation in the data due to random error ("noise").

10.2.9 The ANOVA Partition of the Variation in the Data

Introduction

We can think of *between-groups* variation in the data as arising primarily from the nonrandom effect of the factor (that is, from differences among the population means $\mu_1, \mu_2, \ldots, \mu_i$) and *within-groups* variation as arising purely from random error (heterogeneity among individuals within the k groups). We'll see in a bit that if we were to lump the observations from all k groups together into one big sample, the two types of variation (between- and within-groups) would account for *all* of the variation in the data.

Total Variation

To see what this means, consider combining the k groups into one big, overall sample. Then the *total variation* in the data can be measured by the variance (squared standard deviation) of the overall sample. This variance reflects the sizes of the deviations of the N individual observations away from the overall mean $\bar{Y}$, and is written using double summation notation as

$$S^2 \;=\; \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y})^2}{N-1}\,.$$

The double summation is used because we're summing the squared deviations for all N individuals, that is, for the n individuals ($j = 1, 2, \ldots, n$) in every one of the k groups ($i = 1, 2, \ldots, k$).

The numerator of this overall sample variance is called the ***total sum of squares***, denoted **SSTo**.

Total Sum of Squares:

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

The total sum of squares measures *total* variation in the data. It will be large if either there are large differences *between* the groups *or* there's substantial variation of observations *within* each group. Therefore, SSTo reflects both *between–groups* variation *and within-groups* variation.

Partition of the Total Variation

To see what was meant at the beginning of this section by the statement that the *between-groups* variation and *within-groups* variation in the data account for the *total* variation (after lumping the groups together), we have the following fact, which is known as the ***ANOVA partition*** of the total variation in the data.

Fact 10.3 ANOVA Partition: The sums of squares defined previously satisfy the following relation.

$$\text{SSTo} = \text{SSTr} + \text{SSE}. \tag{10.13}$$

This splits the variation in the data as:

$$\text{Total Variation} = \text{Between-Groups Variation} + \text{Within-Groups Variation}$$

A mathematical verification of the ANOVA partition is given in Subsection 10.2.14. The partition is illustrated in the following example.

Example 10.5: Sums of Squares and the ANOVA Partition

For the data on lead measurements at the five labs (Example 10.3), statistical software gives the following values for the sums of squares.

$$\begin{aligned} \text{SSTo} &= 36.758 \\ \text{SSTr} &= 13.813 \\ \text{SSE} &= 22.945. \end{aligned}$$

We see that indeed SSTo = SSTr + SSE since

$$36.758 = 13.813 + 22.945.$$

This indicates that more than a third of the total variation in the lead measurements (13.813 out of 36.758) is the result of differences among the five labs' group means.

10.2.10 Degrees of Freedom

Associated with each sum of squares is a quantity called its ***degrees of freedom*** (or ***df***). The degrees of freedom are the number of deviations, among those used to compute the sum of squares, that are "free to vary." They're important because they'll be used later to obtain p-values for ANOVA F tests.

To illustrate, consider the sum of squares in the numerator of a sample variance S^2 (10.9). We saw in Chapter 3 that for any set of data $X_1, X_2, \ldots, X_n$, the deviations $X_i - \bar{X}$ used to compute that sum of squares always add up to zero. Because they sum to zero, the value of any one of them can be determined from the values of the other $n-1$, so we say only $n-1$ of the deviations are "free to vary." Thus there are $n-1$ *degrees of freedom* associated with the sum of squares in a sample variance.

Here are the degrees of freedom associated with each of the sums of squares SSTo, SSTr, and SSE.

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

$$\begin{aligned} df \text{ for SSTo} &= N-1 \\ df \text{ for SSTr} &= k-1 \\ df \text{ for SSE} &= k(n-1) = N-k \end{aligned}$$

We'll look at these one at a time to see why they make sense. First, for SSTo, the degrees of freedom is $N-1$ because the N deviations $Y_{ij} - \bar{Y}$ used to compute SSTo sum to zero, so only $N-1$ of them are "free to vary." For SSTr, the degrees of freedom is $k-1$ because the k deviations $\bar{Y}_i - \bar{Y}$ used to compute SSTr sum to zero, so only $k-1$ of them are "free to vary." Finally, for SSE, the degrees of freedom is

$k(n-1) = N-k$, because within each of the k groups, the n deviations $Y_{ij} - \bar{Y}_i$ used to compute SSE sum to zero, so only $n-1$ of them (within each group) are "free to vary" Therefore, in all, because there are k groups, only $k(n-1) = N-k$ of the deviations used to compute SSE are "free to vary."

The degrees of freedom, like the associated sums of squares in (10.13), are additive in the following sense.

Fact 10.4 The degrees of freedom given above satisfy the following relation.

$$df \text{ for SSTo } = df \text{ for SSTr} + df \text{ for SSE.} \tag{10.14}$$

This is easily verified by noting that

$$N - 1 \;=\; (k-1) \;+\; (N-k).$$

Example 10.6: Degrees of Freedom and the ANOVA Partition

For the study of lead measurements from five labs, we have $k = 5$, $n = 10$, and $N = 50$, so

$$df \text{ for SSTo } = N - 1 = 49,$$

$$df \text{ for SSTr } = k - 1 = 4,$$

and

$$df \text{ for SSE } = N - k = 45.$$

We see that indeed (10.14) holds, as expected, since $49 = 4 + 45$.

10.2.11 Mean Squares

Introduction

The treatment sum of squares SSTr measures *between-groups* variation among group means, and the error sum of squares SSE measures of *within-groups* variation due to random error. But these two measures of variation aren't directly comparable to each other (because they depend differently on n and k). To make them comparable, we convert them to *mean squares*. A ***mean square*** is defined as a sum of squares divided by its degrees of freedom:

$$\text{Mean Square } = \frac{\text{Sum of Squares}}{\text{Degrees of Freedom}}.$$

A familiar example of a mean square is the sample variance S^2 (10.9), which we interpret as an "average" squared deviation away from the sample mean. "Averaging" is done using $n-1$ instead of n to produce a more accurate estimator of the population variance σ^2.

We define the ***mean square for treatments***, denoted **MSTr**, and the ***mean squared error***, denoted **MSE**, as below.

Mean Squares: For one-factor ANOVA, the mean square for treatments and mean squared error

are

$$\text{MSTr} = \frac{\text{SSTr}}{k-1} \tag{10.15}$$

$$\text{MSE} = \frac{\text{SSE}}{N-k}. \tag{10.16}$$

These two measures of between- and within-groups variation are directly comparable to each other and will be used later in the ANOVA F test to test for differences among the k population means.

MSTr and MSE Under H_0 and H_a

The mean square for treatments MSTr measures between-groups variation among the group means $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$, which will vary substantially if the population means $\mu_1, \mu_2, \ldots, \mu_k$ differ, but will also vary to a lesser extent just by chance (sampling error) even when the population means are equal. To test for differences among the population means, we'll need to distinguish between variation in the $\bar{Y}_i$'s that's due purely to sampling error and variation that's due to differences among the population means (*and* sampling error).

The mean squared error MSE measures variation in individual observations within groups. It turns out that when the population means are equal, MSTr and MSE will be approximately equal, but when there are differences among the population means, MSTr will tend to be larger than MSE. This fact, stated formally below, will be used later to test for differences among the population means.

Fact 10.5 Consider random samples from k populations. Suppose that the one-factor ANOVA model (10.2) or (10.5) is appropriate and that the ϵ_{ij}'s are independent and $\epsilon_{ij} \sim \text{N}(0, \sigma)$.

Consider also the equivalent sets of hypotheses:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$	or	$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$
H_a : The μ_i's aren't all equal		H_a : The α_i's don't all equal 0

Then

1. When H_0 is true, both MSE and MSTr are estimators of σ^2.
2. When H_a is true, MSE is still an estimator of σ^2, but MSTr will tend to *overestimate* σ^2.

Comment: Here's why MSTr estimates σ^2 when the population means are identical. We know that the standard error of a sample mean $\bar{Y}$ is $\sigma/\sqrt{n}$, so the *variance* of $\bar{Y}$ is σ^2/n. Now, if the population means are identical, we can think of the sample means $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$ as a random sample from a distribution whose variance is σ^2/n, in which case the sample variance of the $\bar{Y}_i$'s,

$$S^2 = \frac{\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2}{k-1},$$

is an estimate of σ^2/n. In order to get an estimate of σ^2, we'd need to multiply this sample variance by n. Doing so gives exactly the MSTr (see the n in the front of SSTr in (10.10)). This is *why the n appears in the front of SSTr.*

Comment: To see why MSE estimates σ^2, regardless of whether or not the null hypothesis is true, consider the sample variances $S_1^2, S_2^2, \ldots, S_k^2$. Because the population variances are all equal to σ^2, each S_i^2 is an

estimate of σ^2, regardless of whether the null or alternative hypothesis is true. Writing SSE in terms of the S_i^2's as in (10.11), the MSE can be written as

$$\begin{aligned} \text{MSE} &= \frac{(n-1)S_1^2 + (n-1)S_2^2 + \cdots + (n-1)S_k^2}{k(n-1)} \\ &= \frac{S_1^2 + S_2^2 + \cdots + S_k^2}{k}, \end{aligned}$$

the average of the sample variances. Since each S_i^2 estimates σ^2 regardless of whether the null or alternative hypothesis is true, their average, the MSE, does too.

Comment: The reason why MSTr overestimates σ^2 when the alternative hypothesis is true is that differences among the μ_i's will contribute to excess variation among the $\bar{Y}_i$'s, which in turn will inflate the value SSTr and therefore also that of MSTr.

Estimating σ

Because the MSE estimates σ^2, the common population variance (or variance of the N(0, σ) error distribution) regardless of whether the null or alternative hypothesis is true, its square root estimates σ.

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

$$\hat{\sigma} = \sqrt{\text{MSE}}. \tag{10.17}$$

10.2.12 The One-Factor ANOVA *F* Test

Hypotheses and Test Statistic

Recall (Subsection 10.2.5) that the one ***one-factor ANOVA F test*** is a test of

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$	or	$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$
H_a : The μ_i's aren't all equal		H_a : The α_i's don't all equal 0

depending on whether the group means version or the treatment effects version of the ANOVA model is being used to describe the data.

Here's the ***one-factor ANOVA F test statistic***.

One-Factor ANOVA *F* Test Statistic:

$$F = \frac{\text{MSTr}}{\text{MSE}}. \tag{10.18}$$

The following properties of the F statistic will help us interpret its observed value.

- Because MSTr and MSE measure between- and within-groups variation in the data, respectively, we can think of F as
$$F = \frac{\text{Between-Groups Variation}}{\text{Within-Groups Variation}}.$$
- If the null hypothesis was true, we'd expect $F \approx 1$ because MSTr and MSE would both estimate σ^2.
- If the alternative hypothesis was true, we'd expect $F > 1$ because MSE would estimate σ^2 (still) but MSTr would tend to overestimate σ^2.

Therefore,

Large values of F (larger than about 1) provide evidence in favor of H_a : The μ_i's aren't all equal (or H_a : The α_i's don't all equal 0).

The *F* Distribution

To decide whether an observed value of F 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 F Under H_0: Consider random samples from k populations. Suppose that the one-factor ANOVA model (10.2) or (10.5) is appropriate and that the ϵ_{ij}'s are independent and either the ϵ_{ij}'s are N(0, σ) or the k sample sizes n are all large. Then when

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \qquad \text{or} \qquad H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

is true, the F test statistic follows (at least approximately) a distribution called the ***F distribution***, which has two parameters, its $\boldsymbol{k-1}$ ***numerator degrees of freedom*** and its $\boldsymbol{N-k}$ ***denominator degrees of freedom***. We write this as

$$F \;=\; \frac{\text{MSTr}}{\text{MSE}} \;\sim\; F\left(k-1,\; N-k\right).$$

There's a different F *distribution* for each pair of values for its *numerator* and *denominator degrees of freedom*. All F distributions are right skewed and lie entirely to the right of zero. Together, the numerator and denominator degrees of freedom determine center and spread of the distribution. Some F distribution density curves are shown below for various values of the degrees of freedom.

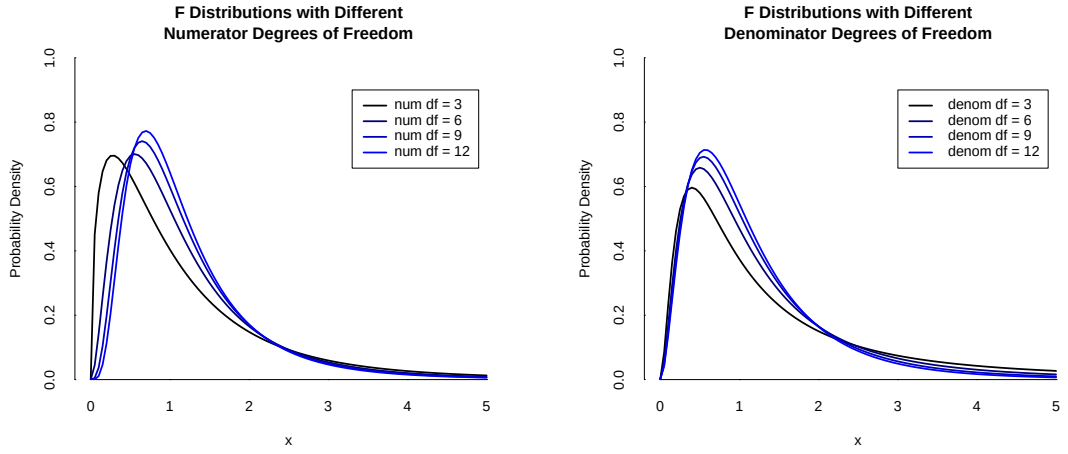


Figure 10.6: F distributions with 10 denominator degrees of freedom, but different values for the numerator degrees of freedom (left). F distributions with 6 numerator degrees of freedom, but different values for the denominator degrees of freedom (right).

P-Values

Because *large* values of the F statistic provide evidence against the null hypothesis, the p-value for the ANOVA F test is the tail probability under the F distribution to the *right* of the observed F value (and

the rejection region, for the rejection region approach, is the rightmost $100\alpha\%$ of the distribution). Fig. 10.7 shows the p-value when the test statistic value is $F = 2.7$ and there are $k = 4$ groups with $n = 5$ observations per group.

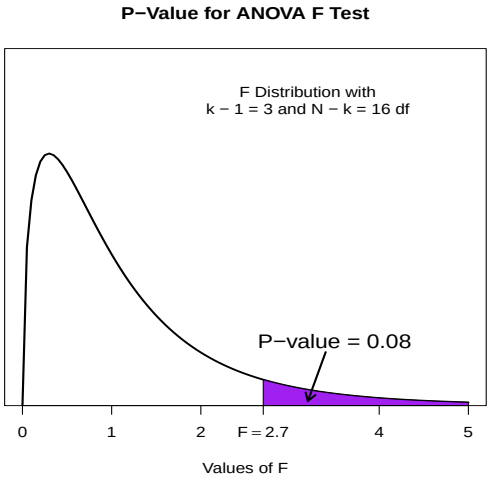


Figure 10.7: P-value for the ANOVA F test, when the test statistic value is $F = 2.7$ and there are $k = 4$ groups with $n = 5$ observations per group, is the tail area to the right of 2.7 under the $F(3, 16)$ distribution.

The One-Factor ANOVA *F* Test Procedure

The one-factor ANOVA F test procedure is summarized in the table below.

One-Factor ANOVA F Test for $\mu_1, \mu_2, \ldots, \mu_k$

Assumptions: Data are random samples from k populations (or responses of individuals to k treatments in a randomized experiment), the one-factor ANOVA model (10.2) or (10.5) is appropriate, the ϵ_{ij}'s are independent and either they follow a N(0, σ) distribution or the k sample sizes n are all large.*

Null hypothesis: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ or $H_0 : \alpha_1 = \alpha_1 = \ldots = \alpha_k = 0$.

Test statistic value: $F = \dfrac{\text{MSTr}}{\text{MSE}}$.

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

Alternative hypothesis	P-value = area under F-distribution with $k-1$ and $N-k$ d.f.:	Rejection region = F values such that:**
H_a : The μ_i's aren't all equal or H_a : The α_i's don't all equal 0	to the right of F	$F \geq F_{\alpha,k-1,N-k}$

* The ANOVA F test can also be carried out (using statistical software) when the k sample sizes $n_1, n_2, \ldots, n_k$ 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.

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

The ANOVA Table

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

One-Factor ANOVA Table:

Source	DF	SS	MS	F	P-value
Factor	$k-1$	SSTr	MSTr=SSTr/$(k-1)$	F=MSTr/MSE	p
Error	$N-k$	SSE	MSE=SSE/ $(N-k)$		
Total	$N-1$	SSTo			

The table is arranged so that the first row (labeled "Factor") pertains to *between-groups* variation, the second row ("Error") to *within-groups* variation, and the last row ("Total") to *total* variation.

Carrying Out the ANOVA F Test

In practice, ANOVA is carried out using statistical software, which will produce the ANOVA table.

Example 10.7: One-Factor ANOVA

For the data from the study of lead measurements made at the five labs (Example 10.3), we want to decide if there are any differences among the five labs' results. So we're testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \qquad \text{or} \qquad H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$
$$H_a : \text{The } \mu_i\text{'s aren't all equal} \qquad H_a : \text{The } \alpha_i\text{'s don't all equal 0}$$

where μ_i is the true (unknown) mean result upon sending a water specimen to the ith lab (and α_i is the true (unknown) *effect* of the ith lab on a lead measurement).

The ANOVA table, obtained using statistical software, is shown below.

Source	DF	SS	MS	F	P-value
Lab	4	13.813	3.453	6.77	0.000
Error	45	22.945	0.510		
Total	49	36.758			

We see that

$$\text{SSTr} = 13.813, \quad \text{SSE} = 22.945, \quad \text{and} \quad \text{SSTo} = 36.758.$$

Also,

$$df \text{ for SSTr} = 4, \quad df \text{ for SSE} = 45, \quad \text{and} \quad df \text{ for SSTo} = 49.$$

We can confirm that the mean squares are equal to the sums of squares divided by their degrees of freedom,

$$\text{MSTr} = 3.453 = \frac{13.813}{4} \qquad \text{MSE} = 0.510 = \frac{22.945}{45},$$

and that the F statistic is the ratio of MSTr to MSE,

$$F = 6.77 = \frac{3.453}{0.510}.$$

The p-value 0.000 is the tail area to the right of $F = 6.77$ under the $F(4, 45)$ distribution curve. Using a level of significance $\alpha = 0.05$, we reject the null hypothesis and conclude that there are statistically significant differences among the group means given in Example 10.3.

We'll see how to check the normality and common standard deviation assumptions required by the F test in Section 10.2.13.

The estimate $\hat{\sigma}$ of the true (unknown) standard deviation σ of the N(0, σ) distribution of the random error is

$$\hat{\sigma} = \sqrt{\text{MSE}} = \sqrt{0.510} = 0.7144.$$

Because we concluded that there are differences among the five labs' means, a logical next question would be "Which labs' means differ from each other?" We'll see a method for answering this question in Section 10.7.

10.2.13 Using Residuals to Check the ANOVA *F* Test Assumptions

The ANOVA F test requires that three conditions be met:

1. The k samples are from normal populations, or equivalently, the errors ϵ_{ij} in the ANOVA model follow a normal distribution.

2. The k populations 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.
3. The k random samples are collected independently of each other or, equivalently, the random errors in the ANOVA model are independent of each other.

The third assumption is usually addressed in the study design. Individual observations should be separated sufficiently in space and time to ensure independence. The other assumptions (normality and common σ) are checked via plots of the residuals.

Checking the Normality Assumption

Although we *could* check normality of the data separately for each group using a histogram or normal probability, these plots can be uninformative when the sample sizes are small. Instead, it's usually preferable to plot the N *residuals* altogether in one histogram or normal probability plot. Since the residuals e_{ij} are approximations of the errors ϵ_{ij} in the ANOVA model, the normality assumption is tenable as long as the residual plot doesn't show strong signs of non-normality.

Checking the Common σ Assumption

The assumption that the k populations share a common standard deviation is tenable if the amount of variation within the groups is about the same from one groups to the next. There are a few ways to check this assumption.

1. **Plot the residuals versus the fitted values**: In a plot of the residuals (y axis) versus the fitted values (group means, x axis), the amount of vertical spread away from a horizontal line at $y = 0$ should be about the same from one group to the next. Fig. 10.9 in Example 10.8 shows the residuals versus the fitted values (group means) for the lead measurements made at the five labs. The amount of vertical spread above and below the line is roughly the same from one group to the next, suggesting that the common population standard deviation assumption is tenable. The five labs are ordered from left to right by their group means (fitted values). It's not uncommon for groups whose means are larger to also have larger standard deviations, so ordering the groups in this way makes it easier to spot violations of the common standard deviation assumption when they exist. In Fig. 10.10 of Example 10.9, variation increases from left to right, suggesting that the common standard deviation assumption isn't tenable.
2. **Compare sample standard deviations**: Another way to check the common σ assumption is to use the following rule of thumb. *If the largest of the k 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. In particular, when the sample sizes are small, the sample standard deviations will tend to vary a lot, and in this case the rule may be a bit conservative.

We'll see what to do when the normality and constant standard deviation assumptions *aren't* met in Sections 11.3 and 10.4.

Example 10.8: Checking Assumptions for One-Factor ANOVA

In Example 10.7, an ANOVA F test found statistically significant differences among the mean lead measurements made at the five labs. To validate this result, we need to check the normality and common standard deviation assumptions.

The residuals, obtained using statistical software, are shown in a histogram and normal probability plot below.

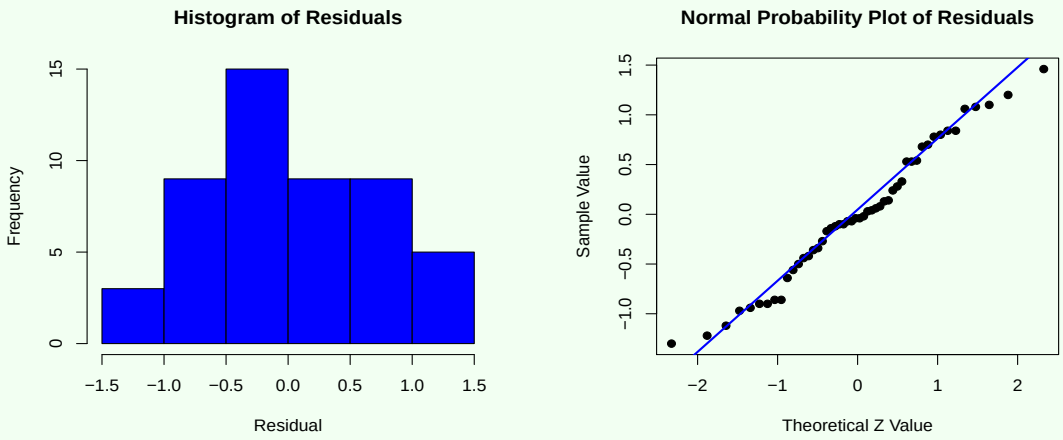


Figure 10.8: Histogram (left) and normal probability plot (right) of the residuals after fitting the one-factor ANOVA model to the lead measurements data.

The plots don't give any indications of non-normality, so the normality assumption appears to be met.

A plot of the residuals versus the fitted values (group means) is shown below.

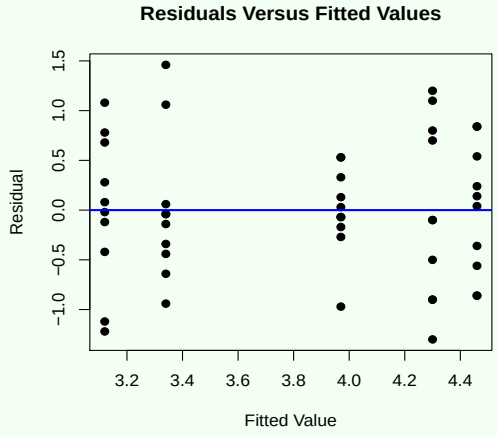


Figure 10.9: Plot of residuals versus fitted values after fitting the one-factor ANOVA model to the lead measurements data.

The amount of (vertical) spread in the points appears to be fairly constant from left to right, suggesting that it's reasonable to assume that the true (unknown) population standard deviation σ is the same from one lab to the next. Furthermore, the largest sample standard deviation, $S_1 = 0.904$, is only about twice as large as the smallest, $S_2 = 0.440$, which doesn't raise any alarms about the population standard deviations being unequal.

Since the normality and common standard deviation assumptions appear to be met, the results of the ANOVA F test performed in Example 10.7 are valid.

10.2.14 Comments on the ANOVA Partition of the Total Variation

We wrap up this section with a look at why the ANOVA partition

$$\text{SSTo} \;=\; \text{SSTr} + \text{SSE} \tag{10.19}$$

holds. To this end, notice that we can write a total deviation $Y_{ij} - \bar{Y}$ as

$$\underbrace{Y_{ij} - \bar{Y}}_{\substack{\text{Total deviation of individual} \\ \text{observation away from} \\ \text{overall mean}}} \;=\; \underbrace{\bar{Y}_i - \bar{Y}}_{\substack{\text{Deviation of treatment} \\ \text{group mean away from} \\ \text{overall mean}}} \;+\; \underbrace{Y_{ij} - \bar{Y}_i}_{\substack{\text{Deviation of individual} \\ \text{observation away from} \\ \text{treatment group mean}}} \tag{10.20}$$

If we square both sides of (10.20) and then sum over all $N = nk$ individuals, it can be shown that we end up with:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y})^2}_{\text{Total sum of squares}} \;=\; \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{Y}_i - \bar{Y})^2}_{\text{Treatment sum of squares}} \;+\; \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2}_{\text{Error sum of squares}} \tag{10.21}$$

The cross products terms $(\bar{Y}_i - \bar{Y})(Y_{ij} - \bar{Y}_i)$ don't show up on the right side of (10.21) because, it turns out, they sum to zero.

The left side of (10.21) is the total sum of squares SSTo. The first term on the right side is the treatment sum of squares SSTr, and the second term is the error sum of squares SSE. Thus (10.21) is equivalent to (10.19).

10.3 Dealing With Unequal Standard Deviations: Transformations and Nonparametric Procedures

As mentioned in Subsection 10.2.13, it's not uncommon for groups whose means are larger to also have larger standard deviations. This is illustrated in Fig. 10.10. When the variation changes with the mean, the common standard deviation assumption required by the ANOVA F test isn't met, and so the results of the F test might not be valid. Instead, there are two options:

1. **Transform the data to equalize the standard deviations**: When the standard deviation increases with the mean, it's sometimes possible to make a transformation of the data, most commonly by taking their logs or their square roots, so that the transformed data have a more constant standard deviation across the groups. Often, a standard deviation that increases with the mean is accompanied by right-skewness of the data, and taking logs (or square roots) corrects for both the non-constant standard deviation *and* the non-normality. Once the data have been transformed, the ANOVA F test can be carried out on the transformed data.
2. **Carry out a nonparametric test**: We could carry out a nonparametric test that doesn't require an assumption of common population standard deviations, such as the *Kruskal-Wallis test* described in Section 10.5.

Example 10.9: Transformation to Stabilize the Standard Deviation

Environmental and ecological studies that focus on direct chemical analysis of water or sediment don't necessarily reveal actual ecological health. It's sometimes preferable, therefore, to instead focus on indicators of ecological health such as species richness and species diversity, among others.

The table below shows data on species richness (number of different species present) in 0.1 m^2 quadrats at 23 locations on the soft bottom of the Saronikos Gulf off the coast of Athens, Greece [12]. Each location is classified as mud, muddy sand, or sandy mud. One outlier has been removed. The sample means and standard deviations are given at the bottom of the table.

Species Richness

Mud	Sandy Mud	Muddy Sand
24	24	20
8	54	20
6	20	36
12	41	71
14	39	48
15		39
22		56
		30
		37
		78
		52
$\bar{Y}_1 = 14.4$	$\bar{Y}_2 = 35.6$	$\bar{Y}_3 = 44.3$
$S_1 = 6.7$	$S_2 = 13.8$	$S_3 = 19.0$

The largest sample standard deviation, $S_2 = 19.0$, is more than twice as large as the smallest, $S_1 = 6.7$, suggesting that the common population standard deviation assumption required for an ANOVA F test isn't tenable. Boxplots of the data and a plot of the residuals versus the fitted values (group means) after fitting the ANOVA model are shown below.

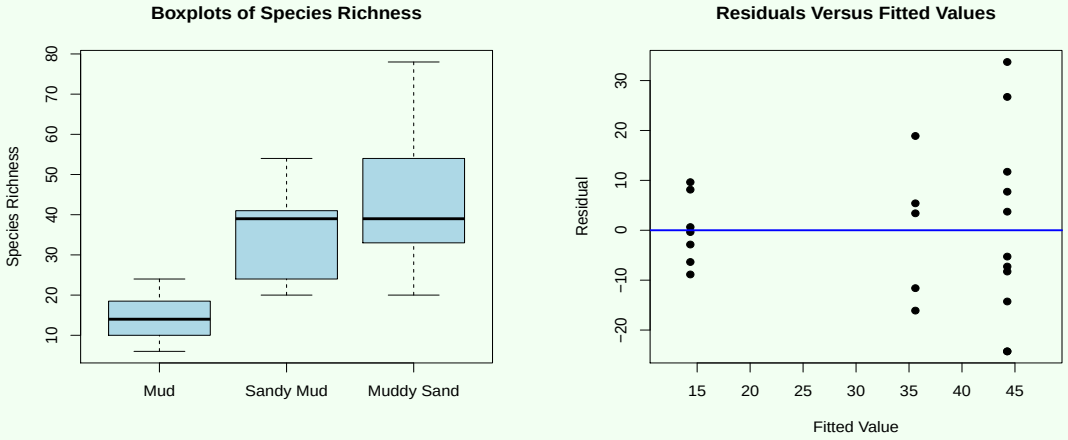


Figure 10.10: Boxplots of the species richness data (left) and plot of the residuals versus fitted values (right).

Both plots show that the standard deviation increases with the mean, suggesting that a log transformation might help stabilize the standard deviation across the groups. Boxplots of the data after taking their logs and a plot of the residuals versus fitted values after fitting the ANOVA model to the logs are shown below. The plots confirm that the standard deviation has become more constant from group to group, so an ANOVA F test using the logs of the data would be justified.

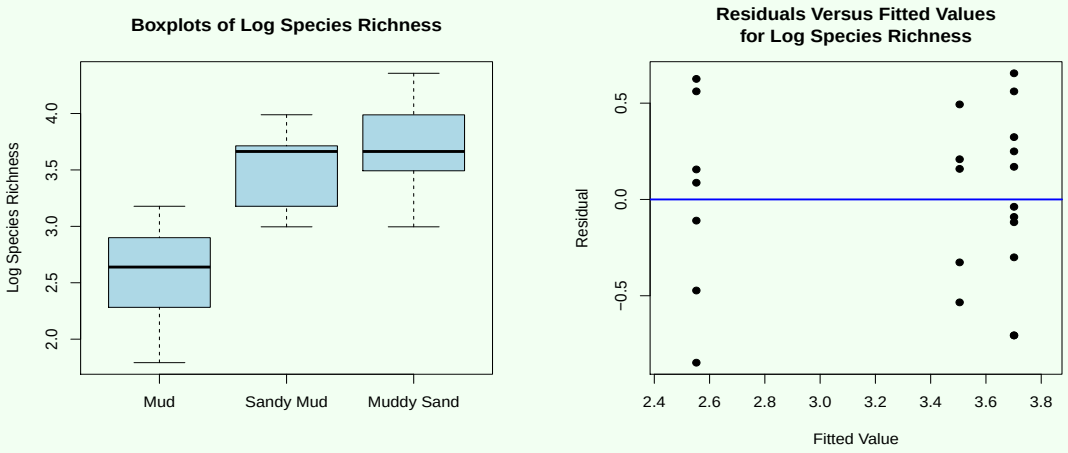


Figure 10.11: Boxplots of the logs of the species richness data (left) and plot of the residuals versus fitted values using the log data (right).

10.4 Dealing With Non-Normal Data: Transformations and Nonparametric Procedures

The ANOVA F test, like the t tests, is a *parametric* test because it requires that the k samples were drawn from normal populations (or equivalently that the error term ϵ in the ANOVA model follows a normal distribution). If this assumption isn't met (and the sample sizes aren't large), there are two options:

1. **Transform the data to normality**: It's sometimes possible to transform all k samples (same transformation on every sample), for example by taking their logs or using another transformation in the Ladder of Powers, so that the transformed values are more normally distributed, and then carry out the ANOVA F test on the transformed data. As mentioned in the previous section, such transformations often make the common standard deviation assumption more tenable too.
2. **Carry out a nonparametric test**: We could carry out a nonparametric test that doesn't rely on an assumption of normality. The *Kruskal-Wallis test* described in the next section is a nonparametric alternative to the ANOVA F test.

A third option, when the data follow a known (but non-normal) distribution (for example when they're Poisson counts) is to use a procedure specifically developed for use with that distribution. These procedures can be found in more specialized statistics books.

10.5 Kruskal-Wallis Test

10.5.1 Introduction

The ***Kruskal-Wallis test*** is a *nonparametric* test for deciding if there are any differences among the means of k populations. Unlike the ANOVA F test, it doesn't require a normality assumption about the populations, so it's a nonparametric alternative to the ANOVA F test.

Example 10.10 presents a study for which the ANOVA F might not be appropriate, but the Kruskal-Wallis test would.

Example 10.10: Kruskal-Wallis Test

The Large River Monitoring Network is part of the U.S. Geological Survey's Biomonitoring of Environmental Status and Trends Program that monitors environmental contaminants in several large U.S. river basins. In each river basin, contaminants and their biological indicators are measured in fish at several monitoring stations.

The table below shows aluminum (Al) concentrations (μg/g wet weight) measured in female common carp in three river basins, the Colorado, Columbia, and Mississippi River basins. (This is a subset of the full set of data that was collected.)

Al in Fish

Colorado River Basin	Columbia River Basin	Mississippi River Basin
32	18.9	54.9
232	64.1	24.1
36	33.2	40.2
73	48.8	52.7
53	13.7	28.9
20	21.1	73.3
28	43.2	26.7
24	66.5	26.6
24		32.6
11		
37		
30		
26		

Side-by-side boxplots and dot plots of the data are shown below.

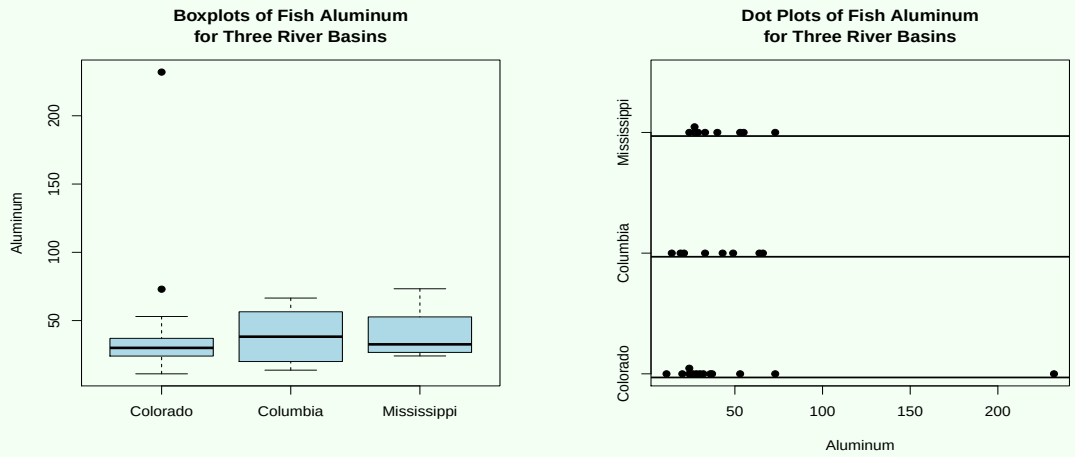


Figure 10.12: Boxplots (left) and dot plots (right) of aluminum in female carp from the Colorado, Columbia, and Mississippi River basins.

We want to decide if there are any differences among the mean Al concentrations for the three basins. The plots give a slight indication that the samples are from right skewed distributions, and in particular, the Colorado's has a large outlier, so because the sample sizes aren't large, a one-factor ANOVA F test might not be valid. Instead, in Example 10.11, we'll conduct a Kruskal-Wallis test.

10.5.2 The Kruskal-Wallis Test Procedure

We'll assume that we have independent samples of sizes $\boldsymbol{n_1, n_2, \ldots n_k}$ (not necessarily the same) from k *continuous* populations that have the *same shape* (but not necessarily normal). We'll want to test

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a\text{: The } \mu_i\text{'s aren't all equal}$$

where $\boldsymbol{\mu_1, \mu_2, \ldots, \mu_k}$ are the (unknown) population means.

The Kruskal-Wallis test is carried out by combining the observations from the k groups, sorting and ranking them from smallest (rank = 1) to largest (rank = N), and then comparing the mean *ranks* for the k groups.

As before, we'll use the notation

$\boldsymbol{Y_{ij}}$ = The jth observation in the ith group.

and

$\boldsymbol{N}$ = The ***overall sample size***, that is, the total number of observations in all k groups combined.

We'll also let

$\boldsymbol{R_{ij}}$ = The *rank* of Y_{ij} in the overall sample (after combining the k groups).

and

$\bar{R}_i$ = The ***ith group mean rank***, defined to be the mean of the *ranks* (in the overall combined sample) of the observations from the ith group, that is,

$$\bar{R}_i \;=\; \frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij}\,.$$

Also, we define

$\bar{R}$ = The ***overall mean rank***, defined as the mean *rank* of the N observations in the combined sample. Because these ranks are $1, 2, \ldots, N$, we have

$$\bar{R} \;=\; \frac{1}{N}\,(1+2+\cdots+N)\,. \tag{10.22}$$

Using the shortcut formula ((8.20) in Chapter 8),

$$1+2+\cdots+N \;=\; \frac{N(N+1)}{2},$$

it's easy to see that a simplified form for $\bar{R}$ is

$$\bar{R} \;=\; \frac{N+1}{2}.$$

Here's how to compute the ***Kruskal-Wallis test statistic***, denoted K_w.

Kruskal-Wallis Test Statistic:

1. Combine the observations from the k groups into an overall sample, keeping track of which group each observation originally belonged to, sort and rank the observations in the overall sample from smallest (rank = 1) to largest (rank = N). 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.
2. Compute the group mean ranks $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ and the overall mean rank $\bar{R}$.
3. The test statistic is
$$K_w \;=\; \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2. \tag{10.23}$$

This test statistic is the sum of (sample size-weighted) squared deviations of group mean ranks away from the overall mean rank, multiplied by the constant $12/(N(N+1))$. The constant has the effect of "averaging" the squared deviations, and is needed in order for K_w to follow a particular sampling distribution under the null hypothesis (see Subsection 10.5.2). The sum of squares in K_w measures variation in the group mean *ranks*, just as the treatment sum of squares SSTr in ANOVA measured variation in the group means of the original observations.

The next example demonstrates the calculation of the test statistic.

Example 10.11: Kruskal-Wallis Test Statistic

For the data on fish aluminum from Example 10.10, we have $k = 3$ river basins (the populations),

and we want to test

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3$$
$$H_a\text{: The } \mu_i\text{'s aren't all equal}$$

where μ_1, μ_2, and μ_3 are the true (unknown) population mean Al concentrations in female carp from the Colorado, Columbia, and Mississippi River basins, respectively.

The three sample sizes are $n_1 = 13$, $n_2 = 8$, and $n_3 = 9$, so $N = 13+8+9 = 30$. The combined samples, sorted and ranked, are given below (sample 1 = Colorado, 2 = Columbia, and 3 = Mississippi, with the different font shades indicating basins).

Sample	**1**	2	2	**1**	2	**1**	**1**	3	**1**
Observation	**11.0**	13.7	18.9	**20.0**	21.1	**24.0**	**24.0**	24.1	**26.0**
Rank	**1**	2	3	**4**	5	**6.5**	**6.5**	8	**9**

3	3	**1**	3	**1**	**1**	3	2	**1**	**1**	3
26.6	26.7	**28.0**	28.9	**30.0**	**32.0**	32.6	33.2	**36.0**	**37.0**	40.2
10	11	**12**	13	**14**	**15**	16	17	**18**	**19**	20

2	2	3	**1**	3	2	2	**1**	3	**1**	
43.2	48.8	52.7	**53.0**	54.9	64.1	66.5	**73.0**	73.3	**232.0**	
21	22	23	**24**	25	26	27	**28**	29	**30**	

The tied observations were assigned the average rank. Notice that the three basins are evenly "intermingled" in the overall sample because there's substantial overlap in their Al concentrations.

The three group mean ranks are:

$$\begin{aligned}\bar{R}_1 &= \frac{1}{13}(1+4+6.5+6.5+9+12+14+15+18+19+24+28+30)\\ &= 14.4.\end{aligned}$$

$$\begin{aligned}\bar{R}_2 &= \frac{1}{8}(2+3+5+17+21+22+26+27)\\ &= 15.4.\end{aligned}$$

$$\begin{aligned}\bar{R}_3 &= \frac{1}{9}(8+10+11+13+16+20+23+25+29)\\ &= 17.2.\end{aligned}$$

The mean ranks are similar in value (due to the "intermingling" of the three basins in the overall sample), and are therefore close to the overall mean rank,

$$\bar{R} = \frac{N+1}{2} = \frac{30+1}{2} = 15.5.$$

The observed test statistic value is

$$\begin{aligned}K_w &= \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i\left(\bar{R}_i - \bar{R}\right)^2\\ &= \frac{12}{30(30+1)}\left(13\cdot(14.4-15.5)^2 + 8\cdot(15.4-15.5)^2 + 9\cdot(17.2-15.5)^2\right)\\ &= 0.54.\end{aligned}$$

This test statistic value is close to zero because there's so little variation in the group mean ranks. We'll determine the p-value in Example 10.12.

When the null hypothesis is true, the k samples come from identical populations (they have the same shapes *and* centers), so observations from the different groups will be evenly "intermingled" in the sorted overall sample. In this case, the group mean ranks $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ won't differ much, except due to chance variation, and therefore won't differ much from the overall mean rank $\bar{R}$, so K_w will be fairly close to zero (or more precisely, as will be seen, close to $k-1$). On the other hand, when the alternative is true, the groups will be "segregated" in the sorted overall sample, leading to large differences among $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ and a large value of K_w. It follows that

Large values of K_w (larger than $k-1$) provide evidence in favor of H_a: Not all μ_i's are equal.

The Chi-Square Distribution

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

Fact 10.6 Consider independent random samples from k continuous populations whose means are $\mu_1, \mu_2, \ldots, \mu_k$. Suppose also that the populations have the same shape and that the sample sizes $n_1, n_2, \ldots, n_k$ are all large. Then when

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

is true, the Kruskal-Wallis test statistic K_w follows (approximately) a distribution called the ***chi-square distribution*** with $\boldsymbol{k-1}$ ***degrees of freedom***, sometimes denoted as the $\boldsymbol{\chi^2(k-1)}$ distribution. We write this as

$$K_w \sim \chi^2(k-1).$$

There's a different chi-square distribution for each value of its one parameter, its ***degrees of freedom***. All chi-square distributions are right skewed and lie entirely to the right of zero. The degrees of freedom determine the center and spread of the distribution. Several chi-square density curves are shown below for various values of its degrees of freedom.

Comment: The sample sizes are considered large enough for K_w to follow a chi-square distribution (under the null hypothesis) as long as they're all five or larger when $k > 3$ and all six or larger if $k = 3$.

Comment: The $k-1$ degrees of freedom associated with the test statistic K_w refers to the fact that only $k-1$ of the deviations $\bar{R}_i - \bar{R}$ used to compute K_w are "free to vary" (because they sum to zero).

The mean and standard error of the chi-square distribution are

Mean and Standard Error of the Chi-Square Distribution: The mean $\boldsymbol{\mu_{K_w}}$ and standard error $\boldsymbol{\sigma_{K_w}}$ of a chi-square distribution with $k-1$ degrees of freedom are

$$\mu_{K_w} = k-1$$

and

$$\sigma_{K_w} = \sqrt{2(k-1)}.$$

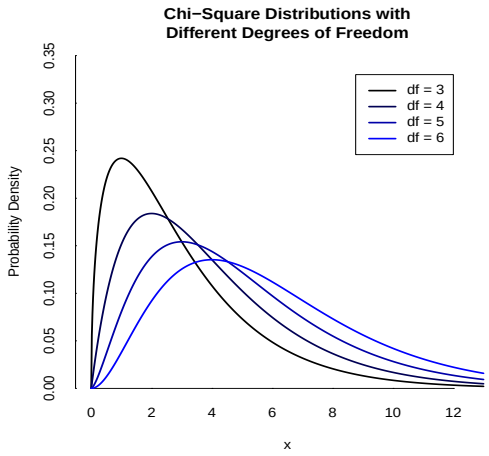


Figure 10.13: Chi-square distributions with different degrees of freedom.

Thus if the null hypothesis was true, we'd expect the value of K_w to be roughly equal to the mean of its sampling distribution, $k-1$. Values larger than this are evidence in favor of the alternative hypothesis.

P-Values

Because *large* values of the test statistic K_w provide evidence against the null hypothesis, the p-value for the Kruskal-Wallis test test is the tail probability under the $\chi^2(k-1)$ distribution to the *right* of the observed K_w value (and the rejection region, for the rejection region approach, is the rightmost $100\alpha\%$ of the distribution). Fig. 10.14 shows the p-value when $K_w = 8.3$ and there are $k = 5$ groups.

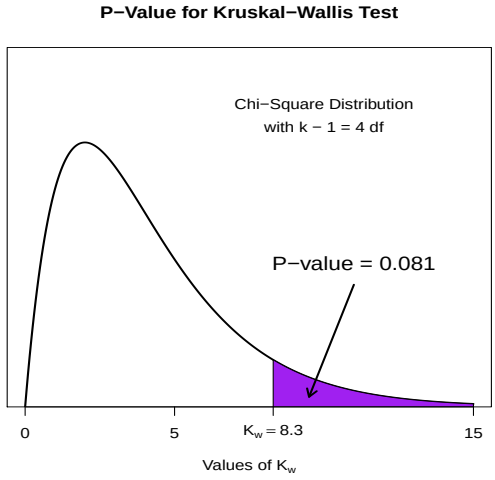


Figure 10.14: P-value for the Kruskal-Wallis test when the observed test statistic value is $K_w = 8.3$ and there are $k = 5$ groups. The p-value is the tail area to the right of 8.3 under the $\chi^2(4)$ distribution.

The Kruskal-Wallis test procedure is summarized in the table below.

Kruskal-Wallis Test for $\mu_1, \mu_2, \ldots, \mu_k$

Assumptions: The data are independent random samples from k continuous populations that differ, if at all, by their means $\mu_1, \mu_2, \ldots, \mu_k$ but not their shapes, and the sample sizes $n_1, n_2, \ldots, n_k$ are all large.*

Null hypothesis: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$.

Test statistic value: $K_w = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2$.

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

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

* The sample sizes are considered to be large when they're all 5 or larger if $k > 3$, and all 6 or larger if $k = 3$. For smaller sample sizes, the test statistic K_w can be compared to a table of tail areas or critical values of the exact sampling distribution of K_w, found, for example, in [5] or [11].

** $\chi^2_{\alpha,k-1}$ is the $100(1-\alpha)$th percentile of the χ^2 distribution with $k-1$ d.f.

Example 10.12: Kruskal-Wallis Test

In Example 10.11, the test statistic for deciding whether there are differences among the mean Al concentrations in fish from three river basins was found to be $K_w = 0.54$. Because it's so close to zero, this provides very little evidence for any differences among the three means.

From a table of upper tail areas of the chi-square distribution with $k - 1 = 2$ degrees of freedom, we find the p-value to be greater than 0.100. Statistical software reports the p-value as 0.7634. Thus, using a level of significance $\alpha = 0.05$, we fail to reject the null hypothesis, and conclude that there are no statistically significant differences among the three means.

10.5.3 What if the Population Distributions Don't All Have the Same Shape?

The Kruskal-Wallis test, as described above, relies on an assumption that the population distributions all have the same shape. But the test *can* still be carried out when they *don't* have the same shape. Recall that the same was true of the rank sum test (Chapter 8). However, in this situation, it doesn't testing for differences among the means $\mu_1, \mu_2, \ldots, \mu_k$. Instead, it tests hypotheses that can be stated in words as

H_0 : The k population distributions are identical (same means and shapes)
H_a : At least one distribution has values that are systematically different from the others

Here "systematically different" is interpreted as "systematically larger or systematically smaller," as defined more formally in Subsection 8.7.4 of Chapter 8, and could result from the populations having different means *or* having different shapes.

10.6 Which Test Should Be Used, the ANOVA F Test or the Kruskal-Wallis Test?

If the k samples are from a non-normal populations or from populations whose standard deviations aren't equal, and we don't want to transform the data, then unless the sample sizes are large, the one-factor ANOVA F test shouldn't be used. In this case we have little choice but to use a nonparametric test such as the Kruskal-Wallis test.

But if the normality assumption *is* met (and the population standard deviations are about the same), we have a choice between the ANOVA F test and the Kruskal-Wallis. It can be shown that when the normality and common standard deviation assumptions are met, the ANOVA F test is *more powerful* than the Kruskal-Wallis (or other nonparametric test). In other words, when the alternative hypothesis is true, the ANOVA F test is less likely to lead to a Type II error and thus more likely to find statistically significant differences or effects. Therefore, whenever we have a choice, *the ANOVA F test is preferred.*

The intuition behind why the ANOVA F test is more powerful is that it's based on the actual numerical values of the data (via the group means $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$) as opposed to just their ranks. Therefore the ANOVA F test makes use of the complete information that's contained in the data, whereas the Kruskal-Wallis test doesn't. As a result, there's some loss of power in the Kruskal-Wallis test. The lack of power in the Kruskal-Wallis test can be particularly severe when the samples sizes are small.

10.7 Multiple Comparison Procedures

10.7.1 Introduction

If H_0 is rejected in an ANOVA F test or a Kruskal-Wallis test, we can conclude that *at least one difference* exists *somewhere* among the k means. But further analysis is needed to determine *which* means differ from each other.

A ***multiple comparison procedure*** is a hypothesis testing procedure, conducted *after* the null is rejected in an F test or a Kruskal-Wallis test, that tests for differences, *one pair of means at a time.* In other words, it's used to performs multiple ***pairwise*** comparisons between the k means.

It can be shown that when there are k groups, the total number of pairwise comparisons that need to be made is

$$\text{Number of pairs } \mu_i \text{ and } \mu_j \text{ to compare} \;=\; \frac{k(k-1)}{2}. \tag{10.24}$$

Example 10.13: Multiple Comparisons

In Example 10.7 we found statistically significant evidence for differences among the $k = 5$ labs' mean lead measurements. The next step is to decide *which* labs differ from each other. To compare the labs to each other one pair at a time, according to (10.24), we need to make

$$\frac{5(5-1)}{2} \;=\; 10$$

comparisons. Here they are listed out:

Lab1 vs Lab2	Lab1 vs Lab3	Lab1 vs Lab4	Lab1 vs Lab5
Lab2 vs Lab3	Lab2 vs Lab4	Lab2 vs Lab5	
Lab3 vs Lab4	Lab3 vs Lab5		
Lab4 vs Lab5			

It may be tempting to simply perform $k(k-1)/2$ separate two-sample t tests (or pooled t tests or rank sum tests), each using a level of significance, say, $\alpha = 0.05$. The problem with this approach is that even though the probability of making a Type I error would be 0.05 on *any given* test, the probability of making *at least one* Type I error *somewhere* among the *set* of tests would be substantially greater than 0.05. For example, suppose the number of groups is $k = 15$. Then the number of pairwise comparisons is 105, and in this case, because a Type I error occurs 5% the time when a null hypothesis is true, we'd expect about 5 or 6 of the 105 tests to erroneously conclude the two means differ even if in reality the k population means were all the same.

The goal of a *multiple comparison procedure* is to keep the probability of making *at least one* Type I error among the *family* of pairwise comparison tests small. When multiple pairwise tests are carried out, we call the probability of making *at least one* Type I error (when the k population means are all the same) the ***familywise Type I error rate***, and denote it by $\boldsymbol{\alpha_f}$. Thus

$$\alpha_f \;=\; P(\text{You reject } H_0 : \mu_i - \mu_j = 0 \text{ for at least one pair, } \mu_i \text{ and } \mu_j)$$

(when in reality $\mu_1 = \mu_2 = \cdots = \mu_k$). The level of significance used on each *particular* pairwise comparison test, which is also the Type I error probability for that test, is called the ***pairwise Type I error rate*** and is denoted $\boldsymbol{\alpha_p}$. In other words, for each i and j,

$$\alpha_p \;=\; P(\text{You reject } H_0 : \mu_i - \mu_j = 0 \text{ for the particular pair, } \mu_i \text{ and } \mu_j)$$

(when in reality $\mu_i = \mu_j$).

10.7.2 The Bonferroni Procedure

Several multiple comparison procedures have been developed, all of which control the familywise Type I error rate at some desired level, such as $\alpha_f = 0.05$. The procedure we'll be looking at, called the ***Bonferroni procedure***, accomplishes this by using a sufficiently small level of significance for each pairwise test of hypotheses of the form

$$\begin{aligned} H_0 : \mu_i - \mu_j &= 0 \\ H_a : \mu_i - \mu_j &\neq 0 \end{aligned} \tag{10.25}$$

More specifically, it divides the overall familywise Type I error rate equally among the pairwise tests, so for example to perform the 10 pairwise tests comparing labs in Example 10.13, we'd use level of significance

$$\alpha_p \;=\; \frac{0.05}{10} \;=\; 0.005$$

for each test. In general, the Bonferroni procedure holds the overall familywise Type I error rate to some desired level α_f by using the following fact.

Fact 10.7 The familywise Type I error rate will be no greater than the desired level α_f if each pairwise test is performed using the ***Bonferroni-corrected level of significance***

$$\alpha_p \;=\; \frac{\alpha_f}{\text{Number of Pairwise Tests}} \;=\; \frac{\alpha_f}{k(k-1)/2}.$$

10.7.3 Bonferroni Procedure After an ANOVA F Test

Suppose in an ANOVA F, the null hypothesis is rejected. To decide *which* means among $\mu_1, \mu_2, \ldots, \mu_k$ differ from each other, because we've already assumed (for the F test) that the populations are normal (or we had large sample sizes), two-sample t tests are appropriate for the pairwise comparisons. In addition,

because we've also assumed that the population standard deviations are all equal to the same value, σ, instead of using the sample standard deviations S_i and S_j to compute the t test statistics, we can increase the power of each t test by ***pooling*** the k samples to get a more accurate, combined estimate of σ. The appropriate pooled estimate, from (10.17), is the square root of the mean squared error, obtained from the ANOVA table.

We now summarize the procedure.

Bonferroni Multiple Comparison Procedure After One-Factor ANOVA: After a one-factor ANOVA F test has rejected H_0, to decide which pairs of means differ while controlling the familywise Type I error rate at α_f, for each pair of means μ_i and μ_j, test the hypotheses

$$\begin{aligned} H_0 : \mu_i - \mu_j &= 0 \\ H_a : \mu_i - \mu_j &\neq 0 \end{aligned}$$

using the ***Bonferroni pairwise t test statistic***

$$t = \frac{\bar{Y}_i - \bar{Y}_j - 0}{\sqrt{\frac{\text{MSE}}{n} + \frac{\text{MSE}}{n}}} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\frac{2 \cdot \text{MSE}}{n}}} \tag{10.26}$$

and decision rule

Reject H_0 if p-value $< \alpha_p$
Fail to reject H_0 if p-value $\geq \alpha_p$.

where $\alpha_p = \alpha_f/(k(k-1)/2)$. When the corresponding H_0 is true, the test statistic (10.26) follows a t distribution with $N - k$ degrees of freedom, from which the p-value for that test is obtained.

Comment: Each pairwise t test uses the two-sided alternative hypothesis because no direction for the difference between the corresponding means was specified prior to performing the ANOVA F test.

Example 10.14: Bonferroni Procedure After an ANOVA *F* Test

Continuing from Example 10.13, we'll use the Bonferroni procedure to decide *which* labs' means differ from each other, while controlling the familywise Type I error rate at $\alpha_f = 0.05$.

For the first of the 10 comparisons listed in Example 10.13, we're testing

$$\begin{aligned} H_0 : \mu_1 - \mu_2 &= 0 \\ H_a : \mu_1 - \mu_2 &\neq 0 \end{aligned}$$

where μ_1 and μ_2 are the true (unknown) population mean lead measurements at Labs 1 and 2, respectively. Since $k = 5$, the Bonferroni-corrected level of significance to use for each pairwise test is

$$\alpha_p = \frac{0.05}{5(5-1)/2} = 0.005,$$

and so the decision rule is

Reject H_0 if p-value < 0.005
Fail to reject H_0 if p-value ≥ 0.005.

The sample means (from Example 10.3) are $\bar{Y}_1 = 4.30$ and $\bar{Y}_2 = 3.97$, the MSE (from the ANOVA table in Example 10.7) is 0.510, and the common sample size is $n = 10$. Thus the observed value of

the test statistic (10.26) is

$$t = \frac{4.30 - 3.97}{\sqrt{\frac{2(0.510)}{10}}} = 1.03,$$

and the p-value, from the t distribution with $N - k = 45$ degrees of freedom, is $2(0.1543) = 0.3086$. Thus we fail to reject the null hypothesis and therefore have no reason to believe that μ_1 and μ_2 differ.

The complete set of test statistic values and p-values for the pairwise tests is shown below. Statistically significant differences (at the Bonferroni-corrected significance level $\alpha_p = 0.005$) are marked with an asterisk.

Pair of Means	t	P-value
Lab1 vs Lab2	1.03	0.3070
Lab1 vs Lab3	-0.50	0.6188
Lab1 vs Lab4	3.69	0.0006*
Lab1 vs Lab5	3.01	0.0043*
Lab2 vs Lab3	-1.53	0.1320
Lab2 vs Lab4	2.66	0.0107
Lab2 vs Lab5	1.97	0.0547
Lab3 vs Lab4	4.20	0.0001*
Lab3 vs Lab5	3.51	0.0010*
Lab4 vs Lab5	-0.69	0.4945

We conclude that Labs 1 and 4 differ, Labs 1 and 5 differ, Labs 3 and 4 differ, and Labs 3 and 5 differ. These results are consistent with the side-by-side boxplots in Example 10.3.

10.7.4 Bonferroni Procedure After a Kruskal-Wallis Test

If the null hypothesis is rejected in a Kruskal-Wallis test and we want to to decide *which* of the means differ from each other, we can perform multiple rank sum tests, each time using the Bonferroni-corrected level of significance.

Bonferroni Multiple Comparison Procedure After a Kruskal-Wallis Test: After a Kruskal-Wallis test has rejected H_0, to decide which pairs of means differ while controlling the familywise Type I error rate at α_f, for each pair of means μ_i and μ_j, test the hypotheses

$$\begin{aligned} H_0 : \mu_i - \mu_j &= 0 \\ H_a : \mu_i - \mu_j &\neq 0 \end{aligned}$$

using a *rank-sum test* with decision rule

Reject H_0 if p-value $< \alpha_p$
Fail to reject H_0 if p-value $\geq \alpha_p$

where $\alpha_p = \alpha_f/(k(k-1)/2)$.

10.7.5 Other Multiple Comparison Procedures

The Bonferroni multiple comparison procedure is conservative, especially when k is large, in the sense that it sometimes doesn't detect differences between population means when in reality those differences exist. In fact, it's possible that the ANOVA F test or Kruskal-Wallis test leads rejection of the null hypothesis, yet the pairwise t tests (or rank sum tests) don't detect *any* differences among the means. If this occurs after an ANOVA F test, it's advisable to turn to one of the other multiple comparison procedures, such as the ***Duncan*** or ***Tukey*** procedures. Details about these can be found in [7].

10.8 Problems

10.1 Explain what's wrong with each of the following statements.

a) The ANOVA F test is a test of the null hypothesis that the k sample means are equal.

b) The mean squares in an ANOVA are additive, that is, MSTo = MSTr + MSE.

c) Within-groups variation in the response variable reflects differences among the k group means.

d) When we reject the null hypothesis in an ANOVA F test we conclude that none of the k population (or treatment) means equal each other.

10.2 A study compared four groups with 11 observations per group. The resulting treatment sum of squares was SSTr = 17.0 and the error sum of squares was SSE = 80.0.

a) Find the values of the mean square for treatments MSTr and the mean squared error MSE.

b) Give the value of the ANOVA F statistic.

c) Give the numerator and denominator degrees of freedom for the distribution of the F statistic under the null hypothesis.

d) Sketch the F distribution of part c, mark the observed value of F on the horizontal axis, and shade the area corresponding to the p-value (which turns out to be 0.0505), as in Fig. 10.7.

10.3 Driving heavy equipment on farmland can compress the soil and hinder future crops by inhibiting root growth. A study was carried out to examine the effects of soil compression on soil penetrability, which determines how much resistance a plant's roots will meet when growing through the soil [13]. Higher penetrability values mean the roots meet less resistance. Soil penetrability was measured on 20 parcels of land under each of three soil compression levels, loose, intermediate and compressed, giving 60 measurements total. A one-factor ANOVA was carried out, and part of the ANOVA table is shown below.

Source	DF	SS	MS	F	P-value
Compression	?	?	9.13	?	0.000
Error	?	?	0.07		
Total	?	?			

Fill in the values that are missing from the ANOVA table.

10.4 In an agricultural experiment designed to compare the effects of four fertilizer treatments on the yield of corn, 40 plots of land were randomly allocated to the treatment groups, with 10 plots being treated by each fertilizer. At the end of the growing season, the corn yield was recorded for each plot and a one-factor ANOVA carried out. Part of the ANOVA table is shown below.

Source	DF	SS	MS	F	P-value
Treatment	?	4751.1	?	?	0.000
Error	?	1173.0	?		
Total	?	?			

Fill in the values that are missing from the ANOVA table.

10.5 Observational studies have found that soils with higher levels of acidity have less healthy microbial communities. In an experiment to decide if it's a cause and effect relationship, nine soil samples were selected from forest sites whose soil pH levels were close to neutral [2]. Each was then treated with one of three levels of acidification (mild, strong, and extreme) using acid water with a pH of 0.5, with three soil specimens assigned randomly to each treatment group. After an incubation period of 80 days, the microbial biomass (microbial carbon as a percent of total organic carbon in the soil) was measured. The table below shows the data.

Microbial Biomass

Mild Acidification	Strong Acidification	Extreme Acidification
1.35	0.45	0.91
1.29	0.36	0.57
1.24	0.48	0.51

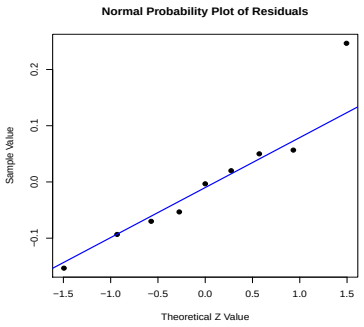
a) Write out the group means version of the ANOVA model for the data, including any assumptions about the random error term ϵ in the model.

b) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) population means μ_1, μ_2 and μ_3.

c) Carry out a one-factor ANOVA and write out the resulting ANOVA table.

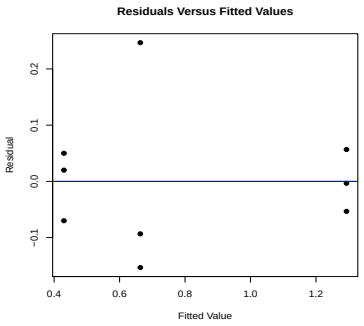
d) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, does acidification have any effect on microbial biomass?

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



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

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



Based on the plot, does the assumption, required by the F test, that the standard deviation σ of the error distribution is the same for the three groups appear to be met?

g) Assuming that σ is the same for the three groups, what's the estimated value of σ?

h) You should have found that there are significant differences among the three mean biomass percents. Carry out a Bonferroni multiple comparison procedure, controlling the overall familywise Type I error rate at 0.05, to determine *which* means differ from each other.

10.6 Refer to the experiment to decide if soil acidification affects microbial community health described in Problem 10.5.

Another variable recorded for each plot at the end of the incubation period was the metabolic quotient, qCO_2, a measure of microbial respiration per unit of microbial biomass. A small metabolic quotient indicates a more energy efficient, and therefore healthier, microbial community. The table below shows the qCO_2 data (μg CO_2-C/mg biomass-C/h).

Metabolic Quotient

Mild Acidification	Strong Acidification	Extreme Acidification
2.20	4.8	2.0
0.78	5.2	3.8
0.86	2.4	5.9

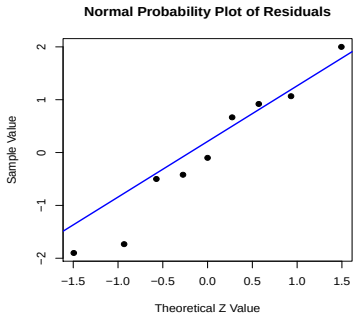
a) Write out the group means version of the ANOVA model for the data, including any assumptions about the random error term ϵ in the model.

b) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) population means μ_1, μ_2 and μ_3.

c) Carry out a one-factor ANOVA and write out the resulting ANOVA table.

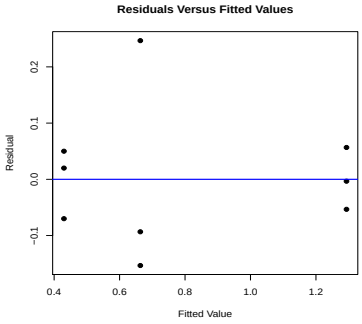
d) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, does acidification have any effect on metabolic quotient?

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



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

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



Based on the plot, does the assumption, required by the F test, that the standard deviation σ of the error distribution is the same for the three groups appear to be met?

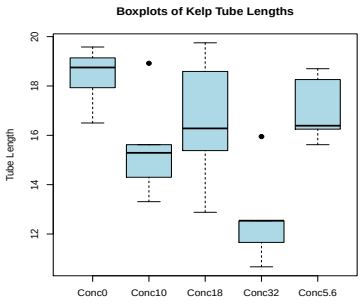
g) Assuming that σ is the same for the three groups, what's the estimated value of σ?

h) You should have found no significant differences among the three mean metabolic quotients. Would it make sense to carry out Bonferroni multiple comparison tests to determine *which* of the three means differ from each other?

10.7 Example 5.1 in Chapter 5 described an experiment to asses the toxic effects of heavy metal pollutants on aquatic life. Giant kelp (*Mactocystis pyrifera*) were exposed to copper (Cu) at five concentrations, and the lengths of their embryonic gametophyte germination tubes were measured. Smaller tube lengths indicate more severe toxic responses and are an indicator of potential toxicity to other aquatic organisms. Five replicate observations were made at each exposure concentration. The table below shows the tube lengths (mm).

Tube Lengths for Five Cu Exposure Concentrations

0.0 μg/L	5.6 μg/L	10.0 μg/L	18.0 μg/L	32.0 μg/L
19.58	18.26	13.31	18.59	12.54
18.75	16.25	18.92	12.88	10.67
19.14	16.39	15.62	16.28	15.95
16.50	18.70	14.30	15.38	12.54
17.93	15.62	15.29	19.75	11.66



Side-by-side boxplots of the tube lengths for the five treatment groups are below.
We want to decide if there are any significant differences among the mean tube lengths for the five Cu concentrations.

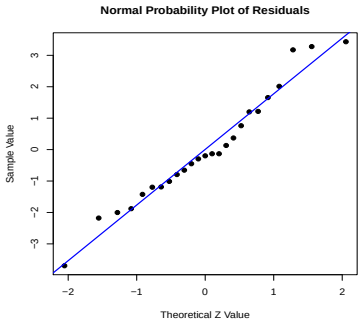
a) Write out the treatment effects version of the ANOVA model for the data, including any assumptions about the random error term ϵ in the model.

b) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) treatment effects $\alpha_1, \alpha_2, \ldots, \alpha_5$.

c) Carry out a one-factor ANOVA and write out the resulting ANOVA table.

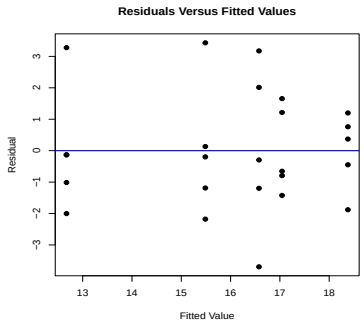
d) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, does copper have any effect on tube length?

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



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

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



Based on the plot, does the assumption, required by the F test, that the standard deviation σ of the error distribution is the same for the five groups appear to be met?

g) Assuming that σ is the same for the five groups, what's the estimated value of σ?

h) You should have found that there are significant differences among the mean tube lengths for the five Cu concentrations. Carry out a Bonferroni multiple comparison procedure, controlling the overall familywise Type I error rate at 0.05, to determine *which* means differ from each other.

10.8 An experiment was carried out to study the effects of two forest insecticides on aquatic insects [14], [10]. Fifteen 5-liter aquariums, each containing 10 aquatic insects in natural water, were randomized to three treatment groups, five aquariums to each group. The first, serving as a control group, received no insecticide. The second received insecticide A and the third insecticide B. After a 28-day period the number of deceased insects was recorded for each aquarium. The data are below.

Number of Deceased Insects

Control	Insecticide A	Insecticide B
2	3	3
2	2	1
1	2	5
3	1	5
0	0	9

We want to decide if there are any differences in the effects of the three treatments on the number of insects that die.

a) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) population mean numbers of insects that die, μ_1, μ_2, and μ_3, when ten of them are exposed to each of the three experimental conditions.

b) Carry out a one-factor ANOVA and write out the resulting ANOVA table.

c) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, are there any differences in the effects of the three experimental conditions on insect mortality?

10.9 In a laboratory quality assurance study, five standard water specimens containing a known concentration 20 μg/L of cadmium (Cd) were sent to each of five labs for analysis [16]. Each lab was also sent five specimens containing 100 μg/L of Cd.

The goal of the study was to find out if there are any differences among the five labs' results for either concentration. The tables below show the data.

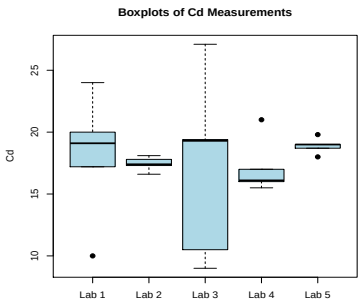
Cd (Concentration = 20)

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
10.0	17.8	27.1	21.0	18.0
20.0	17.3	19.4	16.0	19.0
17.2	16.6	9.0	16.1	19.0
24.0	17.4	10.5	17.0	18.7
19.1	18.1	19.3	15.5	19.8

Cd (Concentration = 100)

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
92.0	90.5	107.4	96.0	91.0
100.0	87.6	108.1	90.7	101.0
97.8	85.6	83.8	89.4	102.0
100.0	89.9	81.9	91.0	92.7
109.0	90.1	94.2	85.9	99.9

In this problem, we'll analyze the 20 μg/L data. (The 100 μg/L data will be analyzed in Problem 10.10.) Side-by-side boxplots of the Cd measurements for the five labs are below.



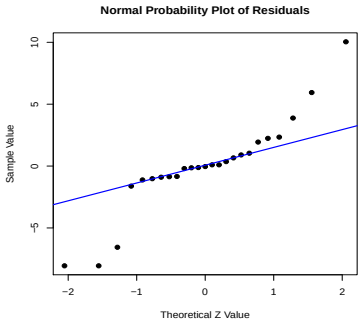
a) Write out the group means version of the ANOVA model for the data, including any assumptions about the random error term ϵ in the model.

b) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) population mean Cd measurements $\mu_1, \mu_2, \ldots, \mu_5$ at the five labs.

c) Carry out a one-factor ANOVA (on the 20 μg/L data) and write out the resulting ANOVA table.

d) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, are there any statistically significant differences among the five labs' results?

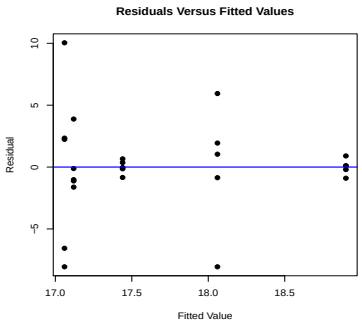
e) A normal probability plot of the residuals is below.



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

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

Based on the plot, does the assumption, required by the F test, that the standard deviation σ of the error distribution is the same for the five labs appear to be met?

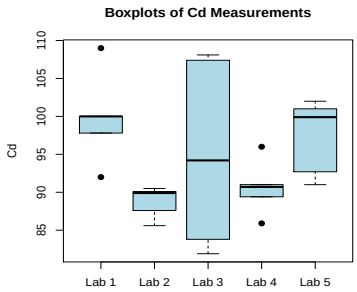


g) Assuming that σ is the same for the five labs, what's the estimated value of σ?

h) You should have found that there are no significant differences among the five labs' results. Would it make sense to carry out Bonferroni multiple comparison tests to determine *which* of the five means differ from each other?

10.10 Refer to the laboratory quality assurance study in which standard water specimens containing known concentrations 20 and 100 μg/L of cadmium (Cd) were sent to each of five labs for analysis, as described in Problem 10.9.

In this problem, we'll analyze the 100 μg/L data. The goal of the study was to find out if there are any differences among the five labs' results. Side-by-side boxplots of the Cd measurements for the five labs are below.



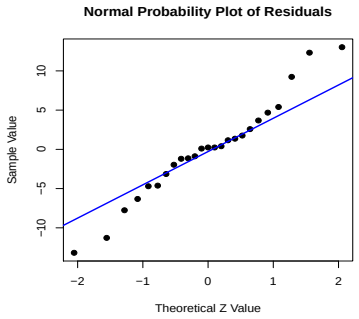
a) Write out the group means version of the ANOVA model for the data, including any assumptions about the random error term ϵ in the model.

b) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) population mean Cd measurements $\mu_1, \mu_2, \ldots, \mu_5$ at the five labs.

c) Carry out a one-factor ANOVA (on the 100 μg/L data) and write out the resulting ANOVA table.

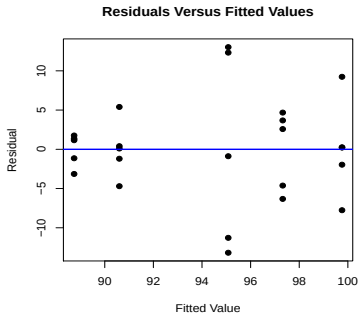
d) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, are there any statistically significant differences among the five labs' results?

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



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

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



Based on the plot, does the assumption, required by the F test, that the standard deviation σ of the error distribution is the same for the five labs appear to be met?

g) Assuming that σ is the same for the five labs, what's the estimated value of σ?

h) You should have found that there are no significant differences among the five labs' results. Would it make sense to carry out Bonferroni multiple comparison tests to determine *which* of the five means differ from each other?

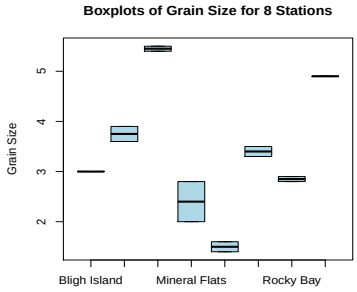
10.11 A study was conducted between 1977 and 1980 to establish baseline levels of various hydrocarbons prior to the start of oil tanker movement through Prince William Sound on the southern coast of Alaska [9]. The hydrocarbons were measured in sediment specimens and mussels at eight stations along the sound. In addition, grain size and other sediment qualities were measured to facilitate comparison to future work.

At each station, two line transects (30 m) were selected parallel to the water line. Along each transect, sediment cores were taken at 10 randomly selected points and then composited (combined). The table below shows the sediment grain size (on the $\log_2$(mm) scale) for each transect in June, 1978.

Grain Size in Prince William Sound

Sampling Station	Grain Size
Bligh Island	3.0
Bligh Island	3.0
Constantine Harbor	3.6
Constantine Harbor	3.9
Dayville Flats	5.4
Dayville Flats	5.5
Mineral Flats	2.0
Mineral Flats	2.8
Naked Island	1.6
Naked Island	1.4
Olsen Bay	3.5
Olsen Bay	3.3
Rocky Bay	2.9
Rocky Bay	2.8
Siwash Bay	4.9
Siwash Bay	4.9

Side-by-side boxplots of the grain sizes for the eight stations are below.



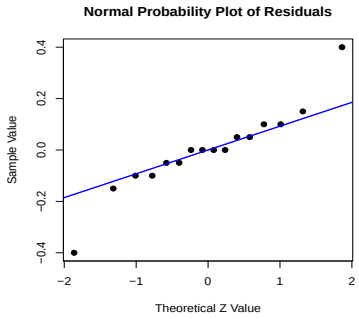
a) Write out the group means version of the ANOVA model for the data, including any assumptions about the random error term ϵ in the model.

b) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) population mean grain sizes $\mu_1, \mu_2, \ldots, \mu_8$ at the eight stations.

c) Carry out a one-factor ANOVA and write out the resulting ANOVA table.

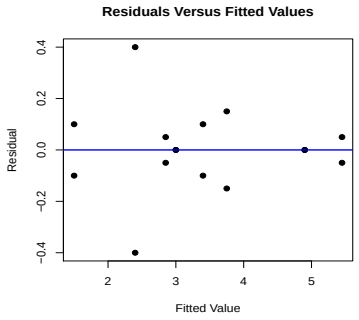
d) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, are there any statistically significant differences among the eight stations' grain sizes?

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



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

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



Based on the plot, does the assumption, required by the F test, that the standard deviation σ of the error distribution is the same for the eight stations appear to be met?

g) Assuming that σ is the same for the eight stations, what's the estimated value of σ?

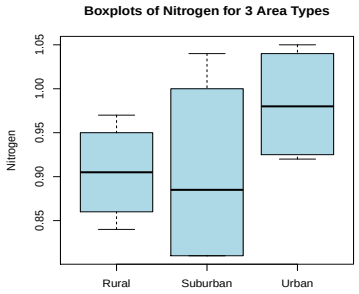
10.12 The decomposition rate of tree leaf litter is determined in part by the litter chemistry, especially nitrogen and lignin concentrations. Because forest stands near urban areas can be highly polluted, plant foliage in these areas may be exposed to O_3, SO_x, and NO_x gases, among other pollutants, and therefore leaf litter decay rates in urban areas may differ from those in suburban and rural areas.

In a study of oak leaf decay rates in urban, suburban, and rural areas around New York City, nitrogen (N, percent) and the carbon-to-nitrogen ratio (C:N) were measured in leaf litter at each of 12 oak plots, four in each area type [15]. The data are below.

Nitrogen Percents and Carbon-Nitrogen Ratios

Plot	Area Type	N	C:N
1	Urban	0.93	55.1
2	Urban	1.03	51.7
3	Urban	0.92	55.0
4	Urban	1.05	51.3
5	Suburban	0.81	65.3
6	Suburban	0.96	55.7
7	Suburban	0.81	63.7
8	Suburban	1.04	50.8
9	Rural	0.84	61.9
10	Rural	0.88	58.5
11	Rural	0.97	54.6
12	Rural	0.93	55.5

In this problem, we'll analyze the N concentrations. (The C:N data will be analyzed in Problem 10.13.) We want to know if there are statistically significant differences among the mean N concentrations for the three area types. Side-by-side boxplots of the data are below.



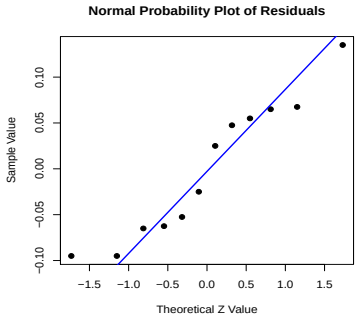
a) Write out the group means version of the ANOVA model for the data, including any assumptions about the random error term ϵ in the model.

b) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) population mean N concentrations μ_1, μ_2, and μ_3 for the three area types.

c) Carry out a one-factor ANOVA and write out the resulting ANOVA table.

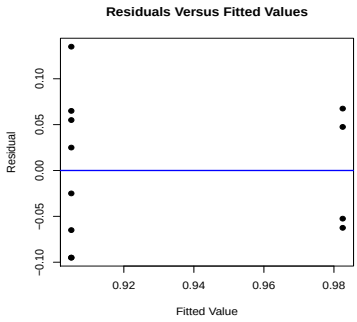
d) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, are there any statistically significant differences among the mean N concentrations for the three area types?

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



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

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



Based on the plot, does the assumption, required by the F test, that the standard deviation σ of the error distribution is the same for the three area types appear to be met?

g) Assuming that σ is the same for the eight stations, what's the estimated value of σ?

h) You should have found no significant differences among the mean N concentrations for the three area types. Would it make sense to carry out Bonferroni multiple comparison tests to determine *which* of the three means differ from each other?

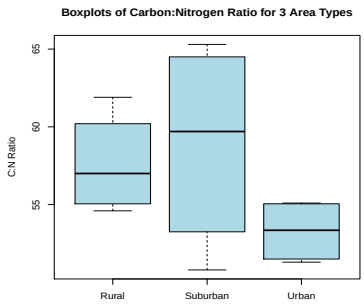
10.13 Refer to the study of oak leaf decay rates in urban, suburban, and rural areas around New York City, in which nitrogen (N) and the carbon-to-nitrogen ratio (C:N) were measured at four plots in each area type, as described in Problem 10.12.

In this problem, we'll analyze the C:N ratios. We want to know if there are statistically significant differences among the mean C:N ratios for the three area types. Side-by-side boxplots of the data are below.

a) Write out the group means version of the ANOVA model for the data, including any assumptions about the random error term ϵ in the model.

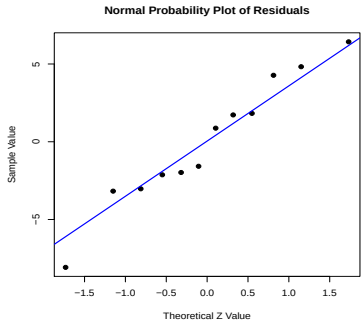
b) State the null and alternative hypotheses for the ANOVA F test in terms of the true (unknown) population mean C:N ratios μ_1, μ_2, and μ_3 for the three area types.

c) Carry out a one-factor ANOVA and write out the resulting ANOVA table.



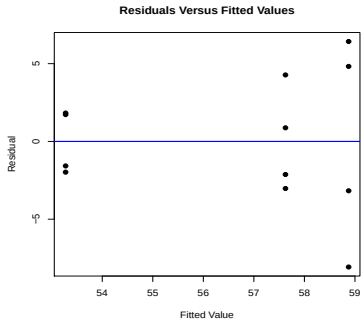
d) State the conclusion of the ANOVA F test using a level of significance $\alpha = 0.05$. Based on the F test, are there any statistically significant differences among the mean C:N ratios for the three area types?

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



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

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



Based on the plot, does the assumption, required by the F test, that the standard deviation σ of the error distribution is the same for the three area types appear to be met?

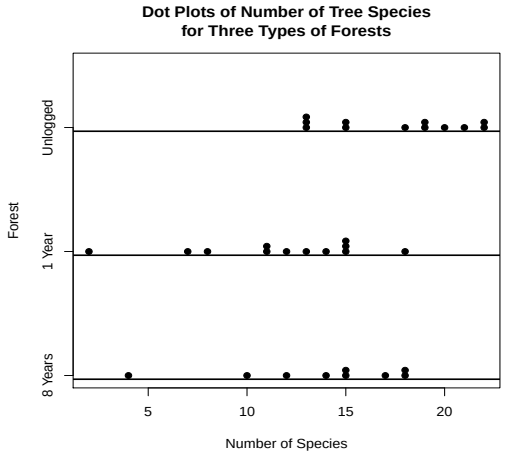
g) Assuming that σ is the same for the eight stations, what's the estimated value of σ?

h) You should have found no significant differences among the mean C:N ratios for the three area types. Would it make sense to carry out Bonferroni multiple comparison tests to determine *which* of the three means differ from each other?

10.14 In a study of the effects of logging in Borneo, data were collected on the number of tree species in 12 unlogged forest plots, 12 similar plots logged just one year earlier, and 9 similar plots logged eight years earlier [4]. The tree-species counts are below.

Number of Species		
Unlogged	Logged 1 Year Ago	Logged 8 Years Ago
22	11	17
18	11	4
22	14	18
20	7	14
15	18	18
21	15	15
13	15	15
13	12	10
19	13	12
13	2	
19	15	
15	8	

We want to decide if there are differences among the true population mean numbers of species per plot for the three types of forests. Dot plots of the data are below.



The dot plots show that two of the three samples have mild outliers on the left side, so rather than carrying out an ANOVA F test, whose results can be influenced by outliers, we'll carry out the nonparametric Kruskal-Wallis test.

a) Carry out a Kruskal-Wallis test to decide if there are differences among the true population mean numbers of species per plot for the three forest types. Use a level of significance $\alpha = 0.05$.

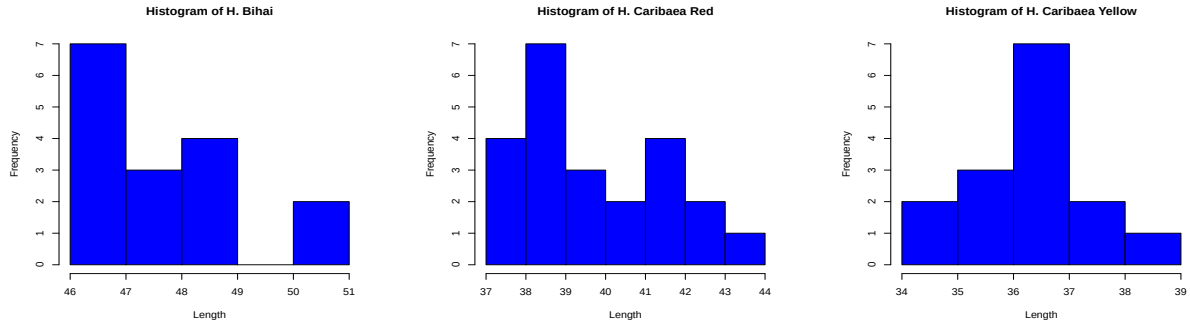
b) In part *a*, you should have found that there are significant differences among three types of forests. Carry out a Bonferroni multiple comparison procedure, controlling the overall familywise Type I error rate at 0.05, by conducting three rank sum tests, one for each of the three pairs of forest types, using the Bonferroni-corrected level of significance. State which of pairs of forest types differ significantly from each other.

10.15 Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the forms of the hummingbirds' beaks have evolved to match each other.

The table below shows data on the lengths (mm) of three varieties of these flowers (*H. bihai*, *H. caribaea red*, and *H. caribaea yellow*) on the island of Dominica in the Caribbean Sea [18].

	Length of Flower (mm)	
H. bihai	H. caribaea red	H. caribaea yellow
47.12	41.90	36.78
46.75	42.01	37.02
46.81	41.93	36.52
47.12	43.09	36.11
46.67	41.47	36.03
47.43	41.69	35.45
46.44	39.78	38.13
46.64	40.57	37.10
48.07	39.63	35.17
48.34	42.18	36.82
48.15	40.66	36.66
50.26	37.87	35.68
50.12	39.16	36.03
46.34	37.40	34.57
46.94	38.20	34.63
48.36	38.07	
	38.10	
	37.97	
	38.79	
	38.23	
	38.87	
	37.78	
	38.01	

We want to decide if there are differences among the true population mean lengths of flowers for the three species. Histograms of the data (below) hint that the lengths of the flowers follow right skewed distributions, so a one-factor ANOVA might not be appropriate.



Instead, we'll carry out a Kruskal-Wallis test.

a) Carry out the Kruskal-Wallis test using a level of significance $\alpha = 0.05$.

b) In part *a*, you should have found that there are significant differences among the three mean flower lengths. Carry out a Bonferroni multiple comparison procedure, controlling the overall familywise Type I error rate at 0.05, by conducting three rank sum tests, one for each of the three pairs of flower varieties, using the Bonferroni-corrected level of significance. State which of pairs of flower varieties differ significantly from each other.

10.16 A study was carried out to investigate the vertical distribution patterns of several species of large zooplankton in the Wilkinson Basin, Gulf of Maine [6].

To collect abundance data on shrimp and gelatinous zooplankton, researchers were lowered in a five foot diameter spherical Plexiglas submersible known as the Johnson-Sea-Link. At the desired depth, the submersible moved forward along a transect at 0.5 knots for four minutes. As it did so, the researchers identified organisms passing through bars demarcating a 2.65 m^2 area on the port side of the submersible. Video recordings were also made for later viewing to help with the zooplankton identification and quantification. This scheme was repeated at three depth categories, shallow (200 - 220 m), medium (220 - 240 m), and deep (240 - 260 m).

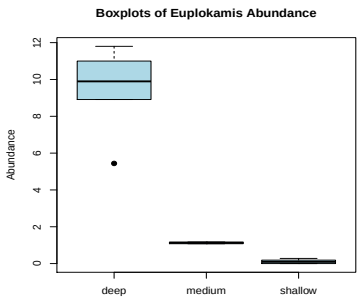
The table below shows the data (number of individuals per m^3) for the genus *Euplokamis* (gelatinous ctenophores), the species *M. norvegica* (a krill), and caridean shrimp.

	Individuals per m^3		
Depth	Euplokamis	M. norvegica	Carideans
Shallow	0.19	0.040	0.000
Shallow	0.00	0.000	0.000
Shallow	0.00	0.000	0.000
Shallow	0.09	0.000	0.000
Shallow	0.28	0.115	0.000
Medium	1.17	0.061	0.018
Medium	1.08	0.061	0.015
Deep	11.80	0.012	0.000
Deep	11.00	0.000	0.024
Deep	8.91	0.015	0.000
Deep	5.44	0.008	0.120
Deep	9.90	0.031	0.012

After analyzing the *Euplokamis* data, the authors of the study concluded that

> "*Euplokamis* remained primarily between 240 and 260 m depth."

We want to decide if the data provide statistically significant evidence to support this claim. Side-by-side boxplots of the *Euplokamis* abundance data for the three depth categories are below.



a) Which test, the one-factor ANOVA F test or the Kruskal-Wallis test, would be most appropriate for deciding if there are any differences among the mean *Euplokamis* abundances for the three depth categories?

b) Carry out the test you chose in part *a*. Use a level of significance $\alpha = 0.05$.

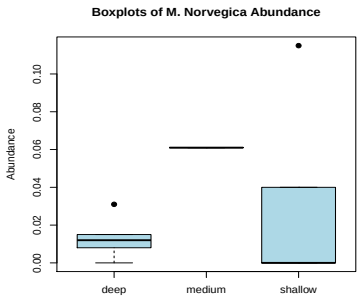
c) If you found significant differences in part *b*, carry out an appropriate multiple comparison procedure to decide which depth categories' means differ from each other.

10.17 Refer to the study to investigate the vertical distribution patterns of zooplankton in the Wilkinson Basin, Gulf of Maine, described in Problem 10.16.

After analyzing the *M. norvegica* data, the authors of the study concluded that

> "*M. norvegica* had a fairly broad daytime depth distribution, being equally abundant between 200 and 260 m depth."

We want to decide if the data provide statistically significant evidence against this claim. Side-by-side boxplots of the *M. norvegica* abundance data for the three depth categories are below.



a) Which test, the one-factor ANOVA F test or the Kruskal-Wallis test, would be most appropriate for deciding if there are any differences among the mean *M. norvegica* abundances for the three depth categories?

b) Carry out the test you chose in part *a*. Use a level of significance $\alpha = 0.05$.

c) If you found significant differences in part *b*, carry out an appropriate multiple comparison procedure to decide which depth categories' means differ from each other.

10.18 Refer to the study to investigate the vertical distribution patterns of zooplankton in the Wilkinson Basin, Gulf of Maine, described in Problem 10.16.

After analyzing the carideans data, the authors of the study concluded that

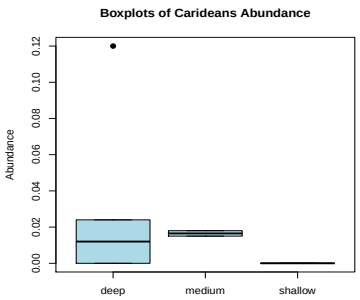
> "The carideans were primarily distributed between 220 and 240 m depth."

We want to decide if the data provide statistically significant evidence to support this claim. Side-by-side boxplots of the carideans abundance data for the three depth categories are below.

a) Which test, the one-factor ANOVA F test or the Kruskal-Wallis test, would be most appropriate for deciding if there are any differences among the mean carideans abundances for the three depth categories?

b) Carry out the test you chose in part *a*. Use a level of significance $\alpha = 0.05$.

c) If you found significant differences in part *b*, carry out an appropriate multiple comparison procedure to decide which depth categories' means differ from each other.



10.19 Dams are built on streams for a variety of reasons, such as flow control, aesthetic appeal in parks, etc., but they can also impact the stream's water quality. Particularly vulnerable is the stream's dissolved oxygen concentration. Pools created by dams often have dissolved oxygen concentrations that fall below desired levels because the physical reaeration capabilities in pools may be much lower than those of free-flowing water.

A study was carried out to determine the effects of dams on dissolved oxygen along a 300 mile stretch of the Illinois Waterway between its confluence with the Mississippi River at Grafton, Illinois to just south of Chicago [3]. For each of seven dams, dissolved oxygen concentrations (mg/L) were measured at several places above and below the dam during summer and fall. Water temperatures (°C) were also measured because temperature had been shown in experiments to be related to aeration, with warmer water reaerating faster than cool water. The table below shows the data and collection dates.

Temperature and Dissolved Oxygen

Dam	Date	Temp Above	Temp Below	DO Above	DO Below
Lockport	9/13/78	27.20	26.00	1.80	2.22
Lockport	10/13/78	20.31	19.80	2.47	2.46
Lockport	8/16/79	26.00	25.81	0.81	0.98
Brandon Road	9/12/78	27.32	27.50	3.18	6.57
Brandon Road	10/12/78	19.55	19.00	3.40	7.15
Brandon Road	8/15/79	24.50	23.80	0.70	6.78
Brandon Road	8/15/79	25.00	23.80	0.56	6.78
Brandon Road	8/15/79	25.00	23.80	0.56	6.88
Brandon Road	8/29/79	26.12	24.80	1.88	6.53
Brandon Road	8/29/79	25.40	24.80	1.52	6.28
Brandon Road	9/11/79	26.00	25.90	1.88	7.42
Brandon Road	9/11/79	26.00	25.90	1.51	7.30
Brandon Road	9/11/79	26.00	25.90	1.55	6.40
Dresden Island	8/25/78	26.23	27.30	5.66	7.24
Dresden Island	8/25/78	26.12	27.30	5.62	7.33
Dresden Island	9/14/78	26.20	27.00	4.50	6.85
Dresden Island	9/14/78	26.20	27.00	4.53	6.58
Dresden Island	9/14/78	26.40	27.20	4.56	6.78
Dresden Island	8/08/79	30.74	30.05	5.49	6.77
Dresden Island	8/08/79	30.37	30.83	5.00	6.60
Dresden Island	8/08/79	30.39	30.17	5.20	6.70
Dresden Island	8/14/79	25.18	24.80	5.67	7.78
Dresden Island	8/14/79	25.04	24.80	5.65	7.60
Dresden Island	8/14/79	24.85	24.80	5.55	7.90
Dresden Island	9/05/79	28.22	27.97	5.76	7.10
Dresden Island	9/05/79	27.85	29.00	5.68	7.10
Marseilles	8/24/78	28.00	27.47	6.60	7.10
Marseilles	9/19/78	25.02	25.00	4.68	6.53
Marseilles	9/19/78	24.65	25.00	4.61	6.43
Marseilles	8/06/79	28.87	29.50	4.70	6.13
Marseilles	8/06/79	28.90	29.50	4.80	6.13
Marseilles	8/06/79	29.08	29.00	5.00	6.23
Marseilles	9/06/79	27.50	27.90	5.51	7.05
Marseilles	9/06/79	27.50	27.90	5.47	7.20
Marseilles	9/06/79	27.50	27.80	5.47	6.80
Marseilles	9/12/79	25.50	25.50	6.33	7.70
Marseilles	9/12/79	25.30	25.40	6.36	7.47
Marseilles	9/12/79	25.37	25.33	6.54	7.57
Starved Rock	8/23/78	25.80	27.20	7.01	7.90
Starved Rock	8/23/78	26.52	27.20	7.21	7.80
Starved Rock	8/23/78	26.44	27.20	7.96	8.00
Starved Rock	8/23/78	25.94	27.20	7.14	7.90
Starved Rock	9/20/78	24.80	25.00	5.90	6.93
Starved Rock	9/20/78	25.52	25.00	5.53	6.50
Starved Rock	8/03/79	28.93	28.20	6.94	6.73
Starved Rock	8/03/79	28.44	27.00	6.47	6.92
Starved Rock	8/03/79	28.17	27.20	6.19	6.93
Starved Rock	9/07/79	26.00	25.50	8.04	7.93
Starved Rock	9/07/79	25.02	25.10	7.18	8.07
Starved Rock	9/07/79	25.11	25.00	7.44	8.10
Starved Rock	9/14/79	23.40	23.10	7.29	8.70
Starved Rock	9/14/79	22.68	22.50	8.13	8.77
Starved Rock	9/14/79	22.52	22.50	8.05	8.67
Peoria	8/12/78	25.21	26.23	6.24	6.55
Peoria	8/12/78	25.20	26.20	6.80	6.55
Peoria	8/12/78	25.19	26.05	6.47	6.62
Peoria	8/12/78	25.24	26.20	6.60	6.74
Peoria	8/12/78	25.66	26.20	6.47	6.71
Peoria	10/02/78	20.09	18.90	6.95	7.42
Peoria	10/02/78	19.34	18.90	7.10	7.42
Peoria	10/02/78	19.17	18.90	7.06	7.40
Peoria	10/02/78	19.21	18.90	7.27	7.39
Peoria	10/02/78	19.16	18.90	7.10	7.43
Peoria	7/27/79	29.71	28.12	6.01	6.30
Peoria	7/27/79	29.47	28.32	6.17	6.17
Peoria	7/27/79	29.08	28.24	6.31	6.43
Peoria	8/01/79	26.59	26.49	6.43	6.45
Peoria	8/01/79	26.69	26.32	6.02	6.64
Peoria	8/01/79	26.85	26.50	6.20	6.70
Peoria	10/10/79	14.47	14.20	9.06	9.61
Peoria	10/10/79	14.50	14.00	8.95	9.74
Peoria	10/10/79	14.46	14.13	9.04	9.75
LaGrange	8/22/79	25.51	26.22	6.28	6.65
LaGrange	8/22/79	25.47	26.87	6.47	6.70
LaGrange	8/22/79	25.37	26.71	6.54	6.75
LaGrange	8/22/79	25.30	26.57	6.47	6.62
LaGrange	8/22/79	25.29	26.51	6.42	6.45
LaGrange	9/21/78	24.07	24.00	2.64	3.16
LaGrange	9/21/78	24.03	24.00	2.52	3.08
LaGrange	9/21/78	23.91	24.00	2.62	3.13
LaGrange	8/07/79	30.07	30.00	5.06	5.22
LaGrange	8/07/79	28.89	30.00	4.86	5.71
LaGrange	8/07/79	29.59	29.50	4.71	5.10
LaGrange	9/18/79	23.56	23.82	6.04	7.45
LaGrange	9/18/79	23.45	23.21	6.10	7.51
LaGrange	9/18/79	23.39	23.02	6.01	7.40
LaGrange	9/26/79	22.73	22.34	6.18	8.08
LaGrange	9/26/79	22.25	21.82	6.58	7.62
LaGrange	9/26/79	21.98	21.48	6.21	7.47
LaGrange	9/26/79	21.98	23.00	6.21	6.10

Although it's to be expected that a dam will cause abrupt changes in a stream's dissolved oxygen concentration, the size of that change may differ depending on the dam's design features (for example the number

and sizes of flow release gates, the waterfall height, pool depth, etc.).

We want to decide if there are differences among the sizes of the seven dams' effects on dissolved oxygen.

a) Compute the differences in dissolved oxygen (above the dam minus below) and make side by side boxplots of the differences for the seven dams.

b) Carry out a one-factor ANOVA to decide if the dissolved oxygen difference (above minus below) depends on the dam. Use a level of significance $\alpha = 0.05$.

c) Check the normality assumption for one-factor ANOVA by making a normal probability plot or a histogram of the residuals.

d) Check the common standard deviation assumption for one-factor ANOVA by making a plot of the residuals versus the fitted values.

Bibliography

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

[2] Evgenia V. Blagodatskaya and Traute-Heidi Anderson. Adaptive responses of soil microbial communities under experimental acid stress in controlled laboratory studies. *Applied Soil Ecology*, 11:207 – 216, 1999.

[3] Thomas A. Butts and Ralph L. Evans. Aeration characteristics of flow release controls on Illinois waterway dams. Technical report, Illinois Institute of Natural Resources, State Water Survey Division at Peoria, Illinois, June 1980. Prepared for the Illinois Division of Water Resources at the request of the U.S. Army Corps of Engineers.

[4] C. H. Cannon, D. R. Peart, and M. Leighton. Tree species diversity in commercially logged Bornean rainforest. *Science*, 281:1366 – 1367, 1998.

[5] W. L. Conover. *Practical Nonparametric Statistics*. John Wiley and Sons, New York, NY, second edition, 1980.

[6] T. M. Frank and E. A. Widder. The correlation of downwelling irradiance and staggered vertical migration patterns of zooplankton in Wilkinson Basin, Gulf of Maine. *Journal of Plankton Research*, 19(12):1975 – 1991, 1997.

[7] D. R. Helsel and R. M. Hirsch. *Statistical Methods in Water Resources*. Techniques of Water-Resources Investigations of the United States Geological Survey, Book 4, Hydrologic Analysis and Interpretation, Chapter A3. U.S. Geological Survey, Sept 2002.

[8] William A. Hopkins et al. Trophic and maternal transfer of selenium in brown house snakes (*Lamprophis fuliginosus*). *Ecotoxicology and Environmental Safety*, 58:285 – 293, 2004.

[9] 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).

[10] D. P. Kreutzweiser. Nontarget effects of neem-based insecticides on aquatic invertebrates. *Ecotoxicology and Environmental Safety*, 36:109 – 117, 1997.

[11] E. L. Lehmann. *Nonparametrics, Statistical Methods Based on Ranks*. Holden-Day, Oakland, CA, 1975.

[12] Simboura N. and A. Zenetos. Benthic indicators to use in Ecological Quality classification of Mediterranean soft bottom marine ecosystems, including a new Biotic Index. *Mediterranean Marine Science*, 3/2:77 – 111, 2002.

[13] S. Gupta Parmeshwar. Reaction of plants to the density of soil. *Journal of Ecology*, 21:452 – 474, 1933.

[14] Douglas G. Pitt and David P. Kreutzweiser. Applications of computer-intensive statistical methods to environmental research. *Ecotoxicology and Environmental Safety*, 39:78 – 97, 1998.

[15] Richard V. Pouyat and Margaret M. Carreiro. Controls on mass loss and nitrogen dynamics of oak leaf litter along an urban-rural land-use gradient. *Oecologia*, 135:288 – 298, 2003.

[16] C.H. Proctor. A simple definition of detection limit. *Journal of Agricultural, Biological, and Environmental Statistics*, 13(1):99–120, March 2008.

[17] Lucie Sliva and D. Dudley Williams. Buffer zone versus whole catchment approaches to studying land use impact on river water quality. *Water Research*, 35(14):3462 – 3472, 2001.

[18] E. J. Temeles and W. J. Kress. Adaptation in a plant-hummingbird association. *Science*, 300:630 – 633, 2003.