Chapter 12
Comparing Two or More Means

12.1 Introduction

In Chapter 8 we considered methods for making inferences about the relationship between two population distributions based on the relationship between the means of these distributions. In many situations interest centers on the relationship among more than two population distributions. Therefore, in this chapter we consider methods of inference for comparing two or more population distributions based on the relationships among the corresponding population means.

We will restrict our attention to situations where the population distributions (density curves) of $k \geq 2$ continuous variables, $Y_1, Y_2, \ldots$, and Y_k, are identical except for their locations on the number line. This generalizes the **shift assumption** of the two population problem to the $k \geq 2$ population problem. Under this shift assumption inferences for comparing the k population distributions reduce to inferences for comparing the k population means. As in the two population case, when the shift assumption is not valid we must be careful about how we interpret an inference about the relationships among the population means.

We will restrict our attention to methods which are appropriate when the data comprise k independent random samples: a random sample of size n_1 from a population with population mean μ_1 (the Y_1 sample); a random sample of size n_2 from a population with population mean μ_2 (the Y_2 sample); $\ldots$, and a random sample of size n_k from a population with population mean μ_k (the Y_k sample). The assumption that these random samples are independent basically means that the method used to select the random sample from a particular population is not influenced by the method used to select the random sample from any other population.

We will use the following small example to clarify the definitions and computations introduced in this chapter. You should use a suitable calculator or computer program to perform these computations.

Example. Potato leafhopper survival. D. L. Dahlman (M.S. thesis, Iowa State University, 1963) studied the survival and behavioral responses of the potato leafhopper *Empoasca Fabae* (Harris) on synthetic media. The data given in Table 1 are survival times (in days) defined as the number of days until 50% of the insects in a cage were dead. This study was conducted using a completely randomized experiment design with two cages (units) assigned to each of four treatment groups ($n_1 = n_2 = n_3 = n_4 = 2$). That is, the 8 cages were randomly assigned to the 4 treatments so that there were two cages in each treatment group. The treatments consisted of four modifications of the basic 2%

agar synthetic feeding medium. The treatments were a control (2% agar), 2% agar plus fructose, 2% agar plus glucose, and 2% agar plus sucrose, respectively.

Table 1. Potato Leafhopper Data.

treatment	survival time
control	2.3
control	1.7
fructose	2.1
fructose	2.3
glucose	3.0
glucose	2.8
sucrose	3.6
sucrose	4.0

We can define the four population means by imagining what would have happened if all of the eight cages were assigned to a particular treatment group. For example, we can define the control population mean $\mu_1 = \mu_C$ as the mean survival time we would have observed if all 8 cages had been assigned to the control group; we can define the fructose population mean $\mu_2 = \mu_F$ as the mean survival time we would have observed if all 8 cages had been assigned to the fructose group; and so on. The notation we will use in the sequel is summarized in Table 2.

Table 2. Potato Leafhopper Population Means.

treatment:	control	fructose	glucose	sucrose
population mean:	μ_C	μ_F	μ_G	μ_S

12.2 Comparing the means of k normal populations

In this section we consider inferences about the relationships among k normal means. First we discuss the analysis of variance (ANOVA) and the overall F–test; then we consider the sequential use of F–tests for comparing nested models; and finally we discuss simultaneous confidence interval estimates for linear combinations of means.

12.2a Assumptions, notation, and the overall F–test

In order to develop inferential methods we need to make an assumption about the form of the population distributions of the $Y's$. We will assume that the k population distributions are normal distributions with a common population variance σ^2 (common population standard deviation σ). Thus, the population distribution of Y_1 is a normal distribution with population mean μ_1 and population variance σ^2; the population distribution of Y_2 is a normal distribution with population mean μ_2 and population variance σ^2; ..., and the population distribution of Y_k is a normal distribution with population mean μ_k and population variance σ^2. As stated in the introduction, we will assume that the data comprise k independent random samples. Notice that we are assuming that the k population variances are equal which, together with the normality assumption, implies that the shift assumption is valid.

First consider the question of whether the k means $\mu_1, \ldots, \mu_k$ are all equal. We can address this question by performing a hypothesis test for the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ versus the alternative hypothesis that at least two of the k means are different ($H_1 :$ it is not true that $\mu_1 = \cdots = \mu_k$). Notice that this alternative hypothesis specifies that the k means are not all equal, it does not specify how the means differ and, in particular, it does not specify that there are k distinct means. We will motivate the method used to perform this hypothesis test about the k means as a comparison of two estimators of the common population variance σ^2.

To make the notation clear we will need to use double subscripts on the observations. As indicated in Table 3, we will let Y_{ij} denote the j^{th} observation in the i^{th} group (i^{th} sample), for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$, and we will let $\overline{Y}_i$ denote the sample mean for the i^{th} group.

Table 3. Notation for the k group (sample) problem.

group	data	sample mean	population mean
group 1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$\overline{Y}_1$	μ_1
group 2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	$\overline{Y}_2$	μ_2
.	.	.	.
.	.	.	.
.	.	.	.
group k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$	$\overline{Y}_k$	μ_k

The pooled estimator S_p^2 of the common variance σ^2 for the model with k population means $\mu_1, \ldots, \mu_k$ is the natural extension of the pooled variance estimator of the two sample case to k samples. That is, S_p^2 is the sum of the squared deviations of the observations from their respective group sample means divided by the appropriate degrees of freedom which is $n - k = (n_1 - 1) + \cdots + (n_k - 1)$, where $n = n_1 + \cdots + n_k$ is the total number of observations. The numerator of S_p^2, denoted by SS(within the k groups), is the sum of squares within the k groups (the sum of squared deviations of the observations within each group from their respective group sample mean). In symbols we have

$$S_p^2 = \frac{\text{SS(within the } k \text{ groups)}}{n-k}, \text{ with}$$

$$\begin{aligned}\text{SS(within the } k \text{ groups)} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_i\right)^2 \\ &= \sum_{j=1}^{n_1}\left(Y_{1j} - \overline{Y}_1\right)^2 + \cdots + \sum_{j=1}^{n_k}\left(Y_{kj} - \overline{Y}_k\right)^2.\end{aligned}$$

The pooled variance estimator S_p^2 is a valid (unbiased) estimator of the common variance σ^2 when the k group means $\mu_1, \ldots, \mu_k$ are distinct and also when some or all of the means are equal. The computations for finding S_p^2 described above are illustrated for the potato leafhopper data in Table 4.

Table 4. Potato leafhopper deviations from treatment means.

treatment	observation	treatment mean	deviation from mean	squared deviation from mean
control	2.3	2.0	.3	.09
control	1.7	2.0	-.3	.09
fructose	2.1	2.2	-.1	.01
fructose	2.3	2.2	.1	.01
glucose	3.0	2.9	.1	.01
glucose	2.8	2.9	-.1	.01
sucrose	3.6	3.8	-.2	.04
sucrose	4.0	3.8	.2	.04
				sum of squared deviations = .3
				$S_p^2 = .3/4 = .075$

Under the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ we can view the k random samples as constituting one random sample of size $n = n_1 + \cdots + n_k$ from a normal population with population variance σ^2. Therefore, when H_0 is true we can estimate the common variance σ^2 using the squared deviations of the observations from the overall sample mean $\overline{Y}$ ($\overline{Y}$ is the average of all n observations and in terms of the k sample means, $\overline{Y}_1, \ldots, \overline{Y}_k$, $\overline{Y} = (n_1\overline{Y}_1 + \cdots + n_k\overline{Y}_k)/n$). The variance estimator S_0^2 is the sum of the squared deviations of the observations from the overall sample mean divided by the appropriate degrees of freedom which is $n-1$. The numerator of S_0^2, denoted by SS(about the overall mean), is the sum of squares about the overall mean (the sum of the squared deviations of the observations from the overall sample mean). In symbols we have

$$S_0^2 = \frac{\text{SS(about the overall mean)}}{n-1}, \text{ with}$$

$$\text{SS(about the overall mean)} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}\right)^2,$$

and we see that S_0^2 is simply the usual one sample estimator of the variance computed ignoring the existence of the k groups.

The variance estimator S_0^2 is a valid (unbiased) estimator of the common variance σ^2 if, and only if, the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ is true. If $H_0 : \mu_1 = \cdots = \mu_k$ is not true, then S_0^2 is positively biased as an estimator of the common variance σ^2, *i.e.*, if H_0 is not true, then S_0^2 tends to systematically overestimate σ^2. The computations for finding S_0^2 described above are illustrated for the potato leafhopper data in the Table 5.

Table 5. Potato leafhopper deviations from overall mean.

treatment	observation	overall mean	deviation from mean	squared deviation from mean
control	2.3	2.725	-.425	.180625
control	1.7	2.725	-1.025	1.050625
fructose	2.1	2.725	-.625	.390625
fructose	2.3	2.725	-.425	.180625
glucose	3.0	2.725	.275	.075625
glucose	2.8	2.725	.075	.005625
sucrose	3.6	2.725	.875	.765625
sucrose	4.0	2.725	1.275	1.625625
				sum of squared deviations = 4.275
				$S_0^2 = 4.275/7 = .6107$

We have defined two estimators S_p^2 and S_0^2 of the common variance σ^2. Both of these estimators are unbiased estimators of σ^2 when $H_0 : \mu_1 = \cdots = \mu_k$ is true. The estimator S_p^2 is unbiased as an estimator of σ^2 even when H_0 is not true; but, S_0^2 is positively biased as an estimator of σ^2 when H_0 is not true. Therefore, we can view an observed value of S_0^2 which is sufficiently large relative to the observed value of S_p^2 as evidence against the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$.

Before we discuss a method for determining whether the observed value of S_0^2 is large relative to S_p^2 we consider a decomposition of the deviation of an observation from the overall mean and a corresponding decomposition of the sum of squares about the overall mean.

The deviation of an observation Y_{ij} from the overall mean $\overline{Y}$ can be expressed as the sum of the deviation of the observation from its group mean $\overline{Y}_i$ and the deviation of its group mean from the overall mean, *i.e.*,

$$Y_{ij} - \overline{Y} = \left(Y_{ij} - \overline{Y}_i\right) + \left(\overline{Y}_i - \overline{Y}\right).$$

Furthermore, it can be shown that, there is a corresponding decomposition of the sum of squares about the overall mean as the sum of the sum of squares within the k groups plus the sum of squares among the k groups, *i.e.*,

$$\text{SS(about the overall mean)} = \text{SS(within the } k \text{ groups)} + \text{SS(among the } k \text{ groups)},$$

where

$$\text{SS(among the } k \text{ groups)} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{k} n_i \left(\overline{Y}_i - \overline{Y}\right)^2.$$

This decomposition is often summarized in a tabular form known as an analysis of variance table or ANOVA table as shown in Table 6.

Table 6. A basic ANOVA table.

source of variation	degrees of freedom	sum of squares
among groups	$k-1$	SS(among the k groups)
within groups	$n-k$	SS(within the k groups)
total	$n-1$	SS(about the overall mean)

Notice that the ANOVA table also indicates the corresponding decomposition of the total degrees of freedom, $n-1$, into the sum of the degrees of freedom among the k groups, $k-1$, and the degrees of freedom within the k groups, $n-k$. You can think of these degrees of freedom as indicating the "amount of information" contained in the corresponding sums of squares. If we use the degrees of freedom to normalize the sum of squares, by dividing the sum of squares by its degrees of freedom, the resulting "average" is known as a mean square, denoted by MS.

From the ANOVA decomposition of the sum of squares about the overall mean we can identify the sum of squares among the k groups, SS(among the k groups), as the term which causes S_0^2 to be positively biased as an estimator of σ^2. Therefore we can determine whether S_0^2 is large relative to S_p^2 by determining whether SS(among the k groups) is large relative to SS(within the k groups). We will base this determination on the ratio of the mean squares corresponding to these sums of squares. The relevant ratio is the F–statistic

$$\begin{aligned} F_{calc} &= \frac{\text{MS(among the } k \text{ groups)}}{\text{MS(within the } k \text{ groups)}} \\ &= \frac{\text{SS(among the } k \text{ groups)}/(k-1)}{\text{SS(within the } k \text{ groups)}/(n-k)}. \end{aligned}$$

When the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ is true this F–statistic follows the F distribution with numerator degrees of freedom $k-1$ and denominator degrees of freedom $n-k$. Sufficiently large values of F_{calc} constitute evidence against $H_0 : \mu_1 = \cdots = \mu_k$. The P–value for this hypothesis test is the probability of observing a value of the F–statistic as large or larger than the calculated value F_{calc}, *i.e.*, the P–value is

$$P\text{–value} = P(F \geq F_{calc}),$$

where F denotes a variable which follows the F distribution with $k-1$ and $n-k$ degrees of freedom. (The F distributions are skewed to the right with density curves which are positive only for positive values of the variable.)

The ANOVA for the potato leafhopper example (including mean squares) is given in Table 7. In this example the calculated F–statistic is $F_{calc} = 1.325/.075 = 17.6667$ and the P–value (computed using the F distribution with 3 and 4 degrees of freedom) is $P(F \geq 17.6667) = .0090$. Since the P–value .0090 is very small, we conclude that there is very strong evidence that diet does have an effect on the survival time of potato leafhoppers in the sense that at least two of the four treatment mean survival times are different.

Table 7. Potato leafhopper ANOVA table.

source of variation	degrees of freedom	sum of squares	mean square
among groups	3	3.975	1.325
within groups	4	.300	.075
total	7	4.275	

12.2b F–tests for comparing nested models

The overall F–test, for $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, developed above is too general to be of much use by itself. This overall F–test only allows us to conclude that either the k group means are all equal or they are not all equal. In many situations, like the potato leafhopper example, there is enough subject matter information to formulate more specific potential restrictions on the k group means. We will now discuss the use of F–tests for sequential comparisons (hypothesis tests) of a nested sequence of candidate models for k group means. We will develop this approach in the context of the potato leafhopper example.

Some natural groupings of the means μ_C, μ_F, μ_G and μ_S of the potato leafhopper example can be formed using the facts that fructose and glucose are 6–carbon sugars while sucrose is a 12–carbon sugar. Consider the following sequence of four nested models for the relationship among these means. These models are nested in the sense that each model in the sequence is a special case (restricted version) of the model that precedes it in the sequence. Thus model (2) is a special case (restricted version) of model (1); model (3) is a special case of model (2); and, model (4) is a special case of model (3).

model (1): The full model with four separate means, μ_C, μ_F, μ_G and μ_S.

model (2): The reduced model with three means, μ_C, μ_S, and the 6-carbon sugar mean μ_6, corresponding to the assumption that there is no difference between the effects of the two 6–carbon sugars in the sense that $\mu_F = \mu_G$.

model (3): The more reduced model with two means, μ_C, and the added sugar mean μ_A, corresponding to the assumption that there is no difference between the effects of the 6–carbon sugars and the 12–carbon sugar in the sense that $\mu_S = \mu_6$.

model (4): The most reduced model with one mean, μ, corresponding to the assumption that there is no difference between the effects of the added sugar diets and the control (no added sugar) diet in the sense that $\mu_C = \mu_A$.

Before we proceed with this example a brief discussion of the hypothesis testing approach to the comparison of a full model with a reduced model is in order. The reduced model is simpler than the full model in the sense that it specifies fewer means. Therefore, unless there is sufficient evidence to the contrary, we would prefer the simpler reduced model over the more complicated full model. This suggests a test of the null hypothesis

$$\begin{aligned}H_0 &: \text{The restrictions which define the reduced model are valid.}\\ &\quad \text{(The reduced model suffices and the full model is not needed.)}\end{aligned}$$

versus the alternative hypothesis

$$\begin{aligned}H_1 &: \text{The restrictions which define the reduced model are not valid.}\\ &\quad \text{(The reduced model does not suffice and the full model is needed.)}\end{aligned}$$

If we find sufficient evidence to reject the null hypothesis, then we conclude that the full model is needed and we abandon the reduced model. But, if we do not find sufficient evidence to reject the null hypothesis, then we conclude that we do not need the full model and the simpler reduced model will suffice.

We will now outline the approach we will use for our analysis of the sequence of four models for the potato leafhopper example.

Step 1: We will first consider a hypothesis test for comparing the full model (1) with the reduced model (2). The full model (1) specifies that there are four means μ_C, μ_F, μ_G, and μ_S. Since the reduced model (2) is obtained from model (1) by imposing the restriction that $\mu_F = \mu_G$, our null hypothesis is

$$H_0 : \mu_F = \mu_G$$

and our alternative hypothesis is

$$H_1 : \mu_F \neq \mu_G.$$

Under H_0 there is a common population mean survival time, μ_6, for the two 6–carbon sugar diets and our model only requires the three means μ_C, μ_S, and μ_6. Under H_1 there are two 6–carbon sugar diet means μ_F and μ_G and our model specifies the four means μ_C, μ_F, μ_G, and μ_S.

If we find sufficient evidence to reject H_0, we will conclude that we cannot reduce the four treatment means to three treatment means as indicated, since $\mu_F \neq \mu_G$ and thus we need four treatment means in our model. If this happens we will stop.

If we are not able to reject H_0 we will conclude that there is no difference between the two 6–carbon sugar treatment means ($\mu_F = \mu_G$) and we only need three treatment means in our model μ_C, μ_S, and the 6-carbon sugar mean μ_6. If this happens we will continue by comparing model (2) (which now plays the role of the full model) with the reduced model (3).

Step 2: If our comparison of model (1) and model (2) (step 1) results in the conclusion that we do not need the four means of model (1), then we will consider a test for comparing the current full model (2) with the reduced model (3). Model (2) specifies that there are three means μ_C, μ_S, and μ_6. Since the reduced model (3) is obtained from model (2) by imposing the restriction that $\mu_S = \mu_6$, our null hypothesis is

$$H_0 : \mu_S = \mu_6$$

and our alternative hypothesis is

$$H_1 : \mu_S \neq \mu_6.$$

Under H_0 there is a common population mean survival time, μ_A, for the three added sugar diets and our model only requires the two means μ_C and μ_A. Under H_1 there are two added sugar diet means, μ_S and μ_6, and our model specifies the three means μ_C, μ_S and μ_6.

If we find sufficient evidence to reject H_0, we will conclude that we cannot reduce the three treatment means to two treatment means as indicated, since $\mu_S \neq \mu_6$, and thus we need three treatment means in our model. If this happens we will stop.

If we are not able to reject H_0 we will conclude that there is no difference between the 6–carbon sugar treatment mean and the sucrose treatment mean ($\mu_S = \mu_6$) and we only need two treatment means in our model μ_C and the added sugar mean μ_A. If this happens we will continue by comparing model (3) (which now plays the role of the full model) with the reduced model (4).

Step 3: If our comparison of model (2) and model (3) (step 2) results in the conclusion that we do not need the three means of model (2), then we will consider a test for comparing the current full model (3) with the reduced model (4). Model (3) specifies that there are two means μ_C and μ_A. Since the reduced model (4) is obtained from model (3) by imposing the restriction that $\mu_C = \mu_A$, our null hypothesis is

$$H_0 : \mu_C = \mu_A$$

and our alternative hypothesis is

$$H_1 : \mu_C \neq \mu_A.$$

Under H_0 there is a common population mean survival time, μ, for all of the diets and our model only requires the one mean μ. Under H_1 there are two means, μ_C and μ_A.

If we find sufficient evidence to reject H_0 we will conclude that we cannot reduce the two treatment means to one treatment mean as indicated since, $\mu_C \neq \mu_A$, and thus we need two treatment means in our model. If this happens we will stop.

If we are not able to reject H_0 we will conclude that there is no difference between the control (no added sugar) treatment mean and the added sugar treatment mean ($\mu_C = \mu_A$) and we only need one treatment mean in our model. If this happens we will stop, since this is the end of our sequence of models.

Now that we have a plan of attack for our comparisons we need to know how to perform an F–test to compare a full model with a reduced model. Consider a full model with a group means and a reduced model with b group means ($b < a$) obtained by restrictions which result in a reduction of the a groups (means) of the full model into the b groups (means) of the reduced model. The sum of squares among the b groups in the reduced model SS(among the b groups) = SS(reduced model) is actually part of the sum of squares among the a groups in the full model SS(among the a groups) = SS(full model). The sum of squares due to the full model after the reduced model SS(full model | reduced model) is defined as the difference,

$$\begin{aligned}\text{SS(full model} \mid \text{reduced model)} &= \text{SS(full model)} - \text{SS(reduced model)} \\ &= \text{SS(among the } a \text{ groups)} - \text{SS(among the } b \text{ groups)},\end{aligned}$$

between the two model sums of squares. The degrees of freedom for this sum of squares is the corresponding difference, df(full model) − df(reduced model) $= a - b$, between the two model degrees of freedom. Partitioning the sum of squares among the a groups of the full model into the sum of squares among the b groups of the reduced model and the sum of squares for the full model after the reduced model yields the ANOVA of Table 8.

Table 8. ANOVA table for model comparison.

source of variation	degrees of freedom	sum of squares
reduced model	$b - 1$	SS(reduced model)
full model after reduced model	$a - b$	SS(full model | reduced model)
within the a groups of the full model	$n - a$	SS(within the a groups)
total	$n - 1$	SS(about the overall mean)

The F–test for comparing these models can be viewed as a test of

$$H_0: \text{ the reduced model with } b \text{ group means will suffice}$$

versus

$$H_1: \text{ the full model with } a \text{ group means is needed.}$$

More formally, the null hypothesis specifies that the restrictions which reduce the a means of the full model to the b means of the reduced model are valid. The F–statistic for this comparison is

$$F_{calc} = \frac{\text{MS(full model} \mid \text{reduced model)}}{\text{MS(within the } a \text{ groups of the full model)}}.$$

If the P–value $P(F \geq F_{calc})$, where F denotes an F variable with $a - b$ and $n - a$ degrees of freedom, is small enough, we reject H_0 and conclude that the reduced model with b group means is not appropriate and we need the full model with a group means. If the P–value is not small enough, we fail to reject H_0 and conclude that the reduced model with b group means is appropriate and we do not need the full model with a group means.

We will now use this method to evaluate the sequence of four models proposed above for the potato leafhopper example. The ANOVA's for models (1) and (2) are given in Tables 9 and 10.

Table 9. ANOVA table for model (1).

source of variation	degrees of freedom	sum of squares	mean square
among the 4 groups	3	3.975	1.325
within the 4 groups	4	.300	.075
total	7	4.275	

Table 10. ANOVA table for model (2).

source of variation	degrees of freedom	sum of squares	mean square
among the 3 groups	2	3.485	1.7425
within the 3 groups	5	.790	.1580
total	7	4.275	

The ANOVA for comparing model (1) and model (2) provided in Table 11 can be constructed from the information in the preceding ANOVA tables. The only computation required is to subtract the reduced model among groups sum of squares from the full model among groups sum of squares to get SS(full model | reduced model) = $3.975 - 3.485 = .49$, with $3 - 2 = 1$ degrees of freedom.

Table 11. ANOVA table for comparing model (1) and model (2).

source of variation	degrees of freedom	sum of squares	mean square
reduced model	2	3.485	1.7425
full model after reduced model	1	.490	.4900
within the 4 groups	4	.300	.0750
total	7	4.275	

The calculated F–statistic for comparing model (1) and model (2) is $F_{calc} = .49/.075 = 6.5333$ with a P–value of .0629. (This P–value is computed using the F distribution with 1 and 4 degrees of freedom.) This P–value is not small enough to allow us to reject the hypothesis that $\mu_F = \mu_G$ so we conclude that the three means (μ_C, μ_S, μ_6) of the reduced model (2) will suffice and we do not need the four means of the full model (1). We now proceed to compare the current full model (2) to the reduced model (3). The ANOVA for model (3) is given in Table 12.

Table 12. ANOVA table for model (3).

source of variation	degrees of freedom	sum of squares	mean square
among the 2 groups	1	1.4017	1.4017
within the 2 groups	6	2.8733	.4789
total	7	4.275	

We can produce the ANOVA for comparing model (2) and model (3) of Table 13 as before. In this case we find that SS(full model | reduced model) $= 3.485 - 1.4017 = 2.0833$, with $2 - 1 = 1$ degrees of freedom.

Table 13. ANOVA table for comparing model (2) and model (3).

source of variation	degrees of freedom	sum of squares	mean square
reduced model	1	1.4017	1.4017
full model after reduced model	1	2.0833	2.0833
within the 3 groups	5	.7900	.1580
total	7	4.275	

The calculated F–statistic for comparing model (2) and model (3) is $F_{calc} = 2.0833/.158 = 13.1854$ with a P–value of .0150. (This P–value is computed using the F distribution with 1 and 5 degrees of freedom.) This P–value is small enough to allow us to reject the hypothesis that $\mu_S = \mu_6$ so we conclude that the three means (μ_C, μ_S, μ_6) of model (2) are needed in the sense that the reduced model (3) with two means (μ_C, μ_A) does not suffice. We will stop at this point and base any further inferences about these diets on the three means of model (2).

Remark: At each stage of our sequential comparison of models for the potato leafhopper example we arrived at the reduced model by combining two groups from the full model which caused the degrees of freedom for the full model after the reduced model to be one in each comparison. It is possible to compare models for which the degrees of freedom for the full model after the reduced model is larger than one. We can demonstrate this by supposing that it had not occurred to us to consider combining the two 6–carbon sugar groups. That is, suppose that our initial comparison had been between model (1), with four separate means, and model (3), with two means (μ_C and μ_A), one for the no added sugar control diet group and one for the added sugar diet group. In this case the reduced model is obtained from the full model by combining the three added sugar groups to get a single added sugar group and the corresponding null hypothesis is $H_0 : \mu_F = \mu_G = \mu_S$. Thus, in this case the full model has 4 means (3 degrees of freedom), the reduced model has 2 means (1 degree of freedom), and the sum of squares for the full model after the reduced model has $3 - 1 = 2$ degrees of freedom. For this comparison we would have SS(full model | reduced model) $= 3.975 - 1.4017 = 2.5733$, with $3 - 1 = 2$ degrees of freedom, SS(within the 4 groups of the full model) $= .3$ with 4 degrees of freedom, and a calculated F–statistic of $F_{calc} = (2.5733/2)/(.3/4) = 17.1553$. If we were to perform this test, the P–value would be computed using the F distribution with 2 and 4 degrees of freedom.

12.2c Confidence intervals for linear combinations of means

A linear combination of the k population means $\mu_1, \ldots, \mu_k$ is a quantity of the form

$$\lambda = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k,$$

where the coefficients, $c_1, \ldots, c_k$, are suitably chosen constants. For example, if we take all of the coefficients in this linear combination to be $1/k$, we obtain the average of the means $(\mu_1 + \mu_2 + \cdots + \mu_k)/k$. If we take one coefficient to be 1, a second to be -1, and the others to be 0, we obtain a difference of two means, *e.g.*, taking $c_1 = -1$, $c_2 = 1$ and the other $c_i = 0$, yields $\mu_2 - \mu_1$.

The obvious estimate of the linear combination $\lambda = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k$ is the corresponding linear combination of the sample means

$$\hat{\lambda} = c_1\overline{Y}_1 + c_2\overline{Y}_2 + \cdots + c_k\overline{Y}_k.$$

In the present context of k independent random samples of sizes $n_1, \ldots, n_k$ with a common population variance σ^2, the population standard error of this estimated linear combination is

$$\text{S.E.}(\hat{\lambda}) = \sqrt{\sigma^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots \frac{c_k^2}{n_k}\right)}$$

which can be estimated, using the pooled estimator $S_p^2 = \text{MS(within)}$ of the common variance, by the sample standard error

$$\widehat{\text{S.E.}}(\hat{\lambda}) = \sqrt{S_p^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots \frac{c_k^2}{n_k}\right)}.$$

A set of confidence intervals is said to form a set of simultaneous 95% confidence intervals if the procedure which yields the set of confidence intervals is such that 95% of the time all of the intervals will contain the corresponding parameters. We can use the

Scheffé method to form simultaneous 95% confidence intervals for linear combinations of k population means. The basic idea of this method is to use a margin of error multiplier which is large enough to insure that the collection of confidence intervals it produces for all possible linear combinations of the k means form a set of simultaneous 95% confidence intervals. The margin of error multiplier for Scheffé's method when there are k means in the model is $\sqrt{(k-1)F_{(k-1,n-k)}(.95)}$, where $F_{(k-1,n-k)}(.95)$ is the 95th percentile of the F distribution with $k-1$ and $n-k$ degrees of freedom. Thus the 95% Scheffé margin of error for $\hat{\lambda} = c_1\overline{Y}_1 + c_2\overline{Y}_2 + \cdots + c_k\overline{Y}_k$ is

$$\text{M.E.}(\hat{\lambda}) = \sqrt{(k-1)[F_{(k-1,n-k)}(.95)]S_p^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots \frac{c_k^2}{n_k}\right)}.$$

We now return to our analysis of the potato leafhopper example for which we have concluded that model (2) with the three means μ_C, μ_S, and μ_6 is the appropriate model. We now need to make some sort of inference about the relationship among these three population mean survival times. We will use selected linear combinations and confidence intervals to explore the relationship among these three population mean survival times.

The sample means for the three diet groups are $\overline{Y}_C = 2$ (based on the $n_C = 2$ control diet observations), $\overline{Y}_S = 3.8$ (based on the $n_S = 2$ sucrose diet observations), and $\overline{Y}_6 = 2.55$ (based on the $n_6 = 4$ 6–carbon sugar diet observations). The pooled estimate of the population variance is $S_p^2 = \text{MS(within)} = .158$ with 5 degrees of freedom. For this model we have $k = 3$ means and $n = 8$ observations; therefore, the Scheffé margin of error multiplier is $\sqrt{2(5.7861)} = 3.4018$ (since the 95th percentile of the F distribution with 2 and 5 degrees of freedom is 5.7861).

We will begin our comparisons among the three means by estimating the three pairwise differences, $\mu_S - \mu_C$, $\mu_6 - \mu_C$, and $\mu_S - \mu_6$. First note that given two sample means $\overline{Y}_1$ and $\overline{Y}_2$ based on n_1 and n_2 observations the estimated standard error of $\overline{Y}_1 - \overline{Y}_2$ is

$$\widehat{\text{SE}}(\overline{Y}_1 - \overline{Y}_2) = \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

The estimates of the three pairwise differences and the corresponding standard errors and simultaneous 95% margins of error are given in the Table 14.

Table 14. Estimates of the pairwise differences.

difference	estimate	standard error	margin of error
$\mu_S - \mu_C$	$\overline{Y}_S - \overline{Y}_C = 1.8$	$\sqrt{.158\left(\frac{1}{2}+\frac{1}{2}\right)} = .3975$	$3.4018(.3975) = 1.3522$
$\mu_6 - \mu_C$	$\overline{Y}_6 - \overline{Y}_C = .55$	$\sqrt{.158\left(\frac{1}{2}+\frac{1}{4}\right)} = .3442$	$3.4018(.3442) = 1.1709$
$\mu_S - \mu_6$	$\overline{Y}_S - \overline{Y}_6 = 1.25$	$\sqrt{.158\left(\frac{1}{2}+\frac{1}{4}\right)} = .3442$	$3.4018(.3442) = 1.1709$

Adding and subtracting these margins of error from the corresponding estimates to get confidence intervals we conclude that we are 95% confident that $.4478 \le \mu_S - \mu_C \le 3.1522$, $-.6209 \le \mu_6 - \mu_C \le 1.7209$, and $.0791 \le \mu_S - \mu_6 \le 2.4209$. These confidence intervals suggest that μ_6 and μ_C are not different and that μ_S is larger than both of the other means. Thus, a confidence interval for $\mu_S - (\mu_C + \mu_6)/2$ would be useful for indicating how much larger μ_S is than the average of the other two means. Since this expression is a linear combination of the three means we can add a confidence interval for this combination to our set of confidence intervals and still have simultaneous confidence of 95%. Our estimate of $\mu_S - (\mu_C + \mu_6)/2$ is $\overline{Y}_S - (\overline{Y}_C + \overline{Y}_6)/2 = 1.525$ with standard error

$$\begin{aligned} \widehat{\mathrm{SE}}\left(\overline{Y}_S - \frac{\overline{Y}_C + \overline{Y}_6}{2}\right) &= \sqrt{S_p^2\left[\frac{1}{n_S} + \frac{1}{4n_C} + \frac{1}{4n_6}\right]} \\ &= \sqrt{.158\left[\frac{1}{2} + \frac{1}{8} + \frac{1}{16}\right]} = .3296 \end{aligned}$$

and margin of error $3.4018(.3296) = 1.1212$. Thus we are 95% confident that

$$.4038 \le \mu_S - (\mu_C + \mu_6)/2 \le 2.6462 \text{ and } -.6209 \le \mu_6 - \mu_C \le 1.7209.$$

Based on these confidence intervals we can conclude that there is no difference between the effects of adding a 6–carbon sugar to the diet or using the standard diet with no added sugar in the sense that the data are consistent with the claim that $\mu_C = \mu_6$. On the other hand, we find that adding the 12–carbon sugar sucrose to the potato leafhopper diet increases the mean survival time by something between .4038 and 2.6462 days over the average of the mean survival times corresponding to a diet with no added sugar or with an added 6–carbon sugar, *i.e.*, we can conclude with 95% confidence that μ_S exceeds the average $(\mu_C + \mu_6)/2$ by at least .4038 days and as much as 2.6462 days.

We will now revisit the fruitfly fecundity example of Chapter 8 and consider an analysis for this example using the methods of the present chapter.

Example. Fecundity of fruitflies (revisited). Sokal, R.R. and Rohlf, F.J. (1969) *Biometry*, W.H. Freeman, p.232, discuss a study conducted to compare the fecundity of three genetic lines of *Drosophila melanogaster*. The data, provided in Table 5 of Chapter 8, consist of per diem fecundities (number of eggs laid per female per day for the first 14 days of life) for 25 females of three lines of *Drosophila melanogaster*. Two of these genetic lines were selected for resistance (RS) and susceptibility (SS) to DDT, the third line is a nonselected control (NS). Recall that the investigator wanted to know if there was any evidence that the population mean fecundities for the two selected lines (μ_{RS} and μ_{SS})

were different. The investigator also wanted to know how the population mean fecundity μ_{NS} for the nonselected line related to the mean fecundities of the selected lines.

When we first considered this example, we found that the data was reasonably modeled as consisting of three independent random samples, each of size 25, from normal distributions with respective population mean fecundities μ_{RS}, μ_{SS}, and μ_{NS} and with common population variance σ^2. We can address the investigator's question about the relationship between the mean fecundities of the selected lines using the following sequence of two nested models.

model (1): The full model with three separate means, μ_{RS} for the resistant line, μ_{SS} for the susceptible line, and μ_{NS} for the nonselected line.

model (2): The reduced model with two means, μ_{NS} for the nonselected line and μ_S for the selected lines corresponding to the assumption that there is no difference between the mean fecundities for the two selected lines in the sense that $\mu_{RS} = \mu_{SS}$.

The ANOVA's for models (1) and (2) are given in Tables 15 and 16 and the ANOVA for comparing these models is given in Table 17.

Table 15. ANOVA table for the full model with 3 lines.

source of variation	degrees of freedom	sum of squares	mean square
among the 3 lines	2	1362.2115	681.1057
within the 3 lines	72	5659.0224	78.5975
total	74	7021.2339	

Table 16. ANOVA table for reduced model with 2 lines.

source of variation	degrees of freedom	sum of squares	mean square
among the 2 lines	1	1329.0817	1329.0817
within the 2 lines	73	5692.1522	77.9747
total	74	7021.2339	

Table 17. ANOVA table for comparing the models.

source of variation	degrees of freedom	sum of squares	mean square
2 line model	1	1329.0817	1329.0817
3 line model after 2 line model	1	33.1298	33.1298
within the 3 lines	72	5659.0224	78.5975
total	74	7021.2339	

The calculated F–statistic for comparing model (1) and model (2) is $F_{calc} = 33.1298/78.5975 = .42$ with a P–value of .5182. (This P–value is computed using the F

distribution with 1 and 72 degrees of freedom.) This P–value is quite large indicating that there is no evidence that μ_{RS} is different from μ_{SS}. We can conclude that the three means (μ_{RS}, μ_{SS}, and μ_{NS}) of the full model (1) are not needed and we are justified in adopting the simplified model (2) with mean fecundity μ_{NS} for the nonselected line and mean fecundity μ_S for the selected lines. The remainder of our analysis will be in terms of this reduced model.

Before we proceed with our analysis of this example it is instructive to compare the ANOVA F–test we just used to test the null hypothesis $H_0 : \mu_{RS} = \mu_{SS}$ versus the alternative hypothesis $H_1 : \mu_{RS} \neq \mu_{SS}$ with the t–test we used in Chapter 8 for this same hypothesis test. The ANOVA F–test is equivalent to a t–test based on the difference $\overline{Y}_{RS} - \overline{Y}_{SS} = 1.628$ and the pooled sample variance MS(within) = 78.5975 with 72 degrees of freedom. This pooled sample variance has 72 degrees of freedom, since it is computed using all three of the samples. The t–test we considered in Chapter 8 was based on the difference $\overline{Y}_{RS} - \overline{Y}_{SS} = 1.628$ and the pooled sample variance S_p^2 with 48 degrees of freedom based on the two samples from the selected lines. Thus, these two t–tests differ because they use different estimated standard errors due to the way in which the population variance is estimated. If the assumption of a common variance for all three lines is reasonable, then the ANOVA F–test is better than the t–test of Chapter 8, since it is based on a better (higher degrees of freedom) estimate of the population variance.

Since there are only two means in the reduced model we can use the overall F–test to compare these means. The calculated F–statistic for testing the null hypothesis $H_0 : \mu_{NS} = \mu_S$ is $F_{calc} = 1329.0817/77.9747 = 17.05$ with a P–value that is less than .0001. (This P–value is computed using the F distribution with 1 and 73 degrees of freedom.) This very small P–value indicates that there is very strong evidence that the mean fecundity for the nonselected line μ_{NS} is not the same as the mean fecundity μ_S for the selected lines. This F–test for comparing these two means is equivalent to the t–test we performed in Chapter 8 in these sense that these two tests give the same P–value. In fact, for the present circumstance of comparing two means (using a model with only two means) the square of the Student's t–statistic is equal to the F–statistic. We can form a 95% confidence interval for the difference $\mu_{NS} - \mu_S$ between these mean fecundities to determine which mean is larger and to get an estimate of the size of this difference. In this example, we are 95% confident that $\mu_{NS} - \mu_S$ is between 4.6192 and 13.241. That is, we are 95% confident that the population mean fecundity (mean number of eggs laid per day for the first 14 days of life) μ_{NS} for the nonselected line exceeds the population mean fecundity μ_S for the selected lines by at least 4.6192 eggs per day and perhaps as much as 13.241 eggs per day.

In conclusion, we have found that the distributions of fruitfly fecundity for two selected populations are identical (since we assumed a common variance and since we failed to reject

$\mu_{RS} = \mu_{SS}$); but, the distribution of fruitfly fecundity for the nonselected population differs from the distribution for the selected population by having a larger (by 4.6192 to 13.241 eggs per day) population mean fecundity.

Before we leave this example we will consider one more approach to its analysis. Suppose that we did not have enough *a priori* information to allow use to confidently propose a reasonable sequence of nested models for our analysis. In this situation we could perform an exploratory analysis by using the Scheffé method to form simultaneous 95% confidence intervals for interesting linear combinations of the three population mean fecundities.

We begin our analysis by considering the three pairwise differences between the population mean fecundities. The estimates of the three pairwise differences and the simultaneous 95% confidence intervals are given in the Table 18.

Table 18. Estimates of the pairwise differences.

difference	estimate	confidence interval
$\mu_{NS} - \mu_{RS}$	8.116	(1.848, 14.384)
$\mu_{NS} - \mu_{SS}$	9.744	(3.476, 16.012)
$\mu_{RS} - \mu_{SS}$	1.628	(-4.640, 7.896)

Based on these simultaneous confidence intervals we can conclude that the population mean fecundities μ_{RS} and μ_{SS} for the selected lines are not different and we can conclude that the population mean fecundity for the nonselected line μ_{NS} is larger than each of the other population mean fecundities. Since we have concluded that the selected line means are not different it would be of interest to also consider a contrast between the nonselected line population mean μ_{NS} and the average $(\mu_{RS} + \mu_{SS})/2$. The estimate of the contrast $\mu_{NS} - (\mu_{RS} + \mu_{SS})/2$ is 8.93 and the Scheffé method gives the confidence interval $(3.5020, 14.3581)$. Thus we can conclude, with 95% confidence that $-4.640 \leq \mu_{RS} - \mu_{SS} \leq 7.896$ and $3.5020 \leq \mu_{NS} - (\mu_{RS} + \mu_{SS})/2 \leq 14.3581$. This allows us to conclude that $\mu_{RS} = \mu_{SS}$ and μ_{NS} exceeds $(\mu_{RS} + \mu_{SS})/2$ by at least 3.5020 and as much as 14.3581 eggs per day.