Inferences for Contrasts and Treatment Means

4

4.1 Introduction

The objective of an experiment is often much more specific than merely determining whether or not all of the treatments give rise to similar responses. For example, a chemical experiment might be run primarily to determine whether or not the yield of the chemical process increases as the amount of the catalyst is increased. A medical experiment might be concerned with the efficacy of each of several new drugs as compared with a standard drug. A nutrition experiment may be run to compare high fiber diets with low fiber diets. Such treatment comparisons are formalized in Sect. 4.2. The purpose of this chapter is to provide confidence intervals and hypothesis tests about treatment comparisons and treatment means. We start, in Sect. 4.3, by considering a single treatment comparison or mean, and then, in Sect. 4.4, we develop the techniques needed when more than one treatment comparison or mean is of interest. The number of observations required to achieve confidence intervals of given lengths is calculated in Sect. 4.5. SAS and R commands for confidence intervals and hypothesis tests are provided in Sects. 4.6 and 4.7, respectively.

4.2 Contrasts

In Chap. 3, we defined a contrast to be a linear combination of the parameters $\tau_1, \tau_2, \ldots, \tau_v$ of the form

$$\sum c_i \tau_i \,, \quad \text{with} \quad \sum c_i = 0 \,.$$

For example, $\tau_u - \tau_s$ is the contrast that compares the effects (as measured by the response variable) of treatments u and s. If $\tau_u - \tau_s = 0$, then treatments u and s affect the response in exactly the same way, and we say that these treatments do not differ. Otherwise, the treatments do differ in the way they affect the response. We showed in Sect. 3.4 that for a completely randomized design and the one-way analysis of variance model (3.3.1), every contrast $\sum c_i \tau_i$ is estimable with least squares estimate

$$\sum c_i \hat{\tau}_i = \sum c_i (\hat{\mu} + \hat{\tau}_i) = \sum c_i \overline{y}_{i\cdot} \tag{4.2.1}$$

A. Dean et al., *Design and Analysis of Experiments*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-52250-0_4

and corresponding least squares estimator $\sum c_i \overline{Y}_{i.}$. The variance of the least squares estimator is

$$\text{Var}\left(\sum c_i \overline{Y}_{i.}\right) = \sum c_i^2 \text{Var}(\overline{Y}_{i.}) = \sum c_i^2 (\sigma^2/r_i) = \sigma^2 \sum (c_i^2/r_i)\,. \tag{4.2.2}$$

The first equality uses the fact that the treatment sample means $\overline{Y}_{i.}$ involve different response variables, which in model (3.3.1) are independent. The error variance σ^2 is generally unknown and is estimated by the unbiased estimate *msE*, giving the estimated variance of the contrast estimator as

$$\widehat{\text{Var}}\left(\sum c_i \overline{Y}_{i.}\right) = msE \sum (c_i^2/r_i).$$

The *estimated standard error* of the estimator is the square root of this quantity, namely,

$$\sqrt{\widehat{\text{Var}}\left(\sum c_i \overline{Y}_{i.}\right)} = \sqrt{msE \sum (c_i^2/r_i)}\,. \tag{4.2.3}$$

Normalized Contrasts

When several contrasts are to be compared, it is sometimes helpful to be able to measure them all on the same scale. A contrast is said to be *normalized* if it is scaled so that its least squares estimator has variance σ^2. From (4.2.2), it can be seen that a contrast $\Sigma c_i \tau_i$ is normalized by dividing it by $\sqrt{\Sigma c_i^2/r_i}$. If we write $h_i = c_i/\sqrt{\Sigma c_i^2/r_i}$, then the least squares estimator $\Sigma h_i \overline{Y}_{i.}$ of the normalized contrast $\Sigma h_i \tau_i$ has the following distribution:

$$\sum h_i \overline{Y}_{i.} \sim N\left(\sum h_i \tau_i, \sigma^2\right), \quad \text{where } h_i = \frac{c_i}{\sqrt{\sum c_i^2/r_i}}\,.$$

Normalized contrasts will be used for hypothesis testing (Sect. 4.3.3).

Contrast Coefficients

It is convenient to represent a contrast by listing only the coefficients of the parameters $\tau_1, \tau_2, \ldots, \tau_v$. Thus, $\sum c_i \tau_i = c_1\tau_1 + c_2\tau_2 + \cdots + c_v\tau_v$ would be represented by the list of *contrast coefficients*

$$[c_1,\ c_2,\ \ldots,\ c_v]\,.$$

Some types of contrasts are used frequently in practice, and these are identified in Sects. 4.2.1–4.2.4.

4.2.1 Pairwise Comparisons

As the name suggests, *pairwise comparisons* are simple differences $\tau_u - \tau_s$ of pairs of parameters τ_u and τ_s ($u \neq s$). These are of interest when the experimenter wishes to compare each treatment with every other treatment. The list of contrast coefficients for the pairwise difference $\tau_u - \tau_s$ is

$$[0, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0]\,,$$

where the 1 and -1 are in positions u and s, respectively. The least squares estimate of $\tau_u - \tau_s$ is obtained from (4.2.1) by setting $c_u = 1$, $c_s = -1$, and all other c_i equal to zero, giving

$$\hat{\tau}_u - \hat{\tau}_s = \overline{y}_{u.} - \overline{y}_{s.}\,,$$

and the corresponding least squares estimator is $\overline{Y}_{u.} - \overline{Y}_{s.}$. Its estimated standard error is obtained from (4.2.3) and is equal to

$$\sqrt{\widehat{\mathrm{Var}}(\overline{Y}_{u.} - \overline{Y}_{s.})} = \sqrt{msE\,((1/r_u) + (1/r_s))}\,.$$

Example 4.2.1 `Battery experiment, continued`

Details for the battery experiment were given in Sect. 2.5.2 (p. 24). The experimenter was interested in comparing the life per unit cost of each battery type with that of each of the other battery types. The average lives per unit cost (in minutes/dollar) for the four batteries, calculated from the data in Table 2.8, p. 27, are

$$\overline{y}_{1.} = 570.75\,, \quad \overline{y}_{2.} = 860.50\,, \quad \overline{y}_{3.} = 433.00\,, \quad \overline{y}_{4.} = 496.25\,.$$

The least squares estimates of the pairwise differences are, therefore,

$$\hat{\tau}_1 - \hat{\tau}_2 = -289.75\,, \quad \hat{\tau}_1 - \hat{\tau}_3 = 137.75\,, \quad \hat{\tau}_1 - \hat{\tau}_4 = 74.50\,,$$
$$\hat{\tau}_2 - \hat{\tau}_3 = 427.50\,, \quad \hat{\tau}_2 - \hat{\tau}_4 = 364.25\,, \quad \hat{\tau}_3 - \hat{\tau}_4 = -63.25\,.$$

The estimated pairwise differences suggest that battery type 2 (alkaline, store brand) is vastly superior to the other three battery types in terms of the mean life per unit cost. Battery type 1 (alkaline, name brand) appears better than types 3 and 4, and battery type 4 (heavy duty, store brand) better than type 3 (heavy duty, name brand). We do, however, need to investigate whether or not these perceived differences might be due only to random fluctuations in the data.

In Example 3.4.2 (p. 40), the error variance was estimated to be $msE = 2367.71$. The sample sizes were $r_1 = r_2 = r_3 = r_4 = 4$, and consequently, the estimated standard error for each pairwise comparison is equal to

$$\sqrt{2367.71\left(\frac{1}{4} + \frac{1}{4}\right)} = 34.41 \text{ min/\$}\,.$$

It can be seen that all of the estimated pairwise differences involving battery type 2 are bigger than four times their estimated standard errors. This suggests that the perceived differences in battery type 2 and the other batteries are of sizable magnitudes and are unlikely to be due to random error. We shall formalize these comparisons in terms of confidence intervals in Example 4.4.3 later in this chapter. □

4.2.2 Treatment Versus Control

If the experimenter is interested in comparing the effects of one special treatment with the effects of each of the other treatments, then the special treatment is called the *control*. For example, a pharmaceutical experiment might involve one or more experimental drugs together with a standard drug that has been on the market for some years. Frequently, the objective of such an experiment is to compare the effect of each experimental drug with that of the standard drug but not necessarily with the effects of any of the other experimental drugs. The standard drug is then the control. If we code the control as level 1, and the experimental drugs as levels $2, 3, \ldots, v$, respectively, then the contrasts of interest are $\tau_2 - \tau_1, \tau_3 - \tau_1, \ldots, \tau_v - \tau_1$. These contrasts are known as *treatment versus control* contrasts. They form a subset of the pairwise differences, so we can use the same formulae for the least squares estimate and the estimated

standard error. The contrast coefficients for the contrast $\tau_i - \tau_1$ are $[-1, 0, \ldots, 0, 1, 0, \ldots, 0]$, where the 1 is in position i.

4.2.3 Difference of Averages

Sometimes the levels of the treatment factors divide naturally into two or more groups, and the experimenter is interested in the *difference of averages* contrast that compares the average effect of one group with the average effect of the other group(s). For example, consider an experiment that is concerned with the effect of different colors of exam paper (the treatments) on students' exam performance (the response). Suppose that treatments 1 and 2 represent the pale colors, white and yellow, whereas treatments 3, 4, and 5 represent the darker colors, blue, green and pink. The experimenter may wish to compare the effects of light and dark colors on exam performance. One way of measuring this is to estimate the contrast $\frac{1}{2}(\tau_1 + \tau_2) - \frac{1}{3}(\tau_3 + \tau_4 + \tau_5)$, which is the difference of the average effects of the light and dark colors. The corresponding contrast coefficients are

$$\left[\frac{1}{2}, \frac{1}{2}, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}\right].$$

From (4.2.1) and (4.2.3), the least squares estimate would be

$$\frac{1}{2}\overline{y}_{1.} + \frac{1}{2}\overline{y}_{2.} - \frac{1}{3}\overline{y}_{3.} - \frac{1}{3}\overline{y}_{4.} - \frac{1}{3}\overline{y}_{5.}$$

with estimated standard error

$$\sqrt{msE\left(\frac{1}{4r_1} + \frac{1}{4r_2} + \frac{1}{9r_3} + \frac{1}{9r_4} + \frac{1}{9r_5}\right)}.$$

Example 4.2.2 `Battery experiment, continued`

In the battery experiment of Sect. 2.5.2, p. 24, battery types 1 and 2 were alkaline batteries, while types 3 and 4 were heavy duty. In order to compare the running time per unit cost of these two types of batteries, we examine the contrast $\frac{1}{2}(\tau_1 + \tau_2) - \frac{1}{2}(\tau_3 + \tau_4)$. The least squares estimate is

$$\frac{1}{2}(570.75 + 860.50) - \frac{1}{2}(433.00 + 496.25) = 251.00 \text{ min/\$},$$

suggesting that the alkaline batteries are more economical (on average by over four hours per dollar spent). The associated standard error is $\sqrt{msE(4/16)} = 24.32$ min/\$, so the estimated difference in running time per unit cost is over ten times larger than the standard error, suggesting that the observed difference is not just due to random fluctuations in the data. □

4.2.4 Trends

Trend contrasts may be of interest when the levels of the treatment factor are quantitative and have a natural ordering. For example, suppose that the treatment factor is temperature and its selected levels are 50 °C, 75 °C, 100 °C, coded as 1, 2, 3, respectively. The experimenter may wish to know whether

the value of the response variable increases or decreases as the temperature increases and, if so, whether the rate of change remains constant. These questions can be answered by estimating linear and quadratic trends in the response.

The trend contrast coefficients for v *equally spaced* levels of a treatment factor and *equal sample sizes* are listed in Table A.2 for values of v between 3 and 7. For v treatments, trends up to $(v-1)$th order can be measured. Experimenters rarely use more than four levels for a quantitative treatment factor, since it is unusual for strong quartic and higher-order trends to occur in practice, especially within the narrow range of levels considered in a typical experiment.

Table A.2 does not tabulate contrast coefficients for unequally spaced levels or for unequal sample sizes. The general method of obtaining the coefficients of the trend contrasts involves fitting a regression model to the noncoded levels of the treatment factor. It can be shown that the linear trend contrast coefficients can easily be calculated as

$$c_i \;=\; r_i(x_i - \overline{x}_{\cdot}), \;\text{ where } \overline{x}_{\cdot} = (\Sigma r_i x_i)/n\,, \tag{4.2.4}$$

where r_i is the number of observations taken on the ith uncoded level x_i of the treatment factor, and $n = \Sigma r_i$ is the total number of observations. We are usually interested only in whether or not the linear trend is likely to be negligible, and to make this assessment, the contrast estimate is compared with its standard error. Consequently, we may multiply or divide the calculated coefficients by any integer without losing any information. When the r_i are all equal, the coefficients listed in Appendix A.2 are obtained, possibly multiplied or divided by an integer. Expressions for quadratic and higher-order trend coefficients are more complicated (see Draper and Smith 1998, Chap. 22).

Example 4.2.3 `Heart-lung pump experiment, continued`

The experimenter who ran the heart–lung pump experiment of Example 3.4.1, p. 37, expected to see a linear trend in the data, since he expected the flow rate to increase as the number of revolutions per minute (rpm) of the pump head was increased. The plot of the data in Fig. 3.1 (p. 38) shows the observed flow rates at the five different levels of rpm. From the figure, it might be anticipated that the linear trend is large but higher-order trends are very small.

The five levels of rpm observed were 50, 75, 100, 125, 150, which are equally spaced. Had there been equal numbers of observations at each level, then we could have used the contrast coefficients $[-2, -1,\ 0,\ 1,\ 2\,]$ for the linear trend contrast and $[\,2, -1, -2, -1,\ 2\,]$ for the quadratic trend contrast as listed in Table A.2 for $v = 5$ levels of the treatment factor. However, here the sample sizes were $r_1 = r_3 = r_5 = 5$, $r_2 = 3$ and $r_4 = 2$. The coefficients for the linear trend are calculated via (4.2.4). Now $n = \Sigma r_i = 20$, and

$$(\Sigma r_i x_i)/n \;=\; 20^{-1} \times (5(50) + 3(75) + 5(100) + 2(125) + 5(150)) \;=\; 98.75\,.$$

So, we have

x_i	$r_i(x_i - \overline{x}_{\cdot})$	
50	$5 \times (50 - 98.75) =$	-243.75
75	$3 \times (75 - 98.75) =$	-71.25
100	$5 \times (100 - 98.75) =$	6.25
125	$2 \times (125 - 98.75) =$	52.50
150	$5 \times (150 - 98.75) =$	256.25

If the coefficients are multiplied by 4, they are then integers each divisible by 5 so rather than using the calculated coefficients $[-243.75,\ -71.25,\ 6.25,\ 52.50,\ 256.25]$, we can multiply them by 4/5 and use the linear trend coefficients $[-195, -57, 5, 42, 205]$. The average flow rates (l/min) were calculated as

$$\overline{y}_{1.} = 1.1352\,,\ \overline{y}_{2.} = 1.7220\,,\ \overline{y}_{3.} = 2.3268\,,\ \overline{y}_{4.} = 2.9250\,,\ \overline{y}_{5.} = 3.5292\,.$$

The least squares estimate $\Sigma c_i \overline{y}_{i.}$ of the linear contrast is then

$$-195\overline{y}_{1.} - 57\overline{y}_{2.} + 5\overline{y}_{3.} + 42\overline{y}_{4.} + 205\overline{y}_{5.} = 538.45$$

l/min. The linear trend certainly appears to be large. However, before drawing conclusions, we need to compare this trend estimate with its corresponding estimated standard error. The data give $\sum\sum y_{it}^2 = 121.8176$, and we calculate the error sum of squares (3.4.5), p. 39, as $ssE = 0.0208$, giving an unbiased estimate of σ^2 as

$$msE = ssE/(n - v) = 0.0208/(20 - 5) = 0.001387\,.$$

The estimated standard error of the linear trend estimator is then

$$\sqrt{msE\left(\frac{(-195)^2}{5} + \frac{(-57)^2}{3} + \frac{(5)^2}{5} + \frac{(42)^2}{2} + \frac{(205)^2}{5}\right)} = 4.988\,.$$

Clearly, the estimate of the linear trend is extremely large compared with its standard error.

Had we normalized the contrast, the linear contrast coefficients would each have been divided by

$$\sqrt{\sum c_i^2/r_i} = \sqrt{\frac{(-195)^2}{5} + \frac{(-57)^2}{3} + \frac{(5)^2}{5} + \frac{(42)^2}{2} + \frac{(205)^2}{5}} = 134.09,$$

and the normalized linear contrast estimate would have been 4.0156. The estimated standard error of all normalized contrasts is $\sqrt{msE} = 0.03724$ for this experiment, so the normalized linear contrast estimate remains large compared with the standard error. □

4.3 Individual Contrasts and Treatment Means

4.3.1 Confidence Interval for a Single Contrast

In this section, we obtain a formula for a confidence interval for an individual contrast. If confidence intervals for more than one contrast are required, then the multiple comparison methods of Sect. 4.4 should be used instead. We give the formula first, and the derivation afterwards. A $100(1-\alpha)\%$ confidence interval for the contrast $\Sigma c_i \tau_i$ is

$$\begin{aligned}\sum c_i \overline{y}_{i.} - t_{n-v,\alpha/2}\sqrt{msE\sum c_i^2/r_i} \ \le\ \sum c_i\tau_i \\ \le \sum c_i \overline{y}_{i.} + t_{n-v,\alpha/2}\sqrt{msE\sum c_i^2/r_i}\,.\end{aligned} \tag{4.3.5}$$

We can write this more succinctly as

$$\sum c_i \tau_i \ \in \ \left(\sum c_i \overline{y}_{i.} \pm t_{n-v,\alpha/2} \sqrt{msE \sum c_i^2 / r_i} \right), \tag{4.3.6}$$

where the symbol $\pm$, which is read as "plus or minus," denotes that the upper limit of the interval is calculated using $+$ and the lower limit using $-$. The symbols "$\Sigma c_i \tau_i \ \in$" mean that the interval includes the true value of the contrast $\Sigma c_i \tau_i$ with $100(1-\alpha)\%$ confidence. For future reference, we note that the general form of the above confidence interval is

$$\sum c_i \tau_i \ \in \ \left(\sum c_i \hat{\tau}_i \pm t_{df,\alpha/2} \sqrt{\widehat{\text{Var}}(\Sigma c_i \hat{\tau}_i)} \right), \tag{4.3.7}$$

where *df* is the number of degrees of freedom for error.

To derive the confidence interval (4.3.5), we will need to use some results about normally distributed random variables. As we saw in the previous section, for the completely randomized design and one-way analysis of variance model (3.3.1), the least squares estimator of the contrast $\sum c_i \tau_i$ is $\sum c_i \overline{Y}_{i.}$, which has variance $\text{Var}(\Sigma c_i \overline{Y}_{i.}) = \sigma^2 \sum c_i^2 / r_i$. This estimator is a linear combination of normally distributed random variables and therefore also has a normal distribution. Subtracting the mean and dividing by the standard deviation gives us a random variable

$$\frac{D}{\sigma} = \frac{\sum c_i \overline{Y}_{i.} - \sum c_i \tau_i}{\sigma \sqrt{\sum c_i^2 / r_i}}, \tag{4.3.8}$$

which has a $N(0, 1)$ distribution. We estimate the error variance, σ^2, by *msE*, and from Sect. 3.4.6, p. 39, we know that

$$MSE/\sigma^2 = SSE/(n-v)\sigma^2 \sim \chi^2_{n-v}/(n-v).$$

It can be shown that the random variables D and MSE are independent (see Graybill, 1976), and the ratio of a normally distributed random variable and a chi-squared random variable that are independent has a t-distribution with the same number of degrees of freedom as the chi-squared distribution. Hence, the ratio $D/\sqrt{MSE}$ has a t distribution with $n-v$ degrees of freedom. Using the expression (4.3.8), we can now write down the following probability statement about $D/\sqrt{MSE}$:

$$P\left(-t_{n-v,\alpha/2} \ \le \ \frac{\sum c_i \overline{Y}_{i.} - \sum c_i \tau_i}{\sqrt{MSE \sum c_i^2 / r_i}} \ \le \ t_{n-v,\alpha/2} \right) = 1 - \alpha,$$

where $t_{n-v,\alpha/2}$ is the percentile of the t_{n-v} distribution corresponding to a probability of $\alpha/2$ in the right-hand-tail, the value of which can be obtained from Table A.4. Manipulating the two inequalities, the probability statement becomes

$$P\left(\sum c_i \overline{Y}_{i.} - t_{n-v,\alpha/2} \sqrt{MSE \sum c_i^2 / r_i} \ \le \ \sum c_i \tau_i \right. \tag{4.3.9}$$

$$\left. \le \ \sum c_i \overline{Y}_{i.} + t_{n-v,\alpha/2} \sqrt{MSE \sum c_i^2 / r_i} \right) = 1 - \alpha.$$

Then replacing the estimators by their observed values in this expression gives a $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$ as in (4.3.5).

Example 4.3.1 `Heart-lung pump experiment, continued`

Consider the heart–lung pump experiment of Examples 3.4.1 and 4.2.3, p. 37 and 73. The least squares estimate of the difference in fluid flow at 75 rpm and 50 rpm (levels 2 and 1 of the treatment factor, respectively) is

$$\Sigma c_i \overline{y}_{i.} = \overline{y}_{2.} - \overline{y}_{1.} = 0.5868$$

l/min. Since there were $r_2 = 5$ observations at 75 rpm and $r_1 = 3$ observations at 50 rpm, and $msE =$ 0.001387, the estimated standard error of this contrast is

$$\sqrt{msE\ \Sigma c_i^2/r_i} = \sqrt{0.001387\ \left(\frac{1}{3} + \frac{1}{5}\right)} = 0.0272\ \text{l/min}\,.$$

Using this information, together with $t_{15,0.025} = 2.131$, we obtain from (4.3.6) a 95% confidence interval (in units of l/min) for $\tau_2 - \tau_1$ as

$$(0.5868 \pm (2.131)(0.0272)) \quad = \quad (0.5288, 0.6448)\,.$$

This tells us that with 95% confidence, the fluid flow at 75 rpm of the pump is between 0.53 and 0.64 liters per minute greater than at 50 rpm. □

Confidence bounds, or one-sided confidence intervals, can be derived in the same manner as two-sided confidence intervals. For the completely randomized design and one-way analysis of variance model (3.3.1), a $100(1-\alpha)\%$ *upper confidence bound* for $\sum c_i \tau_i$ is

$$\sum c_i \tau_i < \sum c_i \overline{y}_{i.} + t_{df,\alpha} \sqrt{msE \sum c_i^2/r_i}\,, \tag{4.3.10}$$

and a $100(1-\alpha)\%$ *lower confidence bound* for $\sum c_i \tau_i$ is

$$\sum c_i \tau_i > \sum c_i \overline{y}_{i.} - t_{df,\alpha} \sqrt{msE \sum c_i^2/r_i}\,, \tag{4.3.11}$$

where $t_{df,\alpha}$ is the percentile of the t distribution with df degrees of freedom and probability α in the right-hand tail.

4.3.2 Confidence Interval for a Single Treatment Mean

For the one-way analysis of variance model (3.3.1), the true mean response $\mu + \tau_s$ of the sth level of a treatment factor was shown in Sect. 3.4 to be estimable with least squares estimator $\overline{Y}_{s.}$. Although one is unlikely to be interested in only one of the treatment means, we can obtain a confidence interval as follows.

Since $\overline{Y}_{s.} \sim N(\mu + \tau_s, \sigma^2/r_s)$ for model (3.3.1), we can follow the same steps as those leading to (4.3.6) and obtain a $100(1-\alpha)\%$ confidence interval for $\mu + \tau_s$ as

$$\mu + \tau_s \in (\overline{y}_{s.} \pm t_{df,\alpha/2} \sqrt{msE/r_s})\,. \tag{4.3.12}$$

Example 4.3.2 `Heart-lung pump experiment, continued`

Suppose that the experimenter had required a 99% confidence interval for the true average fluid flow ($\mu + \tau_3$) for the heart–lung pump experiment of Example 3.4.1, p. 37, when the revolutions per minute of the pump are set to 100 rpm. Using (4.3.12) and $r_3 = 5$, $\overline{y}_{3.} = 2.3268$, $msE = 0.001387$, $n - v = 20 - 5$, and $t_{15,0.005} = 2.947$, the 99% confidence interval for $\mu + \tau_3$ is

$$\mu + \tau_3 \quad \in \quad (2.3268 \pm (2.947)(0.01666)) \quad = \quad (2.2777, 2.3759)\,.$$

So, with 99% confidence, the true average flow rate at 100 rpm of the pump is believed to be between 2.28 and 2.38 l/min. □

4.3.3 Hypothesis Test for a Single Contrast or Treatment Mean

The outcome of a hypothesis test can be deduced from the corresponding confidence interval in the following way. The null hypothesis $H_0 : \Sigma c_i \tau_i = h$ will be rejected at significance level α in favor of the two-sided alternative hypothesis $H_A : \Sigma c_i \tau_i \neq h$ if the corresponding confidence interval for $\Sigma c_i \tau_i$ fails to contain h. For example, the 95% confidence interval for $\tau_2 - \tau_1$ in Example 4.3.1 does not contain zero, so the hypothesis $H_0 : \tau_2 - \tau_1 = 0$ (that the flow rates are the same at 50 and 75 rpm) would be rejected at significance level $\alpha = 0.05$ in favor of the alternative hypothesis (that the flow rates are not equal).

We can make this more explicit, as follows. Suppose we wish to test the hypothesis $H_0 : \Sigma c_i \tau_i = 0$ against the alternative hypothesis $H_A : \Sigma c_i \tau_i \neq 0$. The interval (4.3.6) fails to contain 0 if the absolute value of $\Sigma c_i \overline{y}_{i.}$ is bigger than $t_{n-v,\alpha/2}\sqrt{msE\,\Sigma c_i^2/r_i}$. Therefore, the rule for testing the null hypothesis against the alternative hypothesis is

$$\text{reject } H_0 \text{ if } \left| \frac{\sum c_i \overline{y}_{i.}}{\sqrt{msE \sum c_i^2/r_i}} \right| > t_{n-v,\alpha/2}\,, \tag{4.3.13}$$

where | | denotes absolute value. We call such rules *decision rules*. If H_0 is rejected, then H_A is automatically accepted. The test statistic can be squared, so that the decision rule becomes

$$\text{reject } H_0 \text{ if } \frac{\left(\sum c_i \overline{y}_{i.}\right)^2}{msE \sum c_i^2/r_i} > t^2_{n-v,\alpha/2} = F_{1,n-v,\alpha}\,,$$

and the F distribution can be used instead of the t distribution. Notice that the test statistic is the square of the normalized contrast estimate divided by msE. We call the quantity

$$ssc = \frac{\left(\sum c_i \overline{y}_{i.}\right)^2}{\sum c_i^2/r_i} \tag{4.3.14}$$

the *sum of squares for the contrast*, or *contrast sum of squares* (even though it is the "sum" of only one squared term). The decision rule can be more simply expressed as

$$\text{reject } H_0 \text{ if } \frac{ssc}{msE} > F_{1,n-v,\alpha}\,. \tag{4.3.15}$$

For future reference, we can see that the general form of *ssc/msE* is

$$\frac{ssc}{msE} = \frac{\left(\sum c_i \hat{\tau}_i\right)^2}{\widehat{\text{Var}}\left(\sum c_i \hat{\tau}_i\right)} . \tag{4.3.16}$$

The above test is a two-tailed test, since the null hypothesis will be rejected for both large and small values of the contrast. One-tailed tests can be derived also, as follows.

The decision rule for the test of $H_0 : \Sigma c_i \tau_i = 0$ against the one-sided alternative hypothesis $H_A : \sum c_i \tau_i > 0$ is

$$\text{reject } H_0 \text{ if } \frac{\sum c_i \overline{y}_{i.}}{\sqrt{msE \sum c_i^2 / r_i}} > t_{n-v,\alpha} . \tag{4.3.17}$$

The outcome of this test can be deduced from the appropriate one-sided confidence bound. In particular, the null hypothesis will be rejected at significance level α if the corresponding $100(1-\alpha)\%$ lower confidence bound for $\sum c_i \tau_i$ in Eq. (4.3.11) is above zero so excludes zero.

Similarly, for the one-sided alternative hypothesis $H_A : \sum c_i \tau_i < 0$, the decision rule is

$$\text{reject } H_0 \text{ if } \frac{\sum c_i \overline{y}_{i.}}{\sqrt{msE \sum c_i^2 / r_i}} < -t_{n-v,\alpha} . \tag{4.3.18}$$

Here the null hypothesis will be rejected at significance level α if the corresponding $100(1-\alpha)\%$ upper confidence bound for $\sum c_i \tau_i$ in Eq. (4.3.10) is below zero so excludes zero.

If the hypothesis test concerns a single treatment mean, for example, $H_0 : \mu + \tau_s = 0$, then the decision rules (4.3.13)–(4.3.18) are modified by setting c_s equal to one and all the other c_i equal to zero.

Example 4.3.3 `Filter experiment`

Lorenz et al. (1982) describe an experiment that was carried out to determine the relative performance of seven membrane filters in supporting the growth of bacterial colonies. The seven filter types are regarded as the seven levels of the treatment factor and are coded $1, 2, \ldots, 7$. Filter types 1, 4, and 7 were received presterilized. Several different types of data were collected, but the only data considered here are the colony counts of fecal coliforms from a sample of Olentangy River water (August 1980) that grew on each filter. Three filters of each type were observed and the average colony counts[1] were

$$\overline{y}_{1.} = 36.0,\ \overline{y}_{2.} = 18.0,\ \overline{y}_{3.} = 27.7,\ \overline{y}_{4.} = 28.0,\ \overline{y}_{5.} = 28.3,\ \overline{y}_{6.} = 37.7,\ \overline{y}_{7.} = 30.3 .$$

The mean squared error was $msE = 21.6$. Suppose we wish to test the hypothesis that the presterilized filters do not differ from the nonpresterilized filters in terms of the average colony counts, against a two-sided alternative hypothesis that they do differ. The hypothesis of interest involves a difference of averages contrast, that is,

$$H_0 : \frac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \frac{1}{4}(\tau_2 + \tau_3 + \tau_5 + \tau_6) = 0 .$$

[1]Reprinted from Journal AWWA, Vol. 74, No. 8 (August 1982), by permission. Copyright © 1982, American Water Works Association.

From (4.3.15), the decision rule is to reject H_0 if

$$\frac{ssc}{msE} = \frac{\left[\frac{1}{3}(\overline{y}_{1.} + \overline{y}_{4.} + \overline{y}_{7.}) - \frac{1}{4}(\overline{y}_{2.} + \overline{y}_{3.} + \overline{y}_{5.} + \overline{y}_{6.})\right]^2}{msE\left[\frac{(\frac{1}{3})^2}{3} + \frac{(\frac{1}{3})^2}{3} + \frac{(\frac{1}{3})^2}{3} + \frac{(-\frac{1}{4})^2}{3} + \frac{(-\frac{1}{4})^2}{3} + \frac{(-\frac{1}{4})^2}{3} + \frac{(-\frac{1}{4})^2}{3}\right]} > F_{1,14,\alpha}\,.$$

Selecting a probability of a Type I error equal to $\alpha = 0.05$, this becomes

$$\text{reject } H_0 \text{ if } \frac{(3.508)^2}{(21.6)(0.1944)} = 2.931 > F_{1,14,0.05}\,.$$

Since $F_{1,14,0.05} = 4.6$, there is not sufficient evidence to reject the null hypothesis, and we conclude that the presterilized filters do not differ significantly from the nonpresterilized filters when α is set at 0.05.

Notice that the null hypothesis would be rejected if the probability of a Type I error is set a little higher than $\alpha = 0.10$, since $F_{1,14,0.10} = 3.10$. Thus, if these experimenters are willing to accept a high risk of incorrectly rejecting the null hypothesis, they would be able to conclude that there is a difference between the presterilized and the nonpresterilized filters.

A 95% confidence interval for this difference can be obtained from (4.3.6) as follows:

$$\frac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \frac{1}{4}(\tau_2 + \tau_3 + \tau_5 + \tau_6) \in \left(3.508 \pm t_{14,0.025}\sqrt{(21.6)(0.1944)}\right)\,,$$

and since $t_{14,0.025} = 2.145$, the interval becomes

$$(3.508 \pm (2.145)(2.0492)) = (-0.888, 7.904)\,,$$

where the measurements are average colony counts. The interval contains zero, which agrees with the hypothesis test at $\alpha = 0.05$. □

4.3.4 Equivalence of Tests and Confidence Intervals (Optional)

There is a stronger relationship between hypothesis tests and confidence intervals (including both 1- and 2-sided confidence intervals) than was described in Sect. 4.3.3. As already discussed, the outcome of a hypothesis test at significance level α can be deduced from the corresponding $100(1-\alpha)\%$ confidence interval. Correspondingly, though less well known, one can conclude from a hypothesis test that the true value of the parameter is in the corresponding confidence interval, by virtue of rejecting all values outside the interval, providing more specific test conclusions than simply whether or not one rejects the null hypothesis and so believes the alternative.

To illustrate this, consider a two-tailed level-α test of the null hypothesis $H_0 : \sum c_i\tau_i = 0$ against the alternative hypothesis $H_A : \sum c_i\tau_i \neq 0$. Under standard practice, only the null hypothesis H_0 is tested at significance level α. If H_0 is rejected in favor of H_A, one simply eliminates zero as a possible value of the treatment contrast, and the hypothesis testing procedure guarantees that the probability of making a mistake by rejecting $H_0 : \sum c_i\tau_i = 0$ when it is true is at most α.

Expanding upon standard practice, suppose one not only tests H_0; rather, suppose one conducts a standard two-tailed level-α test of the null hypothesis $H_{0b} : \sum c_i\tau_i = b$ against the alternative

hypothesis $H_{Ab} : \sum c_i \tau_i \neq b$ for *each* real number b. Then the probably of rejecting $H_0 : \sum c_i \tau_i = 0$ if it is true is still controlled to be α. Moreover, even though this expanded testing procedure involves conducting an infinite number of tests rather than only one, the probability of making *any* false rejections—namely, of falsely rejecting any true null hypothesis H_{0b}—is still at most α. This follows from the *partitioning principle*, (see Finner and Strassburger 2002, and references therein). In particular, because the sets $\{b\}$ partition the set of real numbers, H_{0b} is only true for exactly one value of b, b^* say. So, one can only make a mistake by rejecting the only true null hypothesis H_{0b^*}, and the probability of rejecting H_{0b^*} is α. Thus, in terms of error rates, there is no additional cost in testing infinitely many hypothesis H_{0b} instead of just one.

Furthermore, as we know, the null hypothesis H_{0b} will be rejected at level-α precisely for those values b outside the $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$. In other words, all values of $\sum c_i \tau_i$ outside of the $100(1-\alpha)\%$ confidence interval are rejected at simultaneous significance level α. Hence, one can conclude from this extended test that the true value of $\sum c_i \tau_i$ is in the corresponding $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$. This is true whether or not one rejects H_0, providing a more specific conclusion than simply rejecting the null hypothesis or not.

For example, if one does reject $H_0 : \sum c_i \tau_i = 0$ at significance level α, then one can conclude with Type I error probability, α not only that $\sum c_i \tau_i \neq 0$ but also more specifically that the true value of $\sum c_i \tau_i$ is in the corresponding $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$, where this confidence interval will consist only of positive values if H_0 is rejected and $\sum c_i \hat{\tau}_i > 0$, or only of negative values if H_0 is rejected and $\sum c_i \hat{\tau}_i < 0$. On the other hand, if one fails to reject H_0, one can still conclude that the true value of $\sum c_i \tau_i$ is in the corresponding $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$, but this confidence interval will include zero as a possible value of the treatment contrast.

The analogous equivalence exists between one-sided tests and corresponding confidence bounds. Consider for example the standard level-α test of $H_0 : \Sigma c_i \tau_i = 0$ against the one-sided alternative hypothesis $H_A : \sum c_i \tau_i > 0$. More broadly, one can conduct a standard one-tailed α-level test of the null hypothesis $H_{0b} : \sum c_i \tau_i = b$ against the alternative hypothesis $H_{Ab} : \sum c_i \tau_i > b$ for each real number b. In so doing, $H_{0b} : \sum c_i \tau_i = b$ will be rejected for exactly those values of b that are below the $100(1-\alpha)\%$ lower confidence bound for $\sum c_i \tau_i$ given in Eq. (4.3.11). In other words, the values of $\sum c_i \tau_i$ rejected at level α are exactly the values outside of the $100(1-\alpha)\%$ (one-sided) confidence interval. Consequently, whether or not H_0 is rejected, one can conclude that the true value of $\sum c_i \tau_i$ is above the $100(1-\alpha)\%$ lower confidence bound for $\sum c_i \tau_i$, and one will reject H_0 and conclude $\sum c_i \tau_i > 0$ exactly when the lower confidence bound is positive.

Similarly, for testing $H_0 : \Sigma c_i \tau_i = 0$ against $H_A : \sum c_i \tau_i < 0$ at level α, one can expand this by conducting a standard one-tailed level-α test of $H_{0b} : \sum c_i \tau_i = b$ against $H_{Ab} : \sum c_i \tau_i < b$ for each real number b. Then the values of $\sum c_i \tau_i$ rejected at level α are exactly the values above the $100(1-\alpha)\%$ upper confidence bound given in Eq. (4.3.10). Consequently, whether or not H_0 is rejected, one can conclude that the true value of $\sum c_i \tau_i$ is below its $100(1-\alpha)\%$ upper confidence bound. Also, one will reject H_0 and conclude $\sum c_i \tau_i < 0$ exactly when the upper confidence bound is negative.

The equivalence between testing and confidence intervals illustrated above applies quite broadly, including for example to one-step multiple comparison procedures such as those considered in the next section (as discussed by Voss 2008, 2010). The partitioning principle also facilitates the construction of more complicated confidence sets corresponding to stepwise multiple tests (see Stefansson et al. 1988).

4.4 Methods of Multiple Comparisons

4.4.1 Multiple Confidence Intervals

Often, the most useful analysis of experimental data involves the calculation of a number of different confidence intervals, one for each of several contrasts or treatment means. The confidence level for a single confidence interval is based on the probability, like (4.3.9), that the random interval will be "correct" (meaning that the random interval will contain the true value of the contrast or function).

It is shown below that when several confidence intervals are calculated, the probability that they are all simultaneously correct can be alarmingly small. Similarly, when several hypotheses are to be tested, the probability that at least one hypothesis is incorrectly rejected can be uncomfortably high. Much research has been done over the years to find ways around these problems. The resulting techniques are known as *methods of multiple comparison*, the intervals are called *simultaneous confidence intervals*, and the tests are called *simultaneous hypothesis tests*.

Suppose an experimenter wishes to calculate m confidence intervals, each having a $100(1-\alpha^*)\%$ confidence level. Then each interval will be individually correct with probability $1-\alpha^*$. Let S_j be the event that the jth confidence interval will be correct and $\overline{S}_j$ the event that it will be incorrect ($j = 1, \ldots, m$). Then, using the standard rules for probabilities of unions and intersections of events, it follows that

$$P(S_1 \cap S_2 \cap \cdots \cap S_m) = 1 - P(\overline{S}_1 \cup \overline{S}_2 \cup \cdots \cup \overline{S}_m)\,.$$

This says that the probability that all of the intervals will be correct is equal to one minus the probability that at least one will be incorrect. If $m = 2$,

$$\begin{aligned} P(\overline{S}_1 \cup \overline{S}_2) &= P(\overline{S}_1) + P(\overline{S}_2) - P(\overline{S}_1 \cap \overline{S}_2) \\ &\le P(\overline{S}_1) + P(\overline{S}_2)\,. \end{aligned}$$

A similar result, which can be proved by mathematical induction, holds for any number m of events, that is,

$$P(\overline{S}_1 \cup \overline{S}_2 \cup \cdots \cup \overline{S}_m) \le \sum_j P(\overline{S}_j)\,,$$

with equality if the events $\overline{S}_1, \overline{S}_2, \ldots, \overline{S}_m$ are mutually exclusive. Consequently,

$$P(S_1 \cap S_2 \cap \cdots \cap S_m) \ge 1 - \sum_j P(\overline{S}_j) = 1 - m\alpha^*\,; \tag{4.4.19}$$

that is, the probability that the m intervals will simultaneously be correct is at least $1 - m\alpha^*$. The probability $m\alpha^*$is called the *overall significance level* or *experimentwise error rate*. A typical value for α^* for a single confidence interval is 0.05, so the probability that six confidence intervals each calculated at a 95% individual confidence level will simultaneously be correct is at least 0.7. Although "at least" means "bigger than or equal to," it is not known in practice how much bigger than 0.7 the probability might actually be. This is because the degree of overlap between the events $\overline{S}_1, \overline{S}_2, \ldots, \overline{S}_m$ is generally unknown. The probability "at least 0.7" translates into an *overall confidence level* of "at least 70%" when the responses are observed. Similarly, if an experimenter calculates ten confidence intervals each having individual confidence level 95%, then the simultaneous confidence level for the

ten intervals is at least 50%, which is not very informative. As m becomes larger the problem becomes worse, and when $m \geq 20$, the overall confidence level is at least 0%, clearly a useless assertion!

Similar comments apply to the hypothesis testing situation. If hypotheses for m different contrasts are to be tested, each at significance level α^*, then the probability that at least one hypothesis is incorrectly rejected is at most $m\alpha^*$.

Various methods have been developed to ensure that the overall confidence level is not too small and the overall significance level is not too high. Some methods are completely general, that is, they can be used for any set of estimable functions, while others have been developed for very specialized purposes such as comparing each treatment with a control. Which method is best depends on which contrasts are of interest and the number of contrasts to be investigated. In this section, four methods are discussed that control the overall confidence level and overall significance level. The terms *preplanned contrasts* and *data snooping* occur in the summary of methods and the subsequent subsections. These have the following meanings. Before the experiment commences, the experimenter will have written out a checklist, highlighted the contrasts and/or treatment means that are of special interest, and designed the experiment in such a way as to ensure that these are estimable with as small variances as possible. These are the preplanned contrasts and means. After the data have been collected, the experimenter usually looks carefully at the data to see whether anything unexpected has occurred. One or more unplanned contrasts may turn out to be the most interesting, and the conclusions of the experiment may not be as anticipated. Allowing the data to suggest additional interesting contrasts is called data snooping.

The following summary is written in terms of confidence intervals, but it also applies to hypothesis tests. A shorter confidence interval corresponds to a more powerful hypothesis test. The block designs mentioned in the summary will be discussed in Chaps. 10 and 11.

Summary of Multiple Comparison Methods

1. **Bonferroni method for preplanned comparisons**
   Applies to any m preplanned estimable contrasts or functions of the parameters. Gives shorter confidence intervals than the other methods listed if m is small. Can be used for any design. Cannot be used for data snooping.

2. **Scheffé method for all comparisons**
   Applies to any m estimable contrasts or functions of the parameters. Gives shorter intervals than Bonferroni's method if m is large. Allows data snooping. Can be used for any design.

3. **Tukey method for all pairwise comparisons**
   Best for all pairwise comparisons. Can be used for completely randomized designs, randomized block designs, and balanced incomplete block designs. Is believed to be applicable (conservative) for other designs as well. Can be extended to include all contrasts, but Scheffé's method is generally better for these.

4. **Dunnett method for treatment-versus-control comparisons**
   Best for all treatment-versus-control contrasts. Can be used for completely randomized designs, randomized block designs, and balanced incomplete block designs.

Details of confidence intervals obtained by each of the above methods are given in Sects. 4.4.2–4.4.6. The terminology "a set of simultaneous $100(1-\alpha)\%$ confidence intervals" will always refer to the fact that the *overall* confidence level for a set of contrasts or treatment means is (at least) $100(1-\alpha)\%$. Each of the four methods discussed gives confidence intervals of the form

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \hat{\tau}_i \pm w \sqrt{\widehat{\text{Var}}(\Sigma c_i \hat{\tau}_i)} \right), \tag{4.4.20}$$

where w, which we call the *critical coefficient*, depends on the method, on v, on the number of confidence intervals calculated, and on the number of error degrees of freedom. The term

$$msd = w \sqrt{\widehat{\text{Var}}(\Sigma c_i \hat{\tau}_i)}\,,$$

which is added and subtracted from the least squares estimate in (4.4.20), is called the *minimum significant difference*, because if the estimate is larger than *msd*, the confidence interval excludes zero, and the contrast is significantly different from zero.

4.4.2 Bonferroni Method for Preplanned Comparisons

The inequality (4.4.19) shows that if m simultaneous confidence intervals are calculated for preplanned contrasts, and if each confidence interval has confidence level $100(1-\alpha^*)\%$, then the overall confidence level is greater than or equal to $100(1-m\alpha^*)\%$. Thus, an experimenter can ensure that the overall confidence level is at least $100(1-\alpha)\%$ by setting $\alpha^* = \alpha/m$. This is known as the Bonferroni method for simultaneous confidence intervals. Replacing α by α/m in the formula (4.3.6), p. 75, for an individual confidence interval, we obtain a formula for a set of simultaneous $100(1-\alpha)\%$ confidence intervals for m preplanned contrasts $\Sigma c_i \tau_i$ in a completely randomized design with the one-way analysis of variance model (3.3.1), as

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \overline{y}_{i\cdot} \pm t_{n-v,\alpha/(2m)} \sqrt{msE \sum_i c_i^2 / r_i} \right), \tag{4.4.21}$$

where the critical coefficient, w_B, is

$$w_B = t_{n-v,\alpha/(2m)}\,.$$

Since $\alpha/(2m)$ is likely to be an atypical value, the percentiles $t_{n-v,\alpha/(2m)}$ may need to be obtained by use of a computer package, or by approximate interpolation between values in Table A.4, or by using the following approximate formula due to Peiser (1943):

$$t_{df,\alpha/(2m)} \approx z_{\alpha/(2m)} + (z^3_{\alpha/(2m)} + z_{\alpha/(2m)})/(4(df))\,, \tag{4.4.22}$$

where *df* is the error degrees of freedom (equal to $n-v$ in the present context), and where $z_{\alpha/(2m)}$ is the percentile of the standard normal distribution corresponding to a probability of $\alpha/(2m)$ in the right hand tail. The standard normal distribution is tabulated in Table A.3 and covers the entire range of values for $\alpha/(2m)$. When m is very large, $\alpha/(2m)$ is very small, possibly resulting in extremely wide simultaneous confidence intervals. In this case, the Scheffé or Tukey methods described in the following subsections would be preferred.

If some of the m simultaneous intervals are for true mean responses $\mu + \tau_s$, then the required intervals are of the form (4.3.12), p. 76, with α replaced by α/m, that is,

$$\mu + \tau_s \in \left(\overline{y}_{s\cdot} \pm t_{n-v,\alpha/(2m)} \sqrt{msE/r_s} \right). \tag{4.4.23}$$

Similarly, replacing α by α/m in (4.3.15), a set of m null hypotheses, each of the form

$$H_0 : \sum_{i=1}^{v} c_i \tau_i = 0 ,$$

can be tested against their respective two-sided alternative hypotheses at overall significance level α using the set of decision rules each of the form

$$\text{reject } H_0 \text{ if } \frac{ssc}{msE} > F_{1,df,\alpha/m} . \tag{4.4.24}$$

Each null hypothesis is rejected if the corresponding confidence interval (4.4.21) excludes zero, and each confidence interval consists of exactly those values that would not be rejected by a two-tailed test.

Note that Bonferroni's method can be use only for *preplanned* contrasts and means. An experimenter who looks at the data and then proceeds to calculate simultaneous confidence intervals for the few contrasts that look interesting has effectively calculated a very large number of intervals. This is because the interesting contrasts are usually those that seem to be significantly different from zero, and a rough mental calculation of the estimates of a large number of contrasts has to be done to identify these interesting contrasts. Scheffé's method should be used for contrasts that were selected after the data were examined.

Example 4.4.1 `Filter experiment, continued`

The filter experiment was described in Example 4.3.3, p. 78. Suppose that before the data had been collected, the experimenters had planned to calculate a set of simultaneous 90% confidence intervals for the following $m = 3$ contrasts. These contrasts have been selected based on the details of the original study described by Lorenz et al. (1982).

(i) $\frac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \frac{1}{4}(\tau_2 + \tau_3 + \tau_5 + \tau_6)$. This contrast measures the difference in the average effect of the presterilized and the nonpresterilized filter types. This was used in Example 4.3.3 to illustrate a hypothesis test for a single contrast.

(ii) $\frac{1}{2}(\tau_1 + \tau_7) - \frac{1}{5}(\tau_2 + \tau_3 + \tau_4 + \tau_5 + \tau_6)$. This contrast measures the difference in the average effects of two filter types with gradated pore size and five filter types with uniform pore size.

(iii) $\frac{1}{6}(\tau_1 + \tau_2 + \tau_4 + \tau_5 + \tau_6 + \tau_7) - \tau_3$. This contrast is the difference in the average effect of the filter types that are recommended by their manufacturers for bacteriologic analysis of water and the single filter type that is recommended for sterility testing of pharmaceutical or cosmetic products.

From Example 4.3.3, we know that

$$\bar{y}_{1.} = 36.0, \quad \bar{y}_{2.} = 18.0, \quad \bar{y}_{3.} = 27.7, \quad \bar{y}_{4.} = 28.0, \quad \bar{y}_{5.} = 28.3,$$
$$\bar{y}_{6.} = 37.7, \quad \bar{y}_{7.} = 30.3, \quad r_i = 3, \quad msE = 21.6.$$

The formula for each of the three preplanned simultaneous 90% confidence intervals is given by (4.4.21) and involves the critical coefficient $w_B = t_{14,(0.1)/6} = t_{14,0.0167}$, which is not available in Table A.4. Either the value can be calculated from a computer program, or an approximate value can be obtained from formula (4.4.22) as

$$t_{14,0.0167} \approx 2.128 + (2.128^3 + 2.128)/(4 \times 14) = 2.338.$$

The minimum significant difference for each of the three simultaneous 90% confidence intervals is

$$msd = 2.338\sqrt{(21.6)\sum c_i^2/3} = 6.2735\sqrt{\sum c_i^2}\,.$$

Thus, for the first interval, we have

$$msd = 6.2735\sqrt{3\left(\frac{1}{9}\right)+4\left(\frac{1}{16}\right)} = 4.791\,,$$

giving the interval as

$$\frac{1}{3}(\tau_1+\tau_4+\tau_7)-\frac{1}{4}(\tau_2+\tau_3+\tau_5+\tau_6) \in (3.508 \pm 4.791) = (-1.283,\ 8.299)\,.$$

Calculating the minimum significant differences separately for the other two confidence intervals leads to

$$\frac{1}{2}(\tau_1+\tau_7)-\frac{1}{5}(\tau_2+\tau_3+\tau_4+\tau_5+\tau_6) \in (-0.039,\ 10.459)\,;$$

$$\frac{1}{6}(\tau_1+\tau_2+\tau_4+\tau_5+\tau_6+\tau_7)-\tau_3 \in (-4.759,\ 8.793)\,.$$

Notice that all three intervals include zero, although the second is close to excluding it. Thus, at overall significance level $\alpha = 0.10$, we would fail to reject the hypothesis that there is no difference in average colony counts between the presterilized and nonpresterilized filters, nor between filter 3 and the others, nor between filters with gradated and uniform pore sizes. At a slightly higher significance level, we would reject the hypothesis that the filters with gradated pore size have the same average colony counts as those with uniform pore size. The same conclusion would be obtained if (4.4.24) were used to test simultaneously, at overall level $\alpha = 0.10$, the hypotheses that each of the three contrasts is zero. The confidence interval, whether utilized directly or obtained as the conclusion of the test, has the added benefit that it provides more specific conclusions. For example, we can say with overall 90% confidence that on average, the filters with gradated pore size give rise to colony counts up to 10.4 greater than the filters with uniform pore sizes. □

4.4.3 Scheffé Method of Multiple Comparisons

The main drawbacks of the Bonferroni method of multiple comparisons are that the m contrasts to be examined must be preplanned and the confidence intervals can become very wide if m is large. Scheffé's method, on the other hand, provides a set of simultaneous $100(1-\alpha)\%$ confidence intervals whose widths are determined only by the number of treatments and the number of observations in the experiment, no matter how many contrasts are of interest. The two methods are compared directly later in this section.

Scheffé's method is based on the fact that every possible contrast $\Sigma c_i \tau_i$ can be written as a linear combination of the set of $(v-1)$ treatment versus control contrasts, $\tau_2-\tau_1$, $\tau_3-\tau_1$, …, $\tau_v-\tau_1$. (We leave it to the reader to check that this is true.) Once the experimental data have been collected, it is possible to find a $100(1-\alpha)\%$ *confidence region* for these $v-1$ treatment-versus-control contrasts.

The confidence region not only determines confidence bounds for each treatment-versus-control contrast, it determines bounds for *every* possible contrast $\Sigma c_i \tau_i$ and, in fact, for *any number* of contrasts, while the overall confidence level remains fixed. The mathematical details are given by Scheffé (1959).

For v treatments in a completely randomized design and the one-way analysis of variance model (3.3.1), a set of simultaneous $100(1-\alpha)\%$ confidence intervals for all contrasts $\Sigma c_i \tau_i$ is given by

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \overline{y}_{i\cdot} \pm \sqrt{(v-1) F_{v-1,n-v,\alpha}} \sqrt{msE \sum_i c_i^2 / r_i} \right) . \tag{4.4.25}$$

Notice that this is the same form as the general formula (4.4.20), p. 83, where the critical coefficient w is

$$w_S = \sqrt{(v-1) F_{v-1,n-v,\alpha}} \ .$$

If confidence intervals for the treatment means $\mu + \tau_i$ are also of interest, the critical coefficient w_S needs to be replaced by

$$w_S^* = \sqrt{v F_{v,n-v,\alpha}} \, .$$

The reason for the increase in the numerator degrees of freedom is that any of the functions $\mu + \tau_i$ can be written as a linear combination of the $v-1$ treatment versus control contrasts and one additional function $\mu + \tau_1$. For the completely randomized design and model (3.3.1), a set of simultaneous $100(1-\alpha)\%$ confidence intervals for any number of true mean responses and contrasts is therefore given by

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \overline{y}_{i\cdot} \pm \sqrt{v F_{v,n-v,\alpha}} \sqrt{msE \sum_i c_i^2 / r_i} \right)$$

together with

$$\mu + \tau_s \in \left(\overline{y}_{i\cdot} \pm \sqrt{v F_{v,n-v,\alpha}} \sqrt{msE / r_s} \right) . \tag{4.4.26}$$

Example 4.4.2 `Filter experiment, continued`

If we look at the observed average colony counts,

$$\begin{aligned} &\overline{y}_{1\cdot} = 36.0, \quad \overline{y}_{2\cdot} = 18.0, \quad \overline{y}_{3\cdot} = 27.7, \quad \overline{y}_{4\cdot} = 28.0, \\ &\overline{y}_{5\cdot} = 28.3, \quad \overline{y}_{6\cdot} = 37.7, \quad \overline{y}_{7\cdot} = 30.3, \end{aligned}$$

for the filter experiment of Examples 4.3.3 and 4.4.1 (p. 78 and 84), filter type 2 appears to give a much lower count than the other types. One may wish to recalculate each of the three intervals in Example 4.4.1 with filter type 2 excluded. It might also be of interest to compare the filter types 1 and 6, which showed the highest average colony counts, with the other filters. These are *not preplanned* contrasts. They have become interesting only after the data have been examined, and therefore we need to use Scheffé's method of multiple comparisons. In summary, we are interested in the following twelve contrasts:

$$\begin{gathered} \tfrac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \tfrac{1}{3}(\tau_3 + \tau_5 + \tau_6) \, , \qquad \tfrac{1}{2}(\tau_1 + \tau_7) - \tfrac{1}{4}(\tau_3 + \tau_4 + \tau_5 + \tau_6) \, , \\ \tfrac{1}{5}(\tau_1 + \tau_4 + \tau_5 + \tau_6 + \tau_7) - \tau_3 \, , \\ \tau_1 - \tau_3 \, , \qquad \tau_1 - \tau_4 \, , \qquad \tau_1 - \tau_5 \, , \qquad \tau_1 - \tau_6 \, , \qquad \tau_1 - \tau_7 \, , \\ \tau_6 - \tau_3 \, , \qquad \tau_6 - \tau_4 \, , \qquad \tau_6 - \tau_5 \, , \qquad \tau_6 - \tau_7 \, . \end{gathered}$$

The formula for a set of Scheffé 90% simultaneous confidence intervals is given by (4.4.25) with $\alpha = 0.10$. Since $v = 7$, $n = 21$, and $msE = 21.6$ for the filter experiment, the minimum significant difference for each interval becomes

$$msd = \sqrt{6F_{6,14,0.10}}\sqrt{21.6\ \Sigma c_i^2/3} = 9.837\sqrt{\Sigma c_i^2}\,.$$

The twelve simultaneous 90% confidence intervals are then

$$\begin{aligned}\frac{1}{3}(\tau_1+\tau_4+\tau_7) - \frac{1}{3}(\tau_3+\tau_5+\tau_6) &\\ \in \left((31.43-31.23) \pm 9.837\sqrt{3\left(\frac{1}{9}\right)+3\left(\frac{1}{9}\right)}\right) &\\ = (-7.83, 8.23)\,, &\end{aligned}$$

$$\frac{1}{2}(\tau_1+\tau_7) - \frac{1}{4}(\tau_3+\tau_4+\tau_5+\tau_6) \in (-5.79, 11.24)\,,$$

$$\frac{1}{5}(\tau_1+\tau_4+\tau_5+\tau_6+\tau_7) - \tau_3 \in (-6.42, 15.14)\,,$$

$$\begin{aligned}\tau_1-\tau_3 &\in (-5.61, 22.21)\,, & \tau_6-\tau_3 &\in (-3.91, 23.91)\,,\\ \tau_1-\tau_4 &\in (-5.91, 21.91)\,, & \tau_6-\tau_4 &\in (-4.21, 23.61)\,,\\ \tau_1-\tau_5 &\in (-6.21, 21.61)\,, & \tau_6-\tau_5 &\in (-4.51, 23.31)\,,\\ \tau_1-\tau_6 &\in (-15.61, 12.21)\,, & \tau_6-\tau_7 &\in (-6.51, 21.31)\,,\\ \tau_1-\tau_7 &\in (-8.21, 19.61)\,. & &\end{aligned}$$

These intervals are all fairly wide and all include zero. Consequently, at overall error rate $\alpha = 0.1$, we are unable to infer that any of the contrasts are significantly different from zero. □

Relationship Between Analysis of Variance and the Scheffé Method

The analysis of variance test and the Scheffé method of multiple comparisons are equivalent in the following sense. The analysis of variance test will reject the null hypothesis $H_0 : \tau_1 = \tau_2 = \cdots = \tau_v$ at significance level α if there is at least one confidence interval among the infinite number of Scheffé simultaneous $100(1-\alpha)\%$ confidence intervals for all contrasts $\Sigma c_i \tau_i$ that excludes zero. However, the intervals that exclude zero may not be among those for the interesting contrasts being examined.

Other methods of multiple comparisons do not relate to the analysis of variance test in this way. It is possible when using one of the other multiple comparison methods that one or more intervals in a simultaneous $100(1-\alpha)\%$ set may exclude 0, while the analysis of variance test of H_0 is *not* rejected at significance level α. Hence, if specific contrasts of interest have been identified in advance of running the experiment and a method of multiple comparisons other than Scheffé's method is to be used, then it is sensible to analyze the data using only the multiple comparison procedure.

4.4.4 Tukey Method for All Pairwise Comparisons

In some experiments, confidence intervals may be required only for pairwise difference contrasts. Tukey, in 1953, proposed a method that is specially tailored to handle this situation and that gives shorter intervals for pairwise differences than do the Bonferroni and Scheffé methods.

For the completely randomized design and the one-way analysis of variance model (3.3.1), Tukey's simultaneous confidence intervals for all pairwise comparisons $\tau_i - \tau_s$, $i \neq s$, with overall confidence level at least $100(1-\alpha)\%$ is given by

$$\tau_i - \tau_s \in \left((\overline{y}_{i.} - \overline{y}_{s.}) \pm w_T \sqrt{msE\left(\frac{1}{r_i} + \frac{1}{r_s}\right)} \right), \tag{4.4.27}$$

where the critical coefficient w_T is

$$w_T = q_{v,n-v,\alpha}/\sqrt{2}\,,$$

and where $q_{v,n-v,\alpha}$ is tabulated in Appendix A.8. When the sample sizes are equal ($r_i = r$; $i = 1, \ldots, v$), the overall confidence level is *exactly* $100(1-\alpha)\%$. When the sample sizes are unequal, the confidence level is *at least* $100(1-\alpha)\%$.

The derivation of (4.4.27) is as follows. For equal sample sizes, the formula for Tukey's simultaneous confidence intervals is based on the distribution of the statistic

$$Q = \frac{\max\{T_i\} - \min\{T_i\}}{\sqrt{MSE/r}}\,,$$

where $T_i = \overline{Y}_{i.} - (\mu + \tau_i)$ for the one-way analysis of variance model (3.3.1), and where $\max\{T_i\}$ is the maximum value of the random variables $T_1, T_2, \ldots, T_v$ and $\min\{T_i\}$ the minimum value. Since the $\overline{Y}_{i.}$'s are independent, the numerator of Q is the range of v independent $N(0, \sigma^2/r)$ random variables, and is standardized by the estimated standard deviation. The distribution of Q is called the *Studentized range distribution*. The percentile corresponding to a probability of α in the right-hand tail of this distribution is denoted by $q_{v,n-v,\alpha}$, where v is the number of treatments being compared, and $n - v$ is the number of degrees of freedom for error. Therefore,

$$P\left(\frac{\max\{T_i\} - \min\{T_i\}}{\sqrt{MSE/r}} \leq q_{v,n-v,\alpha}\right) = 1 - \alpha\,.$$

Now, if $\max\{T_i\} - \min\{T_i\}$ is less than or equal to $q_{v,n-v,\alpha}\sqrt{MSE/r}$, then it must be true that $|T_i - T_s| \leq q_{v,n-v,\alpha}\sqrt{MSE/r}$ for *every* pair of random variables T_i, T_s, $i \neq s$. Using this fact and the above definition of T_i, we have

$$\begin{aligned} 1 - \alpha = P\,\big(-q_{v,n-v,\alpha}\sqrt{MSE/r} &\leq (\overline{Y}_{i.} - \overline{Y}_{s.}) - (\tau_i - \tau_s) \\ &\leq q_{v,n-v,\alpha}\sqrt{MSE/r}\,, \text{ for all } i \neq s\big)\,. \end{aligned}$$

Replacing $\overline{Y}_{i.}$ by its observed value $\overline{y}_{i.}$, and MSE by the observed value msE, a set of simultaneous $100(1-\alpha)\%$ confidence intervals for all pairwise differences $\tau_i - \tau_s$, $i \neq s$, is given by

$$\tau_i - \tau_s \in \left((\overline{y}_{i.} - \overline{y}_{s.}) \pm q_{v,n-v,\alpha}\sqrt{msE/r} \right),$$

which can be written in terms of the critical coefficient as

$$\tau_i - \tau_s \in \left((\overline{y}_{i.} - \overline{y}_{s.}) \pm w_T \sqrt{msE\left(\frac{1}{r} + \frac{1}{r}\right)} \right). \tag{4.4.28}$$

More recently, Hayter (1984) showed that the same form of interval can be used for unequal sample sizes as in (4.4.27), and that the overall confidence level is then at least $100(1-\alpha)\%$.

Example 4.4.3 `Battery experiment, continued`

In the battery experiment of Example 4.2.1 (p. 71), we considered the pairwise differences in the life lengths per unit cost of $v = 4$ different battery types, and we obtained the least squares estimates

$$\begin{aligned} \hat{\tau}_1 - \hat{\tau}_2 &= -289.75\,, \quad \hat{\tau}_1 - \hat{\tau}_3 = 137.75\,, \quad \hat{\tau}_1 - \hat{\tau}_4 = 74.50\,, \\ \hat{\tau}_2 - \hat{\tau}_3 &= 427.50\,, \quad \hat{\tau}_2 - \hat{\tau}_4 = 364.25\,, \quad \hat{\tau}_3 - \hat{\tau}_4 = -63.25\,. \end{aligned}$$

The standard error was $\sqrt{msE(\frac{1}{4}+\frac{1}{4})} = 34.41$, and the number of error degrees of freedom was $n - v = (16-4) = 12$. From Table A.8, $q_{4,12,0.05} = 4.20$, so $w_T = 4.20/\sqrt{2}$, and the minimum significant difference is

$$msd = (4.20/\sqrt{2})\,(34.41) = 102.19\,.$$

Therefore, using Tukey's method, the simultaneous 95% confidence intervals for the pairwise comparisons of lifetimes per unit cost of the different battery types are

$$\begin{aligned} \tau_1 - \tau_2 &\in (-289.75 \pm 102.19) = (-391.94, -187.56), \\ \tau_1 - \tau_3 &\in (137.75 \pm 102.19) = (35.56, 239.94), \\ \tau_1 - \tau_4 &\in (-27.69, 176.69), \qquad \tau_2 - \tau_3 \in (325.31, 529.69), \\ \tau_2 - \tau_4 &\in (262.06, 466.44), \qquad \tau_3 - \tau_4 \in (-165.44, 38.94). \end{aligned}$$

Four of these intervals exclude zero, and one can conclude (at an overall 95% confidence level) that battery type 2 (alkaline, store brand) has the highest lifetime per unit cost, and battery type 3 (heavy duty, name brand) has lower lifetime per unit cost than does battery type 1 (alkaline, name brand). The intervals show us that with overall 95% confidence, battery type 2 is between 188 and 391 minute per dollar better than battery type 1 (the name brand alkaline battery) and even more economical than the heavy-duty brands. □

Example 4.4.4 `Bonferroni, Scheffé and Tukey methods compared`

Suppose that $v = 5$, $n = 35$, and $\alpha = 0.05$, and that only the 10 pairwise comparisons $\tau_i - \tau_s$, $i \neq s$, are of interest to the experimenter and these were specifically selected prior to the experiment (i.e., were preplanned). If we compare the critical coefficients for the three methods, we obtain

$$\begin{aligned} \text{Bonferroni}: \quad w_B &= t_{30,.025/10} &&= 3.02, \\ \text{Scheffé}: \quad w_S &= \sqrt{4\,F_{4,30,.05}} &&= 3.28, \\ \text{Tukey}: \quad w_T &= \tfrac{1}{\sqrt{2}} q_{5,30,.05} &&= 2.91. \end{aligned}$$

Since w_T is less than w_B, which is less than w_S for this example, the Tukey intervals will be shorter than the Bonferroni intervals, which will be shorter than the Scheffé intervals. □

4.4.5 Dunnett Method for Treatment-Versus-Control Comparisons

In 1955, Dunnett developed a method of multiple comparisons that is specially designed to provide a set of simultaneous confidence intervals for preplanned treatment-versus-control contrasts $\tau_i - \tau_1$ ($i = 2, \ldots, v$), where level 1 corresponds to the control treatment. The intervals are shorter than those given by the Scheffé, Tukey, and Bonferroni methods, but the method should not be used for any other type of contrasts.

The formulae for the simultaneous confidence intervals are based on the joint distribution of the estimators $\overline{Y}_{i.} - \overline{Y}_{1.}$ of $\tau_i - \tau_1$ ($i = 2, \ldots, v$). This distribution is a special case of the multivariate t distribution and depends on the correlation between $\overline{Y}_{i.} - \overline{Y}_{1.}$ and $\overline{Y}_{s.} - \overline{Y}_{1.}$. For the completely randomized design, with equal numbers of observations $r_2 = \cdots = r_v = r$ on the experimental treatments and $r_1 = c$ observations on the control treatment, the correlation is

$$\rho = r/(c + r)\,.$$

In many experiments, the same number of observations will be taken on the control and experimental treatments, in which case $\rho = 0.5$. However, the shortest confidence intervals for comparing $v - 1$ experimental treatments with a control treatment are generally obtained when c/r is chosen to be close to $\sqrt{v-1}$. Since we have tabulated the multivariate t-distribution only with correlation $\rho = 0.5$, we will discuss only the case $c = r$. Other tables can be found in the book of Hochberg and Tamhane (1987), and intervals can also be obtained via some computer packages (see Sects. 4.6.2 and 4.7.2 for the SAS and R software, respectively).

If the purpose of the experiment is to determine which of the experimental treatments give a significantly higher response than the control treatment, then one-sided confidence bounds should be used. For a completely randomized design with equal sample sizes and the one-way analysis of variance model (3.3.1), Dunnett's simultaneous one-sided $100(1-\alpha)\%$ confidence bounds for treatment-versus-control contrasts $\tau_i - \tau_1$ ($i = 2, 3, \ldots, v$) are

$$\tau_i - \tau_1 \geq (\overline{y}_{i.} - \overline{y}_{1.}) - w_{D1}\sqrt{msE\left(\frac{2}{r}\right)}\,, \tag{4.4.29}$$

where the critical coefficient is

$$w_{D1} = t^{(0.5)}_{v-1,n-v,\alpha}$$

and where $t^{(0.5)}_{v-1,n-v,\alpha}$ is the percentile of the maximum of a multivariate t-distribution with common correlation 0.5 and $n - v$ degrees of freedom, corresponding to a Type I error probability of α in the right-hand tail. The critical coefficient is tabulated in Table A.9. If the right hand side of (4.4.29) is positive, we infer that the ith experimental treatment gives a larger response than the control.

If the purpose is to determine which of the experimental treatments give a significantly lower response than the control, then the inequality is reversed, and the confidence bound becomes

$$\tau_i - \tau_1 \leq (\overline{y}_{i.} - \overline{y}_{1.}) + w_{D1}\sqrt{msE\left(\frac{2}{r}\right)}\,. \tag{4.4.30}$$

If the right-hand side is negative, we infer that the ith experimental treatment gives a smaller response than the control.

To determine which experimental treatments are better than the control *and* which ones are worse, two-sided intervals of the general form (4.4.20) are used as for the other multiple comparison methods. For the completely randomized design, one-way analysis of variance model (3.3.1), and equal sample sizes, the formula is

$$\tau_i - \tau_1 \in \left(\overline{y}_{i\cdot} - \overline{y}_{1\cdot} \pm w_{D2} \sqrt{msE\left(\frac{2}{r}\right)} \right), \tag{4.4.31}$$

where the critical coefficient is

$$w_{D2} = |t|^{(0.5)}_{v-1,n-v,\alpha}$$

and is the upper critical value for the maximum of the absolute values of a multivariate t-distribution with correlation 0.5 and $n - v$ error degrees of freedom, corresponding to the chosen value of α in the right-hand tail. The critical coefficients for equal sample sizes are provided in Table A.10.

For future reference, the general formula for Dunnett's two-sided simultaneous $100(1-\alpha)\%$ confidence intervals for treatment versus control contrasts $\tau_i - \tau_1$ $(i = 2, 3, \ldots, v)$ is

$$\tau_i - \tau_1 \in \left((\hat{\tau}_i - \hat{\tau}_1) \pm w_{D2} \sqrt{\widehat{\mathrm{Var}}(\hat{\tau}_i - \hat{\tau}_1)} \right), \tag{4.4.32}$$

and, for one-sided confidence bounds, we replace w_{D2} by w_{D1} and replace "$\in$" by "$\leq$" or "$\geq$." The critical coefficients are

$$w_{D2} = |t|^{(0.5)}_{v-1,df,\alpha} \text{ and } w_{D1} = t^{(0.5)}_{v-1,df,\alpha}$$

for two-sided and one-sided intervals, respectively, where *df* is the number of error degrees of freedom.

Example 4.4.5 `Soap experiment, continued`

Suppose that as a preplanned objective of the soap experiment of Sect. 2.5.1, p. 20, the experimenter had wanted simultaneous 99% confidence intervals comparing the weight losses of the deodorant and moisturizing soaps (levels 2 and 3) with that of the regular soap (level 1). Then it is appropriate to use Dunnett's method as given in (4.4.31). From Sect. 3.7.2, $r_1 = r_2 = r_3 = 4$, $msE = 0.0772$, $\hat{\tau}_2 - \hat{\tau}_1 = 2.7350$, and $\hat{\tau}_3 - \hat{\tau}_1 = 2.0275$. From Table A.10, $w_{D2} = |t|^{(0.5)}_{v-1,n-v,\alpha} = |t|^{(0.5)}_{2,9,0.01} = 3.63$, so the minimum significant difference is

$$msd = 3.63\sqrt{msE(2/4)} = 0.713\,.$$

Hence, the simultaneous 99% confidence intervals are

$$\tau_2 - \tau_1 \in (2.7350 \pm 0.713) \approx (2.022, 3.448)$$

and

$$\tau_3 - \tau_1 \in (2.0275 \pm 0.713) \approx (1.314, 2.741)\,.$$

One can conclude from these intervals (with overall 99% confidence) that the deodorant soap (soap 2) loses between 2 and 3.4 g more weight on average than does the regular soap, and the moisturizing soap loses between 1.3 and 2.7 g more weight on average than the regular soap. We leave it to the reader to verify that neither the Tukey nor the Bonferroni method would have been preferred for these contrasts (see Exercise 7). □

4.4.6 Combination of Methods

The Bonferroni method is based on the fact that if m individual confidence intervals are obtained, each with confidence level $100(1-\alpha^*)\%$, then the overall confidence level is at least $100(1-m\alpha^*)\%$. The same fact can be used to combine the overall confidence levels arising from more than one multiple comparison procedure.

In Example 4.4.1 (p. 84), the Bonferroni method was used to calculate simultaneous 90% confidence intervals for $m = 3$ preplanned contrasts. In Example 4.4.2 (p. 86), the analysis was continued by calculating simultaneous 90% Scheffé intervals for twelve other contrasts. The overall error rate for these two sets of intervals combined is therefore at most $0.1 + 0.1 = 0.2$, giving an overall, or "experimentwise," confidence level of at least $100(1-0.2)\% = 80\%$ for all fifteen intervals together.

Different possible strategies for multiple comparisons should be examined when outlining the analysis at step (g) of the checklist (Sect. 2.2, p. 7). Suppose that in the above example the overall level for all intervals (both planned and otherwise) had been required to be at least 90%. We examine two possible strategies that could have been used. First, the confidence levels for the Bonferroni and Scheffé contrasts could have been adjusted, dividing $\alpha = 0.10$ into two pieces, α_1 for the preplanned contrasts and α_2 for the others, where $\alpha_1 + \alpha_2 = 0.10$. This strategy would have resulted in intervals that were somewhat wider than the above for all of the contrasts. Alternatively, Scheffé's method could have been used with $\alpha = 0.10$ for all of the contrasts including the three preplanned contrasts. This strategy would have resulted in wider intervals for the three preplanned contrasts but not for the others. Both strategies would result in an overall, or experimentwise, confidence level of 90% instead of 80%.

4.4.7 Methods Not Controlling Experimentwise Error Rate

We have introduced four methods of multiple comparisons, each of which allows the experimenter to control the overall confidence level, and the same methods can be used to control the experimentwise error rate when multiple hypotheses are to be tested. There exist other multiple comparison procedures that are more powerful (i.e., that more easily detect a nonzero contrast) but do not control the overall confidence level nor the experimentwise error rate. While some of these are used quite commonly, we do not advocate their use. Such procedures include Duncan's multiple range test, Fisher's protected LSD procedure, and the Newman–Keuls method. (For more details, see Hsu 1996.)

4.5 Sample Sizes

Before an experiment can be run, it is necessary to determine the number of observations that should be taken on each level of each treatment factor (step (h) of the checklist in Sect. 2.2, p. 7). In Sect. 3.6.2, a method was presented to calculate the sample sizes needed to achieve a specified power of the test of the hypothesis $H_0 : \tau_1 = \cdots = \tau_v$. In this section we show how to determine the sample sizes to achieve confidence intervals of specified lengths.

The lengths of confidence intervals decrease as sample sizes increase. Consequently, if the length of an interval is specified, it should be possible to calculate the required sample sizes, especially when these are equal. However, there is a problem. Since the experimental data have not yet been collected, the value of the mean squared error is not known. As in Sect. 3.6.2, if the value of the mean squared error can be reasonably well be guessed at, either from previous experience or from a pilot study, then a trial and error approach to the problem can be followed, as illustrated in the next example.

Example 4.5.1 `Bean-soaking experiment`

Suppose we were to plan an experiment to compare the effects of $v = 5$ different soaking times on the growth rate of mung bean seeds. The response variable will be the length of a shoot of a mung bean seed 48 hours after soaking. Suppose that a pilot experiment has indicated that the mean square for error is likely to be not more than 10 mm^2, and suppose that we would like a set of 95% simultaneous confidence intervals for pairwise differences of the soaking times, with each interval no wider than 6 mm (that is, the half width or minimum significant difference should be no greater than 3 mm).

The formula for each of the simultaneous confidence intervals for pairwise comparisons using Tukey's method of multiple comparisons is given by (4.4.27) p. 88. For equal sample sizes, the interval half width, or minimum significant difference, is required to be at most 3 mm; that is, we require

$$msd = w_T \sqrt{10\left(\frac{1}{r} + \frac{1}{r}\right)} \leq 3\,,$$

where $w_T = q_{5,5r-5,.05}/\sqrt{2}$ or, equivalently,

$$q^2_{5,5r-5,.05} \leq 0.9r\,.$$

Adopting a trial-and-error approach, we guess a value for r, say $r = 10$. Then, from Table A.8, we find $q^2_{5,45,.05} \approx 4.03^2 = 16.24$, which does not satisfy the requirement that $q^2 \leq 0.9r = 9$. A larger value for r is needed, and we might try $r = 20$ next. The calculations are most conveniently laid out in table form, as follows.

r	$5r-5$	$q^2_{5,5r-5,0.05}$	$0.9r$	*Action*
10	45	$4.03^2 = 16.24$	9.00	Increase r
20	95	$3.95^2 = 15.60$	18.00	Decrease r
15	70	$3.97^2 = 15.76$	13.50	Increase r
18	85	$3.96^2 = 15.68$	16.20	Decrease r
17	80	$3.96^2 = 15.68$	15.30	

If $r = 17$ observations are taken on each of the five soaking times, and if the mean square for error is approximately 10 mm^2 in the main experiment, then the 95% Tukey simultaneous confidence intervals for pairwise comparisons will be a little over the required 6 mm in length. If $r = 18$ observations are taken, the interval will be a little shorter than the 6 mm required. If the cost of the experiment is high, then $r = 17$ would be selected; otherwise, $r = 18$ might be preferred.

Trial and error procedures such as that illustrated in Example 4.5.1 for Tukey's method of multiple comparisons can be used for any of the other multiple comparison methods to obtain the approximate sample sizes required to meet the objectives of the experiment. The same type of calculation can be done for unequal sample sizes, provided that the relative sizes are specified, for example $r_1 = 2r_2 = 2r_3 = 2r_4$.

Unless more information is desired on some treatments than on others, or unless costs or variances are unequal, it is generally advisable to select equal sample sizes whenever possible. Choosing equal sample sizes produces two benefits! Confidence intervals for pairwise comparisons are all the same length, which makes them easier to compare, and the multiple comparison and analysis of variance procedures are less sensitive to an incorrect assumption of normality of the error variables.

Quite often, the sample size calculation will reveal that the required number of observations is too large to meet the budget or the time restrictions of the experiment. There are several possible remedies:

(a) Refine the experimental procedure to reduce the likely size of *msE*,

(b) Omit one or more treatments,

(c) Allow longer confidence intervals,

(d) Allow a lower confidence level.

4.6 Using SAS Software

In this section we illustrate how to use the SAS software to generate information for confidence intervals and hypothesis tests for individual contrasts and means and also for the multiple comparison procedures. We use the data from the battery experiment of Sect. 2.5.2 (p. 24).

A sample SAS program to analyze the data is given in Table 4.1. As in Chap. 3, line numbers have been included for reference but are not part of the SAS program. A data set `BATTERY`, with variables `TYPE`, `LPUC` and `ORDER`, is created from the statements in lines 1–9. The treatment factor is the type of battery `TYPE`, the response variable is the life per unit cost `LPUC`, and the one-way analysis of variance model (3.3.1) was used for the analysis. Lines 10–12 generate the analysis of variance table shown in the top of Fig. 4.1.

4.6.1 Inferences on Individual Contrasts

The SAS statements `ESTIMATE` and `CONTRAST` are part of the `GLM` procedure and are used for making inferences concerning specific contrasts.

The `ESTIMATE` statements (lines 13–15 of Table 4.1) generate information for constructing confidence intervals or conducting hypothesis tests for individual contrasts. Each of the three `ESTIMATE` statements includes a user-selected contrast name in single quotes, together with the name of the factor for which the effects of levels are to be compared, and the coefficients of the contrast to be estimated.

Table 4.1 SAS program for the battery experiment: contrasts and multiple comparisons

```
Line  SAS Program
  1   DATA BATTERY;
  2     INPUT TYPE LPUC ORDER;
  3     LINES;
  4     1 611  1
  5     2 923  2
  6     1 537  3
  7     :   :  :
  8     3 413 16
  9   ;
 10   PROC GLM;
 11     CLASS TYPE;
 12     MODEL LPUC = TYPE;
 13     ESTIMATE 'DUTY'       TYPE  1  1 -1 -1 / DIVISOR = 2;
 14     ESTIMATE 'BRAND'      TYPE  1 -1  1 -1 / DIVISOR = 2;
 15     ESTIMATE 'INTERACTN' TYPE  1 -1 -1  1 / DIVISOR = 2;
 16     CONTRAST 'BRAND'      TYPE  1 -1  1 -1;
 17     LSMEANS TYPE / ADJUST = TUKEY CL PDIFF ALPHA = 0.01;
```

Fig. 4.1 Analysis of variance and output from the CONTRASTS and ESTIMATE statements

Results Viewer - SAS Output

The GLM Procedure

Dependent Variable: LPUC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	427915.2500	142638.4167	60.24	<.0001
Error	12	28412.5000	2367.7083		
Corrected Total	15	456327.7500			

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
BRAND	1	124609.0000	124609.0000	52.63	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
DUTY	251.000000	24.3295516	10.32	<.0001
BRAND	-176.500000	24.3295516	-7.25	<.0001
INTERACTN	-113.250000	24.3295516	-4.65	0.0006

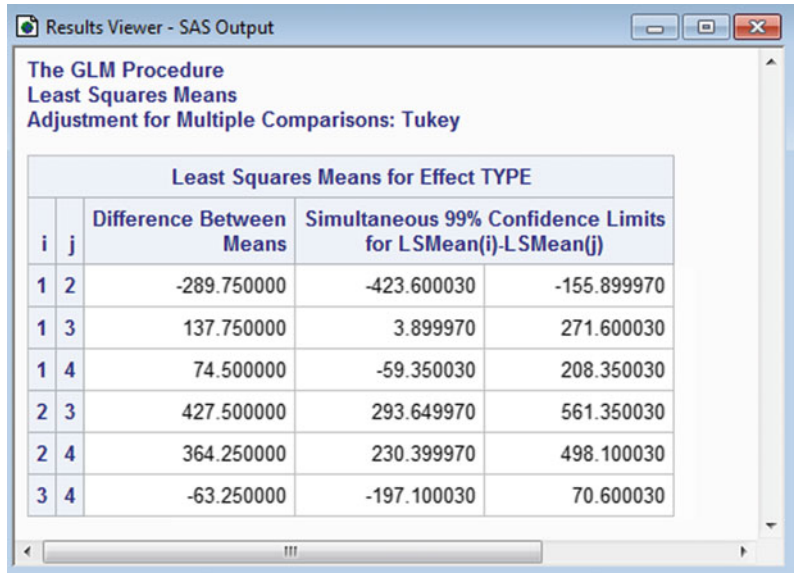
If the contrast coefficients are to be divided by a constant, this is indicated by means of the DIVISOR option. The information generated by these statements is shown in the bottom section of Fig. 4.1. The columns show the contrast name, the contrast estimate $\Sigma c_i \overline{y}_{i\cdot}$, the standard error $\sqrt{msE(\Sigma c_i^2/r_i)}$ for the estimate, the value of the t-statistic for testing the null hypothesis that the contrast is zero (see (4.3.13), p. 77), and the corresponding p-value for a two-tailed test. For each of the contrasts shown in Fig. 4.1, the p-value is at most 0.0006, indicating that all three contrasts are significantly different from zero for any choice of *individual* significance level α^* greater than 0.0006. The overall and individual significance levels should be selected prior to analysis. The parameter estimates and standard errors can be used to construct confidence intervals by hand, using the critical coefficient for the selected multiple comparison methods (see also Sect. 4.6.2).

The CONTRAST statement in line 16 of Table 4.1 generates the information shown in the middle portion of Fig. 4.1 that is needed in (4.3.15), p. 77, for testing the single null hypothesis that the brand contrast is zero versus the alternative hypothesis that it is not zero. The "F Value" of 52.63 is the square of the "t Value" of −7.25 (up to rounding error) for the brand contrast generated by the ESTIMATE statement, the two tests (4.3.13) and (4.3.15) being equivalent.

4.6.2 Multiple Comparisons

The LSMEANS statement (line 17) in the GLM procedure in Table 4.1 can be used to generate the observed least squares means $\overline{y}_{i\cdot}$ for each level of a factor, and to implement the multiple comparisons procedures introduced in Sect. 4.4. Inclusion of the options ADJUST=TUKEY, PDIFF and CL causes the SAS software to use the Tukey method to compare the effects of each pair of levels, providing both p-values for simultaneous testing and confidence limits for simultaneous estimation of all pairwise comparisons. Individual confidence intervals for each treatment mean are also provided as a consequence of the CL option. The option ALPHA=0.01 sets the confidence level at 99%, both for Tukey's

Fig. 4.2 Tukey's method for the battery experiment


Results Viewer - SAS Output

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Least Squares Means for Effect TYPE

i	j	Difference Between Means	Simultaneous 99% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-289.750000	-423.600030	-155.899970
1	3	137.750000	3.899970	271.600030
1	4	74.500000	-59.350030	208.350030
2	3	427.500000	293.649970	561.350030
2	4	364.250000	230.399970	498.100030
3	4	-63.250000	-197.100030	70.600030

method for the simultaneous pairwise comparisons and for the individual confidence intervals for each treatment mean. Part of the corresponding SAS output is given in Fig. 4.2.

Other methods of multiple comparisons can also be requested as options in the `LSMEANS` statement of the `GLM` procedure. For example, the options `ADJUST=BON` and `ADJUST=SCHEFFE` request all pairwise comparisons using the methods of Bonferroni and Scheffé, respectively. The option `ADJUST=DUNNETT` requests Dunnett's 2-sided method of comparing all treatments with a control, the lowest treatment level serving as the control by default. To explicitly specify level 1, say, as the control, replace `PDIFF` with `PDIFF=CONTROL('1')`. Similarly, replacing `PDIFF` with `PDIFF=CONTROLU('1')` requests simultaneous lower bounds for the treatment-versus-control contrasts $\tau_i - \tau_1$ by Dunnett's method and is useful for "upper-tailed"alternative hypotheses—namely, for showing which treatments have a larger effect than the control treatment (coded 1). Likewise, the option `PDIFF=CONTROLL('1')` provides upper bounds useful for "lower-tailed" alternatives—namely, for showing which treatments have a smaller effect than the control treatment (coded 1).

4.7 Using R Software

In this section we illustrate how to use the R software to generate information for confidence intervals and hypothesis tests for individual contrasts and means and also for the multiple comparison procedures. We use the data from the battery experiment of Sect. 2.5.2 (p. 24). The treatment factor is type of battery `Type`, the response variable is the life per unit cost `LPUC`, and the one-way analysis of variance model (3.3.1) was used for the analysis.

A sample R program to analyze the data is given in Table 4.2. Line numbers have been included for the sake of reference but they are not part of the R program. The data are read from file by the statement in line 1 of the R code. In line 2, the data set is augmented with a new variable, `fType`, created by converting the numerical variable `Type` to a factor variable. The `head` command in line 3 displays the first 3 lines of the data set. The function `aov` in line 4 fits the one-way analysis of variance model (3.3.1) to the data, saving the results as the object `model1` for use by subsequent R functions.

Table 4.2 R program for the battery experiment

```
Line  R Code
  1   battery.data = read.table("data/battery.txt", header=T)
  2   battery.data$fType = factor(battery.data$Type)
  3   head(battery.data, 3)

  4   model1 = aov(LPUC ~ fType, data=battery.data) # Fit aov model
  5   anova(model1)  # Display 1-way ANOVA

  6   # Individual contrasts: estimates, CIs, tests
  7   library(lsmeans)
  8   lsmType = lsmeans(model1, ~ fType) # Compute and save lsmeans
  9   levels(battery.data$fType)
 10   summary(contrast(lsmType, list(Duty=c( 1, 1,-1,-1)/2,
 11                                  Brand=c( 1,-1, 1,-1)/2,
 12                                     DB=c( 1,-1,-1, 1)/2)),
 13           infer=c(T,T), level=0.95, side="two-sided")

 14   # Multiple comparisons
 15   confint(lsmType, level=0.90) # Display lsmeans and 90

 16   # Tukey's method
 17   summary(contrast(lsmType, method="pairwise", adjust="tukey"),
 18           infer=c(T,T), level=0.99, side="two-sided")

 19   # Dunnett's method
 20   summary(contrast(lsmType, method="trt.vs.ctrl", adjust="mvt", ref=1),
 21           infer=c(T,T), level=0.99, side="two-sided")
```

The `anova(model1)` function in line 5 generates the analysis of variance data shown in the top of Table 4.3.

4.7.1 Inferences on Individual Contrasts

The `lsmeans` package, loaded in line 7, provides the functionality for computing least squares means and using these for inferences on treatment contrasts. The `lsmeans` statement (line 8) uses the results of the previously fitted model (line 4) to compute least squares means for each battery type (i.e. for each level of `fType`), saving the results as `lsmType`. The `levels` command in line 9 displays the levels of the factor `fType` in order: `"1" "2" "3" "4"`. Using the least squares means saved in line 8, the `summary` and `contrast` functions of the `lsmeans` package (lines 10–13 of Table 4.2) are coupled to generate least squares estimates, tests, and confidence intervals for specified treatment contrasts. For each of these contrasts, the coefficients correspond to the respective levels of `fType` displayed by line 9. In particular, the `contrast` function inputs the least squares means `lsmType` for each battery type plus a list of contrasts, including a name and the coefficients for each, and would generate the information in the middle of Table 4.3 except the confidence limits. The confidence limits are obtained by wrapping the `contrast` function in the `summary` function, for which the option `infer=c(T,T)` requests confidence intervals and tests. The confidence level is optionally specified to be 95% (the default). Two-sided confidence intervals and tests (the default) are also optionally specified, whereas `side="<"` would request upper confidence limits and specify one-sided alternative hypotheses corresponding to the contrasts being less than zero, for example.

Table 4.3 R output: analysis of variance, individual contrasts, and Tukey's method

```
> anova(model1) # Display 1-way ANOVA

Analysis of Variance Table

Response: LPUC
          Df Sum Sq Mean Sq F value  Pr(>F)
fType      3 427915  142638    60.2 1.7e-07
Residuals 12  28412    2368

> # Individual contrasts: estimates, CIs, tests
> library(lsmeans)
> lsmType = lsmeans(model1, ~ fType) # Compute and save lsmeans
> levels(battery.data$fType)

[1] "1" "2" "3" "4"

> summary(contrast(lsmType, list(Duty=c( 1, 1,-1,-1)/2,
+                                Brand=c( 1,-1, 1,-1)/2,
+                                DB=c( 1,-1,-1, 1)/2)),
+         infer=c(T,T), level=0.95, side="two-sided")

 contrast estimate    SE df lower.CL upper.CL t.ratio p.value
 Duty       251.00 24.33 12   197.99   304.01  10.317  <.0001
 Brand     -176.50 24.33 12  -229.51  -123.49  -7.255  <.0001
 DB        -113.25 24.33 12  -166.26   -60.24  -4.655  0.0006

Confidence level used: 0.95

> # Tukey's method
> summary(contrast(lsmType, method="pairwise", adjust="tukey"),
+         infer=c(T,T), level=0.99, side="two-sided")

 contrast estimate     SE df  lower.CL upper.CL t.ratio p.value
 1 - 2     -289.75 34.407 12 -423.6021 -155.898  -8.421  <.0001
 1 - 3      137.75 34.407 12    3.8979  271.602   4.004  0.0082
 1 - 4       74.50 34.407 12  -59.3521  208.352   2.165  0.1882
 2 - 3      427.50 34.407 12  293.6479  561.352  12.425  <.0001
 2 - 4      364.25 34.407 12  230.3979  498.102  10.586  <.0001
 3 - 4      -63.25 34.407 12 -197.1021   70.602  -1.838  0.3035

Confidence level used: 0.99
Conf-level adjustment: tukey method for comparing a family of 4 estimates
P value adjustment: tukey method for comparing a family of 4 estimates
```

Lines 6–13 of the R code are reproduced in the middle of Table 4.3, along with the corresponding output. The output for each listed contrast includes the contrast name, the estimate $\Sigma c_i \overline{y}_{i\cdot}$, the standard error $\sqrt{msE(\Sigma c_i^2/r_i)}$ of the estimate, the number of error degrees of freedom, the 95% confidence interval for the treatment contrast, the value of the t-statistic for testing the null hypothesis that the contrast is zero (see (4.3.13) p. 77), and the corresponding p-value for a two-tailed test. For each of the contrasts shown in Table 4.3, the p-value is less than 0.0006, indicating that all three contrasts are significantly different from zero for any choice of *individual* significance level α^* greater than 0.0006. The overall and individual significance levels should be selected prior to analysis. For multiple

comparisons including non-pairwise comparisons, the contrast estimates and standard errors could be used to construct confidence intervals by hand, using the critical coefficient for the selected multiple comparison methods (see Sect. 4.6.2). Pairwise comparisons will be illustrated in the next section.

4.7.2 Multiple Comparisons

Multiple comparisons procedures introduced in Sect. 4.4 are implemented by the R code in lines 14–21 of Table 4.2. The least squares means package `lsmeans`, loaded in line 7, provides functions to generate the observed least squares means $\overline{y}_{i\cdot}$ for each level of a factor, and also to implement the multiple comparisons procedures introduced in Sect. 4.4. In line 8, `lsmeans` uses the information stored in `model1` to compute the least squares mean for each battery type, saving the least squares means as `lsmType` for subsequent use. The `confint` function in line 15 would display the least squares means and corresponding individual 90% confidence intervals for the treatment means (not shown).

Multiple comparison methods can be implemented by coupling the `summary` and `contrast` functions, as illustrated for Tukey's method in lines 16–18. These code lines and the corresponding output are shown in the bottom of Table 4.3. The option `method="pairwise"` requests all pairwise comparisons. The `contrast` statement embedded in line 17 would apply Tukey's method to test whether each pairwise comparison is zero. One also gets the corresponding Tukey confidence intervals by including the `summary` function and its options, where `infer=c(T,T)` requests confidence intervals as well as tests, `level=0.99` sets the confidence level, and `side="two-sided"` requests two-sided confidence intervals and tests. Tukey's method and two-sided inferences are the defaults for all pairwise comparisons, so `adjust="tukey"` and `side="two-sided"` are redundant here, but one can replace `"tukey"` with `"scheffe"` or `"bonferroni"` to apply the corresponding method, or with `"none"` for no multiple comparisons adjustment. The default confidence level is 95%. Using `method="revpairwise"` reverses the order of the pairwise comparisons, considering $\tau_j - \tau_i$ rather than $\tau_i - \tau_j$.

Implementation of Dunnett's method for all treatment-versus-control comparisons is similar and illustrated by lines 19–21 of Table 4.2. The option `method="trt.vs.ctrl"` yields all treatment-versus-control comparisons (not shown). Dunnett's method uses critical values from the multivariate t-distribution, corresponding to `adjust="mvt"`. These critical values are computed by simulation, so the results vary slightly from run to run unless a simulation seed is specified. Also, if the number of treatments is large, implementation of R functions for the multivariate t-distribution may be slow or simply not work, so the default option `adjust="dunnettx"` provides an approximation of Dunnett's method for two-sided confidence intervals that runs faster and dependably, though it is only applicable when the contrast estimates have pairwise correlations of 0.5 such as in the equireplicate case. The first level of the factor, `"1"`, which happens to be level 1 in this case, is the control by default; the syntax `ref=1` illustrates how to specify the first (or any) level as the control. Also, `"two-sided"` is the default for confidence intervals and tests, but one can specify `side="<"` for the one-sided alternative $H_A : \tau_i < \tau_1$ and the corresponding upper confidence bound for $\tau_i - \tau_1$, or `side=">"` for the alternative $H_A : \tau_i > \tau_1$ and the corresponding lower confidence bound for $\tau_i - \tau_1$.

For additional functionality for multiple comparisons procedures, see the multiple comparisons package `multcomp`.

Exercises

1. **Buoyancy experiment**

   Consider conducting an experiment to investigate the question, "Is the buoyancy of an object in water affected by different concentrations of salt in the water?"

   (a) Complete steps (a)–(d) of the checklist (p. 7) in detail. Specify any preplanned contrasts or functions that should be estimated. State, with reasons, which, if any, methods of multiple comparisons will be used.
   (b) Run a small pilot experiment to obtain a preliminary estimate of σ^2.
   (c) Finish the checklist.

2. **Cotton-spinning experiment, continued**

   For the cotton-spinning experiment of Sect. 2.3, p. 13, identify any contrasts or functions that you think might be interesting to estimate. For any contrasts that you have selected, list the corresponding contrast coefficients.

3. **Meat cooking experiment, continued**

   The meat cooking experiment was described in Exercise 14 of Chap. 3, and the data were given in Table 3.14, p. 68.

   (a) Compare the effects of the six treatments, pairwise, using Scheffé's method of multiple comparisons and a 95% overall confidence level.
   (b) Consider $\mu + (\tau_1 + \tau_4)/2$, $\mu + (\tau_2 + \tau_5)/2$, and $\mu + (\tau_3 + \tau_6)/2$. What do these represent? Make pairwise comparisons of these three expressions, using Scheffé's method of multiple comparisons and a 95% overall confidence level for all treatment contrasts. Interpret the results.

4. **Reaction time experiment**

   (L. Cai, T. Li, Nishant, and A. van der Kouwe, 1996)

   The experiment was run to compare the effects of auditory and visual cues on speed of response of a human subject. A personal computer was used to present a "stimulus" to a subject, and the reaction time required for the subject to press a key was monitored. The subject was warned that the stimulus was forthcoming by means of an auditory or a visual cue. The experimenters were interested in the effects on the subjects' reaction time of the auditory and visual cues and also in different elapsed times between cue and stimulus. Thus, there were two different treatment factors: "cue stimulus" at two levels "auditory" or "visual," and "elapsed time between cue and stimulus" at three levels "five," "ten," or "fifteen" seconds. This gave a total of six treatment combinations, which can be coded as

   $$\begin{array}{ll} 1 = \text{auditory, 5 sec} & 4 = \text{visual, 5 sec} \\ 2 = \text{auditory, 10 sec} & 5 = \text{visual, 10 sec} \\ 3 = \text{auditory, 15 sec} & 6 = \text{visual, 15 sec} \end{array}$$

Table 4.4 Reaction times, in seconds, for the reaction time experiment—(order of collection in parentheses)

Treatments					
1	2	3	4	5	6
0.204 (9)	0.167 (3)	0.202 (13)	0.257 (7)	0.283 (6)	0.256 (1)
0.170 (10)	0.182 (5)	0.198 (16)	0.279 (14)	0.235 (8)	0.281 (2)
0.181 (18)	0.187 (12)	0.236 (17)	0.269 (15)	0.260 (11)	0.258 (4)

The results of a pilot experiment, involving only one subject, are shown in Table 4.4. The reaction times were measured by the computer and are shown in seconds. The order of observation is shown in parentheses.

(a) Identify a set of contrasts that you would find particularly interesting in this experiment. (Hint: A comparison between the auditory treatments and the visual treatments might be of interest). These are your preplanned contrasts.
(b) Plot the data. What does the plot suggest about the treatments?
(c) Test the hypothesis that the treatments do not have different effects on the reaction time against the alternative hypothesis that they do have different effects.
(d) Calculate a set of simultaneous 90% confidence intervals for your preplanned contrasts, using a method or methods of your choice. State your conclusions.

5. **Trout experiment, continued**

Exercise 15 of Chap. 3 (p. 67) concerns a study of the effects of four levels of sulfamerazine (0, 5, 10, 15 g per 100 lb of fish) on the hemoglobin content of trout blood. An analysis of variance test rejected the hypothesis that the four treatment effects are the same at significance level $\alpha = 0.01$.

(a) Compare the four treatments using Tukey's method of pairwise comparisons and a 99% overall confidence level.
(b) Compare the effect of no sulfamerazine on the hemoglobin content of trout blood with the average effect of the other three levels. The overall confidence level of all intervals in parts (a) and (b) should be at least 98%.

6. **Battery experiment, continued**

In Example 4.4.3 (page 89), Tukey's method is used to obtain a set of 95% simultaneous confidence intervals for the pairwise differences $\tau_i - \tau_s$. Verify that this method gives shorter confidence intervals than would either of the Bonferroni or Scheffé methods (for $v = 4$ and $r = 4$).

7. **Soap experiment, continued**

The soap experiment was described in Sect. 2.5.1, p. 20, and an analysis was given in Sect. 3.7.2, p. 50.

(a) Suppose that the experimenter had been interested only in the contrast $\tau_1 - \frac{1}{2}(\tau_2 + \tau_3)$, which compares the weight loss for the regular soap with the average weight loss for the other two soaps. Calculate a confidence interval for this single contrast.

(b) Test the hypothesis that the regular soap has the same average weight loss as the average of the other two soaps. Do this via your confidence interval in part (a) and also via (4.3.13) and (4.3.15).
(c) In Example 4.4.5 (p. 91), Dunnett's method was used for simultaneous 99% confidence intervals for two preplanned treatment-versus-control contrasts. Would either or both of the Bonferroni and Tukey methods have given shorter intervals?
(d) Which method would be the best if all pairwise differences are required? Calculate a set of simultaneous 99% confidence intervals for all of the pairwise differences. Why are the intervals longer than those in part (c)?

8. **Trout experiment, continued**

(a) For the trout experiment in Exercise 15 of Chap. 3 (see p. 67), test the hypotheses that the linear and quadratic trends in hemoglobin content of trout blood due to the amount of sulfamerazine added to the diet is negligible. State the overall significance level of your tests.
(b) Regarding the absence of sulfamerazine in the diet as the control treatment, calculate simultaneous 99% confidence intervals for the three treatment-versus-control comparisons. Which method did you use and why?
(c) What is the overall confidence level of the intervals in part (b) together with those in Exercise 5? Is there a better strategy than using three different procedures for the three sets of intervals? Explain.

9. **Battery experiment, continued**

Suppose the battery experiment of Sect. 2.5.2 (p. 24) is to be repeated. The experiment involved four treatments, and the error standard deviation is estimated from that experiment to be about 48.66 minutes per dollar (minute/dollar).

(a) Calculate a 90% upper confidence limit for the error variance σ^2.
(b) How large should the sample sizes be in the new experiment if Tukey's method of pairwise comparisons is to be used and it is desired to obtain a set of 95% simultaneous confidence intervals of length at most 100 minutes per dollar?
(c) How large should the sample sizes be in the new experiment if Scheffé's method is to be used to obtain a set of 95% simultaneous confidence intervals for various contrasts and if the confidence interval for the duty contrast is to be of length at most 100 minute per dollar?

10. **Trout experiment, continued**

Consider again the trout experiment in Exercise 15 of Chap. 3.

(a) Suppose the experiment were to be repeated. Suggest the largest likely value for the error mean square *msE*.
(b) How many observations should be taken on each treatment so that the length of each interval in a set of simultaneous 95% confidence intervals for pairwise comparisons should be at most 2 g per 100 ml?