# Chapter 5

# Multiple Comparisons

When we make several related tests or interval estimates at the same time, we need to make *multiple comparisons* or do *simultaneous inference.* The issue of multiple comparisons is one of error rates. Each of the individual tests or confidence intervals has a Type I error rate $\mathcal{E}_i$ that can be controlled by the experimenter. If we consider the tests together as a *family,* then we can also compute a combined Type I error rate for the family of tests or intervals. When a family contains more and more true null hypotheses, the probability that one or more of these true null hypotheses is rejected increases, and the probability of any Type I errors in the family can become quite large. Multiple comparisons procedures deal with Type I error rates for families of tests.

Multiple comparisons, simultaneous inference, families of hypotheses

**Example 5.1**

**Carcinogenic mixtures**

We are considering a new cleaning solvent that is a mixture of 100 chemicals. Suppose that regulations state that a mixture is safe if all of its constituents are safe (pretending we can ignore chemical interaction). We test the 100 chemicals for causing cancer, running each test at the 5% level. This is the individual error rate that we can control.

What happens if all 100 chemicals are harmless and safe? Because we are testing at the 5% level, we expect 5% of the nulls to be rejected even when all the nulls are true. Thus, on average, 5 of the 100 chemicals will be declared to be carcinogenic, even when all are safe. Moreover, if the tests are independent, then one or more of the chemicals will be declared unsafe in 99.4% of all sets of experiments we run, even if all the chemicals are safe. This 99.4% is a combined Type I error rate; clearly we have a problem.

## 5.1 Error Rates

Determine error rate to control

When we have more than one test or interval to consider, there are several ways to define a combined Type I error rate for the family of tests. This variety of combined Type I error rates is the source of much confusion in the use of multiple comparisons, as different error rates lead to different procedures. People sometimes ask "Which procedure should I use?" when the real question is "Which error rate do I want to control?". As data analyst, you need to decide which error rate is appropriate for your situation and then choose a method of analysis appropriate for that error rate. This choice of error rate is not so much a statistical decision as a scientific decision in the particular area under consideration.

Data snooping performs many implicit tests

*Data snooping* is a practice related to having many tests. Data snooping occurs when we first look over the data and then choose the null hypotheses to be tested based on "interesting" features in the data. What we tend to do is consider many potential features of the data and discard those with uninteresting or null behavior. When we data snoop and then perform a test, we tend to see the smallest $p$-value from the ill-defined family of tests that we considered when we were snooping; we have not really performed just one test. Some multiple comparisons procedures can actually control for data snooping.

*Simultaneous inference* is deciding which error rate we wish to control, and then using a procedure that controls the desired error rate.

Individual and combined null hypotheses

Let's set up some notation for our problem. We have a set of $K$ null hypotheses $H_{01}, H_{02}, \ldots, H_{0K}$. We also have the "combined," "overall," or "intersection" null hypotheses $H_0$ which is true if *all* of the $H_{0i}$ are true. In formula,

$$H_0 = H_{01} \cap H_{02} \cap \cdots \cap H_{0K}.$$

The collection $H_{01}, H_{02}, \ldots, H_{0K}$ is sometimes called a family of null hypotheses. We reject $H_0$ if any of null hypotheses $H_{0i}$ is rejected. In Example 5.1, $K = 100$, $H_{0i}$ is the null hypothesis that chemical $i$ is safe, and $H_0$ is the null hypothesis that all chemicals are safe so that the mixture is safe.

We now define five combined Type I error rates. The definitions of these error rates depend on numbers or fractions of falsely rejected null hypotheses $H_{0i}$, which will never be known in practice. We set up the error rates here and later give procedures that can be shown mathematically to control the error rates.

The *per comparison error rate* or *comparisonwise error rate* is the probability of rejecting a particular $H_{0i}$ in a single test when that $H_{0i}$ is true. Controlling the per comparison error rate at $\mathcal{E}$ means that the expected fraction of individual tests that reject $H_{0i}$ when $H_0$ is true is $\mathcal{E}$. This is just the usual error rate for a $t$-test or F-test; it makes no correction for multiple comparisons. The tests in Example 5.1 controlled the per comparison error rate at 5%.

Comparisonwise error rate

The *per experiment error rate* or *experimentwise error rate* or *familywise error rate* is the probability of rejecting one or more of the $H_{0i}$ (and thus rejecting $H_0$) in a series of tests when all of the $H_{0i}$ are true. Controlling the experimentwise error rate at $\mathcal{E}$ means that the expected fraction of experiments in which we would reject one or more of the $H_{0i}$ when $H_0$ is true is $\mathcal{E}$. In Example 5.1, the per experiment error rate is the fraction of times we would declare one or more of the chemicals unsafe when in fact all were safe. Controlling the experimentwise error rate at $\mathcal{E}$ necessarily controls the comparisonwise error rate at no more than $\mathcal{E}$. The experimentwise error rate considers all individual null hypotheses that were rejected; if any one of them was correctly rejected, then there is no penalty for any false rejections that may have occurred.

Experimentwise error rate

A statistical discovery is the rejection of an $H_{0i}$. The false discovery fraction is 0 if there are no rejections; otherwise it is the number of false discoveries (Type I errors) divided by the total number of discoveries. The *false discovery rate* (FDR) is the expected value of the false discovery fraction. If $H_0$ is true, then all discoveries are false and the FDR is just the experimentwise error rate. Thus controlling the FDR at $\mathcal{E}$ also controls the experimentwise error at $\mathcal{E}$. However, the FDR also controls at $\mathcal{E}$ the average fraction of rejections that are Type I errors when some $H_{0i}$ are true and some are false, a control that the experimentwise error rate does not provide. With the FDR, we are allowed more incorrect rejections as the number of true rejections increases, but the ratio is limited. For example, with FDR at .05, we are allowed just one incorrect rejection with 19 correct rejections.

False discovery rate

The *strong familywise error rate* is the probability of making any false discoveries, that is, the probability that the false discovery fraction is greater than zero. Controlling the strong familywise error rate at $\mathcal{E}$ means that the probability of making any false rejections is $\mathcal{E}$ or less, regardless of how many correct rejections are made. Thus one true rejection cannot make any false rejections more likely. Controlling the strong familywise error rate at $\mathcal{E}$ controls the FDR at no more than $\mathcal{E}$. In Example 5.1, a strong familywise error rate of $\mathcal{E}$ would imply that in a situation where 2 of the chemicals were carcinogenic, the probability of declaring one of the other 98 to be carcinogenic would be no more than $\mathcal{E}$.

Strong familywise error rate

Simultaneous confidence intervals

Finally, suppose that each null hypothesis relates to some parameter (for example, a mean), and we put confidence intervals on all these parameters. An error occurs when one of our confidence intervals fails to cover the true parameter value. If this true parameter value is also the null hypothesis value, then an error is a false rejection. The *simultaneous confidence intervals* criterion states that all of our confidence intervals must cover their true parameters simultaneously with confidence $1 - \mathcal{E}$. Simultaneous $1 - \mathcal{E}$ confidence intervals also control the strong familywise error rate at no more than $\mathcal{E}$. (In effect, the strong familywise criterion only requires simultaneous intervals for the null parameters.) In Example 5.1, we could construct simultaneous confidence intervals for the cancer rates of each of the 100 chemicals. Note that a single confidence interval in a collection of intervals with simultaneous coverage $1 - \mathcal{E}$ will have coverage greater than $1 - \mathcal{E}$.

More stringent procedures are less powerful

There is a trade-off between Type I error and Type II error (failing to reject a null when it is false). As we go to more and more stringent Type I error rates, we become more confident in the rejections that we do make, but it also becomes more difficult to make rejections. Thus, when using the more stringent Type I error controls, we are more likely to fail to reject some null hypotheses that should be rejected than when using the less stringent rates. In simultaneous inference, controlling stronger error rates leads to less powerful tests.

**Example 5.2**

### Functional magnetic resonance imaging

Many functional Magnetic Resonance Imaging (fMRI) studies are interested in determining which areas of the brain are "activated" when a subject is engaged in some task. Any one image slice of the brain may contain 5000 voxels (individual locations to be studied), and one analysis method produces a $t$-test for each of the 5000 voxels. Null hypothesis $H_{0i}$ is that voxel $i$ is not activated. Which error rate should we use?

If we are studying a small, narrowly defined brain region and are unconcerned with other brain regions, then we would want to test individually the voxels in the brain regions of interest. The fact that there are 4999 other voxels is unimportant, so we would use a per comparison method.

Suppose instead that we are interested in determining if there are any activations in the image. We recognize that by making many tests we are likely to find one that is "significant", even when all nulls are true; we want to protect ourselves against that possibility, but otherwise need no stronger control. Here we would use a per experiment error rate.

Suppose that we believe that there will be many activations, so that $H_0$ is not true. We don't want some correct discoveries to open the flood gates for many false discoveries, but we are willing to live with some false discoveries

as long as they are a controlled fraction of the total made. This is acceptable because we are going to investigate several subjects; the truly activated rejections should be rejections in most subjects, and the false rejections will be scattered. Here we would use the FDR.

Suppose that in addition to expecting true activations, we are also only looking at a single subject, so that we can't use multiple subjects to determine which activations are real. Here we don't want false activations to cloud our picture, so we use the strong familywise error rate.

Finally, we might want to be able to estimate the amount of activation in every voxel, with simultaneous accuracy for all voxels. Here we would use simultaneous confidence intervals.

A *multiple comparisons procedure* is a method for controlling a Type I error rate other than the per comparison error rate.

The literature on multiple comparisons is vast, and despite the length of this Chapter, we will only touch the highlights. I have seen quite a bit of nonsense regarding these methods, so I will try to set out rather carefully what the methods are doing. We begin with a discussion of Bonferroni-based methods for combining generic tests. Next we consider the Scheffé procedure, which is useful for contrasts suggested by data (data snooping). Then we turn our attention to pairwise comparisons, for which there are dozens of methods. Finally, we consider comparing treatments to a control or to the best response.

## 5.2 Bonferroni-Based Methods

The Bonferroni technique is the simplest, most widely applicable multiple comparisons procedure. The Bonferroni procedure works for a fixed set of $K$ null hypotheses to test or parameters to estimate. Let $p_i$ be the $p$-value for testing $H_{0i}$. The Bonferroni procedure says to obtain simultaneous $1 - \mathcal{E}$ confidence intervals by constructing individual confidence intervals with coverage $1 - \mathcal{E}/K$, or reject $H_{0i}$ (and thus $H_0$) if

Ordinary Bonferroni

$$p_i < \mathcal{E}/K \quad .$$

That is, simply run each test at level $\mathcal{E}/K$. The testing version controls the strong familywise error rate, and the confidence intervals are simultaneous. The tests and/or intervals need not be independent, of the same type, or related in any way.

| Reject $H_{0(i)}$ if | Method | Control |
|---|---|---|
| $p_{(i)} < \mathcal{E}/K$ | Bonferroni | Simultaneous confidence intervals |
| $p_{(j)} < \mathcal{E}/(K-j+1)$ for all $j = 1, \ldots, i$ | Holm | Strong familywise error rate |
| $p_{(j)} \leq j\mathcal{E}/K$ for some $j \geq i$ | FDR | False discovery rate; needs independent tests |

**Display 5.1:** Summary of Bonferroni-style methods for $K$ comparisons.

Holm

The *Holm* procedure is a modification of Bonferroni that controls the strong familywise error rate, but does not produce simultaneous confidence intervals (Holm 1979). Let $p_{(1)}, \ldots, p_{(K)}$ be the $p$-values for the $K$ tests sorted into increasing order, and let $H_{0(i)}$ be the null hypotheses sorted along with the $p$-values. Then reject $H_{0(i)}$ if

$$p_{(j)} \leq \mathcal{E}/(K-j+1) \text{ for all } j = 1, \ldots, i.$$

Thus we start with the smallest $p$-value; if it is rejected we consider the next smallest, and so on. We stop when we reach the first nonsignificant $p$-value. This is a little more complicated, but we gain some power since only the smallest $p$-value is compared to $\mathcal{E}/K$.

FDR modification of Bonferroni requires independent tests

The FDR method of Benjamini and Hochberg (1995) controls the False Discovery Rate. Once again, sort the $p$-values and the hypotheses. For the FDR, start with the largest $p$-value and work down. Reject $H_{0i}$ if

$$p_{(j)} \leq j\mathcal{E}/K \text{ for some } j \geq i.$$

This procedure is correct when the tests are statistically independent. It controls the FDR, but not the strong familywise error rate.

The three Bonferroni methods are summarized in Display 5.1. Example 5.3 illustrates their use.

**Example 5.3**

### Sensory characteristics of cottage cheeses

Table 5.1 shows the results of an experiment comparing the sensory characteristics of nonfat, 2% fat, and 4% fat cottage cheese (Michicich 1995). The table shows the characteristics grouped by type and $p$-values for testing the null hypothesis that there was no difference between the three cheeses in the various sensory characteristics. There are 21 characteristics in three groups of sizes 7, 6, and 8.

How do we do multiple comparisons here? First we need to know:

1. Which error rate is of interest?
2. If we do choose an error rate other than the per comparison error rate, what is the appropriate "family" of tests? Is it all 21 characteristics, or separately within group of characteristic?

There is no automatic answer to either of these questions. The answers depend on the goals of the study, the tolerance of the investigator to Type I error, how the results of the study will be used, whether the investigator views the three groups of characteristics as distinct, and so on.

The last two columns of Table 5.1 give the results of the Bonferroni, Holm, and FDR procedures applied at the 5% level to all 21 comparisons and within each group. The $p$-values are compared to the criteria in Display 5.1 using $K = 21$ for the overall family and $K$ of 7, 6, or 8 for by group comparisons.

Consider the characteristic "cheesy flavor" with a .01 $p$-value. If we use the overall family, this is the tenth smallest $p$-value out of 21 $p$-values. The results are

- *Bonferroni* The critical value is $.05/21 = .0024$—not significant.
- *Holm* The critical value is $.05/(21-10+1) = .0042$—not significant.
- *FDR* The critical value is $10 \times .05/21 = .024$—significant.

If we use the flavor family, this is the fourth smallest $p$-value out of six $p$-values. Now the results are

- *Bonferroni* The critical value is $.05/6 = .008$—not significant.
- *Holm* The critical value is $.05/(6-4+1) = .017$ (and all smaller $p$-values meet their critical values)—significant.
- *FDR* The critical value is $4 \times .05/6 = .033$—significant.

**Table 5.1:** Sensory attributes of three cottage cheeses: $p$-values and 5% significant results overall and familywise by type of attribute using the Bonferroni (●), Holm (○), and FDR methods(⋆).

| Characteristic | $p$-value | Overall | By group |
|---|---|---|---|
| | **Appearance** | | |
| White | .004 | ⋆ | ●○⋆ |
| Yellow | .002 | ●○⋆ | ●○⋆ |
| Gray | .13 | | |
| Curd size | .29 | | |
| Size uniformity | .73 | | |
| Shape uniformity | .08 | | |
| Liquid/solid ratio | .02 | ⋆ | ⋆ |
| | **Flavor** | | |
| Sour | .40 | | |
| Sweet | .24 | | |
| Cheesy | .01 | ⋆ | ○⋆ |
| Rancid | .0001 | ●○⋆ | ●○⋆ |
| Cardboard | .0001 | ●○⋆ | ●○⋆ |
| Storage | .001 | ●○⋆ | ●○⋆ |
| | **Texture** | | |
| Breakdown rate | .001 | ●○⋆ | ●○⋆ |
| Firm | .0001 | ●○⋆ | ●○⋆ |
| Sticky | .41 | | |
| Slippery | .07 | | |
| Heavy | .15 | | |
| Particle size | .42 | | |
| Runny | .002 | ●○⋆ | ●○⋆ |
| Rubbery | .006 | ⋆ | ●○⋆ |

These results illustrate that more null hypotheses are rejected considering each group of characteristics to be a family of tests rather than overall (the $K$ is smaller for the individual groups), and fewer rejections are made using the more stringent error rates. Again, the choices of error rate and family of tests are not purely statistical, and controlling an error rate within a group of tests does not control that error rate for all tests.

## 5.3 The Scheffé Method for *All* Contrasts

The Scheffé method is a multiple comparisons technique for contrasts that produces simultaneous confidence intervals for *any* and *all* contrasts, *including contrasts suggested by the data*. Thus Scheffé is the appropriate technique for assessing contrasts that result from data snooping. This sounds like the ultimate in error rate control—arbitrarily many comparisons, even ones suggested from the data! The downside of this amazing protection is low power. Thus we only use the Scheffé method in those situations where we have a contrast suggested by the data, or many, many contrasts that cannot be handled by other techniques. In addition, pairwise comparison contrasts $\overline{y}_{i\bullet} - \overline{y}_{j\bullet}$, even pairwise comparisons suggested by the data, are better handled by methods specifically designed for pairwise comparisons.

Scheffé protects against data snooping, but has low power

We begin with the Scheffé test of the null hypothesis $H_0 : w(\{\alpha_i\}) = 0$ against a two-sided alternative. The Scheffé test statistic is the ratio

$$\frac{SS_w/(g-1)}{MS_E} \ ;$$

we get a $p$-value as the area under an F-distribution with $g-1$ and $\nu$ degrees of freedom to the right of the test statistic. The degrees of freedom $\nu$ are from our denominator $MS_E$; $\nu = N - g$ for the completely randomized designs we have been considering so far. Reject the null hypothesis if this $p$-value is less than our Type I error rate $\mathcal{E}$. In effect, the Scheffé procedure treats the mean square for any single contrast as if it were the full $g-1$ degrees of freedom between groups mean square.

Scheffé F-test

There is also a Scheffé $t$-test for contrasts. Suppose that we are testing the null hypothesis $H_0 : w(\{\alpha_i\}) = \delta$ against a two-sided alternative. The Scheffé $t$-test controls the Type I error rate at $\mathcal{E}$ by rejecting the null hypothesis when

$$\frac{|w(\{\overline{y}_{i\bullet}\}) - \delta|}{\sqrt{MS_E \ \sum_{i=1}^{g} \frac{w_i^2}{n_i}}} > \sqrt{(g-1)F_{\mathcal{E},g-1,\nu}} \ ,$$

where $F_{\mathcal{E},g-1,\nu}$ is the upper $\mathcal{E}$ percent point of an F-distribution with $g-1$ and $\nu$ degrees of freedom. Again, $\nu$ is the degrees of freedom for $MS_E$. For the usual null hypothesis value $\delta = 0$, this is equivalent to the ratio-of-mean-squares version given above.

Scheffé $t$-test

We may also use the Scheffé approach to form simultaneous confidence intervals for any $w(\{\alpha_i\})$:

$$w(\{\overline{y}_{i\bullet}\}) \pm \sqrt{(g-1)F_{\mathcal{E},g-1,\nu}} \times \sqrt{MS_E \ \sum_{i=1}^{g} \frac{w_i^2}{n_i}} \ .$$

Scheffé confidence interval

These Scheffé intervals have simultaneous $1 - \mathcal{E}$ coverage over any set of contrasts, including contrasts suggested by the data.

**Example 5.4** **Acid rain and birch seedlings, continued**

Example 3.1 introduced an experiment in which birch seedlings were exposed to various levels of artificial acid rain. The following table gives some summaries for the data:

| pH | 4.7 | 4.0 | 3.3 | 3.0 | 2.3 |
|---|---|---|---|---|---|
| weight | .337 | .296 | .320 | .298 | .177 |
| n | 48 | 48 | 48 | 48 | 48 |

The $MS_E$ was .0119 with 235 degrees of freedom.

Inspection of the means shows that most of the response means are about .3, but the response for the pH 2.3 treatment is much lower. This suggests that a contrast comparing the pH 2.3 treatment with the mean of the other treatments would have a large value. The coefficients for this contrast are (.25, .25, .25, .25, -1). This contrast has value

$$\frac{.337 + .296 + .320 + .298}{4} - .177 = .1357$$

and standard error

$$\sqrt{.0119\left(\frac{.0625}{48} + \frac{.0625}{48} + \frac{.0625}{48} + \frac{.0625}{48} + \frac{1}{48}\right)} = .0176 \ .$$

We must use the Scheffé procedure to construct a confidence interval or assess the significance of this contrast, because the contrast was suggested by the data. For a 99% confidence interval, the Scheffé multiplier is

$$\sqrt{4 \ F_{.01,4,235}} = 3.688 \ .$$

Thus the 99% confidence interval for this contrast is $.1357 - 3.688 \times .0176$ up to $.1357 + 3.688 \times .0176$, or (.0708, .2006). Alternatively, the $t$-statistic for testing the null hypothesis that the mean response in the last group is equal to the average of the mean responses in the other four groups is $.1357/.0176 = 7.71$. The Scheffé critical value for testing the null hypothesis at the $\mathcal{E} = .001$ level is

$$\sqrt{(g-1)F_{\mathcal{E},g-1,N-g}} = \sqrt{4 \ F_{.001,4,235}} = \sqrt{4 \times 4.782} = 4.37 \ ,$$

so we can reject the null at the .001 level.

Remember, it is not fair to hunt around through the data for a big contrast, test it, and think that you've only done one comparison.

## 5.4 Pairwise Comparisons

A *pairwise comparison* is a contrast that examines the difference between two treatment means $\overline{y}_{i\bullet} - \overline{y}_{j\bullet}$. For $g$ treatment groups, there are

$$\binom{g}{2} = \frac{g(g-1)}{2}$$

different pairwise comparisons. Pairwise comparisons procedures control a Type I error rate at $\mathcal{E}$ for all pairwise comparisons. If we data snoop, choose the biggest and smallest $\overline{y}_{i\bullet}$'s and take the difference, we have not made just one comparison; rather we have made all $g(g-1)/2$ pairwise comparisons, and selected the largest. Controlling a Type I error rate for this greatest difference is one way to control the error rate for all differences.

Tests or confidence intervals

As with many other inference problems, pairwise comparisons can be approached using confidence intervals or tests. That is, we may compute confidence intervals for the differences $\mu_i - \mu_j$ or $\alpha_i - \alpha_j$ or test the null hypotheses $H_{0ij} : \mu_i = \mu_j$ or $H_{0ij} : \alpha_i = \alpha_j$. Confidence regions for the differences of means are generally more informative than tests.

A pairwise comparisons procedure can generally be viewed as a critical value (or set of values) for the $t$-tests of the pairwise comparison contrasts. Thus we would reject the null hypothesis that $\alpha_i - \alpha_j = 0$ if

$$\frac{|\overline{y}_{i\bullet} - \overline{y}_{j\bullet}|}{\sqrt{MS_E}\sqrt{1/n_i + 1/n_j}} > u \ ,$$

Critical values $u$ for $t$-tests

where $u$ is a critical value. Various pairwise comparisons procedures differ in how they define the critical value $u$, and $u$ may depend on several things, including $\mathcal{E}$, the degrees of freedom for $MS_E$, the number of treatments, the number of treatments with means between $\overline{y}_{i\bullet}$ and $\overline{y}_{j\bullet}$, and the number of treatment comparisons with larger $t$-statistics.

An equivalent form of the test will reject if

$$|\overline{y}_{i\bullet} - \overline{y}_{j\bullet}| > u\ \sqrt{MS_E}\sqrt{1/n_i + 1/n_j} = D_{ij} \ \ .$$

Significant differences $D_{ij}$

If all sample sizes are equal and the critical value $u$ is constant, then $D_{ij}$ will be the same for all $i, j$ pairs and we would reject the null if any pair of treatments had mean responses that differed by $D$ or more. This quantity $D$ is called a *significant difference*; for example, using a Bonferroni adjustment to the $g(g-1)/2$ pairwise comparisons tests leads to a Bonferroni significant difference (BSD).

Confidence intervals for pairwise differences $\mu_i - \mu_j$ can be formed from the pairwise tests via

$$(\overline{y}_{i\bullet} - \overline{y}_{j\bullet}) \pm u\sqrt{MS_E}\sqrt{1/n_i + 1/n_j} \ .$$

The remainder of this section presents methods for displaying the results of pairwise comparisons, introduces the Studentized range, discusses several pairwise comparisons methods, and then illustrates the methods with an example.

### 5.4.1 Displaying the results

Underline diagram summarizes pairwise comparisons

Pairwise comparisons generate a lot of tests, so we need convenient and compact ways to present the results. An *underline diagram* is a graphical presentation of pairwise comparison results; construct the underline diagram in the following steps.

1. Sort the treatment means into increasing order and write out treatment labels (numbers or names) along a horizontal axis. The $\overline{y}_{i\bullet}$ values may be added if desired.
2. Draw a line segment under a group of treatments if no pair of treatments in that group is significantly different. Do not include short lines that are implied by long lines. That is, if treatments 4, 5, and 6 are not significantly different, only use one line under all of them—not a line under 4 and 5, and a line under 5 and 6, and a line under 4, 5, and 6.

Here is a sample diagram for three treatments that we label A, B, and C:

C A B

This diagram includes treatment labels, but not treatment means. From this summary we can see that C can be distinguished from B (there is no underline that covers both B and C), but A cannot be distinguished from either B or C (there are underlines under A and C, and under A and B).

Note that there can be some confusion after pairwise comparisons. You must not confuse "is not significantly different from" or "cannot be distinguished from" with "is equal to." Treatment mean A cannot be equal to treatment means B and C and still have treatment means B and C not equal each other. Such a pattern can hold for results of significance tests.

Insignificant difference does not imply equality

There are also several nongraphical methods for displaying pairwise comparisons results. In one method, we sort the treatments into order of increasing means and print the treatment labels. Each treatment label is followed by one or more numbers (letters are sometimes used instead). Any treatments sharing a number (or letter) are not significantly different. Thus treatments sharing common numbers or letters are analogous to treatments being connected by an underline. The grouping letters are often put in parentheses or set as sub- or superscripts. The results in our sample underline diagram might thus be presented as one of the following:

Letter or number tags

| | | | | | |
|---|---|---|---|---|---|
| C (1) | A (12) | B (2) | C (a) | A (ab) | B (b) |
| $C^1$ | $A^{12}$ | $B^2$ | $C^a$ | $A^{ab}$ | $B^b$ |

There are several other variations on this theme.

A third way to present pairwise comparisons is as a table, with treatments labeling both rows and columns. Table elements can flag significant differences or contain confidence intervals for the differences. Only entries above or below the diagonal of the table are needed.

Table of CI's or significant differences

### 5.4.2 The Studentized range

The range of a set is the maximum value minus the minimum value, and *Studentization* means dividing a statistic by an estimate of its standard error. Thus the *Studentized range* for a set of treatment means is

Range, Studentization, and Studentized range

$$\max_i \frac{\overline{y}_{i\bullet}}{\sqrt{MS_E/n}} - \min_j \frac{\overline{y}_{j\bullet}}{\sqrt{MS_E/n}} .$$

Note that we have implicitly assumed that all the sample sizes $n_i$ are the same.

If all the treatments have the same mean, that is, if $H_0$ is true, then the Studentized range statistic follows the Studentized range distribution. Large values of the Studentized range are less likely under $H_0$ and more likely under the alternative when the means are not all equal, so we may use the Studentized range as a test statistic for $H_0$, rejecting $H_0$ when the Studentized

Studentized range distribution

range statistic is sufficiently large. This Studentized range test is a legitimate alternative to the ANOVA F-test.

The Studentized range distribution is important for pairwise comparisons because it is the distribution of the biggest (scaled) difference between treatment means when the null hypothesis is true. We will use it as a building block in several pairwise comparisons methods.

Percent points $q_{\mathcal{E}}(g, \nu)$

The Studentized range distribution depends only on $g$ and $\nu$, the number of groups and the degrees of freedom for the error estimate $MS_E$. The quantity $q_{\mathcal{E}}(g, \nu)$ is the upper $\mathcal{E}$ percent point of the Studentized range distribution for $g$ groups and $\nu$ error degrees of freedom; it is tabulated in Appendix Table D.8.

### 5.4.3 Simultaneous confidence intervals

Tukey HSD or honest significant difference

The Tukey honest significant difference (HSD) is a pairwise comparisons technique that uses the Studentized range distribution to construct simultaneous confidence intervals for differences of all pairs of means. If we reject the null hypothesis $H_{0ij}$ when the (simultaneous) confidence interval for $\mu_i - \mu_j$ does not include 0, then the HSD also controls the strong familywise error rate.

The HSD uses the critical value

$$u(\mathcal{E}, \nu, g) = \frac{q_{\mathcal{E}}(g, \nu)}{\sqrt{2}} \ ,$$

The HSD

leading to

$$HSD = \frac{q_{\mathcal{E}}(g, \nu)}{\sqrt{2}} \sqrt{MS_E} \sqrt{\frac{1}{n} + \frac{1}{n}} = \frac{q_{\mathcal{E}}(g, \nu) \sqrt{MS_E}}{\sqrt{n}} \ .$$

Form simultaneous $1 - \mathcal{E}$ confidence intervals via

$$\overline{y}_{i\bullet} - \overline{y}_{j\bullet} \pm \frac{q_{\mathcal{E}}(g, \nu)}{\sqrt{2}} \ \sqrt{MS_E} \ \sqrt{\frac{1}{n} + \frac{1}{n}} \ .$$

The degrees of freedom $\nu$ are the degrees of freedom for the error estimate $MS_E$.

Strictly speaking, the HSD is only applicable to the equal sample size situation. For the unequal sample size case, the approximate HSD is

$$HSD_{ij} = q_{\mathcal{E}}(g, \nu) \sqrt{MS_E} \sqrt{\frac{1}{2 n_i n_j / (n_i + n_j)}}$$

**Table 5.2:** Total free amino acids in cheeses after 168 days of ripening.

| Strain added | | | |
|---|---|---|---|
| None | A | B | A&B |
| 4.195 | 4.125 | 4.865 | 6.155 |
| 4.175 | 4.735 | 5.745 | 6.488 |

or, equivalently,

Tukey-Kramer form for unequal sample sizes

$$HSD_{ij} = \frac{q_{\mathcal{E}}(g,\nu)}{\sqrt{2}}\sqrt{MS_E}\sqrt{(\frac{1}{n_i}+\frac{1}{n_j})} \ .$$

This approximate HSD, often called the Tukey-Kramer form, tends to be slightly conservative (that is, the true error rate is slightly less than $\mathcal{E}$).

Bonferroni significant difference or BSD

The Bonferroni significant difference (BSD) is simply the application of the Bonferroni technique to the pairwise comparisons problem to obtain

$$\begin{aligned} u = u(\mathcal{E},\nu,K) &= t_{\mathcal{E}/(2K),\nu} \ , \\ BSD_{ij} &= t_{\mathcal{E}/(2K),\nu}\sqrt{MS_E}\sqrt{1/n_i + 1/n_j} \ , \end{aligned}$$

where $K$ is the number of pairwise comparisons. We have $K = g(g-1)/2$ for all pairwise comparisons between $g$ groups. BSD produces simultaneous confidence intervals and controls the strong familywise error rate.

Use HSD when making all pairwise comparisons

When making all pairwise comparisons, the HSD is less than the BSD. Thus we prefer the HSD to the BSD for all pairwise comparisons, because the HSD will produce shorter confidence intervals that are still simultaneous. When only a preplanned subset of all the pairs is being considered, the BSD may be less than and thus preferable to the HSD.

## Example 5.5 Free amino acids in cheese

Cheese is produced by bacterial fermentation of milk. Some bacteria in cheese are added by the cheese producer. Other bacteria are present but were not added deliberately; these are called nonstarter bacteria. Nonstarter bacteria vary from facility to facility and are believed to influence the quality of cheese.

Two strains (A and B) of nonstarter bacteria were isolated at a premium cheese facility. These strains will be added experimentally to cheese to determine their effects. Eight cheeses are made. These cheeses all get a standard

starter bacteria. In addition, two cheeses will be randomly selected for each of the following four treatments: control, add strain A, add strain B, or add both strains A and B. Table 5.2 gives the total free amino acids in the cheeses after 168 days of ripening. (Free amino acids are thought to contribute to flavor.)

Listing 5.1 gives Minitab output showing an Analysis of Variance for these data ①, as well as HSD comparisons (called Tukey's pairwise comparisons) using $\mathcal{E} = .1$ ②; we use the $MS_E$ from this ANOVA in constructing the HSD. HSD is appropriate if we want simultaneous confidence intervals on the pairwise differences. The HSD is

$$\begin{aligned} \frac{q_{\mathcal{E}}(g,\nu)}{\sqrt{2}}\sqrt{MS_E}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} &= \frac{q_{.1}(4,4)}{\sqrt{2}}\sqrt{.1572}\sqrt{\frac{1}{2}+\frac{1}{2}} \\ &= 4.586 \times .3965/1.414 = 1.286 \ . \end{aligned}$$

We form confidence intervals as the observed difference in treatment means, plus or minus 1.286; so for A&B minus control, we have

$$6.322 - 4.185 \pm 1.286 \text{ or } (.851, 3.423) \ .$$

In fact, only two confidence intervals for pairwise differences do not include zero (see Listing 5.1 ②). The underline diagram is:

| C | A | B | A&B |
|---|---|---|---|
| 4.19 | 4.43 | 5.31 | 6.32 |

Note in Listing 5.1 ② that Minitab displays pairwise comparisons as a table of confidence intervals for differences.

### 5.4.4 Strong familywise error rate

Step-down methods work inward from the outside comparisons

A *step-down method* is a procedure for organizing pairwise comparisons starting with the most extreme pair and then working in. Relabel the groups so that the sample means are in increasing order with $\overline{y}_{(1)\bullet}$ smallest and $\overline{y}_{(g)\bullet}$ largest. (The relabeled estimated effects $\widehat{\alpha}_{(i)}$ will also be in increasing order, but the relabeled true effects $\alpha_{[i]}$ may or may not be in increasing order.) With this ordering, $\overline{y}_{(1)\bullet}$ to $\overline{y}_{(g)\bullet}$ is a stretch of $g$ means, $\overline{y}_{(1)\bullet}$ to $\overline{y}_{(g-1)\bullet}$ is a stretch of $g-1$ means, and $\overline{y}_{(i)\bullet}$ to $\overline{y}_{(j)\bullet}$ is a stretch of $j-i+1$ means. In a step-down procedure, all comparisons for stretches of $k$ means use the same critical value, but we may use different critical values for different $k$. This

**Listing 5.1:** Minitab output for free amino acids in cheese.

```
Source     DF        SS        MS        F        P          ①
Trt         3     5.628     1.876    11.93    0.018
Error       4     0.629     0.157
Total       7     6.257

                                 Individual 95% CIs For Mean
                                 Based on Pooled StDev
Level       N      Mean     StDev  ------+---------+---------+---------+
A           2    4.4300    0.4313     (------*-------)
A+B         2    6.3215    0.2355                       (-------*-------)
B           2    5.3050    0.6223              (-------*-------)
control     2    4.1850    0.0141  (-------*-------)
                                 ------+---------+---------+---------+
Pooled StDev =   0.3965              4.0       5.0       6.0       7.0

Tukey's pairwise comparisons                                    ②
    Family error rate = 0.100
Individual error rate = 0.0315

Critical value = 4.59

Intervals for (column level mean) - (row level mean)

           A           A+B         B

 A+B         -3.1784
             -0.6046

 B           -2.1619     -0.2704
              0.4119      2.3034

 control     -1.0419      0.8496     -0.1669
              1.5319      3.4234      2.4069

Fisher's pairwise comparisons                                   ③
    Family error rate = 0.283
Individual error rate = 0.100

Critical value = 2.132

Intervals for (column level mean) - (row level mean)

                   A         A+B           B

 A+B         -2.7369
             -1.0461

 B           -1.7204      0.1711
             -0.0296      1.8619

 control     -0.6004      1.2911      0.2746
              1.0904      2.9819      1.9654
```

has the advantage that we can use larger critical values for long stretches and smaller critical values for short stretches.

$(i)$ and $(j)$ are different if their stretch and all containing stretches reject

Begin with the most extreme pair $(1)$ and $(g)$. Test the null hypothesis that all the means for $(1)$ up through $(g)$ are equal. If you fail to reject, declare all means equal and stop. If you reject, declare $(1)$ different from $(g)$ and go on to the next step. At the next step, we consider the stretches $(1)$ through $(g-1)$ and $(2)$ through $(g)$. If one of these rejects, we declare its ends to be different and then look at shorter stretches within it. If we fail to reject for a stretch, we do not consider any substretches within the stretch. We repeat this subdivision till there are no more rejections. In other words, we declare that means $(i)$ and $(j)$ are different if the stretch from $(i)$ to $(j)$ rejects its null hypothesis and all stretches containing $(i)$ to $(j)$ also reject their null hypotheses.

REGWR is step-down with Studentized range based critical values

The REGWR procedure is a step-down range method that controls the strong familywise error rate without producing simultaneous confidence intervals. The awkward name REGWR abbreviates the Ryan-Einot-Gabriel-Welsch range test, named for the authors who worked on it. The REGWR critical value for testing a stretch of length $k$ depends on $\mathcal{E}$, $\nu$, $k$, and $g$. Specifically, we use

$$u = u(\mathcal{E}, \nu, k, g) = q_{\mathcal{E}}(k, \nu)/\sqrt{2} \quad k = g, g-1,$$

and

$$u = u(\mathcal{E}, \nu, k, g) = q_{k\mathcal{E}/g}(k, \nu)/\sqrt{2} \quad k = g-2, g-3, \ldots, 2.$$

This critical value derives from a Studentized range with $k$ groups, and we use percent points with smaller tail areas as we move in to smaller stretches.

As with the HSD, REGWR error rate control is approximate when the sample sizes are not equal.

**Example 5.6**

### Free amino acids in cheese, continued

Suppose that we only wished to control the strong familywise error rate instead of producing simultaneous confidence intervals. Then we could use REGWR instead of HSD and could potentially see additional significant differences. Listing 5.2 ② gives SAS output for REGWR (called REGWQ in SAS) for the amino acid data.

REGWR is a step-down method that begins like the HSD. Comparing C and A&B, we conclude as in the HSD that they are different. We may now compare C with B and A with A&B. These are comparisons that involve

**Listing 5.2:** SAS output for free amino acids in cheese.

```
              Student-Newman-Keuls test for variable: FAA                ①

                    Alpha= 0.1  df= 4  MSE= 0.157224

           Number of Means          2          3          4
           Critical Range    0.84531 1.1146718 1.2859073

     Means with the same letter are not significantly different.

            SNK Grouping              Mean      N  TRT

                       A            6.3215      2  4

                       B            5.3050      2  3

                       C            4.4300      2  2
                       C
                       C            4.1850      2  1


   Ryan-Einot-Gabriel-Welsch Multiple Range Test for variable: FAA       ②

                    Alpha= 0.1  df= 4  MSE= 0.157224

           Number of Means          2          3          4
           Critical Range  1.0908529 1.1146718 1.2859073

     Means with the same letter are not significantly different.

           REGWQ Grouping              Mean      N  TRT

                       A            6.3215      2  4
                       A
                  B    A            5.3050      2  3
                  B
                  B    C            4.4300      2  2
                       C
                       C            4.1850      2  1
```

stretches of $k = 3$ means; since $k = g - 1$, we still use $\mathcal{E}$ as the error rate. The significant difference for these comparisons is

$$\frac{q_{\mathcal{E}}(k,\nu)}{\sqrt{2}}\sqrt{MS_E}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} = \frac{q_{.1}(3,4)}{\sqrt{2}}\sqrt{.1572}\sqrt{\frac{1}{2}+\frac{1}{2}} = 1.115 \ .$$

Both the B-C and A&B-A differences (1.12 and 1.89) exceed this cutoff, so REGWR concludes that B differs from C, and A differs from A&B. Recall that the HSD did not distinguish C from B.

Having concluded that there are B-C and A&B-A differences, we can now compare stretches of means within them, namely C to A, A to B, and B to A&B. These are stretches of $k = 2$ means, so for REGWR we use the error rate $k\mathcal{E}/g = .05$. The significant difference for these comparisons is

$$\frac{q_{\mathcal{E}/2}(k,\nu)}{\sqrt{2}}\sqrt{MS_E}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} = \frac{q_{.05}(2,4)}{\sqrt{2}}\sqrt{.1572}\sqrt{\frac{1}{2}+\frac{1}{2}} = 1.101 \quad .$$

None of the three differences exceeds this cutoff, so we fail to conclude that those treatments differ and finish. The underline diagram is:

| C | A | B | A&B |
|---|---|---|---|
| 4.19 | 4.43 | 5.31 | 6.32 |

Note in Listing 5.2 ② that SAS displays pairwise comparisons using what amounts to an underline diagram turned on its side, with vertical lines formed by letters.

### 5.4.5 False discovery rate

SNK

The Student-Newman-Keuls (SNK) procedure is a step-down method that uses the Studentized range test with critical value

$$u = u(\mathcal{E},\nu,k,g) = q_{\mathcal{E}}(k,\nu)/\sqrt{2}$$

for a stretch of $k$ means. This is similar to REGWR, except that we keep the percent point of the Studentized range constant as we go to shorter stretches. The SNK controls the false discovery rate, but not the strong familywise error rate. As with the HSD, SNK error rate control is approximate when the sample sizes are not equal.

**Example 5.7**

**Free amino acids in cheese, continued**

Suppose that we only wished to control the false discovery rate; now we would use SNK instead of the more stringent HSD or REGWR. Listing 5.2 ① gives SAS output for SNK for the amino acid data.

SNK is identical to REGWR in the first two stages, so SNK will also get to the point of making the comparisons of the three pairs C to A, A to B, and

B to A&B. However, the SNK significant difference for these pairs is less than that used in REGWR:

$$\frac{q_{\mathcal{E}}(k,\nu)}{\sqrt{2}}\sqrt{MS_E}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} = \frac{q_{.1}(2,4)}{\sqrt{2}}\sqrt{.1572}\sqrt{\frac{1}{2}+\frac{1}{2}} = .845 \ .$$

Both the B-A and A&B-B differences (1.02 and .98) exceed the cutoff, but the A-C difference (.14) does not. The underline diagram for SNK is:

| C | A | B | A&B |
|---|---|---|---|
| 4.19 | 4.43 | 5.31 | 6.32 |

### 5.4.6 Experimentwise error rate

The Analysis of Variance F-test for equality of means controls the experimentwise error rate. Thus investigating pairwise differences only when the F-test has a $p$-value less than $\mathcal{E}$ will control the experimentwise error rate. This is the basis for the Protected least significant difference, or Protected LSD. If the F-test rejects at level $\mathcal{E}$, then do simple $t$-tests at level $\mathcal{E}$ among the different treatments.

Protected LSD uses F-test to control experimentwise error rate

The critical values are from a $t$-distribution:

$$u(\mathcal{E},\nu) = t_{\mathcal{E}/2,\nu} \ ,$$

leading to the significant difference

$$LSD = t_{\mathcal{E}/2,\nu}\sqrt{MS_E}\sqrt{1/n_i + 1/n_j} \ .$$

As usual, $\nu$ is the degrees of freedom for $MS_E$, and $t_{\mathcal{E}/2,\nu}$ is the upper $\mathcal{E}/2$ percent point of a $t$-curve with $\nu$ degrees of freedom.

Confidence intervals produced from the protected LSD do not have the anticipated $1-\mathcal{E}$ coverage rate, either individually or simultaneously. See Section 5.7.

**Free amino acids in cheese, continued** **Example 5.8**

Finally, suppose that we only wish to control the experimentwise error rate. Protected LSD will work here. Listing 5.1 ① shows that the ANOVA F-test is significant at level $\mathcal{E}$, so we may proceed with pairwise comparisons.

Listing 5.1 ③ shows Minitab output for the LSD (called Fisher's pairwise comparisons) as confidence intervals.

LSD uses the same significant difference for all pairs:

$$t_{\mathcal{E}/2,\nu}\sqrt{MS_E}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} = t_{.05,4}\sqrt{.1572}\sqrt{\frac{1}{2}+\frac{1}{2}} = .845 \ .$$

This is the same as the SNK comparison for a stretch of length 2. All differences except A-C exceed the cutoff, so the underline diagram for LSD is:

| C | A | B | A&B |
|---|---|---|---|
| 4.19 | 4.43 | 5.31 | 6.32 |

### 5.4.7 Comparisonwise error rate

LSD

Ordinary $t$-tests and confidence intervals without any adjustment control the comparisonwise error rate. In the context of pairwise comparisons, this is called the least significant difference (LSD) method.

The critical values are the same as for the protected LSD:

$$u(\mathcal{E},\nu) = t_{\mathcal{E}/2,\nu},$$

and

$$LSD = t_{\mathcal{E}/2,\nu}\sqrt{MS_E}\sqrt{1/n_i + 1/n_j} \ .$$

### 5.4.8 Pairwise testing reprise

Choose your error rate, not your method

It is easy to get overwhelmed by the abundance of methods, and there are still more that we haven't discussed. Your anchor in all this is your error rate. Once you have determined your error rate, the choice of method is reasonably automatic, as summarized in Display 5.2. Your choice of error rate is determined by the needs of your study, bearing in mind that the more stringent error rates have fewer false rejections, and also fewer correct rejections.

### 5.4.9 Pairwise comparisons methods that do *not* control combined Type I error rates

There are many other pairwise comparisons methods beyond those already mentioned. In this Section we discuss two methods that are motivated by

| Error rate | Method |
|---|---|
| Simultaneous confidence intervals | BSD or HSD |
| Strong familywise | REGWR |
| False discovery rate | SNK |
| Experimentwise | Protected LSD |
| Comparisonwise | LSD |

**Display 5.2:** Summary of pairwise comparison methods.

completely different criteria than controlling a combined Type I error rate. These two techniques do *not* control the experimentwise error rate or any of the more stringent error rates, and you should not use them with the expectation that they do. You should only use them when the situation and assumptions under which they were developed are appropriate for your experimental analysis.

Duncan's multiple range if there is a cost per error or you believe $H_0$ less likely as $g$ increases

Suppose that you believe *a priori* that the overall null hypothesis $H_0$ is less and less likely to be true as the number of treatments increases. Then the strength of evidence required to reject $H_0$ should decrease as the number of groups increases. Alternatively, suppose that there is a quantifiable penalty for each incorrect (pairwise comparison) decision we make, and that the total loss for the overall test is the sum of the losses from the individual decisions. Under either of these assumptions, the Duncan multiple range (given below) or something like it is appropriate. Note by comparison that the procedures that control combined Type I error rates require more evidence to reject $H_0$ as the number of groups increases, while Duncan's method requires less. Also, a procedure that controls the experimentwise error rate has a penalty of 1 if there are any rejections when $H_0$ is true and a penalty of 0 otherwise; this is very different from the summed loss that leads to Duncan's multiple range.

Duncan's Multiple Range

Duncan's multiple range (sometimes called Duncan's test or Duncan's new multiple range) is a step-down Studentized range method. You specify a "protection level" $\mathcal{E}$ and proceed in step-down fashion using

$$u = u(\mathcal{E}, \nu, k, g) = q_{1-(1-\mathcal{E})^{k-1}}(k, \nu)/\sqrt{2}$$

Experimentwise error rate very large for Duncan

for the critical values. Notice that $\mathcal{E}$ is the comparisonwise error rate for testing a stretch of length 2, and the experimentwise error rate will be $1-(1-\mathcal{E})^{g-1}$, which can be considerably more than $\mathcal{E}$. Thus *fixing Duncan's protection level at $\mathcal{E}$ does **not** control the experimentwise error rate or any more stringent rate*. Do not use Duncan's procedure if you are interested in controlling any of the combined Type I error rates.

Minimize prediction error instead of testing

As a second alternative to combined Type I error rates, suppose that our interest is in predicting future observations from the treatment groups, and that we would like to have a prediction method that makes the average squared prediction error small. One way to do this prediction is to first partition the $g$ treatments into $p$ classes, $1 \leq p \leq g$; second, find the average response in each of these $p$ classes; and third, predict a future observation from a treatment by the observed mean response of the class for the treatment. We thus look for partitions that will lead to good predictions.

One way to choose among the partitions is to use Mallows' $C_p$ statistic:

$$C_p = \frac{SSR_p}{MS_E} + 2p - N \ ,$$

Predictive Pairwise Comparisons

where $SSR_p$ is the sum of squared errors for the Analysis of Variance, partitioning the data into $p$ groups. Partitions with low values of $C_p$ should give better predictions.

This predictive approach makes no attempt to control any Type I error rate; in fact, the Type I error rate is .15 or greater even for $g = 2$ groups! This approach is useful when prediction is the goal, but can be quite misleading if interpreted as a test of $H_0$.

### 5.4.10 Confident directions

All means differ, but their order is uncertain

In our heart of hearts, we often believe that all treatment means differ when examined sufficiently precisely. Thus our concern with null hypotheses $H_{0ij}$ is misplaced. As an alternative, we can make statements of *direction*. After having collected data, we consider $\mu_i$ and $\mu_j$; assume $\mu_i < \mu_j$. We could decide from the data that $\mu_i < \mu_j$, or that $\mu_i > \mu_j$, or that we don't know—that is, we don't have enough information to decide. These decisions correspond in the testing paradigm to rejecting $H_{0ij}$ in favor of $\mu_i < \mu_j$, rejecting $H_{0ij}$ in favor of $\mu_j < \mu_i$, and failing to reject $H_{0ij}$. In the confident directions framework, only the decision $\mu_i > \mu_j$ is an error. See Tukey (1991).

Can only make an error in one direction

*Confident directions* procedures are pairwise comparisons testing procedures, but with results interpreted in a directional context. Confident directions procedures bound error rates when making statements about direction.

If a testing procedure bounds an error rate at $\mathcal{E}$, then the corresponding confident directions procedure bounds a confident directions error rate at $\mathcal{E}/2$, the factor of 2 arising because we cannot falsely reject in the correct direction.

Let us reinterpret our usual error rates in terms of directions. Suppose that we use a pairwise comparisons procedure with error rate bounded at $\mathcal{E}$. In a confident directions setting, we have the following:

Pairwise comparisons can be used for confident directions

| | |
|---|---|
| Strong familywise | The probability of making any incorrect statements of direction is bounded by $\mathcal{E}/2$. |
| FDR | Incorrect statements of direction will on average be no more than a fraction $\mathcal{E}/2$ of the total number of statements of direction. |
| Experimentwise | The probability of making any incorrect statements of direction when all the means are very nearly equal is bounded by $\mathcal{E}/2$. |
| Comparisonwise | The probability of making an incorrect statement of direction for a given comparison is bounded by $\mathcal{E}/2$. |

There is no directional analog of simultaneous confidence intervals, so procedures that produce simultaneous intervals should be considered procedures that control the strong familywise error rate (which they do).

## 5.5 Comparison with Control or the Best

There are some situations where we do not do all pairwise comparisons, but rather make comparisons between a control and the other treatments, or the best responding treatment (highest or lowest average) and the other treatments. For example, you may be producing new standardized mathematics tests for elementary school children, and you need to compare the new tests with the current test to assure comparability of the results. The procedures for comparing to a control or the best are similar.

Comparison with control does not do all tests

### 5.5.1 Comparison with a control

Suppose that there is a special treatment, say treatment $g$, with which we wish to compare the other $g-1$ treatments. Typically, treatment $g$ is a control treatment. The Dunnett procedure allows us to construct simultaneous $1-\mathcal{E}$ confidence intervals on $\mu_i - \mu_g$, for $i = 1, \ldots, g-1$ when all sample sizes are equal via

Two-sided Dunnett

$$\overline{y}_i - \overline{y}_g \pm d_{\mathcal{E}}(g-1,\nu)\ \sqrt{MS_E}\ \sqrt{\frac{1}{n_i}+\frac{1}{n_g}}\ ,$$

where $\nu$ is the degrees of freedom for $MS_E$. The value $d_{\mathcal{E}}(g-1,\nu)$ is tabulated in Appendix Table D.9. These table values are exact when all sample sizes are equal and only approximate when the sizes are not equal.

For testing, we can use

$$u(\mathcal{E},i,j) = d_{\mathcal{E}}(g-1,\nu)\ ,$$

DSD, the Dunnett significant difference

which controls the strong familywise error rate and leads to

$$DSD = d_{\mathcal{E}}(g-1,\nu)\sqrt{MS_E}\ \sqrt{\frac{1}{n_i}+\frac{1}{n_g}}\ ,$$

the Dunnett significant difference. There is also a step-down modification that still controls the strong familywise error rate and is slightly more powerful. We have $g-1$ $t$-statistics. Compare the largest (in absolute value) to $d_{\mathcal{E}}(g-1,\nu)$. If the test fails to reject the null, stop; otherwise compare the second largest to $d_{\mathcal{E}}(g-2,\nu)$ and so on.

One-sided Dunnett

There are also one-sided versions of the confidence and testing procedures. For example, you might reject the null hypothesis of equality only if the noncontrol treatments provide a higher response than the control treatments. For these, test using the critical value

$$u(\mathcal{E},i,j) = d'_{\mathcal{E}}(g-1,\nu)\ ,$$

tabulated in Appendix Table D.9, or form simultaneous one-sided confidence intervals on $\mu_i - \mu_g$ with

$$\overline{y}_i - \overline{y}_g \geq d'_{\mathcal{E}}(g-1,\nu)\ \sqrt{MS_E}\ \sqrt{\frac{1}{n_i}+\frac{1}{n_g}}.$$

For $t$-critical values, a one-sided cutoff is equal to a two-sided cutoff with a doubled $\mathcal{E}$. The same is not true for Dunnett critical values, so that $d'_{\mathcal{E}}(g-1,\nu) \neq d_{2\mathcal{E}}(g-1,\nu)$.

**Example 5.9**

### Alfalfa meal and turkeys

An experiment is conducted to study the effect of alfalfa meal in the diet of male turkey poults (chicks). There are nine treatments. Treatment 1 is a control treatment; treatments 2 through 9 contain alfalfa meal of two different types in differing proportions. Units consist of 72 pens of eight birds each, so there are eight pens per treatment. One response of interest is average daily weight gains per bird for birds aged 7 to 14 days. We would like to know which alfalfa treatments are significantly different from the control in weight gain, and which are not.

Here are the average weight gains (g/day) for the nine treatments:

| | | | | |
|---|---|---|---|---|
| 22.668 | 21.542 | 20.001 | 19.964 | 20.893 |
| 21.946 | 19.965 | 20.062 | 21.450 | |

The $MS_E$ is 2.487 with 55 degrees of freedom. (The observant student will find this degrees of freedom curious; more on this data set later.) Two-sided, 95% confidence intervals for the differences between control and the other treatments are computed using

$$\begin{aligned} d_{\mathcal{E}}(g-1,\nu)\,\sqrt{MS_E}\,\sqrt{\frac{1}{n_i}+\frac{1}{n_g}} &= d_{.05}(8,55)\,\sqrt{2.487}\,\sqrt{\frac{1}{8}+\frac{1}{8}} \\ &= 2.74 \times 1.577/2 \\ &= 2.16\ . \end{aligned}$$

Any treatment with mean less than 2.16 from the control mean of 22.668 is not significantly different from the control. These are treatments 2, 5, 6, and 9.

Give the control more replication

It is a good idea to give the control (treatment $g$) greater replication than the other treatments. The control is involved in every comparison, so it makes sense to estimate its mean more precisely. More specifically, if you had a fixed number of units to spread among the treatments, and you wished to minimize the average variance of the differences $\overline{y}_{g\bullet} - \overline{y}_{i\bullet}$, then you would do best when the ratio $n_g/n_i$ is about equal to $\sqrt{g-1}$.

Personally, I rarely use the Dunnett procedure, because I nearly always get the itch to compare the noncontrol treatments with each other as well as with the control.

### 5.5.2 Comparison with the best

Use MCB to choose best subset of treatments

Suppose that the goal of our experiment is to screen a number of treatments and determine those that give the best response—to pick the winner. The multiple comparisons with best (MCB) procedure produces two results:

- It produces a subset of treatments that cannot be distinguished from the best; the treatment having the true largest mean response will be in this subset with probability $1 - \mathcal{E}$.
- It produces simultaneous $1-\mathcal{E}$ confidence intervals on $\mu_i - \max_{j \neq i} \mu_j$, the difference between a treatment mean and the best of the other treatment means.

The subset selection procedure is the more useful product, so we only discuss the selection procedure.

The best subset consists of all treatments $i$ such that

$$\overline{y}_{i\bullet} > \overline{y}_{j\bullet} - d'_{\mathcal{E}}(g-1,\nu)\ \sqrt{MS_E}\ \sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\ \text{ for all } j \neq i$$

In words, treatment $i$ is in the best subset if its mean response is greater than the largest treatment mean less a one-sided Dunnett allowance. When small responses are good, a treatment $i$ is in the best subset if its mean response is less than the smallest treatment mean plus a one-sided Dunnett allowance.

**Example 5.10**

**Weed control in soybeans**

Weeds reduce crop yields, so farmers are always looking for better ways to control weeds. Fourteen weed control treatments were randomized to 56 experimental plots that were planted in soybeans. The plots were later visually assessed for weed control, the fraction of the plot without weeds. The percent responses are given in Table 5.3. We are interested in finding a subset of treatments that contains the treatment giving the best weed control (largest response) with confidence 99%.

For reasons that will be explained in Chapter 6, we will analyze as our response the square root of percent weeds (that is, 100 minus the percent weed control). Because we have subtracted weed control, small values of the transformed response are good. On this scale, the fourteen treatment means are

| | | | | | | |
|---|---|---|---|---|---|---|
| 1.000 | 2.616 | 2.680 | 2.543 | 2.941 | 1.413 | 1.618 |
| 2.519 | 2.847 | 1.618 | 1.000 | 4.115 | 4.988 | 5.755 |

**Table 5.3:** Percent weed control in soybeans under 14 treatments.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 99 | 95 | 92 | 95 | 85 | 98 | 99 |
| 99 | 92 | 95 | 88 | 92 | 99 | 95 |
| 99 | 95 | 92 | 95 | 92 | 95 | 99 |
| 99 | 90 | 92 | 95 | 95 | 99 | 95 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 95 | 92 | 99 | 99 | 88 | 65 | 75 |
| 85 | 90 | 95 | 99 | 88 | 65 | 50 |
| 95 | 95 | 99 | 99 | 85 | 92 | 72 |
| 97 | 90 | 95 | 99 | 68 | 72 | 68 |

and the $MS_E$ is .547 with 42 degrees of freedom. The smallest treatment mean is 1.000, and the Dunnett allowance is

$$\begin{aligned} d'_{\mathcal{E}}(g-1,\nu)\,\sqrt{MS_E}\,\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} &= d'_{.01}(13,42)\,\sqrt{.547}\,\sqrt{\frac{1}{4}+\frac{1}{4}} \\ &= 3.29 \times .740 \times .707 \\ &= 1.72. \end{aligned}$$

So, any treatment with a mean of $1 + 1.72 = 2.72$ or less is included in the 99% grouping. These are treatments 1, 2, 3, 4, 6, 7, 8, 10, and 11.

## 5.6 Reality Check on Coverage Rates

We already pointed out that the error rate control for some multiple comparisons procedures is only approximate if the sample sizes are not equal or the tests are dependent. However, even in the "exact" situations, these procedures depend on assumptions about the distribution of the data for the coverage rates to hold: for example normality or constant error variance. These assumptions are often violated—data are frequently nonnormal and error variances are often nonconstant.

Violation of distributional assumptions usually leads to true error rates that are not equal to the nominal $\mathcal{E}$. The amount of discrepancy depends on the nature of the violation. Unequal sample sizes or dependent tests are just another variable to consider.

The point is that we need to get some idea of what the true error is, and not get worked up about the fact that it is not *exactly* equal to $\mathcal{E}$.

| In the real world, coverage and error rates are always approximate. |
|---|

## 5.7 A Warning About Conditioning

Except for the protected LSD, the multiple comparisons procedures discussed above do not require the ANOVA F-test to be significant for protection of the experimentwise error rate. They stand apart from the F-test, protecting the experimentwise error rate by other means. In fact, requiring that the ANOVA F-test be significant will alter their error rates.

Requiring the F-test to be significant alters the error rates of pairwise procedures

Bernhardson (1975) reported on how conditioning on the ANOVA F-test being significant affected the per comparison and per experiment error rates of pairwise comparisons, including LSD, HSD, SNK, Duncan's procedure, and Scheffé. Requiring the F to be significant lowered the per comparison error rate of the LSD from 5% to about 1% and lowered the per experiment error rate for HSD from 5% to about 3%, both for 6 to 10 groups. Looking just at those null cases where the F-test rejected, the LSD had a per comparison error rate of 20 to 30% and the HSD per experiment error rate was about 65%—both for 6 to 10 groups. Again looking at just the null cases where the F was significant, even the Scheffé procedure's per experiment error rate increased to 49% for 4 groups, 22% for 6 groups, and down to about 6% for 10 groups.

The problem is that when the ANOVA F-test is significant in the null case, one cause might be an unusually low estimate of the error variance. This unusually low variance estimate gets used in the multiple comparisons procedures leading to smaller than normal HSD's, and so on.

## 5.8 Some Controversy

Simultaneous inference is deciding which error rate to control and then using an appropriate technique for that error rate. Controversy arises because

- Users cannot always agree on the appropriate error rate. In particular, some statisticians (including Bayesian statisticians) argue strongly that the only relevant error rate is the per comparison error rate.

- Users cannot always agree on what constitutes the appropriate family of tests. Different groupings of the tests lead to different results.
- Standard statistical practice seems to be inconsistent in its application of multiple comparisons ideas. For example, multiple comparisons are fairly common when comparing treatment means, but almost unheard of when examining multiple factors in factorial designs (see Chapter 8).

You as experimenter and data analyst must decide what is the proper approach for inference. See Carmer and Walker (1982) for an amusing allegory on this topic.

## 5.9 Further Reading and Extensions

There is much more to the subject of multiple comparisons than what we have discussed here. For example, many procedures for contrasts can be adapted to other linear combinations of parameters, and many of the pairwise comparisons techniques can be adapted to contrasts. A good place to start is Miller (1981), an instant classic when it appeared and still an excellent and readable reference; much of the discussion here follows Miller. Hochberg and Tamhane (1987) contains some of the more recent developments.

The first multiple comparisons technique appears to be the LSD suggested by Fisher (1935). Curiously, the next proposal was the SNK (though not so labeled) by Newman (1939). Multiple comparisons then lay dormant till around 1950, when there was an explosion of ideas: Duncan's multiple range procedure (Duncan 1955), Tukey's HSD (Tukey 1952), Scheffé's all contrasts method (Scheffé 1953), Dunnett's method (Dunnett 1955), and another proposal for SNK (Keuls 1952). The pace of introduction then slowed again. The REGW procedures appeared in 1960 and evolved through the 1970's (Ryan 1960; Einot and Gabriel 1975; Welsch 1977). Improvements in the Bonferroni inequality lead to the modified Bonferroni procedures in the 1970's and later (Holm 1979; Simes 1986; Hochberg 1988; Benjamini and Hochberg 1995).

Curiously, procedures sometimes predate a careful understanding of the error rates they control. For example, SNK has often been advocated as a less conservative alternative to the HSD, but the false discovery rate was only defined recently (Benjamini and Hochberg 1995). Furthermore, many textbook introductions to multiple comparisons procedures do not discuss the different error rates, thus leading to considerable confusion over the choice of procedure.

One historical feature of multiple comparisons is the heavy reliance on tables of critical values and the limitations imposed by having tables only for selected percent points or equal sample sizes. Computers and software remove many of these limitations. For example, the software in Lund and Lund (1983) can be used to compute percent points of the Studentized range for $\mathcal{E}$'s not usually tabulated, while the software in Dunnett (1989) can compute critical values for the Dunnett test with unequal sample sizes. When no software for exact computation is available (for example, Studentized range for unequal sample sizes), percent points can be approximated through simulation (see, for example, Ripley 1987).

Hayter (1984) has shown that the Tukey-Kramer adjustment to the HSD procedure is conservative when the sample sizes are not equal.

## 5.10 Problems

**Exercise 5.1** We have five groups and three observations per group. The group means are 6.5, 4.5, 5.7, 5.6, and 5.1, and the mean square for error is .75. Compute simultaneous confidence intervals (95% level) for the differences of all treatment pairs.

**Exercise 5.2** Consider a completely randomized design with five treatments, four units per treatment, and treatment means

$$3.2892 \quad 10.256 \quad 8.1157 \quad 8.1825 \quad 7.5622 \quad .$$

The MSE is 4.012.

(a) Construct an ANOVA table for this experiment and test the null hypothesis that all treatments have the same mean.

(b) Test the null hypothesis that the average response in treatments 1 and 2 is the same as the average response in treatments 3, 4, and 5.

(c) Use the HSD procedure to compare the means of the five treatments.

**Exercise 5.3** Refer to the data in Problem 3.1. Test the null hypothesis that all pairs of workers produce solder joints with the same average strength against the alternative that some workers produce different average strengths. Control the strong familywise error rate at .05.

**Exercise 5.4** Refer to the data in Exercise 3.1. Test the null hypothesis that all pairs of diets produce the same average weight liver against the alternative that some diets produce different average weights. Control the FDR at .05.

**Exercise 5.5**

Use the data from Exercise 3.3. Compute 95% simultaneous confidence intervals for the differences in response between the the three treatment groups (acid, pulp, and salt) and the control group.

**Problem 5.1**

Use the data from Problem 3.2. Use the Tukey procedure to make all pairwise comparisons between the treatment groups. Summarize your results with an underline diagram.

**Problem 5.2**

In an experiment with four groups, each with five observations, the group means are 12, 16, 21, and 19, and the MSE is 20. A colleague points out that the contrast with coefficients -4, -2, 3, 3 has a rather large sum of squares. No one knows to begin with why this contrast has a large sum of squares, but after some detective work, you discover that the contrast coefficients are roughly the same (except for the overall mean) as the time the samples had to wait in the lab before being analyzed (3, 5, 10, and 10 days). What is the significance of this contrast?

**Problem 5.3**

Consider an experiment taste-testing six types of chocolate chip cookies: 1 (brand A, chewy, expensive), 2 (brand A, crispy, expensive), 3 (brand B, chewy, inexpensive), 4 (brand B, crispy, inexpensive), 5 (brand C, chewy, expensive), 6 (brand D, crispy, inexpensive). We will use twenty different raters randomly assigned to each type (120 total raters). I have constructed five preplanned contrasts for these treatments, and I obtain $p$-values of .03, .04, .23, .47, and .68 for these contrasts. Discuss how you would assess the statistical significance of these contrasts, including what issues need to be resolved.

**Question 5.1**

In an experiment with five groups and 25 degrees of freedom for error, for what numbers of contrasts is the Bonferroni procedure more powerful than the Scheffé procedure?