

# Unit 28

## Scheffe's Pair Wise Comparison of Means

Objectives:

- To understand why pair wise multiple comparison with an overall significance level  $\alpha$  is preferable to individual pair wise comparisons each with the significance level  $\alpha$
- To perform *Scheffe's method of pair wise multiple comparison* when the null hypothesis in a one-way analysis of variance is rejected

When we do not reject the null hypothesis in a one-way ANOVA, no further analysis is called for, since we are concluding that several population means are equal. Since we did not reject the null hypothesis in the one-way ANOVA with the data of Table 27-1 (see Table 27-4), no further analysis is called for. On the other hand, rejecting the null hypothesis in a one-way ANOVA prompts us to do some further analysis in order to identify which population means are different from each other. Identifying significant differences among more than two population means is a type of *multiple comparison*.

One might be tempted to identify significant differences among means by simply applying a two sample  $t$  test on each pair. In doing this, however, the overall significance level will be larger than the individual significance level in each one of the two sample  $t$  tests. This is because the probability of incorrectly rejecting a null hypothesis increases as the number of hypothesis tests performed increases. For instance, if we were to perform three two sample  $t$  tests each with a 0.05 significance level, the overall significance level for all three tests taken together could be as high as 0.14.

To understand why this occurs, consider the probability of observing a head when you flip a fair coin. If you flip the coin only once, you have a  $1/2$  probability of observing a head; however, if you flip the coin three times, the probability that you observe at least one head will be larger than  $1/2$ , and if you flip the coin ten times, it is almost a certainty that you will observe at least one head. This tells us that if we perform enough hypothesis tests simultaneously each with a given significance level (say, 0.05), we can expect a very high probability of incorrectly rejecting at least one null hypothesis. This is often overlooked by researchers who perform several simultaneous hypothesis tests.

Multiple comparisons is used to maintain control over the overall significance level. While there are many multiple comparison methods available, we shall introduce only one method here, known as *Scheffe's method of pair wise multiple comparison of means*. In its most general form, Scheffe's method of multiple comparison is very general, implying that it can be used for the widest variety of different types of comparisons. Here, we shall use it only for pair wise comparison of means.

There are three steps to applying Scheffe's pair wise multiple comparison method. The first step is to obtain the absolute values of pair wise differences between sample means. The second step is to obtain values against which the pair wise differences between sample means are compared and identify the statistically significant pair wise differences. The third and last step is to state the results. Table 28-2 displays in detail these three steps for Scheffe's pair wise multiple comparison method.

To illustrate Scheffe's pair wise multiple comparison method, we shall first use a one-way ANOVA with the data of Table 28-1 to see if there is any evidence of a difference in the mean height of wheat among four different types of soil labeled C, G, H, and T. We shall choose a 0.05 significance level for the one-way ANOVA.

<b>Table 28-1</b>					
<b>Wheat Height (inches) of with</b>					
<b>Four Soil Types</b>					
<u>Soil Type C</u>	35.8	35.9	36.1	36.0	35.7
<u>Soil Type G</u>	35.9	35.8	35.6	35.7	35.8
<u>Soil Type H</u>	35.6	35.5	35.8	35.5	
<u>Soil Type T</u>	36.0	35.9	36.0	36.1	36.2
-----					
<u>Soil Type C</u>	$n_C =$	_____	$\bar{x}_C =$	_____	
<u>Soil Type G</u>	$n_G =$	_____	$\bar{x}_G =$	_____	
<u>Soil Type H</u>	$n_H =$	_____	$\bar{x}_H =$	_____	
<u>Soil Type T</u>	$n_T =$	_____	$\bar{x}_T =$	_____	

**Table 28-2**

**Scheffe’s Method of Pair Wise Multiple Comparison of Means**

- (1) Obtain the absolute values of pair wise differences between sample means  $|\bar{x}_i - \bar{x}_j|$ .
- (2) Identify statistically significant differences by finding those with an absolute value larger than the corresponding comparison value

$$\sqrt{(k - 1)f_{k-1, n-k; \alpha} \text{MSE} (1/n_i + 1/n_j)}$$

- (3) State the significance level chosen and the direction of statistically significant differences between pairs of means.

The first step is to state the null and alternative hypotheses, and choose a significance level. We complete the first step of the hypothesis test as follows:

$$H_0: \mu_C = \mu_G = \mu_H = \mu_T \text{ vs. } H_1: \text{At least one of } \mu_C, \mu_G, \mu_H, \text{ and } \mu_T \text{ is different } (\alpha = 0.05).$$

We complete the second step of the hypothesis test by calculating the *f* test statistic and constructing the ANOVA table. The sample sizes and sample means are displayed at the bottom of Table 28-1. With the appropriate statistical software or programmable calculator, you may verify that these sample means are correct and that the ANOVA table displayed as Table 28-3 is correct. From the ANOVA table, we obtain the following *f* statistic:

$$f_{3, 16} = 9.89.$$

The third step is to define the rejection region, decide whether or not to reject the null hypothesis, and obtain the *p*-value of the test. Our rejection region is defined by the *f*-score above which lies 0.05 of the area under the *f* density curve with 3 numerator degrees of freedom and 16 denominator degrees of freedom; from Table A.4, we see that this *f*-score is  $f_{3, 16; 0.05} = 3.24$ . We then define our rejection region algebraically as

$$f_{3, 16} \geq 3.24.$$

Graphically, we can picture the rejection region as corresponding to the shaded area in the figure at the top of the first page of Table A.4, where the value on the horizontal axis defining the rejection region is  $f_{3, 16; 0.05} = 3.24$ . Since our test statistic  $f_{3, 16} = 9.89$  is in the rejection region, our decision is to reject  $H_0: \mu_C = \mu_G = \mu_H = \mu_T$ ; in other words, our data provides sufficient evidence to suggest at least one difference among the means.

**Table 28-3**

**One-Way ANOVA Table for the Data of Table 28-1**

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>	<i>p-value</i>
Soil Type	0.495	3	0.16500	9.89	< 0.001
Error	0.267	16	0.01669		
Total	0.762	19			

The  $p$ -value of this hypothesis test is the probability of obtaining a test statistic value  $f_{3,16}$  which represents greater differences among the sample means than does the value actually observed  $f_{3,16} = 9.89$ . Graphically, we can picture the  $p$ -value of this hypothesis test as corresponding to the shaded area in the figure at the top of the first page of Table A.4, where the value on the horizontal axis is our observed test statistic value  $f_{3,16} = 9.89$ . By looking at the entries of Table A.4 corresponding to 3 numerator degrees of freedom and 16 denominator degrees of freedom, we find that the observed test statistic  $f_{3,16} = 9.89$  is greater than  $f_{3,16;0.001} = 9.01$ . This tells us that the area above  $f_{3,16} = 9.89$ , which is the  $p$ -value, is less than 0.001. We indicate this by writing  $p\text{-value} < 0.001$ . The fact that  $p\text{-value} < 0.001$  confirms to us that  $H_0$  is rejected with  $\alpha = 0.05$ . However, this also tells us that  $H_0$  would be rejected with  $\alpha = 0.01$  of course with  $\alpha = 0.10$ .

To complete the fourth step of the hypothesis test, we can summarize the results of the hypothesis test as follows:

Since  $f_{3,16} = 9.89$  and  $f_{3,16;0.001} = 3.24$ , we have sufficient evidence to reject  $H_0$ . We conclude that there is at least one difference in the mean height of wheat among the four different types of soil labeled C, G, H, and T ( $p\text{-value} < 0.001$ ). Since  $H_0$  is rejected, we need to use multiple comparison to identify significant differences between the means.

Notice that in the statement of results, the last sentence indicates that further analysis is necessary; specifically, we need to use multiple comparison to identify the soil types for which the mean height is different. We shall use Scheffe's pair wise multiple comparison method.

**Table 28-4**  
**Step 1 of Scheffe's Method for the Data of Table 28-1**

Absolute Values of Pair Wise Differences Between Sample Means $ \bar{x}_i - \bar{x}_j $				
	Type C ( $\bar{x}_C = 35.90$ )	Type G ( $\bar{x}_G = 35.75$ )	Type H ( $\bar{x}_H = 35.60$ )	Type T ( $\bar{x}_T = 36.04$ )
Type C ( $\bar{x}_C = 35.90$ )		0.15	0.30	0.14
Type G ( $\bar{x}_G = 35.75$ )			0.15	0.29
Type H ( $\bar{x}_H = 35.60$ )				0.44
Type T ( $\bar{x}_T = 36.04$ )				

The first step of Scheffe's method is to obtain the absolute values of pair wise differences between sample means. Recall that the sample means are displayed at the bottom of Table 28-1. An easy way to organize these differences is to construct a table of differences as displayed in Table 28-4. The row headings and the column headings are the different soil types. Each entry in the table is the absolute value of the difference between the sample means corresponding to the row and column heading. We have not included diagonal entries in the table, since these would all be equal to zero; we certainly do not need to compare a sample mean with itself. Also, we have included only entries in the upper right triangular portion of the table, since the entries in the lower left triangular portion would just be redundant. Consequently, we have no entries in the "Type C" column and no entries in the "Type T" row; we could have elected to eliminate this row and column altogether. For the sake of convenience, we have included the corresponding sample mean with each row and column heading. Do the calculations to verify that Table 28-4 is correct.

The second step of Scheffe's method is to obtain values against which we can compare the absolute values of differences between sample means. We shall call these the comparison values. An easy way to organize these comparison values is to construct a table organized in the same way as the table of absolute differences between sample means. This has been done in Table 28-5. The row headings and the column headings are the different soil types. Each entry in the comparison table is calculated from the formula

displayed in Step 2 of Table 28-2. The  $(k - 1)$  in this formula is the between samples degrees of freedom. The  $f_{n_* - k, k - 1; \alpha}$  in this formula is the  $f$ -score with  $\alpha$  of the area above it, or, in other words, the  $f$ -score which defines the rejection region in our hypothesis test. The  $MSE$  in the formula is, of course, the mean square error. The last factor under the square root in the formula is the sum of the inverses of the corresponding sample sizes.

**Table 28-5**  
**Step 2 of Scheffe's Method for the Data of Table 28-1**

Comparison Values $\sqrt{(4 - 1)f_{4 - 1, 20 - 4; \alpha} MSE (1/n_i + 1/n_j)}$ (with $\alpha = 0.05$ )				
	Type C ( $n_C = 5$ )	Type G ( $n_G = 6$ )	Type H ( $n_H = 4$ )	Type T ( $n_T = 5$ )
Type C ( $n_C = 5$ )		0.24	0.27	0.25
Type G ( $n_G = 6$ )			0.26	0.24
Type H ( $n_H = 4$ )				0.27
Type T ( $n_T = 5$ )				

Let us illustrate how we would calculate the comparison value in the row for soil type C and the column for soil type G. The degrees of freedom for soil types is  $k - 1 = 4 - 1 = 3$ , the  $f$ -score with 0.05 of the area above it was found earlier to be  $f_{3, 16; 0.05} = 3.24$ , and the mean square error from the ANOVA table (Table 28-3) is 0.01669. For the sake of convenience, we have included the corresponding sample size with each row and column heading in Table 28-5. The sample sizes corresponding to soil types C and G are respectively 5 and 6 (which is easily seen from the data displayed in Table 28-1). We see then that the comparison value in the row for soil type C and the column for soil type G is

$$\sqrt{(4 - 1)(3.24)(0.01669)(1/5 + 1/6)} = 0.24 .$$

Do the calculations to obtain this value in order to verify that the corresponding entry in Table 28-5 is correct.

You should realize that when calculating the other comparison values, only the sample sizes in the formula will change. Do all of the calculations to obtain the other comparison values, and verify that Table 28-5 is correct. As in Table 28-4, we do not need to compare a sample mean with itself, and therefore we have not included diagonal entries in Table 28-5; also, we have included only entries in the upper right triangular portion of the table, since the entries in the lower triangular portion would just be redundant. As in Table 28-5, the "Type C" column and the "Type T" row could have been eliminated.

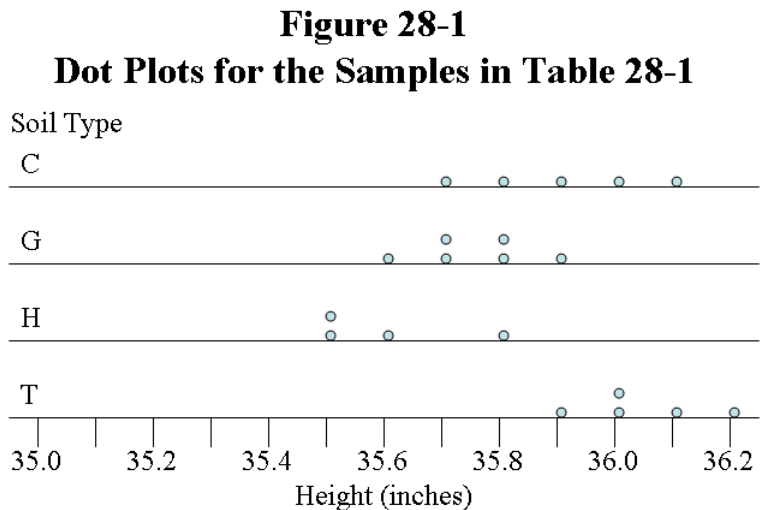
We complete the second step by identifying the statistically significant pair wise differences. Each of the absolute differences in Table 28-4 is compared to the corresponding value in Table 28-5; those absolute differences which are larger are declared as statistically significant at the  $\alpha = 0.05$  level, and those absolute differences which are smaller are not statistically significant at the  $\alpha = 0.05$  level. Compare Tables 28-4 and 28-5, and circle each absolute difference in Table 28-4 which is larger than the corresponding value in Table 28-5. (You should circle the absolute difference between sample means for types C and H, types T and G, and types T and H.)

The third and last step of Scheffe's method is to state the results, including the significance level and the directions of differences. After checking the direction of the significant differences, you should verify that we can summarize the results of Scheffe's method as follows:

With  $\alpha = 0.05$ , we conclude that the mean height of wheat is  
 larger with soil type C than with soil type H,  
 larger with soil type T than with soil type G,  
 and  
 larger with soil type T than with soil type H.

Box plots, one for each soil type, might be used to graphically display the results of the one-way ANOVA and Scheffe's method. However, because of the small sample sizes, we have decided to use the contiguous dot plots displayed as Figure 28-1. (Since a box plot depends on the five-number summary, we prefer to have at least 5 observations per sample in order to construct a box plot.)

We should point out one embarrassing feature about multiple comparison. Although it does not happen very often, it is possible to reject the null hypothesis in a one-way ANOVA but then not find any significant differences when multiple comparison is used. This could happen if the  $p$ -value in the one-way ANOVA  $f$  test is just below the significance level, which is an indication that the evidence against the null hypothesis was just barely strong enough to lead us to reject the null hypothesis.



**Self-Test Problem 28-1.** A 0.05 significance level is chosen for a hypothesis test to see if there is any evidence of a difference in mean weekly TV hours of voters among the rural, suburban, and urban areas in a particular state. The individuals selected for the SURVEY DATA, displayed as Set 1-1 at the end of Unit 1, are treated as comprising three random samples: one from the rural area, one from the suburban area, and one from the urban area.

- Explain how the data for this hypothesis test is appropriate for a one-way ANOVA.
- Complete the four steps of the hypothesis test by completing the table titled *Hypothesis Test for Self-Test Problem 28-1*. As part of the second step, complete the construction of the one-way ANOVA table, displayed as Table 28-6, where you should find that  $SSB = 231.8$ ,  $SSE = 466.5$ , and  $f_{2,27} = 6.71$ .
- If multiple comparison is necessary, apply Scheffe's method and state the results; if multiple comparison is not necessary, explain why not.
- Decide which of the following would be best as a graphical display for the data and say why: (i) three pie charts, (ii) three scatter plots, (iii) three box plots.
- Considering the results of the hypothesis test, decide which of the Type I or Type II errors is possible, and describe this error.
- Decide whether  $H_0$  would have been rejected or would not have been rejected with each of the following significance levels: (i)  $\alpha = 0.01$ , (ii)  $\alpha = 0.10$ .
- What would the presence of one or more outliers in the data suggest about using the  $f$  statistic?

**Hypothesis Test for Self Test Problem 28-1**

Step 1  $H_0$ : \_\_\_\_\_  $H_1$ : \_\_\_\_\_  $\alpha =$  \_\_\_\_\_

Step 2

Step 3

Step 4

**Table 28-6**

**One-Way ANOVA Table for Self-Test Problem 28-1**

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>	<i>p-value</i>
Error					
Total					

### Answers to Self-Test Problems

- 28-1** (a) The data consists of independent random samples selected from more than two populations, and the purpose of a one-way ANOVA is to compare means of such data.
- (b) Step 1:  $H_0: \mu_R = \mu_S = \mu_U$ ,  $H_1$ : At least one of  $\mu_R$ ,  $\mu_S$ , and  $\mu_U$  is different, ( $\alpha = 0.05$ )  
Step 2:  $n_R = 10$ ,  $\bar{x}_R = 8.6$ ,  $n_S = 10$ ,  $\bar{x}_S = 13.0$ ,  $n_U = 10$ ,  $\bar{x}_U = 15.3$ ,  $f_{2,27} = 6.71$ .  
Step 3: The rejection is  $f_{2,27} \geq 3.35$ .  $H_0$  is rejected;  $0.001 < p\text{-value} < 0.01$ .  
Step 4: Since  $f_{2,27} = 6.71$  and  $f_{2,27;0.05} = 3.35$ , we have sufficient evidence to reject  $H_0$ . We conclude that there is no difference in mean weekly TV hours of voters among the rural, suburban, and urban areas in a the state ( $0.001 < p\text{-value} < 0.01$ ). Since  $H_0$  is rejected, we need to use multiple comparison to identify significant differences between the means.
- In Table 28-6, the first label in the *Source* column should be “Area of Residence”; the entries in the *SS* column should respectively be 231.8, 466.5, and 698.3; the entries in the *df* column should respectively be 2, 27, and 29; the entries in the *MS* column should respectively be 115.9 and 17.2778; the entry in the *f* column should be 6.71; the entry in the *p*-value column should be either  $0.001 < p\text{-value} < 0.01$  or an exact *p*-value, obtained from a calculator or computer.
- (c) The absolute differences between sample means are as follows:  
4.4 for rural and suburban, 6.7 for rural and urban, 2.3 for suburban and urban.  
The comparison value is 4.8 for each pair.  
With  $\alpha = 0.05$ , we conclude that the mean weekly TV hours of voters is larger in the urban area than in the rural area.
- (d) Since weekly TV hours is a quantitative variable, three box plots, one for each area of residence category, is an appropriate graphical display.
- (e) Since  $H_0$  is rejected, the Type I error is possible, which is concluding that at least one mean is different when in reality the means are all equal.
- (f)  $H_0$  would have been rejected with both  $\alpha = 0.01$  and  $\alpha = 0.10$ .
- (g) The presence of one or more outliers in the data would suggest that the *f* statistic may not be appropriate.

### Summary

When we do not reject the null hypothesis in a one-way ANOVA, no further analysis is called for, since we are concluding that several population means are equal. When we do reject the null hypothesis in a one-way ANOVA, further analysis is desirable in order to identify which population means are different from each other. Identifying significant differences among more than two population means is a type of *multiple comparison*. Multiple comparison is used to maintain control over the overall significance level. If we perform enough hypothesis tests simultaneously with a given significance level without using multiple comparison, we can expect a very high probability of incorrectly rejecting at least one null hypothesis.

Scheffe's *method of pair wise multiple comparison* is one of many multiple comparison methods available. This method consists of the three steps displayed in Table 28-2. Although it does not happen very often, it is possible to reject the null hypothesis in a one-way ANOVA but then not find any significant differences when multiple comparison is used. This could happen if the *p*-value in the one-way ANOVA *f* test is just below the significance level, which is an indication that the evidence against the null hypothesis was just barely strong enough to lead us to reject the null hypothesis.