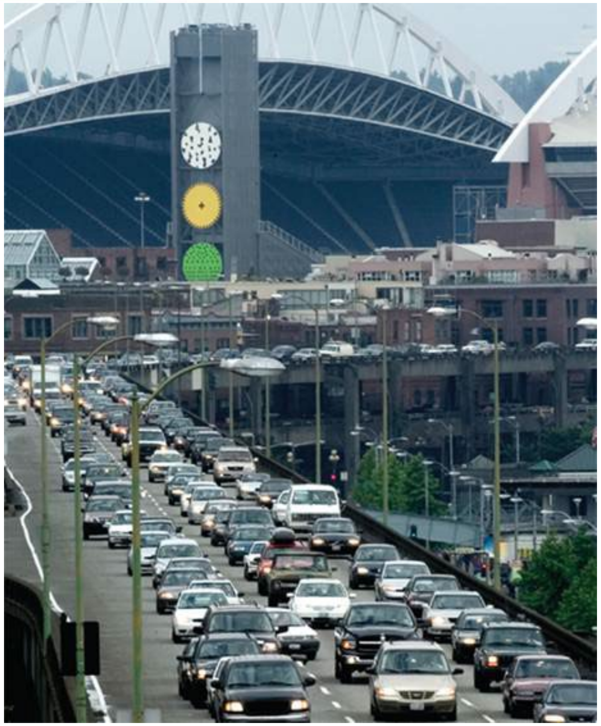
12 Analysis of Variance



12.1 Introduction to the Analysis of Variance Technique

Introduction to the Analysis of Variance Technique

The **analysis of variance** technique **(ANOVA)**, which we are about to explore, will be used to test a null hypothesis about several means, for example,

$$H_o: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

By using our former technique for hypotheses about two means, we could test several hypotheses if each stated a comparison of two means. For example, we could test

$H_1: \mu_1 = \mu_2$ $H_2: \mu_1 = \mu_3$ $H_3: \mu_1 = \mu_4$ $H_4: \mu_1 = \mu_5$ $H_5: \mu_2 = \mu_3$

$H_6: \mu_2 = \mu_4$ $H_7: \mu_2 = \mu_5$ $H_8: \mu_3 = \mu_4$ $H_9: \mu_3 = \mu_5$ $H_{10}: \mu_4 = \mu_5$

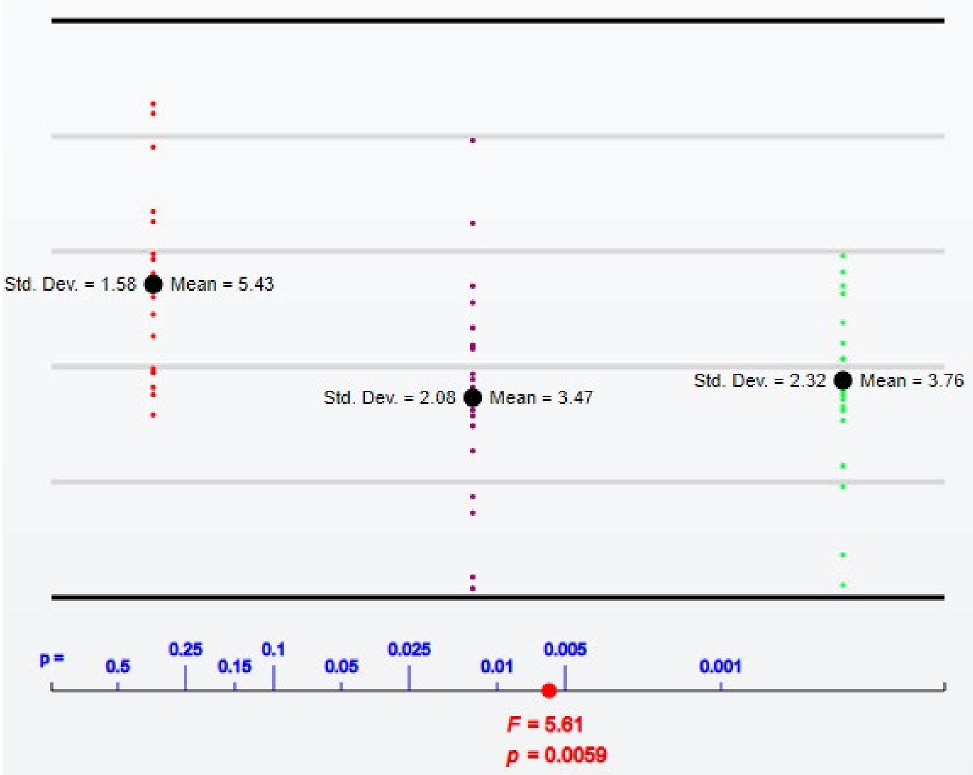
Introduction to the Analysis of Variance Technique

To test the null hypothesis, H_o, that all five means are equal, we would have to test each of these 10 hypotheses using our former technique. Rejection of any one of the 10 hypotheses about two means would cause us to reject the null hypothesis that all five means are equal.

If we failed to reject all 10 hypotheses, we would fail to reject the main null hypothesis. By testing in this manner, the overall type I error rate would become much larger than the value of α associated with a single test.

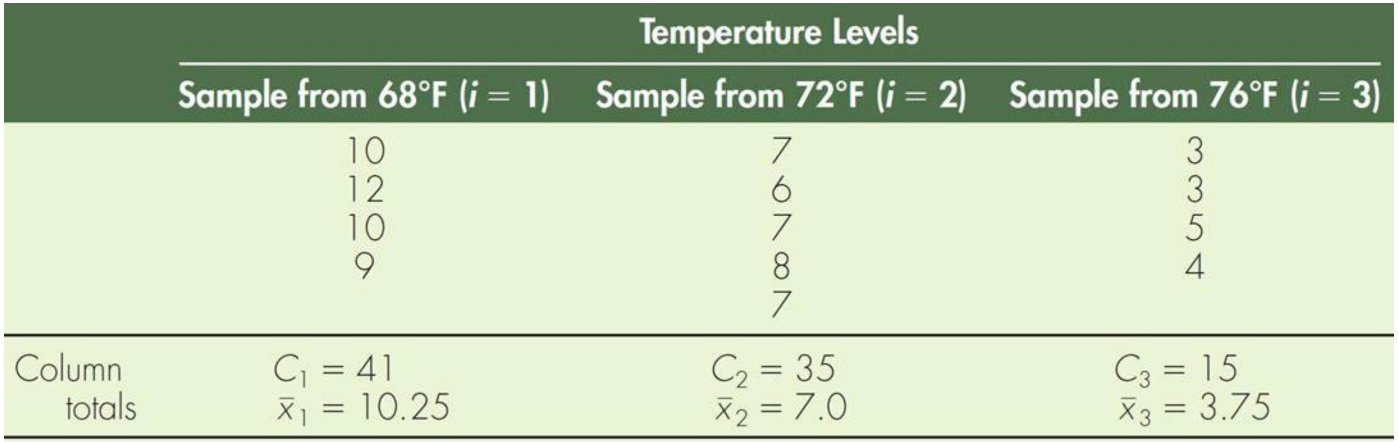
Introduction to the Analysis of Variance Technique

The ANOVA techniques thus allow us to test the null hypothesis (all means are equal) against the alternative hypothesis (at least one mean value is different) with a specified value of α.



Example 1 – *Hypothesis Test for Several Means*

The temperature at which a manufacturing plant is maintained is believed to affect the rate of production in the plant. The data is the number of units produced in 1 hour when the production process in the plant was operating at each of three temperature levels*.*



	Temperature Levels		
	Sample from 68°F ($i = 1$)	Sample from 72°F ($i = 2$)	Sample from 76°F ($i = 3$)
	10	7	3
	12	6	3
	10	7	5
	9	8	4
		7	
Column totals	$C_1 = 41$ $\bar{x}_1 = 10.25$	$C_2 = 35$ $\bar{x}_2 = 7.0$	$C_3 = 15$ $\bar{x}_3 = 3.75$

$k_1=4$	$k_2=5$	$k_3=4$
μ_1	μ_2	μ_3

Example 1 – *Hypothesis Test for Several Means* cont’d

The data values from repeated samplings are called **replicates**. Four replicates, or data values, were obtained for two of the temperatures and five were obtained for the third temperature.

Do these data suggest that temperature has a significant effect on the production level at $\alpha = 0.05$?

The level of production is measured by the mean value; $\bar{x}_i$ indicates the observed production mean at level *i*, where *i* = 1, 2, and 3 correspond to temperatures of 68°F, 72°F, and 76°F, respectively.

Example 1 – *Hypothesis Test for Several Means* cont'd

There is a certain amount of variation among these means. Since sample means are not necessarily the same when repeated samples are taken from a population, some variation can be expected, even if all three population means are equal.

We will next pursue the question: "Is this variation among the $\bar{x}$'s due to chance, or is it due to the effect that temperature has on the production rate?"

Example 1 – *Solution*

Step 1 **a. Parameter of interest:** The “mean” at each *level of the test factor* is of interest: the mean production rate at 68°F, μ_{68}; the mean production rate at 72°F, μ_{72}; and the mean production rate at 76°F, μ_{76}. The factor being tested, plant temperature, has three levels: 68°F, 72°F, and 76°F.

b. Statement of hypotheses:

$$H_o: \mu_{68} = \mu_{72} = \mu_{76}$$

That is, the true production mean is the same at each temperature level tested.

Example 1 – *Solution*

cont'd

In other words, the temperature does not have a significant effect on the production rate. The alternative to the null hypothesis is

H_a: Not all temperature level means are equal.

Thus, we will want to reject the null hypothesis if the data show that one or more of the means are significantly different from the others.

Example 1 – *Solution*

cont'd

Step 2 **a. Assumptions:** The data were randomly collected and are independent of each other.

Equal Variances: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

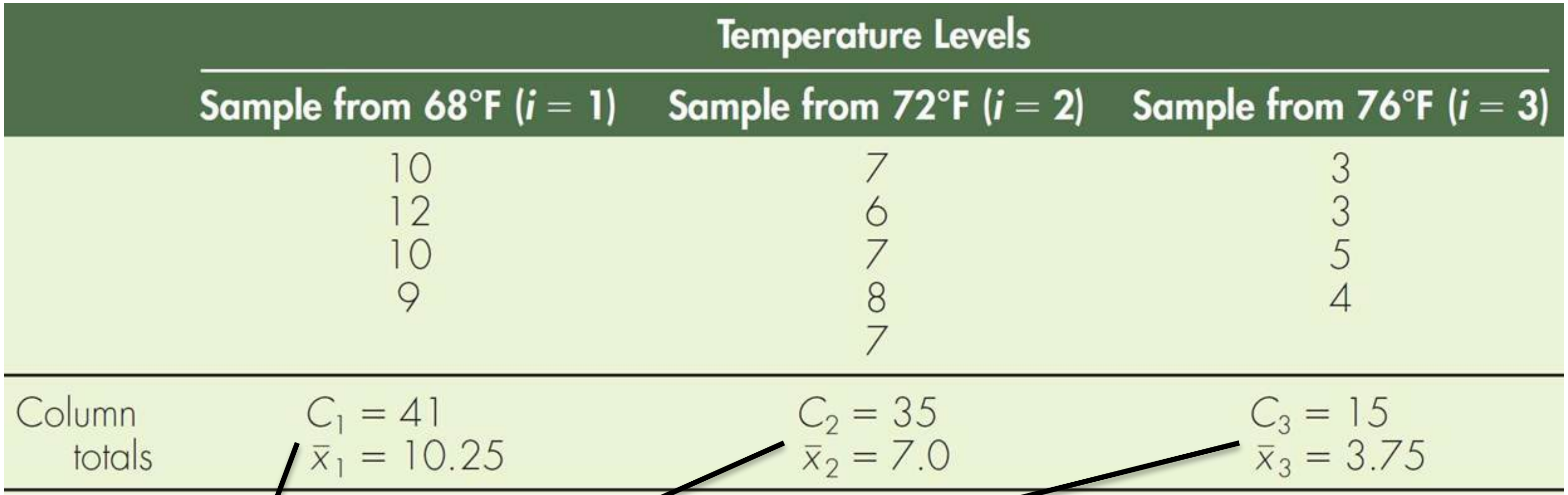
b. Test statistic: We will make the decision to reject H_o or fail to reject H_o by using the *F*-distribution and an *F*-test statistic.

c. Level of significance: $\alpha = 0.05$ (given in the statement of the problem)

Example 1 – *Solution*

cont'd

Step 3 **a. Sample information:** See Table 12.2.



	Temperature Levels		
	Sample from 68°F ($i = 1$)	Sample from 72°F ($i = 2$)	Sample from 76°F ($i = 3$)
	10 12 10 9	7 6 7 8 7	3 3 5 4
Column totals	$C_1 = 41$ $\bar{x}_1 = 10.25$	$C_2 = 35$ $\bar{x}_2 = 7.0$	$C_3 = 15$ $\bar{x}_3 = 3.75$

Sample Results on Temperature and Production

Table 12.2

- **Sample information**
  Next, we can calculate the test statistics.

Example 1 – *Solution*

cont'd

b. Calculated test statistic: We have know that the calculated value of *F* is the ratio of two variances.

The analysis of variance procedure will separate the variation among the entire set of data into two categories.

To accomplish this separation, we first work with the numerator of the fraction used to define **sample variance**, formula (2.5):

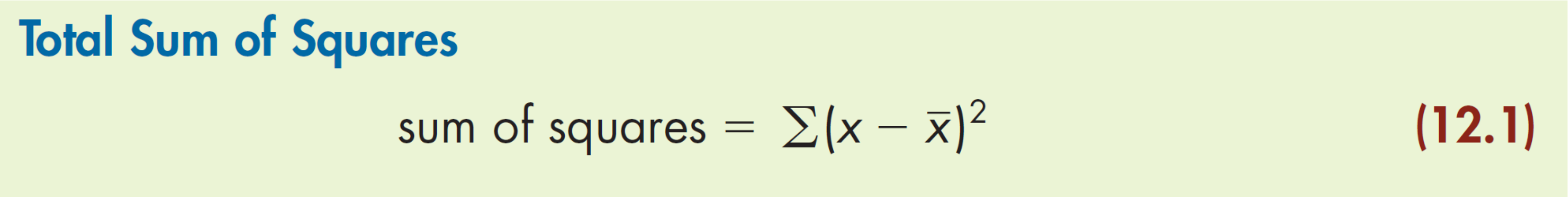
$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

ANOVA applet

Example 1 – *Solution*

cont'd

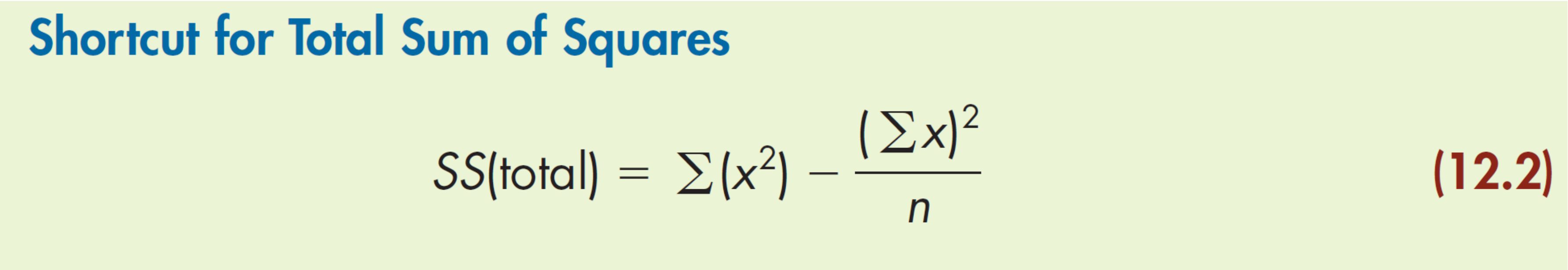
The numerator of this fraction is called the **sum of squares:**



Total Sum of Squares

$$\text{sum of squares} = \sum (x - \bar{x})^2 \qquad \textbf{(12.1)}$$

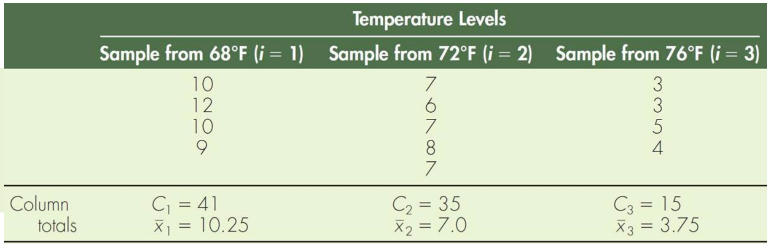
We calculate the **total sum of squares**, ***SS*(total)**, for the total set of data by using a formula that is equivalent to formula (12.1) but does not require the use of $\bar{x}$.



Shortcut for Total Sum of Squares

$$SS(\text{total}) = \sum (x^2) - \frac{(\sum x)^2}{n} \qquad \textbf{(12.2)}$$

Example 1 – *Solution*



	Temperature Levels		
	Sample from 68°F ($i = 1$)	Sample from 72°F ($i = 2$)	Sample from 76°F ($i = 3$)
	10	7	3
	12	6	3
	10	7	5
	9	8	4
		7	
Column totals	$C_1 = 41$ $\bar{x}_1 = 10.25$	$C_2 = 35$ $\bar{x}_2 = 7.0$	$C_3 = 15$ $\bar{x}_3 = 3.75$

Shortcut for Total Sum of Squares

$$SS(\text{total}) = \sum(x^2) - \frac{\left(\sum x\right)^2}{n} = 94 \qquad \textbf{(12.2)}$$

Now we can find *SS*(total) for our example by using the formula (12.2). First,

$$\Sigma(x^2) = 10^2 + 12^2 + 10^2 + 9^2 + 7^2 + 6^2 + 7^2 + 8^2 + 7^2 + 3^2 + 3^2 + 5^2 + 4^2$$
$$= 731$$

$$\Sigma x = 10 + 12 + 10 + 9 + 7 + 6 + 7 + 8 + 7 + 3 + 3 + 5 + 4$$
$$= 91$$

Then, using formula (12.2), we have

$$SS(\text{total}) = \sum(x^2) - \frac{\left(\sum x\right)^2}{n}: \quad SS(\text{total}) = 731 - \frac{(91)^2}{13} = 94$$

Example 1 – *Solution*

cont'd

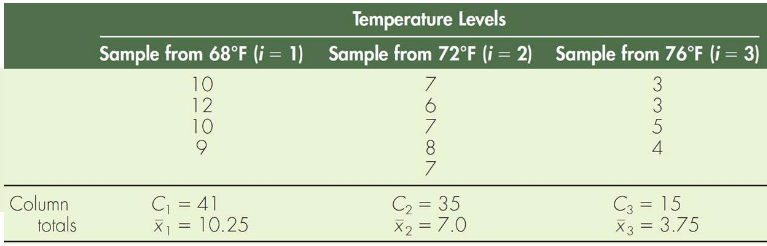
Next, 94, *SS*(total), must be separated into two parts: the sum of squares due to temperature levels, *SS*(temperature), and the sum of squares due to experimental error of replication, *SS*(error).

This splitting is often called **partitioning**, since *SS*(temperature) + *SS*(error) = *SS*(total); that is, in our example, *SS*(temperature) + *SS*(error) = 94.

$$SS(\text{total}) = SS(\text{factor}) + SS(\text{error})$$

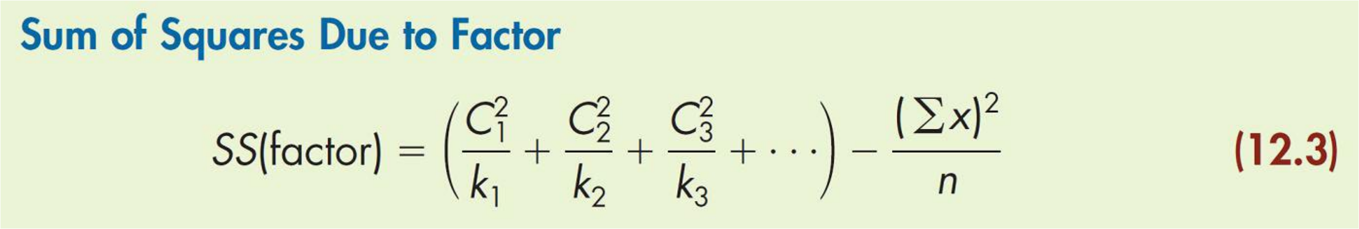
ANOVA Variations applet

Example 1 – *Solution*



	Temperature Levels		
	Sample from 68°F (i = 1)	Sample from 72°F (i = 2)	Sample from 76°F (i = 3)
	10	7	3
	12	6	3
	10	7	5
	9	8	4
		7	
Column totals	$C_1 = 41$ $\bar{x}_1 = 10.25$	$C_2 = 35$ $\bar{x}_2 = 7.0$	$C_3 = 15$ $\bar{x}_3 = 3.75$

The sum of squares, ***SS*(factor)** [*SS*(temperature) for our example], that measures the **variation between the factor levels** (temperatures) is found by using formula (12.3):



Sum of Squares Due to Factor

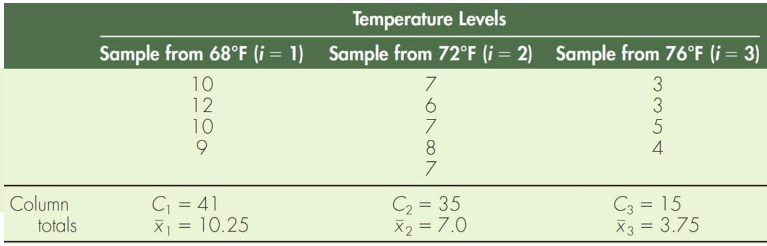
$$SS(\text{factor}) = \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \cdots\right) - \frac{(\Sigma x)^2}{n} \qquad (12.3)$$

where C_i represents the column total, k_i represents the number of replicates at each level of the factor, and n represents the total sample size ($n = \Sigma k_i$)

$$SS(\text{temperature}) = \left(\frac{41^2}{4} + \frac{35^2}{5} + \frac{15^2}{4}\right) - \frac{(91)^2}{13}$$

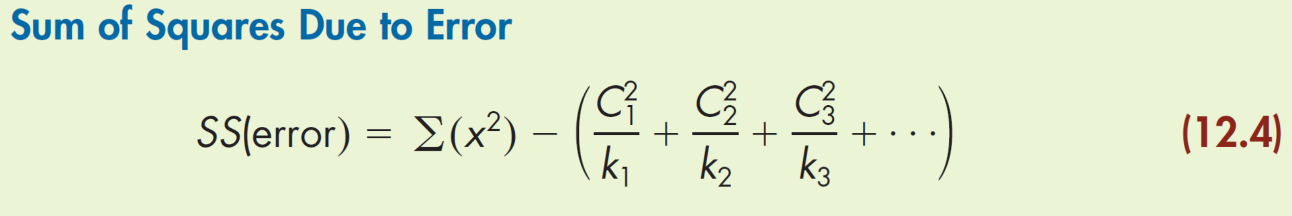
$$= (420.25 + 245.00 + 56.25) – 637.0$$
$$= 721.5 – 637.0$$
$$= 84.5$$

Example 1 – *Solution*



	Temperature Levels		
	Sample from 68°F ($i = 1$)	Sample from 72°F ($i = 2$)	Sample from 76°F ($i = 3$)
	10	7	3
	12	6	3
	10	7	5
	9	8	4
		7	
Column totals	$C_1 = 41$ $\bar{x}_1 = 10.25$	$C_2 = 35$ $\bar{x}_2 = 7.0$	$C_3 = 15$ $\bar{x}_3 = 3.75$

The sum of squares, ***SS*(error)**, that measures the **variation within the rows** is found by using formula (12.4):



Sum of Squares Due to Error

$$SS(\text{error}) = \sum(x^2) - \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \cdots\right) \qquad \textbf{(12.4)}$$

The *SS*(error) for our example can now be found. First,

$$\Sigma(x^2) = 731 \qquad \text{(found previously)}$$

Then, using formula (12.4), we have

$$SS(\text{error}) = \Sigma(x^2) - \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \cdots\right)$$

$$= 731.0 - 721.5$$

$$= 9.5$$

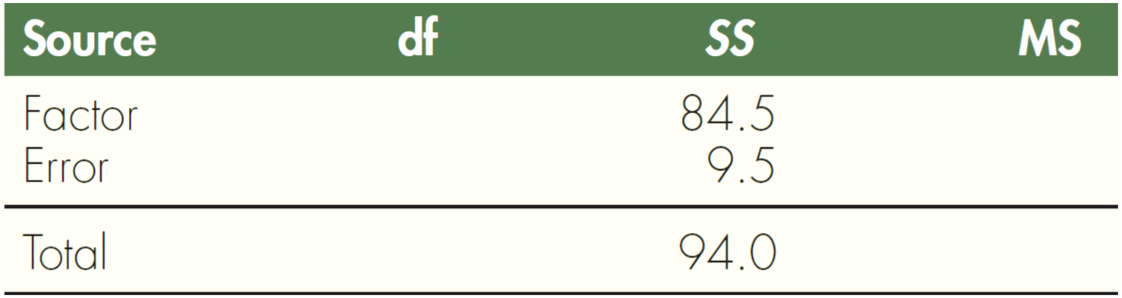
Example 1 – *Solution*

cont’d

Note

SS(total) = *SS*(factor) + SS(error). Inspection of formulas (12.2), (12.3), and (12.4) will verify this.

For convenience we will use an ANOVA table to record the sums of squares and to organize the rest of the calculations. The format of an ANOVA table is shown in Table 12.3.



Source	df	SS	MS
Factor		84.5	
Error		9.5	
Total		94.0	

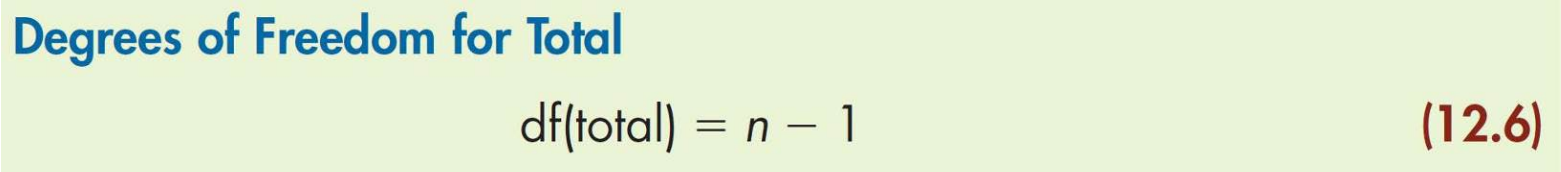
Format for ANOVA Table

Table 12.3

Example 1 – *Solution*

cont’d

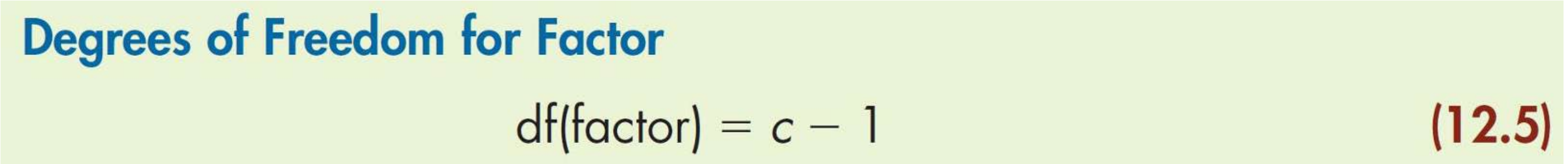
1. df(total) is 1 less than the total number of data:



Degrees of Freedom for Total

$$\text{df(total)} = n - 1 \qquad (12.6)$$

2. df(factor) is 1 less than the number of levels (columns) for which the factor is tested:

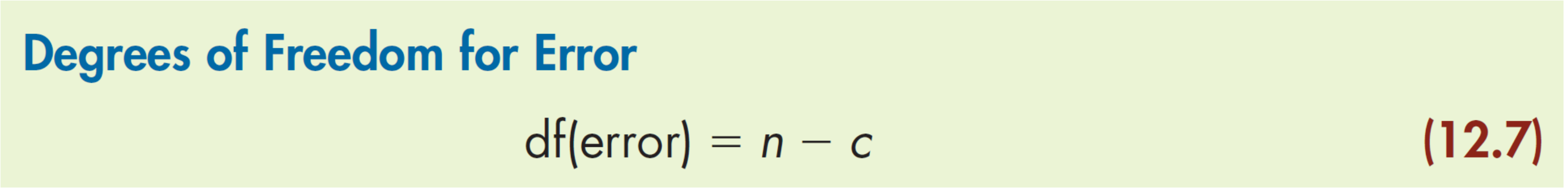


Degrees of Freedom for Factor

$$\text{df(factor)} = c - 1 \qquad (12.5)$$

where *c* is the number of *levels for which the factor is being tested* (number of columns on the data table)

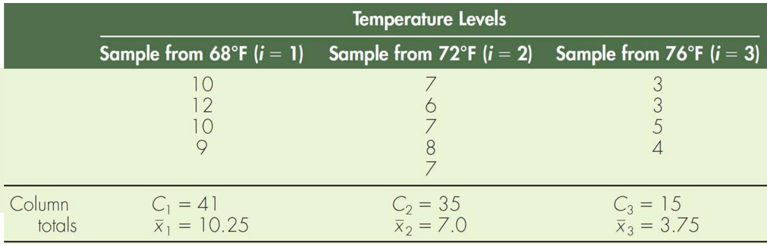
3. df(error) is *c* less than the total number of data



Degrees of Freedom for Error

$$\text{df(error)} = n - c \qquad (12.7)$$

Example 1 – *Solution*



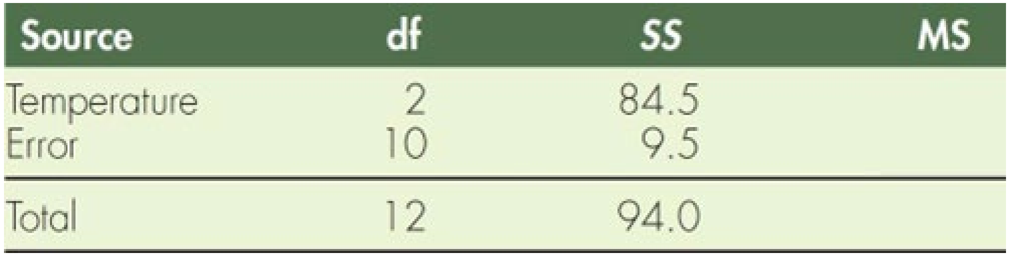
	Temperature Levels		
	Sample from 68°F ($i = 1$)	Sample from 72°F ($i = 2$)	Sample from 76°F ($i = 3$)
	10 12 10 9	7 6 7 8 7	3 3 5 4
Column totals	$C_1 = 41$ $\bar{x}_1 = 10.25$	$C_2 = 35$ $\bar{x}_2 = 7.0$	$C_3 = 15$ $\bar{x}_3 = 3.75$

The degrees of freedom for our illustration are

df(total) = n – 1 = 13 – 1 = 12

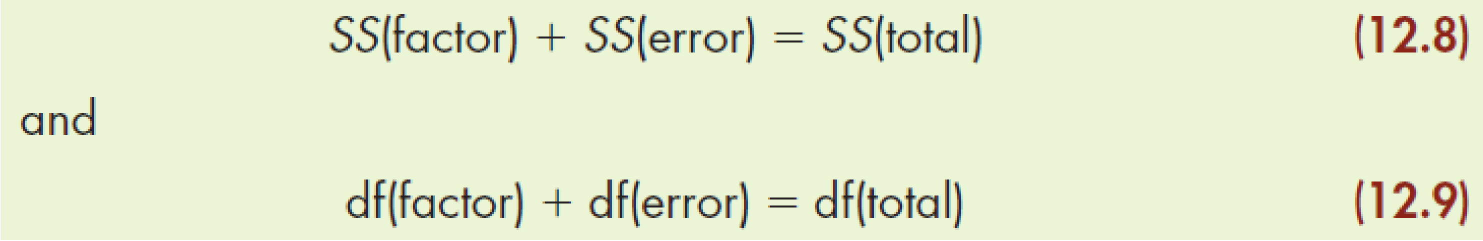
df(temperature) = c – 1 = 3 – 1 = 2

df(error) = n – c = 13 – 3 = 10



Source	df	SS	MS
Temperature	2	84.5	
Error	10	9.5	
Total	12	94.0	

The sums of squares and the degrees of freedom must check; that is,



$$SS(\text{factor}) + SS(\text{error}) = SS(\text{total}) \qquad \textbf{(12.8)}$$

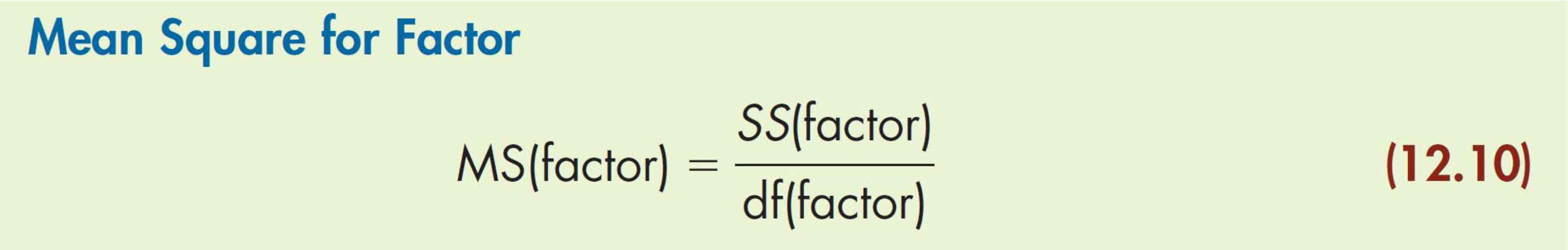
and

$$\text{df(factor)} + \text{df(error)} = \text{df(total)} \qquad \textbf{(12.9)}$$

Example 1 – *Solution*

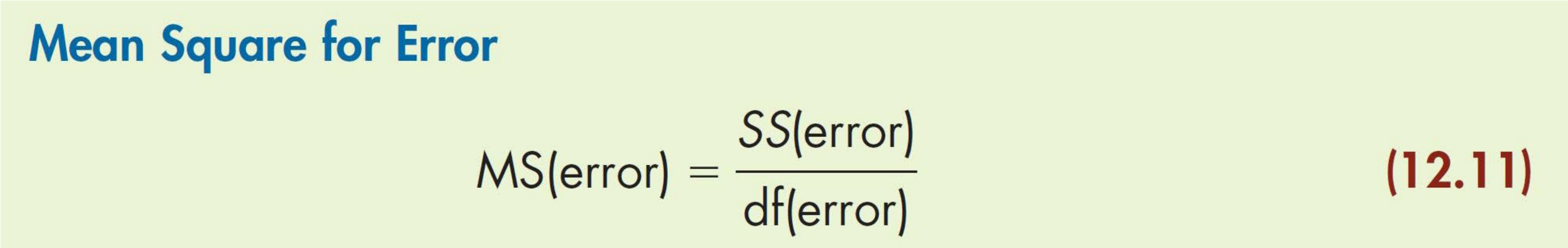
cont’d

The **mean square** for the factor being tested, **MS(factor)**, and for error, **MS(error)**, are obtained by dividing the sum-of-squares value by the corresponding number of degrees of freedom:



Mean Square for Factor

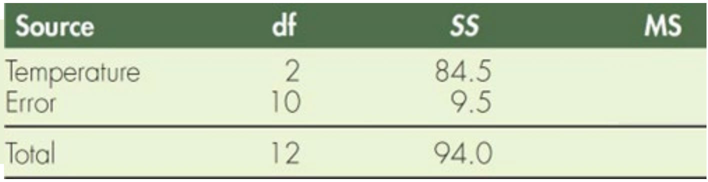
$$\text{MS(factor)} = \frac{SS(\text{factor})}{\text{df(factor)}} \qquad \textbf{(12.10)}$$



Mean Square for Error

$$\text{MS(error)} = \frac{SS(\text{error})}{\text{df(error)}} \qquad \textbf{(12.11)}$$

Example 1 – *Solution*



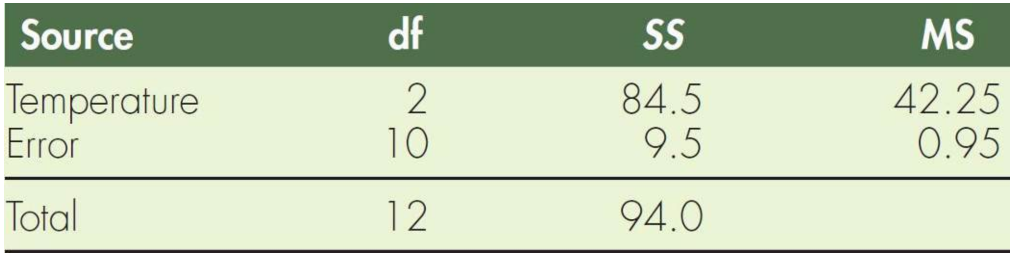
Source	df	*SS*	MS
Temperature	2	84.5	
Error	10	9.5	
Total	12	94.0	

The mean squares for our example are

$$\text{MS(temperature)} = \frac{SS(\text{temperature})}{\text{df(temperature)}} = \frac{84.5}{2} = 42.25$$

$$\text{MS(error)} = \frac{SS(\text{error})}{\text{df(error)}} = \frac{9.5}{10} = 0.95$$

The complete ANOVA table appears as



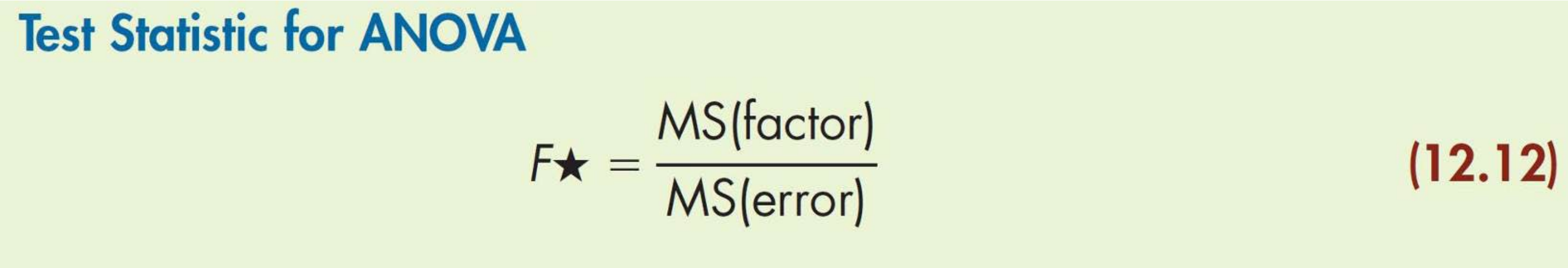
Source	df	*SS*	MS
Temperature	2	84.5	42.25
Error	10	9.5	0.95
Total	12	94.0	

Example 1 – *Solution*

cont'd

The hypothesis test is now completed using the two mean squares as the measures of variance.

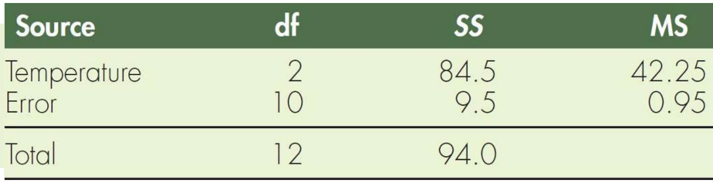
The calculated value of the test statistic, $F★$, is found by dividing the MS(factor) by the MS(error):



Test Statistic for ANOVA

$$F★ = \frac{\text{MS(factor)}}{\text{MS(error)}} \qquad (12.12)$$

Example 1 – *Solution*



Source	df	SS	MS
Temperature	2	84.5	42.25
Error	10	9.5	0.95
Total	12	94.0	

The calculated value of *F* for our example is found by using formula (12.12):

$$F\bigstar = \frac{\mathrm{MS(factor)}}{\mathrm{MS(error)}}:\ F\bigstar = \frac{\mathrm{MS(temperature)}}{\mathrm{MS(error)}} = \frac{42.25}{0.95}$$

$$= 44.47$$

Note

Since the calculated value of *F*, $F\bigstar$, is found by dividing MS(temperature) by MS(error), the number of degrees of freedom for the numerator is df(temperature) = 2 and the number of degrees of freedom for the denominator is df(error) = 10.

Example 1 – *Solution*

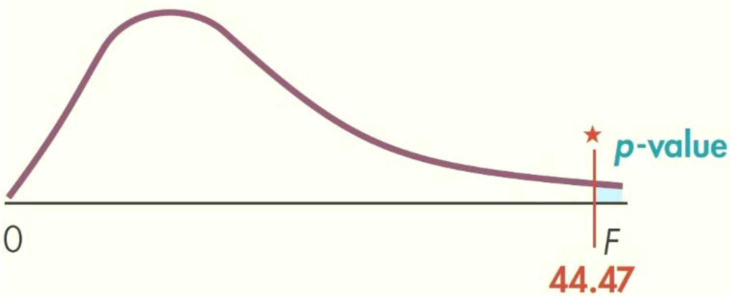
cont'd

Step 4 Probability Distribution:

***p*-Value:**

Use the right-hand tail because larger values of $F★$ indicate "not all equal" as expressed by H_a,

P = $P(F★ > 44.47 \mid \text{df}_n = 2, \text{df}_d = 10)$, as shown in the figure.



Example 1 – *Solution*

cont'd

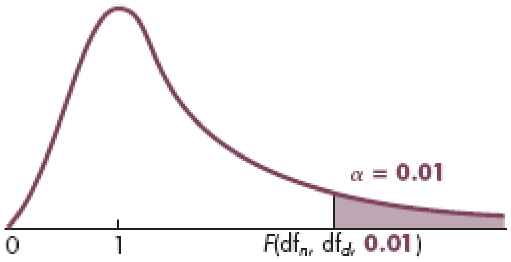
To find the *p*-value, you have two options:

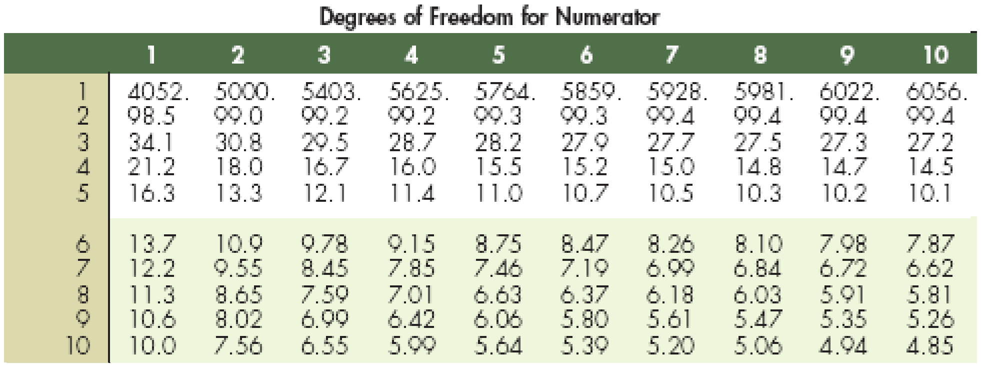
1. Use Table 9C (Appendix B)

to place bounds on the *p*-value:

P = $P(F★ > 44.47 \mid df_n = 2, df_d = 10)$,

P < 0.01.





Degrees of Freedom for Numerator

	1	2	3	4	5	6	7	8	9	10
1	4052.	5000.	5403.	5625.	5764.	5859.	5928.	5981.	6022.	6056.
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85

2. Use a computer or calculator to find the *p*-value:

   P = 0.00001.

b. The *p*-value is smaller than the level of significance, α(0.05).

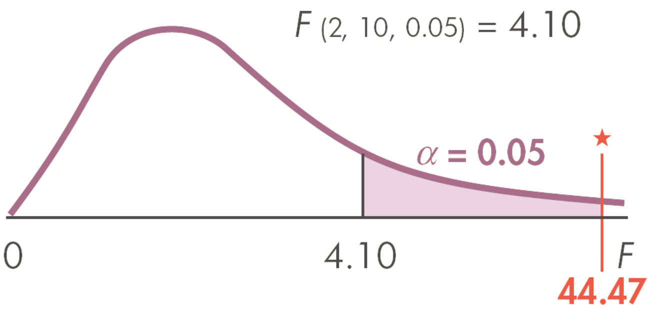
Example 1 – *Solution*

cont'd

Classical:

a. The critical region is the right-hand tail because larger values of $F★$ indicate "not all equal" as expressed by H_a, $\text{df}_n = 2$ and $\text{df}_d = 10$. The critical value is obtained from Table 9A.

b. $F★$ is in the critical region, as shown in **red** in the figure.



Example 1 – *Solution*

cont'd

Step 5. **a. Decision: Reject H_o.**

b. Conclusion: At least one of the room temperatures does have a significant effect on the production rate. The differences in the mean production rates at the tested temperature levels were found to be significant.

Example 1 – *Solution*

cont'd

The mean at 68°F is certainly different from the mean at 76°F because the sample means for these levels are the largest and the smallest, respectively.

Whether any other pairs of means are significantly different cannot be determined from the ANOVA procedure alone.

THE END

- ANY QUESTION?