CAUTION When we conclude that there is sufficient evidence to reject the claim of equal population means, we cannot conclude from ANOVA that any particular mean is different from the others. (There are several other methods that can be used to identify the specific means that are different, and some of them are discussed in Part 2 of this section.)

How Is the *P*-Value Related to the Test Statistic? *Larger* values of the test statistic result in *smaller* *P*-values, so the ANOVA test is right-tailed. Figure 12-2 shows the relationship between the F test statistic and the P-value. Assuming that the populations have the same variance σ^2 (as required for the test), the F test statistic is the ratio of these two estimates of σ^2: (1) variation *between* samples (based on variation among sample means); and (2) variation *within* samples (based on the sample variances).

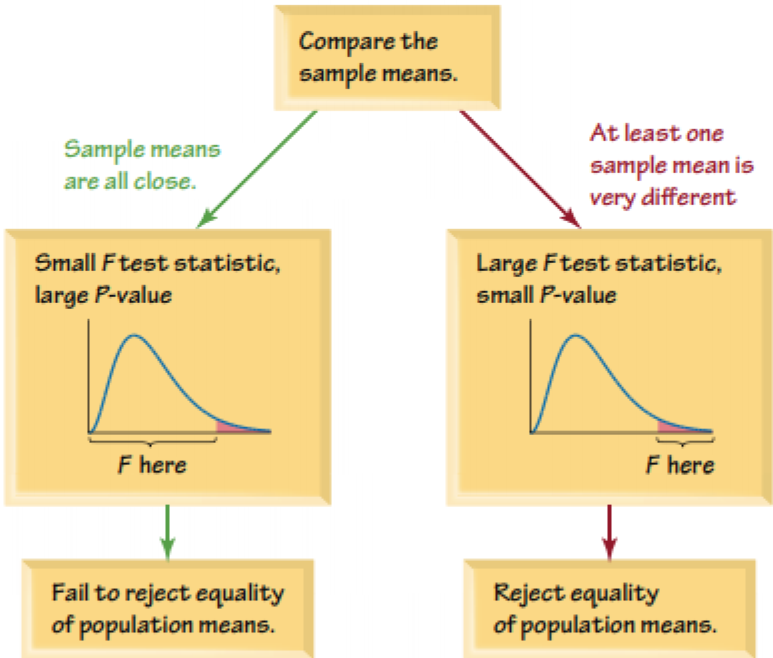


Figure 12-2 Relationship between the *F* Test Statistic and the *P*-Value

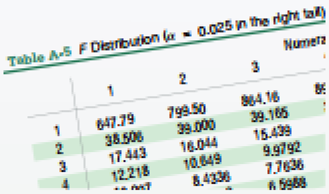
Test Statistic for One-Way ANOVA: $F = \dfrac{\text{variance between samples}}{\text{variance within samples}}$

The numerator of the F test statistic measures variation between sample means. The estimate of variance in the denominator depends only on the sample variances and is not affected by differences among the sample means. Consequently, sample means that are close in value result in a small F test statistic and a large P-value, so we conclude that there are no significant differences among the sample means. Sample means that are very far apart in value result in a large F test statistic and a small P-value, so we reject the claim of equal means.

Why Not Just Test Two Samples at a Time? If we want to test for equality among three or more population means, why do we need a new procedure when we can test for equality of two means using the methods presented in Section 9-3? For example, if we want to use the sample data from Table 12-1 to test the claim that the three

Why 0.05?

In 1925, R. A. Fisher published a book that introduced the method of analysis of variance, and he needed a table of critical values based on numerator degrees of freedom and denominator degrees of freedom, as in Table A-5 in Appendix A. Because the table uses two different degrees of freedom, it becomes very long if many different critical values are used, so Fisher included a table using 0.05 only. In a later edition he also included the significance level of 0.01.



Stephen Stigler, a notable historian of statistics, wrote in *Chance* magazine that the choice of a significance level of 0.05 is a convenient round number that is somewhat arbitrary. Although it is arbitrary, the choice of 0.05 accomplishes the following important goals. (1) The value of a 0.05 significance level results in sample sizes that are reasonable and not too large. (2) The choice of 0.05 is large enough to give us a reasonable chance of identifying important effects (by correctly rejecting a null hypothesis of no effect when there really is an effect). (3) The choice of 0.05 is not so small that it forces us to miss important effects (by making the mistake of failing to reject a null hypothesis of no effect when there really is an effect).