UNIT-IV

ANALYSIS OF VARIANCE

F TEST

In previous chapters, the null hypothesis has been tested with a t ratio. In the two-sample case, t reflects the ratio between the observed difference between the two sample means in the numerator and the estimated standard error in the denominator. For three or more samples, the null hypothesis is tested with a new ratio, the F ratio. Essentially, F reflects the ratio of the observed differences between all sample means (measured as variability between groups) in the numerator and the estimated error term or pooled variance estimate (measured as variability within groups) in the denominator term, that is,

fratio

$$F = \frac{\text{VARIABILITY BETWEEN GROUPS}}{\text{VARIABILITY WITHIN GROUPS}}$$

Like t, F has its own family of sampling distributions that can be consulted, as described in Section 16.6, to test the null hypothesis. The resulting test is known as an F test.

An F test of the null hypothesis is based on the notion that if the null hypothesis is true, both the numerator and the denominator of the F ratio would tend to be about the same, but if the null hypothesis is false, the numerator would tend to be larger than the denominator

F TEST

If Null Hypothesis Is True

If the null hypothesis is true (because there is no treatment effect due to different sleep deprivation periods), the two estimates of variability (between and within groups) would reflect only random error. In this case,

$$F = \frac{\textit{Random error}}{\textit{Random error}}$$

Except for chance, estimates in both the numerator and the denominator are similar, and generally, F varies about a value of 1. If Null Hypothesis Is False If the null hypothesis is false (because there is a treatment effect due to different sleep deprivation periods), both estimates still would reflect random error, but the estimate for between groups would also reflect the treatment effect. In this case,

$$F = \frac{\textit{Random error + Treatment effect}}{\textit{Random error}}$$

When the null hypothesis is false, the presence of a treatment effect tends to cause a chain reaction: The observed differences between group means tend to be large, as does the variability between groups. Accordingly, the numerator term tends to exceed the denominator term, producing an F whose value is larger than 1. When the null hypothesis is false because of a large treatment effect, there is an even more pronounced chain reaction, beginning with very large observed differences between group means and ending with an F whose value tends to be considerably larger than 1.

F TEST

Except for the fact that measures are repeated, the new F test is essentially the same as that described before. The statistical hypotheses still are:

H0: $\mu 0 = \mu 24 = \mu 48$

H1:H0 is false

where μ0, μ24, and μ48 represent the mean aggression scores for the single population of subjects who are deprived of sleep for 0, 24, and 48 hours. Once again, rejection of the null hypothesis implies that sleep deprivation influences aggressive behavior. As before, the F test of the null hypothesis is based on the notion that if the null hypothesis really is true, both the numerator and denominator of the F ratio will tend to be about the same, but if the null hypothesis is false, the numerator (still MS between) will tend to be larger than the denominator (now MS error). As implied, the new denominator term, MS error, tends to be smaller than that for the original one-factor ANOVA because of the elimination of variability due to individual differences in repeated-measures ANOVA.

The hypothesis test for the sleep deprivation experiment with repeated measures, as summarized in the accompanying box, will be discussed later in more detail. Notice the huge increase in the value of F from 7.36 (for the original one-factor ANOVA) to 27 (for the repeated-measures ANOVA), even though both F ratios are based on the same set of nine scores (and the same differences between the three deprivation conditions). The relatively large individual differences in the current example illustrate

TWO COMPLICATIONS

HYPOTHESIS TEST SUMMARY

REPEATED-MEASURES *F* TEST
(Sleep Deprivation Experiment)

Research Problem

On average, are subjects' aggression scores in a controlled social situation affected by sleep deprivation periods of 0, 24, and 48 hours, where each subject experiences all three periods?

Statistical Hypothesis

$$H_0: \mu_0 = \mu_{24} = \mu_{48}$$

H_1: H_0 is false

Decision Rule

Reject H_0 at the .05 level of significance if $F \geq 6.94$ (from Table C in Appendix C, given $df_{between} = 2$ and $df_{error} = 4$).

Calculations

$F = 27$ (See Tables 17.3 and 17.5 for more details.)

Decision

Reject H_0 at the .05 level of significance because $F = 27$ exceeds 6.94.

Interpretation

Mean aggression scores in a controlled social situation are affected by sleep deprivation when subjects experience all three levels of deprivation.

the beneficial effects of repeated-measures ANOVA. In practice, the net effect of a repeated-measures experiment might not be as dramatic.

TWO COMPLICATIONS

The same two complications exist for repeated-measures ANOVA as for the repeated measures t test (see Section 15.1). Presumably, in the sleep deprivation experiment, sufficient time elapses between successive sessions to eliminate any lingering effects due to earlier deprivation periods. If there is any concern that earlier effects of the independent variable linger during subsequent sessions, do not use repeated measures.

An extension of counterbalancing can be used to eliminate any potential bias in favor of one condition merely because of the order in which it was experienced. Presumably, in the sleep deprivation experiment, each of the three subjects has been randomly assigned to undergo a different one of three possible orders of deprivation sequences—either 0, 24, and 48 hours; or 48, 0, and 24 hours; or 24, 48, and 0 hours that, taken together over all three subjects, equalizes the number of times a particular deprivation level was experienced first, second, or third.

ANOVA SUMMARY TABLES

where MS error reflects the variability among scores of subjects within each treatment group, pooled across all treatments, after the removal of variability attributable to individual differences.

For the sleep deprivation experiment with repeated measures,

$$MS_{error} = \frac{SS_{error}}{DF_{error}} = \frac{4}{4} = 1$$

which, because of the removal of variability due to individual differences, is much smaller than the corresponding error term, MS within, of 3.67 for the independent measures experiment in Chapter 16.

Finally, the F ratio is as follows:

F RATIO (REPEATED MEASURES)

$$F = \frac{MS_{between}}{MS_{error}}$$

For the sleep deprivation experiment with repeated measures, the null hypothesis is very suspect because

$$F = \frac{MS_{between}}{MS_{error}} = \frac{27}{1} = 27$$

ANOVA SUMMARY TABLES

ANOVA results can be summarized as shown in Table 17.5. Ordinarily, the shaded numbers in parentheses do not appear in ANOVA tables, but they show the origin of the two relevant MS terms and the F ratio

Table 17.5
ANOVA TABLE: SLEEP DEPRIVATION EXPERIMENT (REPEATED MEASURES)

SOURCE	SS	df	MS	F
Between	54	2	$\left(\frac{54}{2}=\right)$ 27	$\left(\frac{27}{1}=\right)$ 27*
Within	22	6		
Subject	18	2		
Error	4	4	$\left(\frac{4}{4}=\right)$ 1	
Total	76	8		

ANALYSIS OF VARIANCE (REPEATED MEASURES)

Table 17.6
COMPARISON OF SUMMARY TABLES FOR ONE-FACTOR ANOVA AND REPEATED-MEASURE ANOVA

ANOVA TABLE: ORIGINAL SLEEP DEPRIVATION EXPERIMENT (ONE FACTOR)					ANOVA TABLE: SLEEP DEPRIVATION EXPERIMENT (REPEATED MEASURES)				
SOURCE	*SS*	*df*	*MS*	*F*	SOURCE	*SS*	*df*	*MS*	*F*
Between	54	2	27	7.36*	Between	54	2	27	27**
Within	22	6	3.67		Within	22	6		
					Subject	18	2		
					Error	4	4	1	
Total	76	8			Total	76	8		

Table 17.6 compares the ANOVA summary tables for the two sleep deprivation experiments. Since, to facilitate comparisons, exactly the same nine scores were used for both experiments, the two summary tables possess many similarities. Sums of squares and degrees of freedom are the same for between and total variability. The main difference appears in the denominator terms for the F ratios. The MS error for the repeated measures ANOVA is about one-third as small as the MS within for the one-factor ANOVA.

Why More Powerful?

In applications with real data, MS between also would tend to be smaller in repeated measures ANOVA than in one-factor ANOVA because of the absence of individual differences from variability between treatment groups. But then why can it be claimed that, if the null hypothesis is false, the repeated-measures F will tend to be larger? If the null hypothesis is false, the F ratio will tend to be greater than 1. Therefore, the subtraction of essentially the same amount of variability due to individual differences from both the numerator and denominator of the F ratio causes relatively more shrinkage in the smaller denominator term. To illustrate with a simple numerical example: Given any F greater than 1, say a one-factor F = 8÷ 4 = 2, then, subtract from both numerator and denominator any constant (representing individual differences), say 2, to obtain a larger repeated-measures

ESTIMATING EFFECT SIZE

Whenever F is statistically significant in a repeated-measures ANOVA, a variation on the squared curvilinear correlation coefficient, η2, can be used to estimate effect size. The formula for a repeated-measures η2 differs from that for an independent

measures η2 (Formula 16.8 on page 309) because SStotal– SSsubject replaces SStotal in the denominator.

PROPORTION OF EXPLAINED VARIANCE (REPEATED MEASURES)

$$\eta_p^2 = \frac{SS_{between}}{SS_{total} - SS_{subject}} = \frac{SS_{between}}{(SS_{between} + SS_{subject} + SS_{error}) - SS_{subject}}$$

$$\eta_p^2 = \frac{SS_{between}}{SS_{between} + SS_{error}} \qquad (17.6)$$

correlation, because the effects of individual differences have been eliminated from the reduced or partial total variance. This adjustment reflects the fact that, when measures are repeated, the value of η2 for the treatment variable can't possibly account for that portion of total variability attributable to individual differences, that is, SSsubject. Substituting in values for the SS terms from the ANOVA summary in Table 17.5, we have

$$\eta_p^2 = \frac{54}{76 - 18} = \frac{54}{58} = .93$$

When compared with guidelines for effect sizes in Table 17.7, this estimated effect size of .93 would be spectacularly large, indicating that .93, or 93 percent, of total variance in aggression scores (excluding variance due to individual differences) is explained by differences between 0, 24, and 48 hours of sleep deprivation, while the remaining 7 percent of the variance in aggression scores is not explained by hours of sleep deprivation. This very large value for η_P^2 reflects a number of factors. Identical sets of fictitious data (used for both the independent-measures and the repeated-measures experiments) were selected to dramatize the effects of sleep deprivation. Furthermore, although based on the same data, the value of .93 for the repeated-measures estimate, η_P^2, exceeds that of .71 for the independent-measures estimate, η2, essentially because of the smaller denominator term in η_P^2

MULTIPLE COMPARISONS

Rejection of the overall null hypothesis indicates only that not all population means are equal. To pinpoint the one or more differences between pairs of population means that contribute to the rejection of the overall null hypothesis, use a multiple comparison test, such as Tukey's HSD test. Tukey's test supplies a single critical

value, HSD, for evaluating the significance of each difference for every possible pair of means. The value of HSD can be calculated using the following formula:

TUKEY'S *HSD* TEST (REPEATED MEASURES)

$$HSD = q\sqrt{\frac{MS_{error}}{n}} \qquad (17.7)$$

where HSD is the positive critical value for any difference between two means; q is a value obtained from Table G in Appendix C; MSerror is the error term for the repeated measures ANOVA; and n is the sample size in each treatment group (which in repeated measures ANOVA is simply the number of subjects). To obtain a value for q at the .05 level (light numbers) in Table G, find the cell intersected by k, the number of repeated measures or treatment levels, and dferror, the degrees of freedom for the error term in the repeated-measures ANOVA. Given values of k = 3 and dferror = 4 for the sleep deprivation experiment, the intersected cell shows a value of 5.04 for q at the .05 level. Substituting q = 5.04, MSerror = 1, and n = 3 in Equation 17.7, we can solve for HSD as follows:

$$HSD = q\sqrt{\frac{MS_{error}}{n}} = 5.04\sqrt{\frac{1}{3}} = 5.04\ (.57) = 2.87$$

Interpretation

Table 17.8 shows absolute differences of either 3, 6, or 3 for the three pairs of means in the repeated-measures experiment. (These absolute differences are the same as those for the three pairs of means for the one-factor ANOVA shown in Table 16.8 on page 313.) In the case of the repeated-measures experiment, however, all three of the observed differences exceed the critical HSD value of 2.87. We can conclude, therefore, that the population mean aggression scores become progressively higher as the sleep deprivation period increases from 0 to 24 and then to 48 hours. Because of its smaller error term, the repeated-measures ANOVA resulted in significant differences for all three comparisons, while the one-factor ANOVA resulted in a significant difference for only the one most extreme comparison (between 0 and 48 hours of deprivation). In practice, of course, there is no guarantee that the beneficial effects of repeated measures will always be as dramatic.

Estimating Effect Size

The effect size for any significant difference between pairs of means can be estimated with Cohen's d

STANDARDIZED EFFECT SIZE, COHEN'S *d*
(ADAPTED FOR REPEATED-MEASURES ANOVA)

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_{error}}} \qquad (17.8)$$

Table 17.8
ALL POSSIBLE ABSOLUTE DIFFERENCES BETWEEN PAIRS OF MEANS

SLEEP-DEPRIVATION EXPERIMENT: REPEATED MEASURES

	$\bar{X}_0 = 2$	$\bar{X}_{24} = 5$	$\bar{X}_{48} = 8$
$\bar{X}_0 = 2$	—	3*	6*
$\bar{X}_{24} = 5$		—	3*
$\bar{X}_{48} = 8$			—

A TWO-FACTOR EXPERIMENT:

RESPONSIBILITY IN CROWDS

Often referred to as the "bystander effect," do crowds affect our willingness to assume responsibility for the welfare of others and ourselves? For instance, does the presence of bystanders inhibit our reaction to potentially dangerous smoke seeping from a wall vent? Hoping to answer this question, a social psychologist measures any delay in a subject's alarm reaction (the dependent variable) as smoke fills a waiting room occupied only by the subject, plus "crowds" of either zero, two, or four associates of the experimenter—the first independent variable or factor—who act as regular subjects but, in fact, ignore the smoke. As a second independent variable or factor, the social psychologist randomly assigns subjects to one of two "degrees of danger," that is, the rate at which the smoke enters the room, either non dangerous (slow rate) or dangerous (rapid rate). Using this two factor ANOVA design, the psychologist can test not just two but three null hypotheses, namely, the effect on subjects' reaction times of (1) crowd size, (2) degree of danger and, as a bonus, (3) the combination or interaction of crowd size and degree of danger. For computational simplicity, assume that the social psychologist randomly assigns two

subjects to be tested (one at a time) with crowds of either zero, two, or four people and either the Non dangerous or dangerous conditions. The resulting six groups, each consisting of two subjects, represent all possible combinations of the two factors.*

Tables for Main Effects and Interaction

Table 18.1 shows one set of possible outcomes for the two-factor study. Although, as indicated in Chapter 16, the actual computations in ANOVA usually are based on totals, preliminary interpretations can be based on either totals or means. In Table 18.1, The shaded numbers represent four different types of means:

1. The three column means (9, 12, 15) represent the mean reaction times for each crowd size when degree of danger is ignored. Any differences among these column means not attributable to chance are referred to as the main effect of crowd size on reaction time. In ANOVA, main effect always refers to the effect of a single factor, such as crowd size, when any other factor, such as degree of danger, is ignored.

Table 18.1
OUTCOME OF TWO-FACTOR EXPERIMENT (REACTION TIMES IN MINUTES)

	CROWD SIZE						
DEGREE OF DANGER	**ZERO**		**TWO**		**FOUR**		**ROW MEAN**
Dangerous	8	8	8	7	10	9	8
	8		6		8		
Nondangerous	9	10	15	17	24	21	16
	11		19		18		
Column mean		9		12		15	Grand mean = 12

A TWO-FACTOR EXPERIMENT: RESPONSIBILITY IN CROWDS

2. The two row means (8, 16) represent the mean reaction times for degree of danger when crowd size is ignored. Any difference between these row means not attributable to chance is referred to as the main effect of degree of danger on reaction time.

3. The mean of the reaction times for each group of two subjects yields the six means (8, 7, 9, 10, 17, 21) for each combination of the two factors. Often referred to as cell means or treatment-combination means, these means reflect not only the main

effects for crowd size and degree of danger described earlier but, more importantly, any effect due to the interaction between crowd size and degree of danger, as described below.

4. Finally, the one mean for all three column means—or for both row means— yields the overall or grand mean (12) for all subjects in the study.

Graphs for Main Effects

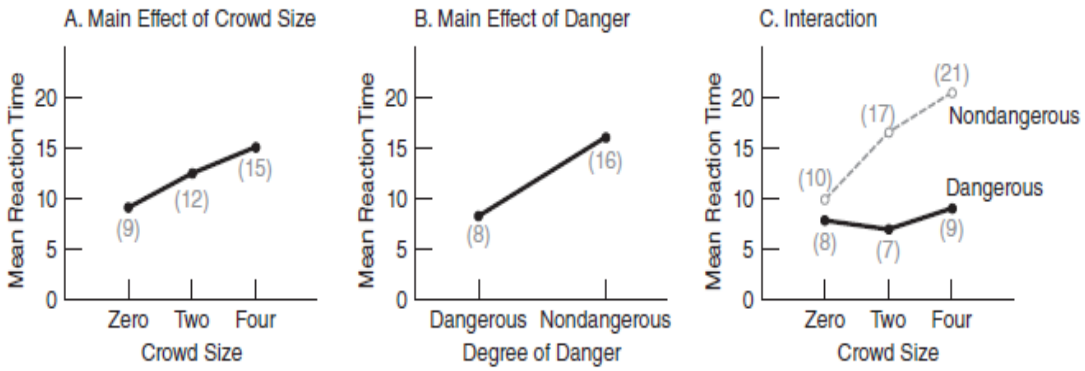
To preview the experimental outcomes, let's look for obvious trends in a series of graphs based on Table 18.1. The slanted line in panel A of Figure 18.1 depicts the large differences between column means, that is, between mean reaction times for subjects, regardless of degree of danger, with crowds of zero, two, and four people. The relatively steep slant of this line suggests that the null hypothesis for crowd size might be rejected. The steeper the slant is, the larger the observed differences between column means and the greater the suspected main effect of crowd size. On the other hand, a fairly level line in panel A of Figure 18.1 would have reflected the relative absence of any main effect due to crowd size. The slanted line in panel B of Figure 18.1 depicts the large difference between row means, that is, between mean reaction times for dangerous and non-dangerous conditions, regardless of crowd size. The relatively steep slope of this line suggests that the null hypothesis for degree of danger also might be rejected; that is, there might be a main effect due to degree of danger.

Graph for Interaction

These preliminary conclusions about main effects must be qualified because of a complication due to the combined effect or interaction of crowd size and degree of danger on reaction time.

Interaction occurs whenever the effects of one factor on the dependent variable

are not consistent for all values (or levels) of the second factor



Panel C of Figure 18.1 depicts the interaction between crowd size and degree of danger. The two nonparallel lines in panel C depict differences between the three cell means in the first row and the three cell means in the second row—that is, between the mean reaction times for the dangerous condition for different crowd sizes and the mean reaction times for the Non dangerous condition for different crowd sizes. Although the line for the dangerous conditions remains fairly level, that for the Non dangerous conditions is slanted, suggesting that the reaction times for the Non dangerous conditions, but not those for the dangerous conditions, are influenced by crowd size. Because the effect of crowd size is not consistent for the Non dangerous and dangerous conditions— portrayed by the apparent Non parallelism between the two lines in panel C of Figure 18.1—the null hypothesis (that there is no interaction between the two factors) might be rejected. Section 18.3 contains additional comments about interaction, as well as a more preferred definition of interaction.

Summary of Preliminary Interpretations

To summarize, a nonstatistical evaluation of the graphs of data for the two-factor experiment suggests a number of preliminary interpretations. Each of the three null hypotheses regarding the effects of crowd size, degree of danger, and the interaction of these factors might be rejected. Because of the suspected interaction, however, any generalizations about the main effects of one factor must be qualified in terms of specific levels of the second factor. Pending the outcome of the statistical analysis, you can speculate that the crowd size probably influences the reaction times for the Non dangerous but not the dangerous conditions.

Progress Check A college dietitian wishes to determine whether students prefer a particular pizza topping (either plain, vegetarian, salami, or everything) and one type of crust (either thick or thin). A total of 160 volunteers are randomly assigned to one of the eight cells in this two-factor experiment. After eating their assigned pizza, the 20 subjects in each cell rate their preference on a scale ranging from 0 (inedible) to 10 (the best). The results, in the form of means for cells, rows, and columns, are as follows:

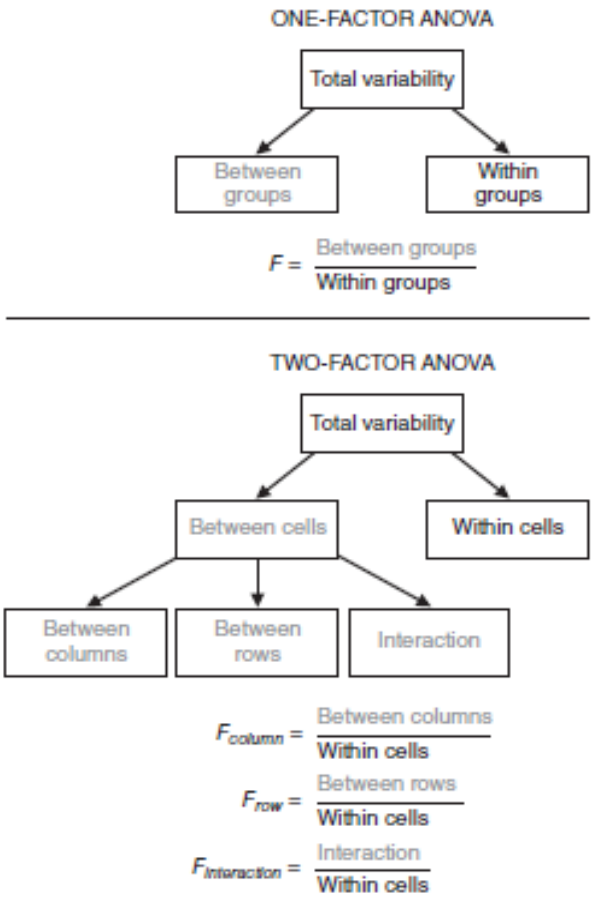
MEAN PREFERENCE SCORES OR PIZZA AS A FUNCTION OF TOPPING AND CRUST

	TOPPING				
CRUST	**PLAIN**	**VEGETARIAN**	**SALAMI**	**EVERYTHING**	**ROW**
Thick	7.2	5.7	4.8	6.1	6.0
Thin	8.9	4.8	8.4	1.3	5.9
Column	8.1	5.3	6.6	3.7	

Construct graphs for each of the three possible effects, and use this information to make preliminary interpretations about pizza preferences. Ordinarily, of course, you would verify these speculations by performing an ANOVA—a task that cannot be performed for these data, since only means are supplied.

THREE F TESTS

As suggested in Figure 18.2, F ratios in both a one- and a two-factor ANOVA always consist of a numerator (shaded) that measures some aspect of variability between



groups or cells and a denominator that measures variability within groups or cells. In a one-factor ANOVA, a single null hypothesis is tested with one F ratio.

In two-factor ANOVA, three different null hypotheses are tested, one at a time, with three *F* ratios: F_{column}, F_{row}, and $F_{interaction}$.

The numerator of each of these three F ratios reflects a different aspect of variability between cells: variability between columns (crowd size), variability between rows (degree of danger), and interaction—any remaining variability between cells not attributable to either variability between columns (crowd size) or rows (degree of danger).

The shaded numerator terms for the three F ratios in the bottom panel of Figure 18.2 estimate random error and, if present, a treatment effect (for subjects treated differently by the investigator). The denominator term always estimates only random error (for subjects treated similarly in the same cell). In practice, a

sufficiently large F value is viewed as rare, given that the null hypothesis is true, and therefore, it leads to the rejection of the null hypothesis. Otherwise, the null hypothesis is retained.

ANALYSIS OF VARIANCE (TWO FACTORS)

Test Results for Two-Factor Experiment As indicated in the boxed summary for the hypothesis test for a smoke alarm experiment, test results agree with our preliminary interpretations based on graphs. Each of the three null hypotheses is rejected at the .05 level of significance. The significant main effects indicate that crowd size and degree of danger, in turn, influence the reaction times of subjects to smoke. The significant interaction, however, indicates that the effect of crowd size on reaction times differs for Non dangerous and dangerous conditions.

INTERACTION

Interaction emerges as the most striking feature of a two-factor ANOVA. As noted previously, two factors interact if the effects of one factor on the dependent variable are not consistent for all of the levels of a second factor. More generally, when two factors are combined, something happens that represents more than a mere composite of their separate effects.

Supplies Valuable Information

Rather than being a complication to be avoided, an interaction often highlights pertinent issues for further study. For example, the interaction between crowd size and degree of danger might encourage the exploration, possibly by interviewing participants, about their reactions to various crowd sizes and degrees of danger. In the process, much might be learned about why some people in groups assume or fail to assume social responsibility.

Other Examples

The combined effect of crowd size and degree of danger could have differed from that described in panel C of Figure 18.1. Examples of some other possible effects are shown in Figure 18.3. The two top panels in Figure 18.3 describe outcomes that, because of their consistency, would cause the retention of the null hypothesis for interaction. The two bottom panels in Figure 18.3 describe outcomes that, because

of their inconsistency, probably would cause the rejection of the null hypothesis for interaction.

Simple Effects

The notion of interaction can be clarified further by viewing each line in Figure 18.3 as a simple effect. A simple effect represents the effect of one factor on the dependent variable at a single level of the second factor. Thus, in panel A, there are two simple effects of crowd size, one for Non dangerous conditions and one for dangerous conditions, and both simple effects are consistent, showing an increase in mean reaction times with larger crowd sizes. Accordingly, the main effect of crowd size can be interpreted without referring to its two simple effects.

Inconsistent Simple Effects

In panel D, on the other hand, the two simple effects of crowd size, one for Non dangerous conditions and one for dangerous conditions, clearly are inconsistent; the simple effect of crowd size for dangerous conditions shows a decrease in mean reaction times with larger crowd sizes, while the simple effect of crowd size for Non dangerous conditions shows just the opposite—an increase in mean reaction times with larger crowd sizes. Accordingly, the main effect of crowd size—assuming one exists— cannot be interpreted without referring to its radically different simple effects.

INTERACTION

HYPOTHESIS TEST SUMMARY
Two-Factor ANOVA (Smoke Alarm Experiment)

Research Problem

Do crowd size and degree of danger, as well as the interaction of these two factors, influence the subjects' mean reaction times to potentially dangerous smoke?

Statistical Hypotheses

H_0: no main effect due to columns or crowd size

(or $\mu_0 = \mu_2 = \mu_4$).

H_0: no main effect due to rows or degree of danger

(or $\mu_{dangerous} = \mu_{nondangerous}$).

H_0: no interaction.

H_1: H_0 is not true.

(Same H_1 accommodates each H_0.)

Decision Rule

Reject H_0 at the .05 level of significance if F_{column} or $F_{interaction} \geq 5.14$ (from Table C in Appendix C, given 2 and 6 degrees of freedom) and if $F_{row} \geq 5.99$ (given 1 and 6 degrees of freedom).

Calculations

$$F_{column} = 6.75$$
$$F_{row} = 36.02$$
$$F_{interaction} = 5.25$$

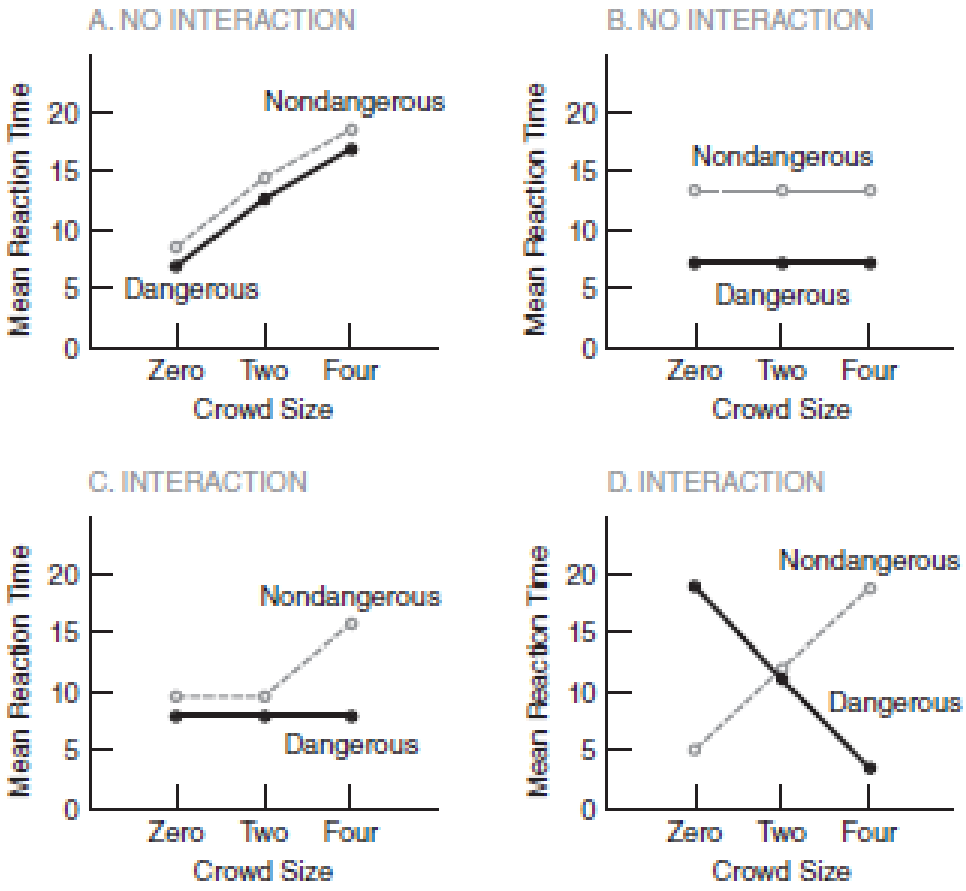
(See Tables 18.3 and 18.6 for more details.)

Decision

Reject all three null hypotheses at the .05 level of significance because $F_{column} = 6.75$ exceeds 5.14; $F_{row} = 36.02$ exceeds 5.99; and $F_{interaction} = 5.25$ exceeds 5.14.

Interpretation

Both crowd size and degree of danger influence the subjects' mean reaction times to smoke. The interaction indicates that the influence of crowd size depends on the degree of danger. It appears that the mean reaction times increase with crowd size for nondangerous but not for dangerous conditions.

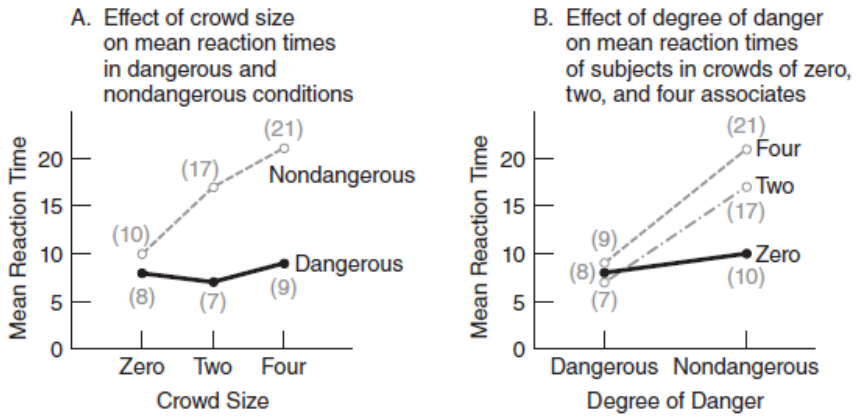


Simple Effects and Interaction

In Figure 18.3, no interaction is present in panels A and B because their respective simple effects are consistent, as suggested by the parallel lines. Interactions could be present in panels C and D because their respective simple effects are inconsistent, as suggested by the diverging or crossed lines. Given the present perspective, interaction can be viewed as the product of inconsistent simple effects

DETAILS: VARIANCE ESTIMATES

TWO VERSIONS OF THE SAME INTERACTION



Describing Interactions

The original interaction between crowd size and degree of danger could have been described in two different ways. First, we could have portrayed the inconsistent simple effects of crowd size for Non dangerous and dangerous conditions by showing panel A of Figure 18.4 (originally shown in panel C of Figure 18.1). Alternately, we could have portrayed the inconsistent simple effects of degree of danger for crowds of zero, two, and four people by showing panel B of Figure 18.4. Although different, the configurations in both panels A and B suggest essentially the same interpretation: Crowd size influences reaction times for Non dangerous but not for dangerous conditions. In cases where one perspective seems to make as much sense as another, it's customary to plot along the horizontal axis the factor with the larger number of levels, as in panel A.

DETAILS: VARIANCE ESTIMATES

Each of the three F ratios in a two-factor ANOVA is based on a ratio involving two variance estimates: a mean square in the numerator that reflects random error plus, if present, any specific treatment effect and a mean square in the denominator that reflects only random error. Before using these mean squares (or variance estimates), we must calculate their sums of squares and their degrees of freedom.

Sums of Squares (Definition Formulas)

Ultimately, the total sum of squares, SStotal, will be divided among its various component sums of squares, that is,

SUMS OF SQUARES (TWO FACTOR)

$$SS_{total} = SS_{column} + SS_{row} + SS_{interaction} + SS_{within} \quad (18.1)$$

ANALYSIS OF VARIANCE (TWO FACTORS)

The computation of the various *SS* terms can be viewed as a two-step effort.

1. The two factors and their interaction are ignored, and we calculate the first three sum of squares terms as if the data originated from a one-factor ANOVA where total variability is partitioned into two components: variability between cells and variability within cells, since

$$SS_{total} = SS_{between} + SS_{within}$$

Variability within cells, SS_{within}, which is often referred to as SS_{error}, will serve as the sum of squares portion of the error mean square in the denominator of each of the three *F* ratios.

- As always, the total sum of squares, SS_{total}, equals the sum of squared deviations of all scores, X, about the grand mean for all scores, $\overline{X}_{grand}$, that is,

$$SS_{total} = \Sigma(X - \overline{X}_{grand})^2.$$

- The between-cells (or treatment) sum of squares, $SS_{between}$, equals the sum of squared deviations of all cell (or treatment) means, $\overline{X}_{cell}$, about the grand mean, $\overline{X}_{grand}$. Expressed symbolically, $SS_{between} = n\Sigma\left(\overline{X}_{cell} - \overline{X}_{grand}\right)^2$, where n, the sample size in each cell, adjusts for the fact that the deviation $\overline{X}_{cell} - \overline{X}_{grand}$ is the same for every score in its cell.
- The within-cells (or error) sum of squares, SS_{within}, equals the sum of squared deviations of all scores, X, about their respective cell means, $\overline{X}_{cell}$, that is, $SS_{within} = \Sigma\left(X - \overline{X}_{cell}\right)^2$. Essentially, this expression requires that the sum of squares within each cell be added across all cells as the first step toward a pooled variance estimate of random error.

2. Variability between cells, $SS_{between}$, is partitioned into three additional sums of squares—SS_{column}, SS_{row}, and $SS_{interaction}$—that reflect identifiable sources of treatment variability in the two-factor ANOVA, since

$$SS_{between} = SS_{column} + SS_{row} + SS_{interaction}$$

- The between-columns sum of squares, SS_{column}, equals the sum of squared deviations of column means, $\overline{X}_{column}$, about the grand mean, $\overline{X}_{grand}$. Expressed symbolically, $SS_{column} = rn\Sigma(\overline{X}_{column} - \overline{X}_{grand})^2$, where r equals the number of rows, n equals the sample size in each cell, and rn equals the total sample size for each column. The product rn adjusts for the fact that the mean deviation, $\overline{X}_{column} - \overline{X}_{grand}$, is the same for every score in its column.
- The between-rows sum of squares, SS_{row}, equals the sum of squared deviations of row means, $\overline{X}_{row}$, about the grand mean, $\overline{X}_{grand}$. Expressed symbolically, $SS_{row} = cn\Sigma(\overline{X}_{row} - \overline{X}_{grand})^2$, where c equals the number of columns and cn equals the total sample size in each row. The product cn adjusts for the fact that the mean deviation, $\overline{X}_{row} - \overline{X}_{grand}$, is the same for every score in its row.
- The interaction sum of squares, $SS_{interaction}$, equals the variability between cells, $SS_{between}$, after the removal of variability between columns, SS_{column}, and variability between rows, SS_{row}, that is,

$$SS_{interaction} = SS_{between} - (SS_{column} + SS_{row})$$

Although $SS_{interaction}$ could be expressed more directly by expanding these three SS terms, the result tends to be more cumbersome than enlightening.

DETAILS: VARIANCE ESTIMATES

Table 18.2
WORD, DEFINITION, AND COMPUTATION FORMULAS FOR *SS* TERMS (TWO-FACTOR ANOVA)

For the total sums of squares,

SS_{total} = the sum of squared deviations for raw scores about the grand mean

$$= \Sigma\left(X - \overline{X}_{grand}\right)^2$$

$$SS_{total} = \Sigma X^2 - \frac{G^2}{N}$$, where G is the grand total and N is its sample size

For the between-cells sum of squares,

$SS_{between}$ = the sum of squared deviations for cell means about the grand mean

$$= n\Sigma\left(\overline{X}_{cell} - \overline{X}_{grand}\right)^2$$

$$SS_{between} = \Sigma\frac{T_{cell}^2}{n} - \frac{G^2}{N}$$, where T_{cell} is the cell total and n is its sample size of each cell

For the within-cells sum of squares,

SS_{within} = the sum of squared deviations of raw scores about their respective cell means

$$= \Sigma\left(X - \overline{X}_{cell}\right)^2$$

$$SS_{within} = \Sigma X^2 - \Sigma\frac{T_{cell}^2}{n}$$, where T_{cell} is the cell total and n is the sample size of each cell

For the between-columns sum of squares,

SS_{column} = the sum of squared column means about the grand mean

$$= rn\Sigma\left(\overline{X}_{column} - \overline{X}_{grand}\right)^2$$

$$SS_{column} = \Sigma\frac{T_{column}^2}{rn} - \frac{G^2}{N}$$, where T_{column} is the column total, r is the number of rows, and rn is th sample size of each column

For the between-rows sum of squares,

SS_{row} = the sum of squared of raw means about the grand mean

$$= cn\Sigma\left(\overline{X}_{row} - \overline{X}_{grand}\right)^2$$

$$SS_{row} = \Sigma\frac{T_{row}^2}{cn} - \frac{G^2}{N}$$, where T_{row} is the row total, c is the number of columns, and cn is the sample size of each row

For the interaction sum of squares,

$$SS_{interaction} = SS_{between} - (SS_{columns} + SS_{row})$$

Sums of Squares (Computation Formulas)

Table 18.2 shows the more efficient, computation formulas, where totals replace means. Notice the highly predictable computational pattern first described in Section 16.4. Each entry is squared, and each total, whether for a column, a row, a

cell, or the grand total, is then divided by its respective sample size. Table 18.3 illustrates the application of these formulas to the data for the two-factor experiment.

Table 18.3
CALCULATION OF *SS* TERMS (TWO-FACTOR ANOVA)

A. COMPUTATIONAL SEQUENCE

Find (and circle) each cell total [1].
Find each column and row total and also the grand total [2]
Substitute numbers into computational formula [3] and solve for SS_{total}
Substitute numbers into computational formula [4] and solve for $SS_{between}$
Substitute numbers into computational formula [5] and solve for SS_{within}
Substitute numbers into computational formula [6] and solve for SS_{column}
Substitute numbers into computational formula [7] and solve for SS_{row}
Substitute numbers into formula [8] and solve for $SS_{interaction}$

B. DATA AND COMPUTATIONS

Crowd Size Degree of Danger

	ZERO	TWO	FOUR	2 Row Totals
Dangerous	8, 8	8, 6	10, 8	48
Nondangerous	9, 11	15, 19	24, 18	96
2 Column Totals →	36	48	60	2 Grand Total = 144

3 $SS_{total} = \Sigma X^2 - \frac{G^2}{N}$

$= (8)^2 + (8)^2 + \ldots + (24)^2 + (18)^2 - \frac{(144)^2}{12} = 2080 - 1728 = 352$

4 $SS_{between} = \Sigma \frac{T^2_{cell}}{n} - \frac{G^2}{N}$

$= \frac{(16)^2}{2} + \frac{(20)^2}{2} + \frac{(14)^2}{2} + \frac{(34)^2}{2} + \frac{(18)^2}{2} + \frac{(42)^2}{2} - \frac{(144)^2}{12} = 2048 - 1728 = 320$

5 $SS_{within} = \Sigma X^2 - \Sigma \frac{T^2_{cell}}{n}$

$= (8)^2 + (8)^2 + \ldots + (24)^2 + (18)^2 - \left[\frac{(16)^2}{2} + \frac{(20)^2}{2} + \frac{(14)^2}{2} + \frac{(34)^2}{2} + \frac{(18)^2}{2} + \frac{(42)^2}{2}\right] = 2080 - 2048 = 32$

6 $SS_{column} = \Sigma \frac{T^2_{column}}{m} - \frac{G^2}{N}$

$= \left[\frac{(36)^2}{4} + \frac{(48)^2}{4} + \frac{(60)^2}{4}\right] - \frac{(144)^2}{12} = 1800 - 1728 = 72$

7 $SS_{row} = \Sigma \frac{T^2_{row}}{cn} - \frac{G^2}{N}$

$= \left[\frac{(48)^2}{6} + \frac{(96)^2}{6}\right] - \frac{(144)^2}{12} = 1920 - 1728 = 192$

8 $SS_{interaction} = SS_{between} - (SS_{column} + SS_{row})$

$= 320 - (72 + 192) = 56$

DETAILS: MEAN SQUARES (MS) AND F RATIOS

Table 18.4
FORMULAS FOR *df* TERMS: TWO-FACTOR ANOVA

$df_{total} = N - 1$, that is, the number of all scores $- 1$
$df_{column} = c - 1$, that is, the number of columns $- 1$
$df_{row} = r - 1$, that is, the number of rows $- 1$
$df_{interaction} = (c - 1)(r - 1)$, that is, the product of df_{row} and df_{column}
$df_{within} = N - (c)(r)$, that is, the number of all scores $-$ the number of cells

Degrees of Freedom (*df*)

The number of degrees of freedom must be determined for each *SS* term in a two-factor ANOVA, and for convenience, the various *df* formulas are listed in **Table 18.4**. The $(c - 1)(r - 1)$ degrees of freedom for $df_{interaction}$ reflect the fact that, from the perspective of degrees of freedom, the original matrix with $(c)(r)$ cells shrinks to $(c - 1)(r - 1)$ cells for $df_{interaction}$. One row and one column of cell totals in the original matrix are not free to vary because of the restriction that all cell totals in each column and all cell totals in each row must sum to fixed totals in the margins (associated with column and row factors.) The $N - (c)(r)$ *degrees of freedom for* df_{within} reflect the fact that the N scores within all cells must sum to the fixed totals in their respective cells, causing one degree of freedom to be lost in each of the $(c)(r)$ cells.

The *df* values for the present study are:

$$df_{total} = N - 1 = 12 - 1 = 11$$
$$df_{column} = c - 1 = 3 - 1 = 2$$
$$df_{row} = r - 1 = 2 - 1 = 1$$
$$df_{interaction} = (c - 1)(r - 1) = (3 - 1)(2 - 1) = 2$$
$$df_{within} = N - (c)(r) = 12 - (3)(2) = 6$$

Check for Accuracy

Recall the general rule that the degrees of freedom for SS_{total} equal the combined degrees of freedom for all remaining *SS* terms, that is,

DEGREES OF FREEDOM (TWO FACTOR)

$$df_{total} = df_{column} + df_{row} + df_{interaction} + df_{within} \qquad (18.2)$$

This formula can be used to verify that the correct number of degrees of freedom has been assigned to each *SS* term.

DETAILS: MEAN SQUARES (MS) AND F RATIOS

Having found values for the various SS terms and their df, we can determine values for the corresponding MS terms and then calculate the three F ratios using the formulas in Table 18.5. Notice that MSwithin appears in the denominator of each of these three F ratios. MSwithin is based on the variability among scores of subjects who are treated

Table 18.5
FORMULAS FOR MEAN SQUARES (*MS*) AND *F* RATIOS

SOURCE	*MS*	*F*
Column	$MS_{column} = \frac{SS_{column}}{df_{column}}$	$F_{column} = \frac{MS_{column}}{MS_{within}}$
Row	$MS_{row} = \frac{SS_{row}}{df_{row}}$	$F_{row} = \frac{MS_{row}}{MS_{within}}$
Interaction	$MS_{interaction} = \frac{SS_{interaction}}{df_{interaction}}$	$F_{interaction} = \frac{MS_{interaction}}{MS_{within}}$
Within	$MS_{within} = \frac{SS_{within}}{df_{within}}$	

similarly, within each cell, pooled across all cells. Regardless of whether any treatment effect is present, it measures only random error. The ANOVA results for the two-factor study are summarized in Table 18.6. The shaded numbers (which ordinarily don't appear in ANOVA summary tables) indicate the origin of each MS term and of each F.

Other Labels

Other labels also might have appeared in Table 18.6. For instance, "Column" and "Row" might have been replaced by descriptions of the treatment variables, in this case, "Crowd Size" and "Degree of Danger." Similarly, "Interaction" might have been replaced by "Crowd Size Degree of Danger," by "Crowd Size * Degree of Danger," or by some abbreviation, such as "CS DD," and "Within" might have been replaced by "Error."

Table 18.6
ANOVA TABLE (TWO-FACTOR EXPERIMENT)

SOURCE	*SS*	*df*	*MS*	*F*
Column	72	2	$\left(\frac{72}{2}=\right)$ 36	$\left(\frac{36}{5.33}=\right)$ 6.75*
Row	192	1	$\left(\frac{192}{1}=\right)$ 192	$\left(\frac{192}{5.33}=\right)$ 36.02*
Interaction	56	2	$\left(\frac{56}{2}=\right)$ 28	$\left(\frac{28}{5.33}=\right)$ 5.25*
Within	32	6	$\left(\frac{32}{6}=\right)$ 5.33	
Total	352	11		

ESTIMATING EFFECT SIZE

TABLE FOR THE F DISTRIBUTION

Each of the three F ratios in Table 18.6 exceeds its respective critical F ratio. To obtain critical F ratios from the F sampling distribution, refer to Table C in Appendix C. Follow the usual procedure, described in Section 16.6, to verify that when 2 and 6 degrees of freedom are associated with Fcolumn and Finteraction, the critical F equals 5.14, and that when 1 and 6 degrees of freedom are associated with Frow, the critical F equals 5.99.

Progress Check A school psychologist wishes to determine the effect of TV violence on disruptive behavior of first graders in the classroom. Two first graders are randomly assigned to each of the various combinations of the two factors: the type of violent TV program (either cartoon or real life) and the amount of viewing time (either 0, 1, 2, or 3 hours). The subjects are then observed in a controlled classroom setting and assigned a score, reflecting the total number of disruptive class behaviors displayed during the test period.

AGGRESSION SCORES OF FIRST GRADERS

	VIEWING TIME (HOURS)			
TYPE OF PROGRAM	**0**	**1**	**2**	**3**
Cartoon	0,1	1,0	3,5	6,9
Real life	0,0	1,1	6,2	6,10

(a) Test the various null hypotheses at the .05 level of significance.

(b) Summarize the results with an ANOVA table. Save the ANOVA summary table for use in subsequent questions.

ESTIMATING EFFECT SIZE

In the previous chapter, a version of the squared curvilinear correlation, η_p^2, was used to estimate effect size after variance due to individual differences had been removed. Essentially the same type of analysis can be conducted for *each* significant F in a two-factor ANOVA. Each η_p^2 estimates the proportion of the total variance attributable to either one of the two factors or to the interaction—after excluding from the total known amounts of variance attributable to the remaining treatment components.

In each case, η_p^2 is calculated by dividing the appropriate sum of squares (either SS_{column}, SS_{row}, or $SS_{interaction}$) by the appropriately reduced total sum of squares, that is,

PROPORTION OF EXPLAINED VARIANCE (TWO-FACTOR ANOVA)

$$\eta_p^2(\text{column}) = \frac{SS_{column}}{SS_{total} - \left(SS_{row} + SS_{interaction}\right)} = \frac{SS_{column}}{SS_{column} + SS_{within}}$$

$$\eta_p^2(\text{row}) = \frac{SS_{row}}{SS_{row} + SS_{within}} \qquad (18.3)$$

$$\eta_p^2(\text{interaction}) = \frac{SS_{interaction}}{SS_{interaction} + SS_{within}}$$

where each η_p^2 is referred to as a *partial* η^2 for that component because the effects of the other two treatment components have been eliminated from the reduced or partial total variance.

Substituting values for the *SS* terms from Table 18.6, we have

$$\eta_p^2(\text{column}) = \frac{72}{72+32} = .69$$

$$\eta_p^2(\text{row}) = \frac{192}{192+32} = .86$$

$$\eta_p^2(\text{interaction}) = \frac{56}{56+32} = .64$$

All three of these estimates would be considered spectacularly large since, according to guidelines derived from Cohen, the estimated effect for any factor or interaction is small if η_p^2 approximates .01; medium if η_p^2 approximates .09; and large if η_p^2 approximates .25 or more. For instance, the value of .86 for η_p^2(row) indicates that .86, or 86 percent, of total variance in reaction times (excluding variance due to crowd size and the interaction) is explained by differences between nondangerous and dangerous conditions, while only the remaining 14 percent of the variance in reaction times is not explained by degree of danger.

MULTIPLE COMPARISONS

Tukey's *HSD* test for multiple comparisons can be used to pinpoint important differences between pairs of column or row means whenever the corresponding main effects are statistically significant and *interpretations of these main effects are not compromised by any inconsistencies associated with a statistically significant interaction.*

To determine *HSD*, use the following expression:

TUKEY'S *HSD* TEST (TWO-FACTOR ANOVA)

$$HSD = q\sqrt{\frac{MS_{within}}{n}} \qquad (18.4)$$

where *HSD* is the positive critical value for any difference between two column or row means; q is a value obtained from Table G in Appendix C; MS_{within} is the mean square for within-cells variability in the two-factor ANOVA; and, *if pairs of column means are being compared, n is rn, the sample size for the entire column, or if pairs of row means are being compared, n is cn, the sample size for an entire row.* To determine the value of q at the .05 or .01 level (light or dark numbers, respectively) in Table G, find the cell intersected by c or r (depending on whether column or row means are being compared) and by df_{within}. The df_{within} is the number of degrees of freedom for within-group variability, $N - (c)\,(r)$, which equals the total number of scores in the two-factor ANOVA, N, minus the total number of cells, $(c)\,(r)$.

In the smoke alarm experiment, Tukey's *HSD* test isn't conducted for the one significant main effect, crowd size, with more than two group means because of the presence of a significant interaction that compromises any interpretation of the main effect.

SIMPLE EFFECTS

Progress Check in Question 18.3, the F for the interaction isn't significant, but F for one of the main effects, Viewing Time, is significant. Using the .05 level, calculate the critical value for Tukey's HSD; evaluate the significance of each possible mean difference for Viewing Time; and interpret the results.

SIMPLE EFFECTS

Whenever the interaction is statistically significant, as in the two-factor smoke alarm experiment, we can conduct new Fse tests, where the se subscript stands for "simple effect," to identify the inconsistencies among simple effects that produce the interaction. These new tests require that, by ignoring the second factor, the original

two-factor ANOVA be transformed into several one-factor or simple-effect ANOVAs. Essentially, the Fse test for simple effects tests the effect of one factor on the dependent variable at a single level of another factor. Table 18.7 shows how the totals for the original two-factor experiment can be viewed as two simple effects for crowd size (corresponding to each one of the two rows in the original two-factor matrix) and three simple effects for degree of danger (corresponding to each of the three columns). Inconsistencies among a set of simple effects usually are associated with a mixture of both significant and nonsignificant Fse tests for that set of simple effects. Among the two simple (row) effects for crowd size, the Fse test is nonsignificant (ns) for the dangerous condition but significant ($p < .01$) for the Non dangerous condition, suggesting that reaction times increase with larger crowd sizes for Non dangerous but not for dangerous conditions. Essentially the same conclusion is suggested by the other set of three simple effects. Among these simple (column) effects for degree of danger, the Fse test for crowd sizes of zero is nonsignificant (ns), but it is significant ($p < .01$) for crowd sizes of two and four, suggesting that the reaction times for Non dangerous conditions exceed those for dangerous conditions for crowd sizes of two and four but not for crowd sizes of zero. Ordinarily, you needn't test both sets of simple effects, as was done above for the sake of completeness. Instead, test only one set, preferably the one that seems to describe best the significant interaction.

Table 18.7
SIMPLE EFFECTS FOR SMOKE ALARM EXPERIMENT (TOTALS)

TWO-FACTOR EXPERIMENT

DEGREE OF DANGER	CROWD SIZE ZERO	TWO	FOUR	ROW TOTAL
Dangerous	16	14	18	48
Nondangerous	20	34	42	96
Column Total	36	48	60	

SIMPLE EFFECT OF CROWD SIZE AT					
DANGEROUS	16	14	18	48	(*ns*)
NONDANGEROUS	20	34	42	96	($p < .01$)

Simple Effect of Degree of Danger at

Zero	Two	Four
16	14	18
20	34	42
36	48	60
(*ns*)	($p < .01$)	($p < .01$)

F_{se} Test for Simple Effects

The new F_{se} test for any simple effect is very similar to the F test for the corresponding main effect in a two-factor ANOVA. The degrees of freedom term is the same, as is the term in the denominator, MS_{within}. Only the term in the numerator, MS_{se}, must be adjusted to estimate the variability associated with just one row or one column. The ratio for any simple effect, F_{se} reads:

F_{se} RATIO (SIMPLE EFFECT)

$$F_{se} = \frac{MS_{se}}{MS_{within}} \tag{18.5}$$

where MS_{se} represents the mean square for the variation of every cell mean in a single row (or a single column) about the overall or grand mean for that entire row (or column) and MS_{within} represents the mean square for the variation of all scores about their cell means for the entire two-factor matrix. (Being based on all scores, MS_{within} serves in the denominator term as the best estimate of random error.)

The degrees of freedom for the numerator of the simple-effect F_{se} ratio is the same as that for the corresponding main-effect F ratio, namely, $df_{between}$, which equals either $c - 1$ *or* $r - 1$, the number of groups (that is, the number of columns or of rows) in the simple effect minus one. The degrees of freedom for the denominator of the simple-effect F_{se} ratio is the same as that for any two-factor F ratio, namely, df_{within}, which equals $N - (c)(r)$: the total number of scores, N, minus the total number of cells, $(c)(r)$, in the two-factor ANOVA.

Calculating F_{se}

Let's calculate F_{se} for the simple effect of crowd size for the nondangerous condition (that is, for the second row in Table 18.7). Since MS_{within} already has been calculated and, as shown in Table 18.6, equals 5.33, we can concentrate on calculating SS_{se}, which, when divided by its degrees of freedom, gives a value for MS_{se}. The computational formula for SS_{se} reads:

SUM OF SQUARES (SIMPLE EFFECT)

$$SS_{se} = \Sigma \frac{T_{se}^2}{n} - \frac{G_{se}^2}{N_{se}} \qquad (18.6)$$

where SS_{se} now signifies the sum of squares for the simple effect; T^2_{se} represents the squared total for each cell in a single row (or a single column); G^2_{se} represents the grand total for all cells in the entire row (or column); n equals the sample size for each cell; and N_{se} equals the total sample size for the entire row (or column).

When totals from the second row in Table 18.7 are substituted into Equation 18.6, it reads:

$$SS_{se}(crowd\ size\ at\ nondangerous) = \frac{(20)^2}{2} + \frac{(34)^2}{2} + \frac{(42)^2}{2} - \frac{(96)^2}{6} = 124$$

Given that the degrees of freedom for MS_{se} (*crowd size at nondangerous*) equals the number of columns minus one, that is, $c - 1 = 3 - 1 = 2$, then

$$MS_{se}(crowd\ size\ at\ nondangerous) = \frac{SS_{se}(crowd\ size\ at\ nondangerous)}{df_{column}} = \frac{124}{2} = 62$$

SIMPLE EFFECTS

$$F_{se}(\textit{crowd size at nondangerous}) = \frac{MS_{se}(\textit{crowd size at nondangerous})}{MS_{within}} = \frac{62}{5.33} = 11.63$$

which is significant ($p < .01$) since 11.63 exceeds the value of 10.92 in Table C in Appendix C for the .01 level of significance, given 2 and 6 degrees of freedom. Using essentially the same procedure, we also can establish that the simple effect of *crowd size at dangerous* (that is, the first row in Table 18.7), is nonsignificant (*ns*) since F_{se} (*crowd size at dangerous*) = 0.38, again with 2 and 6 degrees of freedom. The different test results—one significant, the other nonsignificant—provide statistical support for an important result of the smoke alarm study, namely, that the reaction times tend to increase with crowd size for nondangerous but not for dangerous conditions.

Tukey's *HSD* Test for Multiple Comparisons

If a simple effect is significant and involves more than two groups, Tukey's *HSD* test, as defined in Equation 18.4, can be used to identify pairs of cell means that differ significantly. When using Equation 18.4, the n in the denominator always refers to the sample size of the means being compared, that is, in the case of a simple effect, the sample size for each cell mean. Otherwise, all substitutions are the same whether you're testing a simple effect or a main effect.

Since the simple effect for crowd size at nondangerous is significant and involves more than two groups, Tukey's *HSD* test can be used. Consult Table G in Appendix C, given c (or k) = 3 for the three cells in the simple effect and $df_{within} = 6$, to find the value of q for the .01 level. Substituting values for $q = 6.33$, $MS_{within} = 5.33$, and $n = 2$ into Equation 18.4:

$$HSD = q\sqrt{\frac{MS_{within}}{n}} = 6.33\sqrt{\frac{5.33}{2}} = 6.33(1.63) = 10.32$$

Given values of 10, 17, and 21 for the *mean* reaction times for the nondangerous condition with crowds of zero, two, and four people, respectively, a significant difference ($p < .01$) occurs between the mean reaction times in the nondangerous condition for crowds of zero and four people because the observed mean difference, $21 - 10 = 11$, exceeds the *HSD* value of 10.32. A "borderline" significant difference (within a rounding margin of $p < .05$) occurs between the mean reaction times in the nondangerous condition for crowds of zero and two people because the observed mean difference, $17 - 10 = 7$, is only slightly less than the *HSD* value of 7.07. (This value is obtained from the *HSD* equation above, given $q = 4.34$ from the .05 level of Table G.)

Estimating Effect Size

As we have seen, a significant simple effect with more than two groups can be analyzed further with Tukey's *HSD* test. A significant difference between pairs of means can, in turn, have its effect size estimated with Cohen's *d*, as defined in Equation 16.10 on page 313, that is,

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_{within}}}$$

where d is an estimate of the standardized effect size; $\bar{X}_1$ and $\bar{X}_2$ are the pair of significantly different means; and $\sqrt{MS_{within}}$, the square root of the within-group mean square for the two-factor ANOVA, represents the sample standard deviation.

To estimate the standardized effect size for the significant difference between means for the nondangerous condition with crowd sizes of zero and four, enter $\bar{X}_4 - \bar{X}_0 = 11$ and $MS_{within} = 5.33$ in the above equation and solve for d:

$$d\left(\bar{X}_4, \bar{X}_0\right) = \frac{11}{\sqrt{5.33}} = \frac{11}{2.31} = 4.76$$

which is a very large effect, equivalent to almost five standard deviations. To estimate the standardized effect size for the significant difference between means for the nondangerous condition with crowds of zero and two people, enter $\bar{X}_2 - \bar{X}_0 = 7$ and $MS_{within} = 5.33$ in the above equation and solve for d:

$$d\left(\bar{X}_2, \bar{X}_0\right) = \frac{7}{\sqrt{5.33}} = \frac{7}{2.31} = 3.03$$

which also is a very large effect, equivalent to three standard deviations. Ordinarily, such large values of d, as well as the large values for η_p^2 in the current example, wouldn't be obtained with real data. The fictitious data for the smoke alarm experiment were selected to dramatize various effects in two-factor ANOVA, including an interaction with a significant simple effect, using very small sample sizes.

OVERVIEW: FLOW CHART

FOR TWO-FACTOR ANOVA

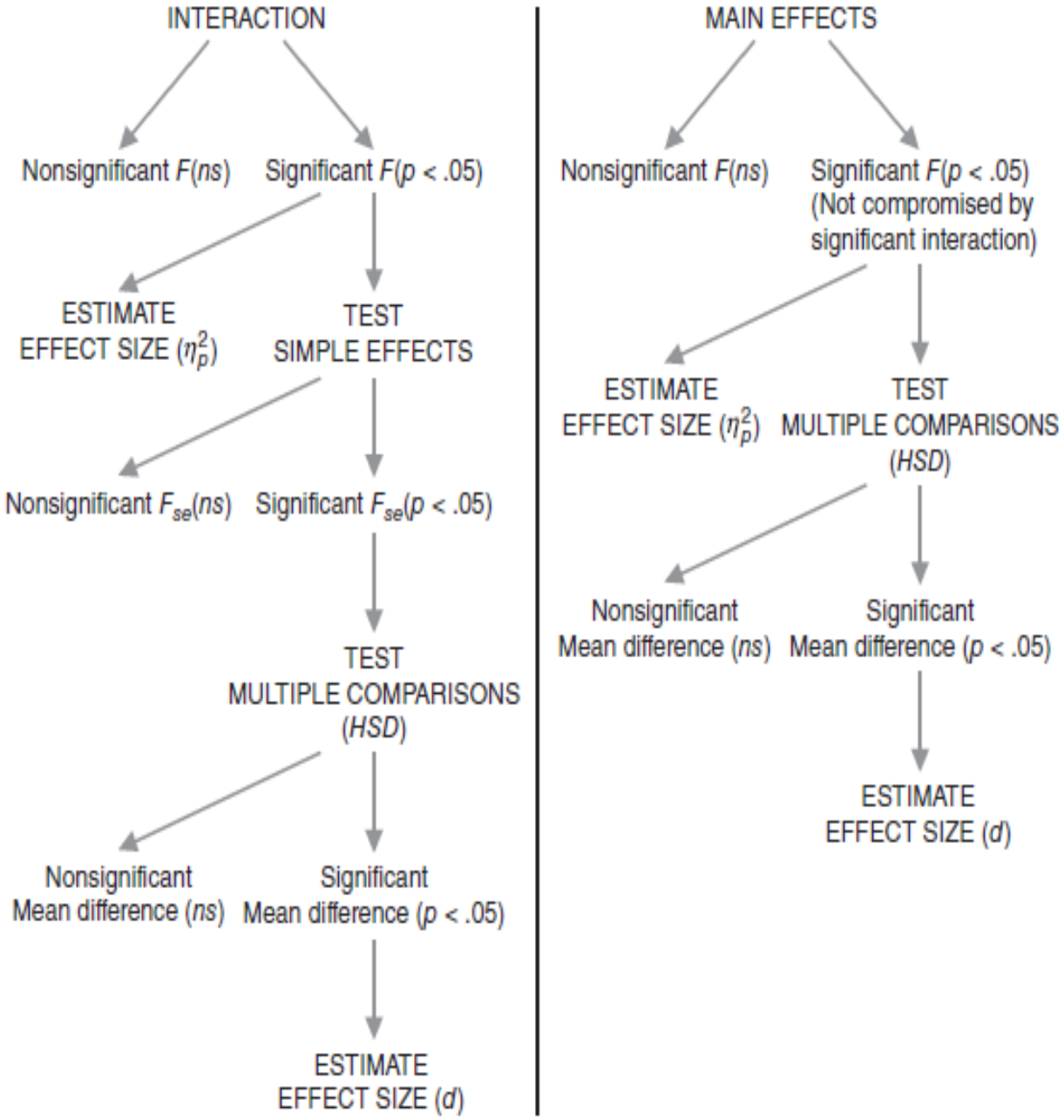
Figure 18.5 shows the steps to be taken when you are analyzing data for a two-factor ANOVA. Once an ANOVA summary table has been obtained, focus on the left-hand panel of Figure 18.5 for the interaction. If the interaction is significant, estimate its effect size with η_p^2 and conduct F_{se} tests for at least one set of simple effects. Ordinarily, the significant interaction will translate into a mix of significant and nonsignificant simple effects. Further, analyze any significant simple effect with *HSD* tests and any significant *HSD* test with an estimate of its effect size, *d*.

Next, focus on the right-hand panel for the main effects. Proceed with additional estimates for η_p^2 and *d*, and with the *HSD* test, only if the interpretation of the significant main effect isn't compromised by a significant interaction.

REPORTS IN THE LITERATURE

Test results for the smoke alarm experiment might be reported as follows: The following table shows the mean reaction times to smoke for subjects as a function of crowd size (zero, two, and four people) and degree of danger (Non dangerous and dangerous):

REPORTS IN THE LITERATURE



Flow chart for two-factor ANOVA.

	ZERO	TWO	FOUR	DEGREE OF DANGER
Dangerous	8	7	9	8
Nondangerous	10	17	21	16
Crowd Size	9	12	15	12

Mean reaction times increase with crowd size [$F(2, 6) = 6.75$, $MSE = 5.33$, $p < .05$, $\eta_p^2 = .69$]; they are larger for nondangerous than for dangerous conditions [$F((1, 6) = 36.02$, $p < .01$, $\eta_p^2 = .86$]; but these findings must be qualified because of the significant interaction [$F(2, 6) = 5.25$, $p < .05$, $\eta_p^2 = .64$]. An analysis of simple effects for crowd size confirms that reaction times increase with crowd size for nondangerous conditions [$F_{se}(2, 6) = 11.63$, $p < .01$] but not for dangerous conditions [$F_{se}(2, 6) = 0.38$, ns]. Furthermore, compared with the mean reaction time of 10 in the nondangerous condition with zero people, the mean reaction time of 17 in the nondangerous condition with two people is significantly longer ($HSD = 7.07$, $p = .05$, $d = 3.03$), and the mean reaction time of 21 in the nondangerous condition with four people also is significantly longer ($HSD = 10.32$, $p < .01$, $d = 4.76$). To summarize, the mean reaction times increase in the presence of crowds of two or four people in nondangerous but not in dangerous conditions.

ANALYSIS OF VARIANCE (TWO FACTORS)

This report reflects the prevalent use of approximate *p*-values rather than a fixed level of significance. The error (or within-group) mean square, $MSE = 5.33$, appears only for the initial F test since it's the same for all remaining F tests. The expression $p = .05$ (rather than $p < .05$) reflects the previously mentioned borderline significance of $\overline{X}_2 - \overline{X}_0 = 7$, given a critical value of $HSD = 7.07$.

ASSUMPTIONS

The assumptions for F tests in a two-factor ANOVA are similar to those for a one factor ANOVA. All underlying populations (for each treatment combination or cell) are assumed to be normally distributed, with equal variances. As with the one-factor ANOVA, you need not be too concerned about violations of these assumptions, particularly if all cell sizes are equal and each cell is fairly large (greater than about

10). Otherwise, in the unlikely event that you encounter conspicuous departures from normality or equality of variances, consult a more advanced statistics book.*

Importance of Equal Sample Sizes

As far as possible, all cells in two-factor studies should have equal sample sizes.Otherwise, to the degree that sample sizes are unequal and the resulting design lacks balance, not only are any violations of assumptions more serious, but problems of interpretation can occur. If you must analyze data based on unequal sample sizes— possibly because of missing subjects, equipment breakdowns, or recording errors —consult a more advanced statistics book.*

OTHER TYPES OF ANOVA

One- and two-factor studies do not exhaust the possibilities for ANOVA. For instance, you could use ANOVA to analyze the results of a three-factor study with three independent variables, three 2-way interactions, and one 3-way interaction. Furthermore, regardless of the number of factors, each subject might be measured repeatedly along all levels of one or more factors. Although the basic concepts described in this book transfer almost intact to a wide assortment of more intricate research designs, computational procedures grow more complex, and the interpretation of results often is more difficult. Intricate research designs, requiring the use of complex types of ANOVA, provide the skilled investigator with powerful tools for evaluating complicated situations. Under no circumstances, however, should a study be valued simply because of the complexity of its design and statistical analysis. Use the least complex design and analysis that will answer your research questions.

CHI-SQUARE (χ2) TEST FOR QUALITATIVE (NOMINAL) DATA

ONE-VARIABLE χ 2 TEST

SURVEY OF BLOOD TYPES

Your blood belongs to one of four genetically determined types: O, A, B, or AB. A bulletin issued by a large blood bank claims that these four blood types are distributed in the U.S. population according to the following proportions: .44 are type 0, .41 are type A, .10 are type B, and .05 are type AB. Let's treat this claim as a null hypothesis to be tested with a random sample of 100 students from a large university.

A Test for Qualitative (Nominal) Data

When observations are merely classified into various categories—for example, as blood types: O, A, B, and AB; as political affiliations: Republican, Democrat, and independent; as ethnic backgrounds: African-American, Asian-American, European-American, etc., the data are qualitative and measurement is nominal, as discussed in Chapter 1. Hypothesis tests for qualitative data require the use of a new test known as the chi-square test (symbolized as χ2 and pronounced "ki square").

One-Variable versus Two-Variable

When observations are classified in only one way, that is, classified along a single qualitative variable, as with the four blood types, the test is a **one-variable χ 2 test.** Designed to evaluate the adequacy with which observed frequencies are described by hypothesized or expected frequencies, a one-variable χ2 test is also referred to as a goodness-of-fit test. Later, when observations are classified in two ways, that is, cross-classified according to two qualitative variables, the test is a two-variable χ2.

STATISTICAL HYPOTHESES

Null Hypothesis For the one-variable χ2 test, the null hypothesis makes a statement about two or more population proportions whose values, in turn, generate the hypothesized or expected frequencies for the statistical test. Sometimes these population proportions are specified directly, as in the survey of blood types:

$$H_0: P_{\mathrm{O}} = .44; P_{\mathrm{A}} = .41; P_{\mathrm{B}} = .10; P_{\mathrm{AB}} = .05$$

where PO refers to the hypothesized proportion of students with type O blood in the population from which the sample was taken, and so forth. Notice that the values of population proportions always must sum to 1.00.

Other Examples At other times, you will have to infer the values of population proportions from verbal statements. For example, the null hypothesis that artists are equally likely to be left-handed or right-handed translates into

$$H_0: P_{left} = P_{right} = .50 \text{ (or } 1/2)$$

where Pleft represents the hypothesized proportion of left-handers in the population of artists.

One-Variable χ 2 Test

Evaluates whether observed frequencies for a single qualitative variable are adequately described by hypothesized or expected frequencies.

DETAILS: CALCULATING χ2

The hypothesis that voters are equally likely to prefer any one of four different candidates (coded 1, 2, 3, and 4) translates into

$$H_0: P_1 = P_2 = P_3 = P_4 = .25 \text{ (or } 1/4)$$

where P1 represents the hypothesized proportion of voters who prefer candidate 1 in the population of voters, and so forth.

Alternative Hypothesis Because the null hypothesis will be false if population proportions deviate in any direction from that hypothesized, the alternative or research hypothesis can be described simply as

$$H_1: H_0 \text{ is false}$$

As usual, the alternative hypothesis indicates that, relative to the null hypothesis, something special is happening in the underlying population, such as, for instance, a tendency for artists to be left-handed or for voters to prefer one or two candidates.

Progress Check * Specify the null hypothesis for each of the following situations.

(Remember, the null hypothesis usually represents a negation of the researcher's hunch or

hypothesis.)

(a) A political scientist wants to determine whether voters prefer candidate A more than candidate

B for president.

(b) A biologist suspects that, upon being released 10 miles south of their home roost, migratory

birds are more likely to fly toward home (north) rather than in any of the three remaining directions (east, south, or west).

(c) A sociologist believes that crimes are not committed with equal likelihood on each of the seven days of the week.

(d) Another sociologist suspects that proportionately more crimes are committed during the two days of the weekend (Saturday and Sunday) than during the five other days of the week. Hint: There are just two (unequal) proportions: one representing the two weekend days and the other representing the five weekdays.

Answers :

(a) $H_0: P_A = P_B = \frac{1}{2}$
(b) $H_0: P_{north} = P_{east} = P_{south} = P_{west} = \frac{1}{4}$
(c) $H_0: P_{Mon} = P_{Tue} = P_{Wed} = P_{Thu} = P_{Fri} = P_{Sat} = P_{Sun} = \frac{1}{7}$
(d) $H_0: P_{weekday} = \frac{5}{7}; P_{weekend} = \frac{2}{7}$

DETAILS: CALCULATING χ 2

If the null hypothesis is true, then, except for chance, hypothetical or expected frequencies (generated from the hypothetical proportions) should describe the observed frequencies in the sample. For example, when testing the blood bank's claim with a sample of 100 students, 44 students should have type O (from the product of .44 and 100); 41 should have type A; 10 should have type B; and only 5 should have type AB. In Table 19.1, each of these numbers is referred to as an expected frequency, fe, that is, the hypothesized frequency for each category of the

qualitative variable if, in fact, the null hypothesis is true. An expected frequency is compared with its

observed frequency, fo, that is, the frequency actually obtained in the sample for each category.

CHI-SQUARE (χ^2) TEST FOR QUALITATIVE (NOMINAL) DATA

Table 19.1
OBSERVED AND EXPECTED FREQUENCIES: BLOOD TYPES OF 100 STUDENTS

	BLOOD TYPE				
FREQUENCY	**O**	**A**	**B**	**AB**	**TOTAL**
Observed (f_o)	38	38	20	4	100
Expected (f_e)	44	41	10	5	100

To find the expected frequency for any category, multiply the hypothesized or expected proportion for that category by the total sample size, namely,

EXPECTED FREQUENCY (ONE-VARIABLE χ^2 TEST)

$$f_e = (expected\ proportion)(total\ sample\ size) \qquad (19.1)$$

where *fe* represents the expected frequency.

Evaluating Discrepancies

It's most unlikely that a random sample—because of its inevitable variability—will exactly reflect the characteristics of its population. Even though the null hypothesis is true, discrepancies will appear between observed and expected frequencies, as in Table 19.1.

The crucial question is whether the discrepancies between observed and expected frequencies are small enough to be regarded as a common outcome, given that the null hypothesis is true. If so, the null hypothesis is retained. Otherwise, if the

discrepancies are large enough to qualify as a rare outcome, the null hypothesis is rejected.

Computing χ 2

To determine whether discrepancies between observed and expected frequencies qualify as a common or rare outcome, a value is calculated for χ 2 and compared with its hypothesized sampling distribution. To calculate χ 2, use the following expression:

χ^2 RATIO

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \qquad (19.2)$$

category of the qualitative variable. Table 19.2 illustrates how to use Formula 19.2 to calculate χ 2 for the present example.

Some Properties of χ 2

Notice several features of Formula 19.2. The larger the discrepancies are between the observed and expected frequencies, *fo* − *fe*, the larger the value of χ 2 and, therefore, as will be seen, the more suspect the null hypothesis will be. Because of the squaring of each discrepancy, negative discrepancies become positive, and the value of χ 2 never

TABLE FOR THE χ2 DISTRIBUTION

Table 19.2
CALCULATION OF χ^2 (ONE-VARIABLE TEST)

A. COMPUTATIONAL SEQUENCE

Find an expected frequency for each expected proportion 1.
List observed and expected frequencies 2.
Substitute numbers in formula 3 and solve for χ^2.

B. DATA AND COMPUTATIONS

1 f_e = (expected proportion)(sample size)

$f_e(O) = (.44)(100) = 44$
$f_e(A) = (.41)(100) = 41$
$f_e(B) = (.10)(100) = 10$
$f_e(AB) = (.05)(100) = 5$

2 Frequency	O	A	B	AB	Total
f_o	38	38	20	4	100
f_e	44	41	10	5	100

3 $$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(38-44)^2}{44} + \frac{(38-41)^2}{41} + \frac{(20-10)^2}{10} + \frac{(4-5)^2}{5}$$

$$= \frac{(-6)^2}{44} + \frac{(-3)^2}{41} + \frac{(10)^2}{10} + \frac{(-1)^2}{5}$$

$$= \frac{36}{44} + \frac{9}{41} + \frac{100}{10} + \frac{1}{5}$$

$$= .82 + .22 + 10.00 + .20$$

$$= 11.24$$

isolation, but relative to the size of expected frequencies. For example, a discrepancy of 5 looms more importantly (and translates into a larger value of χ 2)

relative to an expected frequency of 10 than relative to an expected frequency of 100.

TABLE FOR THE χ 2 DISTRIBUTION

Like *t* and *F*, χ 2 has not one but a family of distributions. Table D in Appendix C supplies critical values from various χ 2 distributions for hypothesis tests at the .10, .05, .01, and .001 levels of significance. To locate the appropriate row in Table D, first identify the correct number of degrees of freedom. For the one-variable test, the degrees of freedom for χ 2 can be obtained from the following expression:

DEGREES OF FREEDOM (ONE-VARIABLE χ^2 TEST)

$$df = c - 1 \qquad (19.3)$$

where *c* refers to the total number of categories of the qualitative variable.

CALCULATION OF χ 2 (ONE-VARIABLE TEST)

A. COMPUTATIONAL SEQUENCE

Find an expected frequency for each expected proportion 1. List observed and expected frequencies 2. Substitute numbers in formula 3 and solve for χ 2.

B. DATA AND COMPUTATIONS

1 *fe* = (expected proportion)(sample size)

fe(O) = (.44)(100) = 44

fe(A) = (.41)(100) = 41

fe(B) = (.10)(100) = 10

fe(AB) = (.05)(100) = 5

2 Frequency O A B AB Total

fo 38 38 20 4 100

fe 44 41 10 5 100

3

2

2

2 2 2 2

2 2 2 2

( )

(38 44) (38 41) (20 10) (4 5)

44 41 10 5

(6) (3) (10) (1)

44 41 10 5

36 9 100 1

44 41 10 5

.82 .22 10.00 .20

11.24

o e

e

f f

f

CHI-SQUARE (χ2) TEST FOR QUALITATIVE (NOMINAL) DATA

Lose One Degree of Freedom

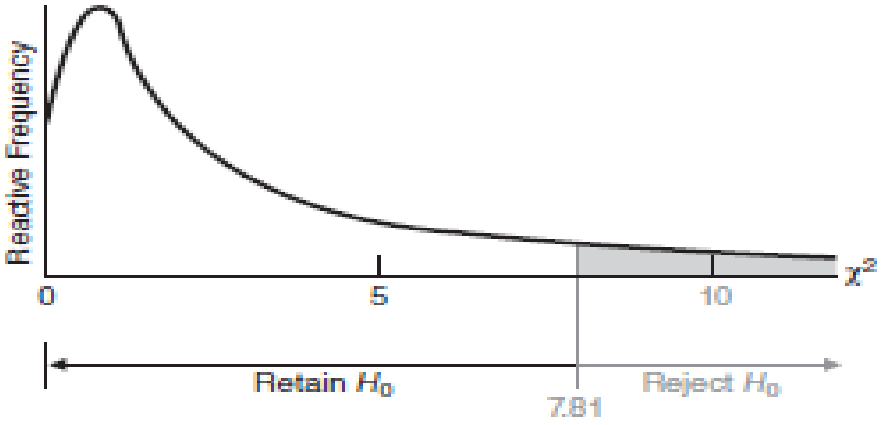
To understand Formula 19.3, focus on the set of observed frequencies for 100 students in Table 19.1. In practice, of course, the observed frequencies for the four © categories have equal status, and any combination of four frequencies that sums to 100 is possible. From the more abstract perspective of degrees of freedom, however, only three (c – 1) of these frequencies are free to vary because of the mathematical restriction that, when calculating χ 2 for the present data, all observed (or expected) frequencies must sum to 100. Although the observed frequencies of

any three of the four categories are free to vary, the frequency of the fourth category must be some number that, when combined with the other three frequencies, will yield a sum of 100. Similarly, if there had been five categories, the frequencies of any four categories would have been free to vary, but not that of the fifth category. For the one-variable test, the number of degrees of freedom always equals one less than the total number of categories (*c*), as indicated in Formula 19.3. In the present example, in which the categories consist of the four blood types, *df* 4 1 3 To find the critical χ 2 for a hypothesis test at the .05 level of significance, locate the cell in Table D, Appendix C, intersected by the row for 3 degrees of freedom and the column for the .05 level of significance. This cell lists a value of 7.81 for the critical χ 2.

χ 2 TEST

Following the usual procedure, assume the null hypothesis to be true and view the observed χ 2 within the context of its hypothesized distribution shown in Figure 19.1

If, because the discrepancies between observed and expected frequencies are relatively small, the observed χ 2 appears to emerge from the dense concentration of possible χ 2 values smaller than the critical χ 2, the observed outcome would be viewed as a common occurrence, on the assumption that the null hypothesis is true. Therefore, the null hypothesis would be retained. On the other hand, if, because the discrepancies between observed and expected frequencies are relatively large, the observed χ 2 appears to emerge from the sparse concentration of possible values equal to or greater than the critical χ 2, the observed outcome would be viewed as a rare occurrence, and the null hypothesis would be rejected. In fact, because the observed χ 2 of 11.24 is larger than the critical χ 2 of 7.81, the null hypothesis should be rejected: There is evidence that the distribution of blood types in the student population differs from that claimed for the U.S. population.



Hypothesized sampling distribution of χ^2 (3 degrees of freedom).

χ^2 TEST

As can be seen in Table 19.1, the present survey contains an unexpectedly large number of students with type B blood. A subsequent investigation revealed that the sample included a large number of Asian-American students, a group that has an established high incidence of type B blood. This might explain why the hypothesized distribution of blood types fails to describe that for the population of students from which the random sample was taken. Certainly, a random sample should be taken from a much broader spectrum of the general population before questioning the blood bank's claim about the distribution of blood types in the U.S. population.

Progress Check * A random sample of 90 college students indicates whether they most desire love, wealth, power, health, fame, or family happiness.

(a) Using the .05 level of significance and the following results, test the null hypothesis that, in the underlying population, the various desires are equally popular.

(b) Specify the approximate *p*-value for this test result.

DESIRES OF COLLEGE STUDENTS

FREQUENCY	LOVE	WEALTH	POWER	HEALTH	FAME	FAMILY HAP.	TOTAL
Observed (f_o)	25	10	5	25	10	15	90

ANSWERS:

(a) *Research Problem*

The attribute most desired by a population of college students is equally distributed among various possibilities.

Statistical Hypotheses

H_0: $P_{love} = P_{wealth} = P_{power} = P_{health} = P_{fame} = P_{family\ happiness} = \frac{1}{6}$

H_1: H_0 is false.

Decision Rule

Reject H_0 at the .05 level of significance if $\chi^2 \geq 11.07$, given $df = 5$.

Calculations

$$\chi^2 = \frac{(25-15)^2}{15} + \frac{(10-15)^2}{15} + \frac{(5-15)^2}{15} + \frac{(25-15)^2}{15} + \frac{(10-15)^2}{15} + \frac{(15-15)^2}{15} = 23.33$$

Decision

Reject H_0 at the .05 level of significance because $\chi^2 = 23.33$ exceeds 11.07.

Interpretation

The attribute most desired by a population of college students is not equally distributed among various possibilities.

(b) $p < .001$

HYPOTHESIS TEST SUMMARY
ONE-VARIABLE χ^2 TEST
(Survey of Blood Types)

Research Problem

Does the distribution of blood types in a population of college students comply with that described in a blood bank bulletin for the U.S. population?

Statistical Hypotheses

$$H_0: P_O = .44;\ P_A = .41;\ P_B = .10;\ P_{AB} = .05$$

(where P_O is the proportion of type O blood in the population, etc.)

$$H_1: H_0 \text{ is false}$$

Decision Rule

Reject H_0 at the .05 level of significance if $\chi^2 \geq 7.81$ (from Table D in Appendix C, given $df = c - 1 = 4 - 1 = 3$).

Calculations

$$\chi^2 = 11.24 \text{ (see Table 19.2.)}$$

Decision

Reject H_0 at the .05 level of significance because $\chi^2 = 11.24$ exceeds 7.81.

Interpretation

The distribution of blood types in a student population differs from that claimed for the U.S. population.

CHI-SQUARE (χ_2) TEST FOR QUALITATIVE (NOMINAL) DATA

χ_2 Test Is Nondirectional

The χ_2 test is nondirectional, as all discrepancies between observed and expected frequencies are squared. All squared discrepancies have a cumulative positive effect on the value of the observed χ_2 and thereby ensure that χ_2 is a nondirectional test, even

though, as illustrated in Figure 19.1, only the upper tail of its distribution contains the rejection region.

TWO-VARIABLE χ^2 TEST

So far, we have considered the case where observations are classified in terms of only one qualitative variable. Now let's deal with the case where observations are cross-classified in terms of two qualitative (nominal) variables.

LOST LETTER STUDY

Viewing the return rate of lost letters as a measure of social responsibility in neighborhoods, a social psychologist intentionally "loses" self-addressed, stamped envelopes near mailboxes.* Furthermore, to determine whether *social responsibility,* as inferred from the mailed return rates, varies with the *type of neighborhood,* lost letters are scattered throughout three different neighborhoods: downtown, suburbia, and a college campus.

Letters are "lost" in *each* of the three types of neighborhoods according to procedures that control for possible contaminating factors, such as the density of pedestrian traffic and mailbox accessibility. (Ordinarily, the social psychologist would probably scatter equal numbers of letters among the three neighborhoods, but to maximize the generality of the current example, we will assume that a total of 200 letters were scattered as follows: 60 downtown, 70 in suburbia, and 70 on campus.) Each letter is crossclassified on the basis of the type of neighborhood where it was lost and whether or not it was returned, as shown in Table 19.3. For instance, of the 60 letters lost downtown, 39 were returned, while of the 70 letters lost in suburbia, 30 were returned. When observations are cross-classified according to two qualitative variables, as with the lost letter study, the test is a **two-variable χ^2 test**.

Table 19.3
OBSERVED FREQUENCIES OF RETURNED LETTERS

	NEIGHBORHOOD			
RETURNED LETTERS	DOWNTOWN	SUBURBIA	CAMPUS	TOTAL
Yes	39	30	51	120
No	21	40	19	80
Total	60	70	70	200

STATISTICAL HYPOTHESES

Null Hypothesis

For the two-variable χ^2 test, the null hypothesis always makes a statement about the lack of relationship between two qualitative variables in the underlying population. In the present case, it states that there is no relationship—that is, no predictability—between type of neighborhood and whether or not letters are returned. This is the same as claiming that the proportions of returned letters are the same for all three types of neighborhoods. Accordingly, the two-variable χ^2 test often is referred to as a *test of independence* for the two qualitative variables.

Although symbolic statements of the null hypothesis are possible, it is much easier to use word descriptions such as

H_0: Type of neighborhood and return rate of lost letters are independent.

or as another example,

H_0:Gender and political preference are independent

If these null hypotheses are true, then among the population of lost letters, the type of neighborhood should not change the probability that a randomly selected lost letter is returned, and among the population of voters, gender should not change the probability that a randomly selected voter prefers the Democrats. Otherwise, if these null hypotheses are false, type of neighborhood should change the probability that a randomly selected lost letter is returned, and gender should change the probability that a randomly selected voter prefers the Democrats.

Alternative Hypothesis

The alternative or research hypothesis always takes the form

H_1: H_0 is false

Progress Check * Specify the null hypothesis for each of the following situations:

(a) A political scientist suspects that there is a relationship between the educational level of adults (grade school, high school, college) and whether or not they favor right-to-abortion legislation.

(b) A marital therapist believes that groups of clients and nonclients are distinguishable on the

basis of whether or not their parents are divorced.

(c) An organizational psychologist wonders whether employees' annual evaluations, as either satisfactory or unsatisfactory, tend to reflect whether they have fixed or flexible work schedules.

Answers

(a) Educational level and attitude toward right-to-abortion legislation are independent.

(b) Clients and nonclients are not distinguishable on the basis of—or are independent of—whether or not their parents are divorced.

(c) Employees' annual evaluations are independent of whether they have fixed or flexible work schedules

DETAILS: CALCULATING χ^2

As in the one-variable χ^2 test, expected frequencies are calculated on the assumption that the null hypothesis is true, and, depending on the size of the discrepancies between observed and expected frequencies, the null hypothesis is either retained or rejected.

CHI-SQUARE (χ_2) TEST FOR QUALITATIVE (NOMINAL) DATA

Finding Expected Frequencies from Proportions

According to the present null hypothesis, type of neighborhood and return rates are independent. Except for chance, the same proportion of returned letters should be observed for each of the three neighborhoods. Referring to the totals in Table 19.3, you will notice that when all three types of neighborhoods are considered together, 120 of the 200 lost letters were returned. Therefore, if the null hypothesis is true, 120/200, or .60, should describe the proportion of returned letters from *each* of the three neighborhoods. More specifically, among the total of 60 letters lost downtown, .60 of this total, that is (.60)(60), or 36 letters, should be returned, and 36 is the expected frequency of returned letters from downtown, as indicated in Table 19.4. By the same token, among the total of 70 letters lost in suburbia (or on campus), .60 of this total, that is, (.60)(70), or 42, is the expected frequency of returned letters from suburbia (or from the campus).

Table 19.4
OBSERVED AND EXPECTED FREQUENCIES OF RETURNED LETTERS

RETURNED LETTERS		NEIGHBORHOOD			
		DOWNTOWN	SUBURBIA	CAMPUS	TOTAL
Yes	f_o	39	30	51	120
	f_e	36	42	42	
No	f_o	21	40	19	80
	f_e	24	28	28	
Total		60	70	70	200

As can be verified in Table 19.4, the expected frequencies for nonreturned letters can be calculated in the same way, after establishing that when all three neighborhoods are considered together, only 80 of the 200 lost letters were not returned. Now, if the null hypothesis is true, 80/200, or .40, should describe the proportion of letters not returned from *each* of the three neighborhoods. For example, among the total of 60 letters lost downtown, .40 of this total, or 24, will be the expected frequency of letters not returned from downtown.

Finding Expected Frequencies from Totals

Expected frequencies have been derived from expected proportions in order to spotlight the reasoning behind the test. In the long run, *it is more efficient to calculate the expected frequencies directly from the various marginal totals*, according to the following formula:

EXPECTED FREQUENCY (TWO-VARIABLE χ^2 TEST)

$$f_e = \frac{(row\ total)(column\ total)}{grand\ total} \qquad (19.4)$$

where f_e refers to the expected frequency for any cell in the cross-classification table; *row total* refers to the total frequency for the row occupied by that cell; *column total* refers to the total frequency for the column occupied by that cell; and *grand total* refers to the total for all rows (or all columns).

DETAILS: CALCULATING χ_2

Using the marginal totals in Table 19.4, we may verify that Formula 19.4 yields the expected frequencies shown in that table. For example, the expected frequency of returned letters from downtown is

$$f_e = \frac{(120)(60)}{200} = \frac{7200}{200} = 36$$

and the expected frequency of returned letters from suburbia is

$$f_e = \frac{(120)(70)}{200} = \frac{8400}{200} = 42$$

Having determined the set of expected frequencies, you may use Formula 19.2 to calculate the value of χ_2, as described in Table 19.5. Incidentally, for computational convenience, all of the fictitious totals in the margins of Table 19.5 were selected to be multiples of 10. In actual practice, the marginal totals are unlikely to be multiples of 10, and consequently, expected frequencies will not always be whole numbers

Table 19.5
CALCULATION OF χ^2 (TWO-VARIABLE TEST)

A. COMPUTATIONAL SEQUENCE

Use formula 1 to obtain all expected frequencies from table of observed frequencies.
Construct a table of observed and expected frequencies. 2
Substitute numbers in formula 3 and solve for χ^2.

B. DATA AND COMPUTATIONS

1 $$f_e = \frac{(\text{column total})(\text{row total})}{\text{grand total}}$$

$$f_e(\text{yes, downtown}) = \frac{(60)(120)}{200} = 36$$

$$f_e(\text{yes, suburbia}) = \frac{(70)(120)}{200} = 42$$

$$f_e(\text{yes, campus}) = \frac{(70)(120)}{200} = 42$$

$$f_e(\text{no, downtown}) = \frac{(60)(80)}{200} = 24$$

$$f_e(\text{no, suburbia}) = \frac{(70)(80)}{200} = 28$$

$$f_e(\text{no, campus}) = \frac{(70)(80)}{200} = 28$$

2

		Downtown	Suburbia	Campus	Total
Yes	f_o	39	30	51	120
	f_e	36	42	42	
No	f_o	21	40	19	80
	f_e	24	28	28	
Total		60	70	70	200

3 $$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(39-36)^2}{36} + \frac{(30-42)^2}{42} + \frac{(51-42)^2}{42} + \frac{(21-24)^2}{24} + \frac{(40-28)^2}{28} + \frac{(19-28)^2}{28}$$

$$= 0.25 + 3.43 + 1.93 + 0.38 + 5.14 + 2.89$$

$$= 14.02$$

CHI-SQUARE (χ_2) TEST FOR QUALITATIVE (NOMINAL) DATA

TABLE FOR THE χ_2 DISTRIBUTION

Locating a critical χ_2 value in Table D, Appendix C, requires that you know the correct number of degrees of freedom. For the two-variable test, degrees of freedom for χ_2 can be obtained from the following expression:

DEGREES OF FREEDOM (TWO-VARIABLE χ^2 TEST)

$$df = (c-1)(r-1) \qquad (19.5)$$

where *c* equals the number of categories for the column variable and *r* equals the number of categories for the row variable. In the present example, which has three columns (downtown, suburbia, and campus) and two rows (returned and not returned),

df =(3-1)(2-1) =(2)(1)= 2

To find the critical χ^2 for a test at the .05 level of significance, locate the cell in Table D intersected by the row for 2 degrees of freedom and the column for the .05 level. In this case, the value of the critical χ^2 equals 5.99.

Explanation for Degrees of Freedom

To understand Formula 19.5, focus on the set of observed frequencies in Table 19.3. In practice, of course, the observed frequencies for the six cells (within the table) have equal status, and any combination of six frequencies that sum to the various marginal totals is possible. However, only two of these frequencies are free to vary. One row and one column of cell frequencies in the original matrix are not free to vary because of the restriction that all cell frequencies in each column and in each row must sum to fixed totals in the margins. From the more abstract perspective of degrees of freedom, the original matrix with $c \times r$ or 3×2 cells shrinks to $(c - 1)(r - 1)$ or $(3 - 1)(2 - 1)$ cells and $df = 2$.

χ^2 TEST

Because the calculated χ^2 of 14.02 exceeds the critical χ^2 of 5.99, the null hypothesis should be rejected: There is evidence that the type of neighborhood is not independent of the return rate of lost letters. Knowledge about the type of neighborhood supplies extra information about the likelihood that lost letters will be returned. A comparison of observed and expected frequencies in Table 19.5 suggests that lost letters are more likely to be returned from either downtown or the campus than from suburbia.

Progress Check * An investigator suspects that there might be a relationship, possibly based on genetic factors, between hair color and susceptibility to poison oak. Three hundred volunteer subjects are exposed to a small amount of poison oak and then classified according to their susceptibility (rash or no rash) and their hair color (red, blond, brown, or

black), yielding the following frequencies:

HAIR COLOR AND SUSCEPTIBILITY TO POISON OAK

	HAIR COLOR				
SUSCEPTIBILITY	**RED**	**BLOND**	**BROWN**	**BLACK**	**TOTAL**
Rash	10	30	60	80	180
No rash	20	30	30	40	120
Total	30	60	90	120	300

ESTIMATING EFFECT SIZE

(a) Test the null hypothesis at the .01 level of significance.

(b) Specify the approximate *p*-value for this test result.

Answers

(a) *Research Problem*

Is hair color related to susceptibility to poison oak?

Statistical Hypotheses

H_0: Hair color and susceptibility to poison oak are independent.

H_1: H_0 is false.

Decision Rule

Reject H_0 at the .01 level if $\chi^2 \geq 11.34$ given $df = (2-1)(4-1) = 3$.

Calculations

$$\chi^2 = \frac{(10-18)^2}{18} + \frac{(30-36)^2}{36} + \frac{(60-54)^2}{54} + \frac{(80-72)^2}{72} + \frac{(20-12)^2}{12}$$

$$+ \frac{(30-24)^2}{24} + \frac{(30-36)^2}{36} + \frac{(40-48)^2}{48} = 15.28$$

Decision

Reject H_0 at the .01 level of significance because $\chi^2 = 15.28$ exceeds 11.34.

Interpretation

There is a relationship between hair color and susceptibility to poison oak.

(b) $p < .01$

HYPOTHESIS TEST SUMMARY
TWO-VARIABLE χ^2 TEST
(Lost Letter Study)

Research Problem
Is there a relationship between the type of neighborhood and the return rate of lost letters?

Statistical Hypotheses

H_0: Type of neighborhood and return rates of lost letters are independent.

H_1: H_0 is false.

DECISION RULE
Reject H_0 at the .05 level of significance if $\chi^2 \geq 5.99$ [from Table D, Appendix C, given that $df = (c - 1)(r - 1) = (3 - 1)(2 - 1) = 2$].

Calculations

$\chi^2 = 14.02$ (See Table 19.5.)

Decision
Reject H_0 at the .05 level of significance because $\chi^2 = 14.02$ exceeds 5.99.

Interpretation
Type of neighborhood and return rate of lost letters are not independent.

ESTIMATING EFFECT SIZE

One way to check the importance of a statistically significant two-variable χ^2 is to use a measure analogous to the squared curvilinear correlation coefficient, η^2, known as the **squared Cramer's phi coefficient** and symbolized as ϕ_c^2. Being independent of sample size (unlike χ^2), ϕ_c^2 *very roughly estimates the proportion of explained variance (or predictability) between two qualitative variables.*

Squared Cramer's Phi Coefficient (ϕ_c^2)

Solve for the squared Cramer's phi coefficient using the following formula:

PROPORTION OF EXPLAINED VARIANCE (TWO-VARIABLE χ^2)

$$\phi_c^2 = \frac{\chi^2}{n(k-1)} \qquad (19.6)$$

where χ^2 is the obtained value of the statistically significant two-variable test, n is the sample size (total observed frequency), and k is the smaller of either the c columns or the r rows (or the value of either if they are the same).

For the lost letter study, given a significant χ^2 of 14.02, $n = 200$, and $k = 2$ (from $r = 2$), we can calculate

$$\phi_c^2 = \frac{14.02}{200(2-1)} = .07$$

One guideline, suggested by Cohen for correlations and listed in **Table 19.6**, is that the strength of the relationship between the two variables is small if ϕ_c^2 approximates .01, medium if ϕ_c^2 approximates .09, and large if ϕ_c^2 approximates or exceeds .25.* Using these guidelines, the estimated strength of the relationship between type of neighborhood and return rate is medium, since $\phi_c^2 = .07$.

Consider calculating ϕ_c^2 whenever a statistically significant two-variable χ^2 has been obtained. However, do not apply these guidelines about the strength of a relationship without regard to special circumstances that could give considerable importance to even a very weak relationship, as suggested in the next section.

Progress Check * Given the significant χ_2 in Exercise 19.4, use Formula 19.6 to estimate whether the strength of the relationship between hair color and susceptibility to poison oak is small, medium, or large.

Answer

$$\phi_c^2 = \frac{15.28}{300(2-1)} = .05$$ (between a small and a medium effect, according to Cohen's guidelines)

ODDS RATIOS

A widely publicized report in *The New England Journal of Medicine* (January 28, 1988) described the incidence of heart attacks (the dependent variable) among over 22,000 physicians who took either an aspirin or a placebo (the independent variable) every other day for the duration of the study. Although a statistical analysis of the results, shown in panel B of **Table 19.7**, yields a highly significant chi square [χ^2 (1, $n = 22{,}071$) = 25.01, $p < .001$, $\phi_c^2 = .001$], the strength of the relationship between these two qualitative variables is very weak, as indicated by the minuscule value of only .001 for Cramer's phi coefficient, ϕ_c^2. (Verify, if you wish, using Formula 19.6.) Sometimes the importance of a seemingly weak relationship can be appreciated more fully by calculating an odds ratio. *An* **odds ratio (OR)** *indicates the relative occurrence of one value of the dependent variable* (occurrence of heart attacks) *across the two categories of the independent variable* (aspirin or a placebo).

Calculating the Odds Ratio

First, find the ***odds*** (*defined as the ratio of frequencies of occurrence to nonoccurrence*)

of the dependent variable for each value of the independent variable.

Referring to the data in Table 19.7, find the odds of a heart attack among physicians in the placebo group by dividing frequencies of their occurrence (189) by their nonoccurrence (10,845) to obtain odds of .0174 to 1 (or, after multiplying by one thousand, 17.4 to 1,000.) Likewise, find the odds of a heart attack among physicians in the aspirin group by dividing frequencies of their occurrence (104) by their nonoccurrence (10,933) to obtain odds of .0095 to 1.

Next, *calculate the ratio of these two odds.* Divide .0174 by .0095 to obtain an odds ratio of 1.83. A physician in the placebo group is 1.83 times more likely

ODDS RATIOS

Table 19.7
CALCULATING THE ODDS RATIO

A. COMPUTATIONAL SEQUENCE

Enter data in 2 × 2 table. 1
Calculate odds for each group by dividing the frequency of occurrence by the frequency of nonoccurrence. 2
Calculate the odds ratio by dividing odds for the two groups. 3
Interpret the odds ratio. 4

B. DATA AND COMPUTATIONS

1

Effect of Aspirin on Heart Attacks

	Heart Attack	No Heart Attack	2 Odds	3 Odds Ratio
Placebo	189	10,845	189/10,845 = .0174 (to 1)	$\frac{.0174}{.0095} = 1.83$
Aspirin	104	10,933	104/10,933 = .0095 (to 1)	

4 A physician in the placebo group is 1.83 times more likely to have a heart attack than one in the aspirin group.

almost twice as likely(!), to have a heart attack as a physician in the aspirin group. (Conversely, if the division of odds is reversed, the odds ratio of .55 signifies that a physician in the aspirin group is .55 times less likely, or about only half as likely, to have a heart attack as a physician in the placebo group.) Given the seriousness of even one heart attack, especially if it happens to you or to someone you know, odds ratios reaffirm the importance of these findings in spite of the very small value of .001 for

$_{2\,c}$. Subsequently, the investigators discontinued this study and recommended aspirin therapy for all high-risk individuals in the population.

Consider calculating an odds ratio to clarify further the importance of a significant χ^2. A 95 percent confidence interval for an odds ratio also can be constructed by using procedures, such as Minitab's *Binary Logistic Regression,* not discussed in this book.

Progress Check * Odds ratios can be calculated for larger cross-classification tables, and one way of doing this is by reconfiguring into a smaller 2 . 2 table. The 2 . 3 table for the lost letter study, Table 19.4, could be reconfigured into a 2 . 2 table if, for example, the investigator is primarily interested in comparing return rates of lost letters only for campus and off-campus locations (both suburbia and downtown), that is,

	NEIGHBORHOOD		
RETURNED LETTERS	OFF-CAMPUS	CAMPUS	TOTAL
YES	69	51	120
NO	61	19	80
TOTAL	130	70	200

(a) Given $\chi^2(1, n = 200) = 7.42$, $p < .01$, $\phi_c^2 = .037$ for these data, calculate and interpret the odds ratio for a returned letter from campus.

(b) Calculate and interpret the odds ratio for a returned letter from off-campus.

Answers

(a) Odds ratio for returned letters from campus

$$OR = \frac{51/19}{69/61} = \frac{2.68}{1.13} = 2.37$$

A returned letter is 2.37 times more likely to come from campus than from off-campus.

(b) $$OR = \frac{69/61}{51/19} = \frac{1.13}{2.68} = .42$$

A returned letter is .42 times less likely to come from off-campus than from campus.

REPORTS IN THE LITERATURE

A report of the original lost letter study might be limited to an interpretative comment, plus

a parenthetical statement that summarizes the statistical analysis and includes a *p*-value

and an estimate of effect size. For example, an investigator might report the following:

There is evidence that the return rate of lost letters is related to the type of neighborhood [$\chi^2(2, n = 200) = 14.02, p < .001, \phi_c^2 = .07$].

The parenthetical statement indicates that a χ^2 based on 2 degrees of freedom and a sample size, n, of 200 was found to equal 14.02. The test result has an approximate p-value less than .001 because, as can be seen in Table D, Appendix C, the observed χ^2 of 14.02 is larger than the critical χ^2 of 13.82 for the .001 level of significance. Furthermore, since a p-value of less than .001 is a very rare event, given that the null hypothesis is true, it supports the research hypothesis, as implied in the interpretative statement. Finally, a value of .07 for ϕ_c^2 indicates that approximately 7 percent of the variance in returned letters is attributable to differences among the three neighborhoods.

SOME PRECAUTIONS

Avoid Dependent Observations

The valid use of χ_2 requires that *observations be independent of one another*. One observation should have no influence on another. For instance, when tossing a pair of dice, the appearance of a six spot on one die has no influence on the number of spots displayed on the other die. A violation of independence occurs whenever a single subject contributes more than one observation (or in the two-variable case, more than one pair of observations). For example, it would have occurred in a preference test for four brands of soda if each subject's preference had been counted more than once, possibly because of a series of taste trials. When considering the use of χ_2, *the total for all observed frequencies never should exceed the total number of subjects*.

Avoid Small Expected Frequencies

The valid use of χ_2 also requires that expected frequencies not be too small. A conservative rule specifies that *all expected frequencies be 5 or more*. Small expected frequencies need not necessarily lead to a statistical dead end; sometimes it is possible to create a larger expected frequency from the combination of smaller expected frequencies (see Review Question 19.16). Otherwise, avoid small expected frequencies by using a larger sample size.

Avoid Extreme Sample Sizes

As discussed in previous chapters, avoid either very small or very large samples. An unduly small sample size produces a test that tends to miss even a seriously false null hypothesis. (By avoiding small expected frequencies, you will automatically protect

the χ^2 test from the more severe cases of small sample size.) An excessively large sample size produces a test that tends to detect small, unimportant departures from null hypothesized values. A power analysis, similar to that described in Section 11.11, could be used—with the aid of software, such as G*Power at http://www.gpower.hhu.de/—to identify a sample size with a reasonable detection rate for the smallest important departure from the null hypothesis.

COMPUTER OUTPUT

Table 19.8 shows an SAS output for the return rates of lost letters in the three neighborhoods. Compare these results—both observed frequencies and the value of χ^2—with those shown in Table 19.5.

Progress Check * Referring to the SAS output, identify

(a) the observed frequency of returned letters in suburbia.

(b) the set of three percents (inside the 2×3 box) that can be most meaningfully compared with the three *total* percents of 30.00, 35.00, and 35.00, for downtown, suburbia, and campus, respectively.

(c) the value of Cramer's squared phi, ϕ_c^2 (Be careful!)

Answers

(a) 30

(b) Either the next-to-last set of percents (designated as Row Pct because they sum to 100 percent in each row) for Yes or returned letters, that is, 32.50, 25.00, and 42.50, or the same set of percents for the No or unreturned letters, that is, 26.25, 50.00, and 23.75. When compared with the total percents, that is, 30.00, 35.00, and 35.00, either set of percents spotlights the relatively low rate of returns in suburbia and the relatively high rates on campus.

(c) Square "Cramer's V (phi)," that is, (.265)(.265) = .07.

Table 19.8
SAS OUTPUT: TWO-VARIABLE χ^2 TEST FOR LOST LETTER DATA

THE SAS SYSTEM
12:45 TUESDAY, JANUARY 5, 2016
TABLE OF RETURNED BY NEIGHBORHOOD

RETURNED Frequency Percent Row Pct Col Pct	NEIGHBORHOOD Downtown	Suburbia	Campus	Total
Yes	39 19.50 32.50 65.00	30 15.00 25.00 42.86	51 25.50 42.50 72.86	120 60.00
No	21 10.50 26.25 35.00	40 20.00 50.00 57.14	19 9.50 23.75 27.14	80 40.00
Total	60 30.00	70 35.00	70 35.00	200 100.00

STATISTICS FOR TABLE OF RETURNED BY NGHBRHD

Statistic	DF	Value	Prob
Chi-Square	2	14.018	0.0009
Likelihood Ratio Chi-Square	2	14.049	0.0009
Mantel-Haenszel Chi-Square	1	13.059	0.0003
Phi Coefficient		0.265	
Contingency Coefficient		0.256	
1 Cramer's V		0.265	

Comment: