"rejection" and "nonrejection.") The decision to reject or not is made by comparing the observed value of F with the value of F located at the critical point of transition. Symbolically, the rules are stated as

Reject H_0 when $F_{observed} \geq F_{(\alpha)}(m, n)$; otherwise, do not reject H_0. (4-4)

In this statement, α refers to the significance level, and m and n to the *df*'s associated with the numerator and denominator of the F ratio, respectively.

There is often some confusion concerning the exact wording of the decision rules, stemming largely from the fact that we cannot *prove* a hypothesis, only *disprove* it. When we say that a particular hypothesis is "accepted," we do *not* mean that it has been proved—just that it is consistent with the facts. Thus, if we reject H_0, this means that the results of the experiment are consistent with the alternative hypothesis that the treatment means are different; in this sense, then, we accept H_1. By the same token, if we do not reject H_0, this means that we consider the results of the experiment consistent with the hypothesis that the treatment means are equal. It is in this sense, too, that we are accepting the null hypothesis.

We will consider two examples of the use of the rejection rule. The first involves an evaluation of the F we calculated earlier in this chapter (see Table 4-2, p. 56). In this example, $m = 2$ and $n = 12$. If we set $\alpha = .05$, the critical value of F which we find in the tabled values of the F distribution (Table C-1 of Appendix C) is 3.89. The rejection region consists of all values of F equal to or greater than 3.89. Substituting in Eq. (4-4), the decision rule becomes

Reject H_0 when $F_{observed} \geq 3.89$; otherwise, do not reject H_0.

Since the F obtained in this example exceeded this value ($F = 7.41$), we would conclude that treatment effects were present in this experiment.

For the second example, we will return to Fig. 4-3 (p. 61). If we set $\alpha = .05$, the critical value of F at $m = 2$ and $n = 42$ is approximately 3.23. (Table C-1 does not have a value for this combination of *df*'s; the value we have given here is associated with $m = 2$ and $n = 40$.) The rejection region consists of all values of F equal to or greater than 3.23. Since the area under the curve to the right of an ordinate drawn at this value of F consists of 5 percent of the total area under the curve, the probability of obtaining an F at least as deviant as 3.23 is .05. When we substitute in Eq. (4-4), the decision rule becomes

Reject H_0 when $F_{observed} \geq 3.23$; otherwise, do not reject H_0.

The size of the rejection region each of us adopts is a personal choice. What probability we choose is often dictated by our concern for *failing* to reject the null hypothesis when a real difference among treatment means exists. (More about this in a moment.) In presenting the results of a statistical test, however,

we should remember that not all researchers who will be reading the report will agree with our choice of significance level. Thus, in order to accommodate most of the researchers adopting different rejection regions, we indicate the *smallest* significance level within which the $F_{observed}$ will fall. For example, suppose we obtained an F of 6.33 in the last experiment we have just considered. An inspection of the F table indicates that the critical value of F at $\alpha = .01$ is approximately 5.18. One way to report the results of this test would be to make the following statement:

$$F(2, 42) = 6.33, p < .01,$$

which means that this value of F falls within a rejection region having an α level that is less than a probability (p) of .01. Such a statement would indicate that the null hypothesis would be rejected by all researchers adopting a significance level at least as small as 1 percent. Since few researchers will be more conservative than this, the smallest rejection region that we would need to report in a scientific journal is one at $\alpha = .01$. We must distinguish in this discussion between the transmission of information concerning the significance level of the $F_{observed}$ from our own decision to reject or not to reject the null hypothesis. In this example, our rejection region was $\alpha = .05$, but we reported the probability associated with the F as $p < .01$. Our decision to reject H_0 depends only upon the presence of $F_{observed}$ in *our* rejection region.

SUMMARY We have seen that a statistical test begins with the specification of the null and alternative hypotheses. We then conduct our experiment and calculate an F ratio. Next, we judge whether we have obtained an F which is incompatible with the hypothesis that the means of the treatment populations are equal. Incompatibility is defined arbitrarily ahead of time as an F which would occur on the basis of chance, assuming H_0 is true, a small proportion of the time, e.g., 1 time in 20 (5 percent). If the $F_{observed}$ falls within this region of incompatibility, we reject the null hypothesis; it it falls within the region of compatibility, we do not reject the null hypothesis.

Errors in Hypothesis Testing

The procedures we follow in hypothesis testing do *not* guarantee that a correct inference will be drawn when we apply the decision rules enumerated in Eq. (4-4). On the contrary, whether or not we decide to reject the null hypothesis, we will be making either a correct decision or an incorrect decision, depending upon the state of affairs in the real world—i.e., the population. The two types of errors that we can commit are defined in Table 4-4. There are two states that "reality" can take: either the null hypothesis is true or it is false; and there are two decisions that we may make: either reject H_0 or do not. The four possible combinations of states of reality and types of decisions are enumerated in the table. Inspection reveals two situations in which we will make the *correct decision*, i.e., no error of inference: (1) if we reject H_0 when it is *false* and (2) if we accept H_0

when it is *true*. On the other hand, in two complementary situations we will make an *incorrect decision*, i.e., an error of inference: (1) if we reject H_0 when it is *true* and (2) if we accept H_0 when it is *false*. These errors of inference are called type I and type II errors or α or β errors, respectively.

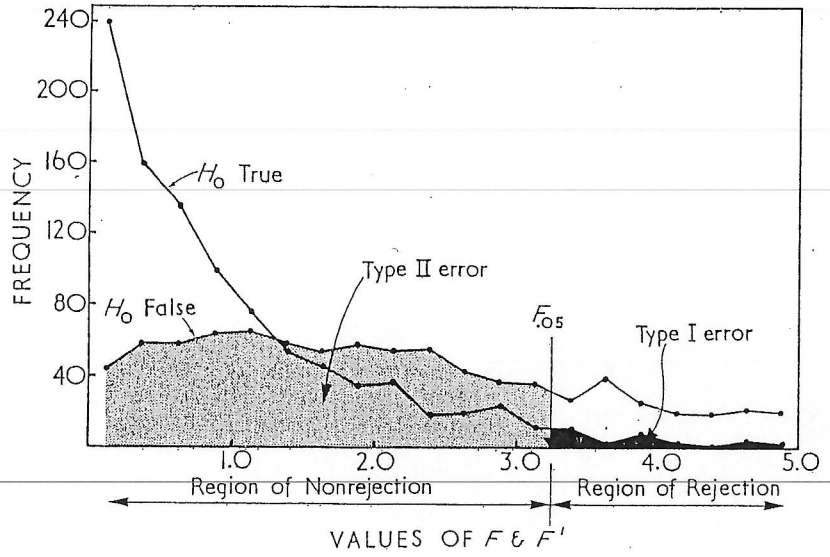

Fig. 4-4 Type I and type II errors illustrated by using empirically determined sampling distributions of F and F'.

To illustrate the two types of errors, we will consider a concrete example. The two sampling distributions from Fig. 4-3 have been reproduced in Fig. 4-4. The curve labeled "H_0 true" is the empirical sampling distribution of $F(2, 42)$

TABLE 4-4 *Errors in Hypothesis Testing*

	Reality	
Decision	H_0 true, H_1 false	H_0 false, H_1 true
Reject H_0 Accept H_1	Incorrect decision: type I error	Correct decision
Accept H_0 Do not accept H_1	Correct decision	Incorrect decision: type II error

when the population means were equal, and the curve labeled "H_0 false" is the empirical sampling distribution of $F'(2, 42)$ when the population means differed in 5-unit steps. These curves indicate the sampling distribution of the F ratio when the null hypothesis is true and when the alternative hypothesis is true. (The distribution for F' plotted in the figure is only one of the permissible alternative hypotheses; one is all that we need to make the point.) One thing is clear in comparing these two sampling distributions: any value of F_{observed} is possible under *both* hypotheses. But we have to hazard a guess about F_{observed} and decide whether it is more likely to have come from the F distribution or from the F' distribution. Since we are testing the null hypothesis, we will have to make a decision concerning values of F that are rare enough—i.e., would occur infrequently enough on the basis of chance—to justify the speculation that F_{observed} came instead from the F' distribution.

Suppose we arbitrarily define rarity in probability terms as $\alpha = .05$. This means that we will consider any F_{observed} that would occur by chance 5 percent of the time or less incompatible with the null hypothesis. The probability is represented by the shaded area under the curve for the F distribution to the *right* of an ordinate erected at $F = 3.23$; this area is the rejection region. This probability also represents the calculated risk we are taking in the decision process, namely, that 5 percent of the time when the null hypothesis is true we will be making the wrong decision—a type I error. Of course, when the alternative hypothesis is true, we make the correct decision. The proportion of the time we can expect to enjoy this state of affairs is represented by the area under the F' curve included in the rejection region; in this example, the proportion is .333.

Now, consider what happens when we do not reject the null hypothesis. (The region of nonrejection includes all values of F which are less than 3.23.) If the null hypothesis happens to be true, we will have made the correct decision by not rejecting H_0. The probability of such a decision is $1 - \alpha = 1 - .05 = .95$. But if the alternative hypothesis is true, we will be in error by failing to reject H_0, and the probability with which this error will be made (a type II error) is represented by the lightly shaded area under the F' curve to the left of $F_{.05}$. In this example, β is .667.

As long as we are committed to make decisions in the face of incomplete knowledge, as every scientist is, we cannot avoid making these errors. We can, however, try to minimize them. We directly control the size of the type I error in our selection of significance level. By setting a region of rejection, we are taking a calculated risk that a certain proportion of the time (for example, $\alpha = .05$), we will obtain F's which fall into this region when the null hypothesis is true. We accept this fact, recognizing that over our lifetimes we will make the wrong decision 5 percent of the time by falsely rejecting the null hypothesis. The size of the type II error is controlled *indirectly*—a point we will discuss in a moment.

There is an obvious reciprocity between these two types of error: if we make one, we cannot make the other. But there is a less obvious relationship. We have

seen that the size of the type I error is under the direct control of the researcher. He sets the α level in the experiment. By decreasing the size of α—i.e., moving to the right the value of F that divides the rejection from the nonrejection region—we will reject *fewer* values of $F_{observed}$ when the null hypothesis is true. However, if the alternative happens to be true, the shaded portion under the F' curve will *increase* and so will our type II error. Consequently, any change in the α level will be accompanied by a change in the opposite direction of the probability of making a type II error. [The relationship between type I and type II errors is developed more fully in Chapter 24.]

How can we control the type II error? In our example, we specified a particular alternative hypothesis, that $\mu_1 = 50$, $\mu_2 = 55$, and $\mu_3 = 60$. We saw that if α is set at .05, we can expect β to equal .667. We could only make this determination by knowing the exact means specified by H_1. In most areas of research in the behavioral sciences, we have no way of offering an *exact* alternative hypothesis such as the one we have been considering. Instead, we settle for the inexact one, namely, that the treatment means are not equal. In some situations it is possible to narrow down the set of reasonable alternative hypotheses, but we will defer a discussion of this possibility to Chapter 24.

There are, however, ways of reducing type II errors. We have already considered an obvious procedure: to *increase* the rejection region. Of course, we do so at the cost of an increase in type I errors! Every researcher must strike a balance between the two types of errors. If it is important to discover *new* facts, then we may be willing to accept more *type I* errors and thus *increase* the rejection region. On the other hand, if it is important not to clog up the literature with *false* facts, then we may be willing to accept more *type II* errors and *decrease* the rejection region. Arguments can be made for both sides of this type I/type II coin; we will discuss these arguments in Chapter 8. Other ways of reducing type II errors will be considered in subsequent chapters (see especially Chapter 24). For the time being, we merely observe that type II errors may be decreased by adding to the number of observations in each treatment condition and by reducing error variance through the design of a more precisely controlled experiment.

ANOTHER NUMERICAL EXAMPLE

Now that we have looked at each step of the one-factor analysis of variance, it is time to work through a numerical example from start to finish. Suppose a researcher is interested in the effect of sleep deprivation on the ability of subjects to perform a vigilance task, such as locating objects moving on a radar screen. He arranged to house the subjects in his laboratory so that he would have control over their sleeping habits. There were $a = 4$ conditions, namely, 4, 12, 20, and 28 hours without sleep. There were $s = 4$ subjects randomly assigned to the

different treatments. The subjects were well trained on the vigilance task before the start of the experiment. They were scored on the number of failures to spot objects on a radar screen during a 30-minute test period. The scores for each subject are presented in Table 4-5.

TABLE 4-5 *Numerical Example*

AS MATRIX

	Hours Without Sleep (Factor *A*)			
	4 hr. a_1	12 hr. a_2	20 hr. a_3	28 hr. a_4
	37	36	43	76
	22	45	75	66
	22	47	66	43
	25	23	46	62
A_i:	106	151	230	247
$\bar{A}_i$:	26.50	37.75	57.50	61.75
$\sum_j^s (AS_{ij})^2$:	2962	6059	13,946	15,825

The analysis begins with the basic summing and squaring operation for the scores in each treatment condition. The results of these calculations are also listed in the table. For the subjects in level a_2, for example,

$$A_2 = \sum AS_{2j} = 36 + 45 + 47 + 23 = 151,$$

$$\bar{A}_2 = \frac{A_2}{s} = \frac{151}{4} = 37.75,$$

and

$$\sum (AS_{2j})^2 = (36)^2 + (45)^2 + (47)^2 + (23)^2 = 6059.$$

We will now calculate the three sums of squares. Substituting the values from Table 4-5 into the computational formulas given in Table 3-2,

$$SS_A = \frac{\sum (A)^2}{s} - \frac{(T)^2}{as}$$

$$= \frac{(106)^2 + (151)^2 + (230)^2 + (247)^2}{4} - \frac{(106 + 151 + 230 + 247)^2}{4(4)}$$

$$= \frac{147{,}946}{4} - \frac{538{,}756}{16}$$

$$= 36{,}986.50 - 33{,}672.25 = 3314.25,$$

$$SS_{S/A} = \sum (AS)^2 - \frac{\sum (A)^2}{s}$$

$$= (2962 + 6059 + 13{,}946 + 15{,}825) - \frac{(106)^2 + (151)^2 + (230)^2 + (247)^2}{4}$$

$$= 38{,}792 - \frac{147{,}946}{4}$$

$$= 38{,}792 - 36{,}986.50 = 1805.50,$$

and

$$SS_T = \sum (AS)^2 - \frac{(T)^2}{as}$$

$$= (2962 + 6059 + 13{,}946 + 15{,}825) - \frac{(106 + 151 + 230 + 247)^2}{4(4)}$$

$$= 38{,}792 - \frac{538{,}756}{16}$$

$$= 38{,}792 - 33{,}672.25 = 5119.75.$$

As a check,

$$SS_A + SS_{S/A} = 3314.25 + 1805.50 = 5119.75 = SS_T.$$

The remainder of the analysis is based upon the formulas listed in Table 4-1 (p. 53). The sums of squares we have just calculated are entered in Table 4-6. The *df*'s associated with the different sums of squares are

$$df_A = a - 1 = 4 - 1 = 3,$$
$$df_{S/A} = a(s - 1) = 4(4 - 1) = 12,$$

and

$$df_T = as - 1 = 4(4) - 1 = 15.$$

TABLE 4-6 *Summary of the Analysis*

Source	*SS*	*df*	*MS*	*F*
A	3314.25	3	1104.75	7.34*
S/A	1805.50	12	150.46	
Total	5119.75	15		

* $p < .01$.

As an arithmetic check,

$$df_A + df_{S/A} = 3 + 12 = 15 = df_T.$$

The between-groups and within-groups mean squares are formed by dividing the relevant sum of squares by the corresponding *df*. Specifically,

$$MS_A = \frac{SS_A}{a - 1} = \frac{3314.25}{3} = 1104.75$$

and

$$MS_{S/A} = \frac{SS_{S/A}}{a(s - 1)} = \frac{1805.50}{12} = 150.46.$$

The *F* ratio is obtained by dividing the first mean square by the second:

$$F = \frac{MS_A}{MS_{S/A}} = \frac{1104.75}{150.46} = 7.34.$$

The results of each of these steps are entered in the summary table. We will assume that the α level has been set at $p = .05$ before the start of the experiment. In order to evaluate the significance of the *F*, we locate the critical value of *F* at $\alpha = .05$ and $df_{\text{num.}} = 3$ and $df_{\text{denom.}} = 12$. From the *F* table (Table C-1 of Appendix C),

$$F(3, 12) = 3.49.$$

The decision rules given in Eq. (4-4) may be stated as follows:

Reject H_0 if $F_{\text{observed}} \geq 3.49$; otherwise, do not reject H_0.

Since F_{observed} exceeds this value, we reject the null hypothesis and conclude that the independent variable produced an effect.

The results of the statistical test are indicated by a footnote in the summary table. In this case, the entry $p < .01$ notes that the *F* we have observed is larger than the value of *F* marking off the 1 percent level of significance; i.e., $F_{(.01)} = 5.95$. As explained earlier, an indication that the obtained *F* falls within the 1 percent rejection region does not necessarily mean that we set α at $p = .01$. This particular statement of the outcome of our analysis is more informative, allowing individual researchers to use their own significance levels in evaluating our finding. Since we adopted the 5 percent level of significance, we would have rejected the null hypothesis in any case.

chapter five

ASSUMPTIONS AND ADDITIONAL CONSIDERATIONS

Certain assumptions concerning the distribution of scores within groups must be met if the analysis of variance is to "work" as described. The values listed in the F table are based on the theoretical F distribution. These values are appropriate for an analysis only when these distribution assumptions are satisfied. If they are not, then we have no simple way of determining whether or not F_{observed} falls within the rejection region of the theoretical sampling distribution of this statistic—whatever it might be with a particular set of violations. The critical question for us, of course, is to see how our *conclusions* are affected by a failure of our experiment to meet these assumptions. Such a consideration is extremely important, since rarely will we find all of the assumptions met in the experiments we conduct. If even the slightest violation can result in a considerable change in the sampling distribution of the F statistic, then we are in trouble.

There has been an important development in the evaluation of the assumptions underlying the analysis of variance. It is a practical approach to the problem—in

a sense, a "user's" approach. Monte Carlo experiments are performed, based on scores drawn at random from populations with characteristics *differing* from those assumed in the analysis. These populations are constructed to have the *same mean* but different shapes and different variances. The resultant sampling distribution of the F statistic, obtained from a large number of these random draws, is compared with the theoretical F distribution. Since the null hypothesis is true in these experiments (the population means are equal), the sampling distribution of the F statistic will equal the theoretical distribution of F only if the violations of the assumptions are *unimportant*. The degree to which the empirically derived sampling distribution deviates from the theoretical distribution provides an assessment of the practical consequences of these violations. As we will discover shortly, the sampling distribution of F is amazingly "robust"; that is, it is insensitive to even flagrant violations of the assumptions.

ASSUMPTIONS UNDERLYING THE SINGLE-FACTOR ANALYSIS OF VARIANCE

Normally Distributed Error Variance

The first assumption states that the individual treatment populations, from which the members of each treatment group are assumed to be randomly drawn, are normally distributed.[1] As a rough test of this assumption, we could look at the distribution of the AS scores within each group and estimate the general shape of the distribution. Or we could check for normality by means of a rather elaborate but objective test (see, for example, Hays, 1963, pp. 586–588). Suppose that the sample distributions are of approximately the same *non*normal shape—or, even worse, that the distributions for the different groups appear to have been drawn from populations with qualitatively different distributions. What does this do to the sampling distribution of F? Apparently, very little, especially if the groups contain equal numbers of subjects. This has been known by statisticians for some time (see Box, 1953). Monte Carlo experiments have also been performed. We will consider one of these studies in some detail.

In the Monte Carlo experiments, attention is focused upon the rejection region of the F distribution. When the sampling is from populations meeting all of the assumptions, the percentage of F ratios which actually fall in the rejection region matches very closely the percentage expected on the basis of the theoretical distribution. This point was demonstrated in Table 4-3, where the probabilities were obtained from a Monte Carlo study in which the scores for each "experiment" were drawn from the same normal distribution. An inspection of the table indicates a close correspondence between the obtained and expected probabilities for a number of possible rejection regions.

[1] This is equivalent to saying that the deviation of the members of the treatment populations from their respective population means (μ_i's) are normally distributed with a mean of zero.

An early Monte Carlo study performed by Norton (1952) is reported in detail by Lindquist (1953, pp. 78–90). Norton drew samples from distributions which were normal, leptokurtic (highly peaked), rectangular, moderately and markedly skewed, and J-shaped. He also conducted Monte Carlo tests where the scores were drawn from distributions having the same shape and from distributions having different shapes. He found that with homogeneous distributions, there was a close matching of empirical and theoretical percentages. The match was not quite as good when the populations were of markedly different form. In this latter case, the discrepancies were of the order of a 2–3 percent overestimation of the 5 percent significance level and of a 1–2 percent overestimation of the 1 percent significance level. This means, for example, that when an experimenter chooses an $\alpha = .05$, his *actual* α level may be as large as $p = .08$.

In short, Norton's study indicates that if we used the 5 percent rejection region, even for the most deviant comparisons, the empirically determined rejection region (the type I error) would be no larger than $\alpha = .08$. (This overestimation would probably have been less if Norton had used larger sample sizes. His sample sizes were 3 or 5.) Norton's and later studies tell us, then, that it is safe to conclude that violations of the normality assumption do not constitute a serious problem, except if the violations are especially severe. Under these circumstances, we need only worry about F's that fall close to the critical value of F defining the start of the rejection region.

Homogeneity of Error Variance

This assumption requires that the variances of the different treatment populations be equal. In terms of an experiment, the within-group mean squares for each group, which provide separate estimates of error variance, should be the same. Three methods are commonly used to test the homogeneity assumption. Two of these are computationally simple, but require equal sample sizes (the Cochran test and the Hartley test), while the other (the Bartlett test) is computationally complex, but accommodates unequal sample sizes.

Homogeneity of variance is tested in the Cochran test by dividing the largest within-group variance by the sum of the individual within-group variances. The Hartley test compares the largest within-group variance against the smallest within-group variance. The outcome of either test is evaluated by means of special tables. The Bartlett test is complicated and is no better than the Cochran and Hartley tests for testing the homogeneity assumption. [Winer (1962, pp. 92–96; 1971, pp. 205–210) discusses and provides examples of all three tests.]

There is a problem associated with these tests, however—a sensitivity to departures from *normality* as well as to the presence of heterogeneity. As a way to avoid this difficulty, Glass (1966) calls attention to a fourth test which is usually *not* sensitive to departures from normality. This test, proposed by Levene (1960), requires a recalculation of the basic sums of squares using

transformed AS scores. (We will consider this test in the final section of this chapter.)

In assessing the significance of a set of differences among group means, we are interested in how seriously the theoretical sampling distribution of F is distorted by the presence of unequal variances, whether these differences are significant or not. As with deviations from normality, even sizable differences among the variances do not appear to distort the F distribution seriously. The work of Box (1954), for example, shows that with equal population means and with variances in the ratio of 1:2:3, the proportion of F ratios falling within the 5 percent rejection region was .058. The results of a number of other studies [e.g., Norton (1952) for the F distribution; Boneau (1960) and Baker, Hardyck, and Petrinovich (1966) for a special two-group case of the F distribution (t)] suggest that the distortion is relatively slight when equal sample sizes are used. (Violations are more serious when unequal sizes are present.) Because of these findings, then, most researchers do not even bother to test the homogeneity assumption with their data.

Independence of Error Components

The deviation of each score from the grand mean of the population ($AS_{ij} - \mu$) is thought to contain two components, a between-group treatment effect ($\mu_i - \mu$) and a within-group deviation ($AS_{ij} - \mu_i$), the latter forming the basis for our estimate of experimental error based on sample scores. A third assumption of the analysis of variance is that the error components are independent—independent within treatment groups as well as independent between treatment groups. Independence here means that each observation is in no way related to any other observation in the experiment. The random assignment of subjects to conditions is the procedure by which we obtain independence. Of course, this is just another way of saying that systematic biases must not be present in the assignment of subjects to conditions. This is not just a statistical assumption, but a basic requirement of experimental design as well. With nonindependence of error components between treatment groups, a confounding of variables is present, and we are unable to make unambiguous inferences concerning the independent influence of our independent variable on the behavior we are studying. This assumption, then, emphasizes the critical importance statistically, as well as experimentally, of ensuring the random assignment of subjects to the treatment groups.

Additivity of Components

The model underlying the analysis we have been discussing assumes that it is proper to view an AS score as the *sum* of effects. If we start with the familiar breakdown of the total deviation into between and within deviations,

$$AS_{ij} - \mu = (\mu_i - \mu) + (AS_{ij} - \mu_i),$$

and move μ from the left side of the equation to the right side, we have

$$AS_{ij} = \mu + (\mu_i - \mu) + (AS_{ij} - \mu_i).$$

That is, an AS_{ij} score is assumed to be made up of three parts: a part representing average performance in the overall population, a deviation reflecting the treatment effect, and a deviation reflecting experimental error.[2] Researchers do not appear to question this assumption of additivity, and little is said in psychological statistics texts about conditions in which the assumption is not tenable or in which a violation of the assumption can be recognized. Thus, it is mentioned here only for the sake of completeness.

UNEQUAL SAMPLE SIZES

As we have noted, most experiments contain an equal number of subjects in each of the treatment conditions. The most obvious reason for this is perhaps the tacit intention to estimate each of the population means with the same degree of precision. There is no compelling need for equal sample sizes in the single-factor analysis of variance, although the same cannot be said without qualification in analyses involving two or more independent variables. In fact, the argument could be made that the individual sample sizes should be chosen with the expectation of precision in mind. That is, if an experimenter had reason to believe that a particular group would be more or less variable than the others (as might be the case in comparing control and experimental groups, for example), he could run a larger number of subjects in the more variable group. [Winer (1962, p. 27; 1971, p. 29) states that the most sensitive design makes sample size proportional to respective population variances.] In this situation, the unequal sample sizes are *planned*.[3] What happens when the inequality is not planned, but occurs because of an inadvertent loss of subjects? We will consider the implications of this latter event in some detail.

Why should subjects fail to complete the experiment? In animal studies, subjects are frequently lost through death and sickness. In human studies, in which testing is to continue over several days, subjects are discarded when they fail to complete the experimental sequence. In a memory study, for instance, some of the subjects may fail to return for their terminal retention test a week later, perhaps because of illness or a conflicting appointment. Subjects also may

[2] An additional assumption that we are making is that $\sum(\mu_i - \mu) = 0$. This assumption is a specification of the *fixed-effects model*, which is appropriate for most of the research in the behavioral sciences. This and additional models are discussed in Chapter 16. For the single-factor experiment, however, there is no difference in the statistical analysis dictated by the different models. Thus, we can ignore the topic at this time.

[3] On the other hand, we noted in the preceding section that violations of the homogeneity assumption are more serious when sample sizes are unequal. Consequently, you should seek expert guidance if you find yourself planning to follow Winer's advice.

be lost when studies require them to reach a performance criterion, such as a certain level of mastery; those who fail to do so are eliminated from the experiment. A third class of situations occurs when some of the subjects fail to produce responses that meet the criteria established for the response measure. Suppose we are interested in the speed with which *correct responses* are made. If a subject fails to give a correct response, he cannot contribute to the analysis. Or suppose we want to analyze the percentage of times *errors* produced on some task are of a particular type. If a subject fails to make any errors, he cannot contribute to the analysis. In such situations, subjects are eliminated from the analysis because they fail to give scorable responses.

It is of *critical* importance to determine the *implication* of these losses. That is, we have assigned our subjects to the experimental conditions in such a way that any differences among the groups at the start of the experiment will be attributed to chance factors. It is this fundamental assumption which allows us to test the null hypothesis. We are not concerned with the loss of subjects per se, but with the question: Has the loss of subjects, for whatever reason, resulted in a loss of *randomness*? If it has, either we must find a way to restore it or simply junk the experiment. No form of statistical juggling will rectify this situation. If randomness may still be safely assumed or has been restored, we can proceed with the statistical analysis of the data.

In each situation, we have to determine whether or not the reason for the subject loss is in any way associated with particular experimental treatments. In animal research, for instance, certain experimental conditions (such as operations, drugs, high levels of food or water deprivation, exhausting training procedures) may actually be responsible for the loss of the subjects. If this were the case, only the strongest and healthiest animals would survive, and the result would be an obvious confounding of subject differences and treatment conditions: the difficult conditions would contain a larger proportion of healthy animals than the less trying conditions. Replacing the lost subjects with new animals drawn from the same population will not provide an adequate solution, since the replacement subjects will not "match" the ones who were lost. If it can be shown that the loss of subjects was approximately the same from all of the conditions or that the loss was not related to the experimental treatments, then we may be able to continue with the analysis.

The same considerations are relevant when human subjects fail to complete the experiment. In the memory study we mentioned, it is likely that more subjects will be lost with the longer retention intervals, where the subjects have a greater "opportunity" to get sick. It is not known whether or not the loss of these subjects affects randomness. A researcher could attempt to see if the subjects who were lost and the subjects who were retained were equivalent in learning, although equality at this point in the experiment does not necessarily mean that the two sets of subjects would have been equivalent at recall. In some experiments, an attempt is made to impose the same subject loss on *all* conditions.

For example, suppose we require all subjects to return for the later retention test, and we follow the rule of discarding any subject who fails to return. The subject who is tested at the long interval and does not return is dropped from the experiment by default. But so is the subject who is tested at a shorter interval and fails to return for the later appointment.

The loss of subjects through failure to reach a criterion of mastery poses similar problems. Clearly, subjects who fail to learn are by definition poorer learners. If one group suffers a greater loss, which may very well happen if the conditions differ in difficulty, the subjects "making" it in the difficult condition represent a greater proportion of fast learners than those completing the training in the easier conditions. The replacement of subjects lost in the difficult condition would not solve this problem, since the replacement subjects would not match in ability the subjects who were discarded. One possibility is to compare the different groups at a lower criterion—one that will allow *all* of the subjects to be included. In this way, *no* subjects will be lost. Such a solution will be adequate if the researcher feels he will obtain the information he needs from the smaller sample of behavior.

Some experimenters solve this problem by artificially imposing a subject selection upon the groups that suffer fewer or no losses. Thus, if it can be assumed that only the poorer subjects were dropped from the more difficult conditions for failure to reach the performance criterion, then it might be possible to drop an equal number of the *poorest* subjects from *all* of the treatment conditions. A similar procedure is sometimes followed when subjects fail to give scorable responses. Suppose, for example, that an investigator is studying the speed of correct response under a number of different treatment conditions. As we have pointed out, subjects may fail to give a correct response and thus not provide a speed score. Some researchers attempt to resolve this difficulty by excluding subjects with the poorest record from the other conditions in order to "restore" equivalence of the groups. In all of these situations, however, it is assumed that the subjects whose data are discarded in the manner described are subjects who would have failed to reach the criterion or to have given any correct responses *if they had been in the condition producing the failures.* This is often a questionable assumption, but it must be made before any meaningful inferences can be drawn from the data so adjusted.

Clearly, then, the loss of subjects is of paramount concern to the experimenter. We have seen that if the loss of subjects is related to the phenomenon under study, randomness is destroyed, and a systematic bias may be added to the differences among the means which cannot be disentangled from the influence of the treatment effects. This is a problem of experimental design which must be solved by the researcher. If he can convince himself (and others) that the subject loss could not have resulted in a bias, there are statistical procedures available which will allow him to analyze his results. We will discuss these methods in Chapter 17.

THE VARIANCE AS A DESCRIPTIVE STATISTIC

For most of us, our research hypotheses are couched in terms of differences that may be observed among the treatment means. Usually, we have little sustained interest in the within-group variances, except with regard to their role in the estimation of experimental error; we look at the variances only to test the homogeneity assumption of the statistical analysis. This does not always have to be the case. Important changes in behavior caused by the different treatments may not be revealed in average performance. Or, if we look for them, systematic differences among the treatment conditions might be reflected in average performance *and* in the variability of the subjects within these conditions.

Examples of the Variance Reflecting Treatment Effects

Suppose, for example, that subjects employ a number of different strategies in performing a particular task. If they are asked to study a prose passage for an eventual test of comprehension, some may try to extract basic idea units while others may attempt to commit the entire passage to memory. The variability of their performance will reflect any differential efficiency that may be associated with these strategies—extracting idea units may be more efficient in general than learning by rote—in addition to any difference in ability that may exist among subjects. Suppose that subjects in a standard, noninstructed condition are contrasted with subjects in a condition in which they are required to perform the task with a particular strategy. Conceivably, such a comparison would not show much of a difference between the means of the two treatment groups, and a statistical analysis might lead to the conclusion that no treatment effect was present. On the other hand, the subjects in the "restricted" or instructed condition might be *more variable* in performance than the subjects in the "free" or uninstructed condition. Subjects forced to abandon their usual strategy might experience great difficulty in switching to the new strategy, which, moreover, might even be incompatible with the old one. Any such negative transfer resulting from a forced switch in strategies would show up as an increase of within-group variability for the instructed condition.

Consider another experiment, focused on the effect of administering a mild electric shock to college students each time they make an error on a motor-tracking task. Suppose that the task consists of tracking a moving object with some sighting device. The score for each subject is the number of errors he makes during a 10-minute tracking period. There are two groups; one receives a shock each time a subject loses track of the object, and one does not. How might the experiment turn out? Subjects differ greatly in how they respond even to the threat of shock—some try harder while others "freeze" and perform poorly. If subjects reacted differentially in this experiment, there should be a marked increase in variance for the shock group, with some subjects reducing

and others increasing their tracking errors. In fact, it is conceivable that the number of subjects responding "positively" to shock would be equal to the number of subjects responding "negatively," the result being no effect upon average tracking errors for the two conditions.

Thus, in some situations a comparison of variances may lead to some interesting speculations about individual differences and the way in which subjects within a group respond to the experimental treatments. These comparisons may reveal important clues as to the processes responsible for whatever effects are observed among the treatment means.

Statistical Analyses

If we are interested in determining whether or not the different treatments affected the variances of the conditions differentially, we are essentially considering the following null and alternative hypotheses:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2,$$

$$H_1: \text{all } \sigma_i^2\text{'s not equal.}$$

We could evaluate the null hypothesis by means of either the Cochran, the Hartley, or the Bartlett test, but, as we have noted, these tests are affected by departures from normality as well as by the presence of heterogeneity of variance. If we are interested only in a rough test of the homogeneity assumption of an analysis of variance, these tests are satisfactory. On the other hand, if we have a systematic interest in the differences among the variances, we want a test which is sensitive primarily to differences among the within-group variances. Such a test has been proposed by Levene (1960).

Briefly, the test consists of an ordinary analysis of variance performed on *transformed AS* scores, which we will refer to as Z scores. A Z score is defined as a within-group deviation score with the sign disregarded (an *absolute* deviation score). That is,

$$Z_{ij} = |AS_{ij} - \bar{A}_i|.$$

We will consider an example in a moment. Levene was able to show with Monte Carlo procedures that an analysis of variance of the Z scores was sensitive to differences in within-group *variances* and for all practical purposes unaffected by deviations from normality. It is interesting to note that this test was developed *empirically*—it evolved from a number of Monte Carlo experiments. Levene turned to the Monte Carlo procedure when the mathematical justification of the test proved to be unmanageable. The important point is that the test does what it was designed to do: it provides a simple test for heterogeneity of variance which is relatively insensitive to violations of the normality assumption. We will now work through an example of the analysis.

The numerical example is based on the data originally given in Table 3-3.

TABLE 5-1 *Numerical Example: The Levene Test for Homogeneity of Variance*

AS MATRIX: BASIC *AS* SCORES AND WITHIN-GROUP DEVIATIONS

a_1		a_2		a_3	
AS_{1j}	$(AS_{1j} - \bar{A}_1)$	AS_{2j}	$(AS_{2j} - \bar{A}_2)$	AS_{3j}	$(AS_{3j} - \bar{A}_3)$
16	16–15	4	4–6	2	2–9
18	18–15	6	6–6	10	10–9
10	10–15	8	8–6	9	9–9
12	12–15	10	10–6	13	13–9
19	19–15	2	2–6	11	11–9

AS MATRIX: TRANSFORMED SCORES, $Z_{ij} = |AS_{ij} - \bar{A}_i|$

	a_1	a_2	a_3
	1	2	7
	3	0	1
	5	2	0
	3	4	4
	4	4	2
$\sum_j^s Z_{ij}$:	16	12	14
$\sum_j^s (Z_{ij})^2$:	60	40	70

SUMMARY OF THE ANALYSIS

Source	Calculations	*SS*	*df*	*MS*	*F*
A	119.20 − 117.60 =	1.60	2	.80	<1
S/A	170 − 119.20 =	50.80	12	4.23	
Total	170 − 117.60 =	52.40	14		

These scores are presented again in the upper portion of Table 5-1. To the right of each *AS* score is listed the deviation of that score from the relevant group mean. (The means for the three groups in order are 15, 6, and 9.) The transformed deviations (Z) are found in the middle portion of the table. For example, for the first subject in the second group:

$$Z_{21} = |AS_{21} - \bar{A}_2| = |4 - 6| = |-2| = 2.$$

The next step is to conduct a one-factor analysis of variance on the Z scores, treating them as we would any other set of scores. For this test, we will use the

following definitions of different quantities:

$$A_i = \sum_j^s Z_{ij} \quad \text{and} \quad T = \sum Z.$$

From the data in Table 5-1,

$$SS_A = \frac{\sum (A)^2}{s} - \frac{(T)^2}{as}$$

$$= \frac{(16)^2 + (12)^2 + (14)^2}{5} - \frac{(16 + 12 + 14)^2}{3(5)},$$

$$SS_{S/A} = \sum (Z)^2 - \frac{\sum (A)^2}{s}$$

$$= [(1)^2 + (3)^2 + \cdots + (4)^2 + (2)^2] - \frac{(16)^2 + (12)^2 + (14)^2}{5},$$

and

$$SS_T = \sum (Z)^2 - \frac{(T)^2}{as}$$

$$= [(1)^2 + (3)^2 + \cdots + (4)^2 + (2)^2] - \frac{(16 + 12 + 14)^2}{3(5)}.$$

The results of these calculations are summarized in the bottom portion of Table 5-1.

The remainder of the analysis is completed in this section of the table. The F ratio is formed in the usual way:

$$F = \frac{MS_A}{MS_{S/A}},$$

and is evaluated in the F tables with the *df* associated with the numerator and denominator terms. This analysis indicates that the variances are homogeneous.