Keppel, G. (1973).

Design and Analysis:
A Researcher's Handbook

Englewood Cliffs, NJ:
Prentice-Hall

123

chapter two

THE LOGIC OF HYPOTHESIS TESTING

In the ideal experiment, we can treat the subjects in the different conditions exactly alike in every respect except for the necessary variation of the independent variable. Unfortunately, this ideal experiment is never performed in real life. That is, it is virtually impossible to conduct an experiment where the *only* difference among treatment groups is the experimental manipulation. Nonetheless, we are still able to conduct experiments and to draw meaningful conclusions from them.

Let us see how this is accomplished. First, certain features can in fact be held constant across the levels of the experiment. All of the testing can be done in the same experimental room, by the same experimenter, and with the same equipment and testing procedures. Second, control of other features of the experiment, though *not* absolute, is sufficiently close to be considered essentially constant. Consider, for example, the mechanical devices that are used to hold various features of the environment constant. A thermostat, for instance, does not achieve an absolute control of the temperature at some fixed value, but it

reduces the variation of the room temperature. An uncontrolled room would be subjected to a wider range of temperatures during the course of an experiment than a controlled room, but a variation will still be present. This variation may be sufficiently small to allow us to view the temperature as constant. Even with these features controlled, however, many variables remain uncontrolled that might influence the behavior we are studying.

We have not mentioned yet a major source of uncontrolled variability present in any experiment, namely, the differences in performance among subjects. One obvious way to hold subject differences constant is to use the *same* subject in each treatment condition—a sort of biological analogue of absolute physical control. Unfortunately, even the same subject is not the same person each time he is tested. Moreover, there are potentially serious carry-over effects from one treatment to another, owing to the successive administration of the different treatments to the same subjects. To avoid this problem, we could try to *match* sets of subjects on important characteristics and then assign one member of each matched set to a different treatment, but matching would never be exact. Thus, neither attempt to control for individual differences among subjects guarantees that the treatment groups will contain subjects of the same average ability.

124

CONTROL BY RANDOMIZATION

This leads us to a third method, one which represents control of a different sort. Specifically, it consists of an elimination of *systematic* differences among the treatment conditions by means of *randomization.* Consider again the control of room temperature. What might we do about controlling the temperature if the room were *not* equipped with a thermostat? We could try to match sets of subjects arriving at different times for the experiment, but for whom the temperature of the room is the same, and then run one of the subjects in one group, one in another group, and so on. But this is an unrealistic and cumbersome procedure. Suppose, instead, that we decide which of the different treatments a subject will receive by some random means at the time of his arrival for the experiment and that we continue to use this method until we have obtained the number of subjects we planned to run in each of the treatment conditions. What happens to the different room temperatures in this case? In a sense, the different temperatures of the experimental room have an equally likely "chance" at the start of each testing session of being assigned to *any one of the treatment levels.* If we follow this procedure with enough subjects, statistical theory tells us that the *average* room temperatures for the treatment groups will be equal. Under these circumstances, then, we will have effected a control of room temperature.

That is fine for temperature, but what about other features of the testing environment which also change from session to session? It may not be

immediately apparent, but once we have controlled *one* environmental feature by randomization, we have controlled *all* other environmental differences as well. Suppose we list some of the characteristics of the testing session present during the very first session in the experiment. The room will be at a certain temperature; there will be a certain humidity; the room illumination will be at a particular level; the noise from the outside filtering into the room will be of a certain intensity; the experiment will be given at a particular time of day, on a particular day, and by a particular experimenter; and so on.

When the experimenter is about to begin, he chooses a particular experimental treatment for the first subject in some random fashion. What this means is that at this point each of the treatment conditions has an equally likely chance of being the one chosen for that particular experimental session. The implication is that the total composite of features which happens to be present at that time has an equally likely chance of being "assigned" to each of the experimental treatments. We come next to the second experimental session. The total composite of features present at the second session will be different than the one present at the first. The room will be at a different temperature, the noise level may not be the same, the session will be at a different time of day, and so on. Before the start of the session, the experimenter again chooses randomly which treatment he will present. As with the first session, the composite of features present this time have an equally likely chance of being associated with each of the treatments.

Suppose this argument is continued until all of the subjects have been assigned to treatment conditions in the experiment. Then each and every feature of the experimental situation, which varies from session to session, has been assigned randomly to the different treatment conditions. There was no systematic bias leading to the running of one condition at the same time of day or only in warm rooms or only when the lights were bright, or whatever. The assignment of the testing sessions to the experimental conditions in a random fashion eliminates from the experiment the possibility of systematic biases involving any of these factors.

Subject differences are also "controlled" by randomization. The subjects who are chosen to participate in an experiment will differ widely on a whole host of characteristics. Some of these will affect the behavior being studied and, hence, must be controlled. Suppose we could give each of our subjects a number which represents his general ability to perform on the sort of task being studied. This number will be a composite score, reflecting the influence of his intelligence, his emotionality, his attitude, his background and training, and so on. Now suppose that we assign the subjects to the different treatment conditions randomly. Subjects with high composite scores are just as likely to be assigned to one of the treatments as to any of the others. The same is true for subjects with low and with medium composite scores. Thus, random assignment of subjects to treatments will ensure in the long run that there will be an equivalence of subjects across the different treatments.

Suppose we take one final step in this argument. Somehow we select the first subject who will be run in the experiment; he may be the first subject who shows up as a volunteer for the experiment, or he may be the rat in the first cage that we come to. When we randomly assign this subject to one of the treatment conditions, we are essentially assigning *jointly* the subject *and* the environmental factors. By assigning him randomly to the treatment conditions, then, we are assigning randomly *all* of the ability and environmental factors as well—whatever the combination of ability and environmental factors may be for this subject. Therefore, randomization of subjects in the assignment to conditions is an indispensable method of guaranteeing that in the long run the treatment conditions will be matched on all environmental factors and subject abilities.

A serious problem presented by this argument has undoubtedly occurred to you. Specifically, we *never* run a sufficiently large number of subjects in our experiment to qualify for the statistician's definition of the "long run." In practice, we are operating in the "short run," meaning that we have no guarantee that our groups will be equivalent with regard to differences in environmental features or to differences in the abilities of subjects. We will return to this problem in a moment.

Methods of Randomization

Because of the fundamental importance of randomization to the design and analysis of experiments, we will consider in detail methods by which randomization may be accomplished. Whatever method we use, we must be able to argue that *all* factors not involved in the manipulation of the independent variable have been neutralized by randomization. As an example, suppose we conduct an experiment with three treatment conditions and we plan to run a total of 30 subjects in the experiment. For the first subject who shows up, we will determine which treatment he receives by some random process.[1] The treatment given to the second subject is determined in the same manner. This procedure is followed until all 30 subjects have served in the experiment. Note that *each subject* is randomly assigned to a treatment and *each testing session* is randomly assigned to a treatment. The critical feature of the random assignment, then, is that each subject-session combination is *equally likely* to be

[1] If there were only two treatment conditions, the treatment selected could be determined by the flip of a coin. If more than two conditions are included in the experiment, we usually give each condition a different number and then refer to tables of random numbers which provide a source of random sequences of digits. Such tables may be found in many statistics texts and in experimental psychology texts. There are also books of random numbers available, such as Moses and Oakford (1963) and a book published by the RAND Corporation (1955). The tables published in Moses and Oakford are especially useful, since they include random permutations of number sets of different sizes. For example, if there are 30 things that we want to randomize, it is far easier to use a random ordering of the numbers 1–50, say, and to select from that ordering the numbers 1–30, than it is to work through a random sequence of digits, two at a time, searching for the first occurrence of each one of these numbers.

assigned to any one of the three treatments. In other words, each of the treatment conditions is equally likely to be assigned to a given subject and to whatever other uncontrolled factors might be present during that period of testing.

In actual practice, we would probably place a *restriction* on this random procedure of assigning treatments to subjects in order to ensure an *equal number* of subjects at each treatment level. (Reasons for this decision are considered in Chapter 5.) When human subjects are appearing in the laboratory at their own convenience, i.e., at a time that they choose, a typical approach is to make the random assignments so that any given treatment selected is not run again until all of the other treatments are represented *once*. In effect, this is a procedure of *sampling without replacement*. In the example, we would decide randomly which of the three treatments to administer to the first subject. For the second subject, we would randomly select the treatment from the two remaining treatments. For the third subject, we must administer the final remaining treatment, since there are only three treatments in the experiment. This completes a *block* of randomized treatments. The treatment given to the fourth subject is decided by selecting randomly from the *total pool* of treatments, i.e., three; the treatment given to the fifth subject is decided by selecting randomly from the remaining two treatments; and so on.

It is generally advisable to work with the smallest possible block, just as we did in the last paragraph. There is a good reason for following such a procedure. We can think of two general classes of variables which must be controlled in any experiment: those which really do fluctuate randomly from session to session and those which do not. We do not have to worry about the first class of variables—even if we run all the subjects in one treatment first and all the subjects in another treatment second, the particular values of these variables at each testing session by definition occur randomly. Thus, we turn to randomization to control the second class of variables, variables which do not fluctuate haphazardly.

We are usually unable to specify ahead of time exactly what the cycles of fluctuation will be; however, we merely assume that they will be present. For example, subjects volunteering for an experiment do not represent a random flow of participants. There are undoubtedly different reasons why a subject volunteers early in the school term rather than late, and these reasons may reflect differences in abilities. The first subjects may be overly anxious or curious or smarter—who knows? The point is that we cannot assume that the flow of volunteers is random. Nor is the fluctuation of room temperature or of time of day or of noise level outside the testing room random. Randomizing in small blocks "helps" this control by ensuring that a block of three subjects, say, representing each treatment once, will not be placed in a room that is too different in temperature. Or, three subjects appearing one after the other are more likely to have the same reason for volunteering at that time than would three subjects who did not.

It is not sufficient, however, just to introduce some sort of randomization in the testing order. To make the randomization "work," we must choose a method which guarantees that features of the experimental situation and differences in the abilities of the subjects are not allowed to exert a *systematic* influence in the experiment. Any factor which does not vary randomly in its "natural state" must be subjected to a process of *neutralization*, consisting in essence of the superimposition of a random process upon the assignment of testing sessions and subjects to the treatment conditions. That is, variables which fluctuate in a systematic fashion during the course of the experiment are transformed into variables which now fluctuate *unsystematically* with respect to their association with the treatment conditions.

Random Assignment Versus Random Sampling

We should say a few words about the distinction between the *random assignment* of subjects to conditions and the *random sampling* of subjects from a known population.

Random sampling requires the specification of a population of subjects and then the assurance that each member of the population has an equally likely chance of being selected for the experiment. If these conditions are met, we will be able to *generalize* the results of our experiment to the population. It should be noted that even if we are able to obtain our subjects by randomly sampling from a population, we will still have to turn to randomization procedures in the assignment of treatments to subjects and to testing sessions. That is, even randomly selected subjects will come to the experiment one at a time and then be given one of the treatment conditions. Who receives which treatment must be determined by chance; otherwise, a systematic bias may result, and this bias will be damaging to any experiment whether the subjects are selected randomly from a population or not.

What about random sampling? Public opinion polls, voter preference polls, marketing research, and television ratings all depend upon random sampling from a known population. Any findings from the sample are then extended to the population. Only rarely will we see random sampling in an experiment, however. And when we do, the population from which the sample was drawn may be so restricted as to be uninteresting in itself, e.g., the rats in a laboratory animal colony, the students at a university taking a course in introductory psychology, or third-grade children in a particular school system. Almost invariably, our subjects are selected out of *convenience*, rather than at random. The failure to sample randomly from a known population means that we are not justified *statistically* in extending our results beyond the bounds of the experiment itself.

Since most researchers accept this "myopic" view of the results of an experiment, how can we ever discover results that *are* generalizable to a meaningful population of organisms? One answer is that past research in a number of

laboratories with subjects chosen from different sources (e.g., different breedi stocks, different suppliers of laboratory animals, and human subjects fro different schools in different sections of the country) have shown that the differences are relatively unimportant in the study of various phenomen Knowing this, an investigator working in this field may feel safe in generalizin his results beyond the single experiment.

The distinction, then, is between a *statistical* generalization, which depend upon random sampling, and a *nonstatistical* generalization, which depend upon knowledge of a particular research area. Cornfield and Tukey (195(make this point quite clear: "In almost any practical situation where analytica statistics is applied, the inference from the observations to the real conclusio has two parts, only the first of which is statistical. A genetic experiment o *Drosophila* will usually involve flies of a certain race of a certain species. Th statistically based conclusions cannot extend beyond this race, yet the geneticis will usually, and often wisely, extend the conclusion to (a) the whole species (b) all *Drosophila*, or (c) a larger group of insects. This wider extension may b implicit or explicit, but it is almost always present" (pp. 912–913).

In short, the generalizability of a given set of results is influenced by statistical considerations, such as the question of random sampling. For most experimenters, however, the extension of a set of findings to a broader class of subjects (or conditions for that matter) is dictated primarily by subject-matter considerations, i.e., what is known in a particular field of research about the *appropriateness* of certain generalizations and the "length" of these generalizations. The availability of this information will depend upon the state of development of the research area and the extent to which extrapolations beyond the particular subjects tested have been successful in the past.

AN INDEX FOR THE EVALUATION OF TREATMENT EFFECTS

We now return to our earlier problem: What can we do about the fact that control by randomization will not be perfect when we are assigning relatively small numbers of subjects to our treatment conditions? Expectations from statistical theory are based upon extremely large sample sizes, much larger than those used in any experiment we will ever analyze. Since we are dealing with the "short run," where we have no assurance that our treatment groups are in fact matched on all relevant factors except the experimental treatments, how can we determine whether the differences we observe in the experiment are due either to the experimental treatments or to these chance differences or to both?

Let us consider, in general terms, a statistical solution to this disturbing problem.

Statistical Hypotheses

Hypothesis testing was described briefly in Chapter 1. At that time, we distinguished between a *research* or *scientific hypothesis* and a *statistical hypothesis*. The former is a fairly general statement about the assumed nature of the world which gets translated into an experiment. (Typically, but not always, a research hypothesis asserts that the treatments *will* produce an effect. If it did not, we would probably not have performed the experiment in the first place!) The latter is a set of precise hypotheses about the parameters of the different treatment populations. Two statistical hypotheses are usually stated, and these are mutually exclusive or incompatible statements about the treatment parameters.

The statistical hypothesis which will be *tested* is called the *null hypothesis*, often symbolized as H_0.[2] The function of the null hypothesis is to specify the values of a particular parameter (the mean, for example) in the different treatment populations. The null hypothesis typically chosen gives the *same* value to the different populations—that is,

$$H_0\colon\ \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_i.$$

This is tantamount to saying that *no* treatment effects are present in the population. If the parameter estimates obtained from the treatment groups are too deviant from those specified by the null hypothesis, H_0 is rejected in favor of the other statistical hypothesis, called the *alternative hypothesis*, H_1. The alternative hypothesis specifies values for the parameter which are *incompatible* with the null hypothesis. Usually, the alternative hypothesis states simply that the values of the parameter in the different treatment populations are *not* equal. Specifically,

$$H_1\colon \text{ all } \mu_i\text{'s not equal.}$$

Stated even more simply, the alternative hypothesis becomes

$$H_1\colon \text{ not } H_0.$$

A decision to reject H_0 implies an acceptance of H_1, which, in essence, constitutes support of our original *research* hypothesis. On the other hand, if the parameter estimates are reasonably close to those specified by the null hypothesis, H_0 is not rejected. This latter decision can be thought of as a failure of the experiment to support the research hypothesis. We will see in a later discussion that a decision to reject or not reject the null hypothesis is not all that simple. Depending upon the true state of the world, i.e., the equality or inequality of the actual population means, we can make an error of inference with *either* decision, rejection or nonrejection. (More will be said about these errors later.)

[2] Authors do not agree in these abbreviations, but the difference is minor and the reader should have no trouble in translating a different designation.

Experimental Error

At the crux of the problem is the fact that we can always attribute some portion of the differences, which we observe among treatment means, to chance factors. All uncontrolled sources of variability in our experiment, which can affect the scores on the response measure, are considered contributors to *experimental error*. As we have noted, the most important uncontrolled source of variability in the behavioral sciences is that due to individual differences. We have also mentioned variations in the various features of the testing environment. Another source of experimental error is what may be called measurement error. A misreading of a dial, a misjudgment that a particular type of behavior had occurred, the variability in reaction time of an experimenter timing a given bit of behavior, and an error in transposing observations recorded in the laboratory to summary worksheets used in performing the statistical analyses are all included in this classification. While not obvious, a given experimental treatment is not exactly the same for each subject serving in that treatment condition; the experimental apparatus cannot be counted on to administer the same treatment for successive subjects. An experimenter cannot construct an identical testing environment (the reading of instructions, the experimenter-subject interaction, and so on) for all subjects in any treatment group. We describe all these different components of experimental error as *unsystematic*, stressing the fact that their influence is *independent* of the treatment effects.

Estimates of Experimental Error

Suppose we were able to estimate the extent to which the differences we observe among the group means are due to experimental error. We would then be in a position to begin to consider the evaluation of the hypothesis that the means of the treatment populations are equal. Consider the scores of subjects in any one of the treatment conditions. We certainly do not expect these scores to be equal. In the *ideal* experiment they would be, with each score reflecting only the effect of the experimental treatment. In an *actual* experiment, of course, all of the sources of uncontrolled variability will also contribute to a subject's score, resulting in a difference in performance for subjects who are administered the same treatment conditions. The variability of subjects treated alike, i.e., within the same treatment level, provides an estimate of experimental error. By the same argument, the variability of subjects within each of the other treatment levels also offers estimates of experimental error. If we assume that experimental error is the same for the different treatment conditions, we can obtain a more stable estimate of this quantity by pooling and averaging these separate estimates.

Assume that we have drawn random samples from a population of subjects, administered the different treatments, recorded the performance of the subjects, and calculated the means of the treatment groups. Further, assume for the moment that the null hypothesis is *true*—that the population means associated

with the treatment conditions are *equal*. Would we expect the *sample* means, the means calculated in the experiment, to be equal? Certainly not. From our discussion of the use of randomization to "control" unwanted factors in our experiment, it should be clear that the means will rarely be equal. If the sample means are not equal, the only reasonable explanation that we can offer for these differences is the operation of experimental error. All of the sources of unsystematic variability, which contribute to the differences among subjects within a given treatment condition, will also be operating to produce differences among the sample means.

Take, for instance, error that results from the random assignment of subjects to treatments. If the procedure is truly random, each subject will have an equal chance of being assigned to any one of the different treatments. But this in no way *guarantees* that the average ability of subjects assigned to these groups is equal. Similarly, for the other contributors to experimental error, there is no reason to expect these uncontrolled sources of error to balance out perfectly across the treatment conditions. In short, then, under these circumstances —an experiment conducted when the null hypothesis is true—differences among the sample means will also reflect the operation of experimental error.

Estimate of Treatment Effects

So far in this discussion we have considered only the case in which the null hypothesis is true. Certainly we hope that we will discover at least a few situations in which the null hypothesis is *false*! Under these circumstances, there are real differences among the means of the treatment populations. Assuming that the subjects in each treatment group are drawn randomly from corresponding treatment populations, the means of the different groups in the experiment should reflect the differences in the population means. The mere fact that the null hypothesis is false does *not* imply that experimental error has vanished, however. Not at all. We will still have differences among subjects who were treated alike within each treatment group, which reflects experimental error, and the influence of experimental error will still be reflected in some of the variation among the group means. The only change is that there is now an additional component contributing to the differences among the means, a systematic component as opposed to an unsystematic one, namely, *treatment effects*.

Thus, differences among treatment means may reflect *two different quantities*: When the population means are equal, the differences among the group means will reflect the operation of experimental error alone, but when the population means are not equal, the differences among the group means will reflect the operation of an unsystematic component and a systematic component, i.e., experimental error and treatment effects, respectively.

EVALUATION OF THE NULL HYPOTHESIS

We have seen that when the null hypothesis is true, we will have two estimates of experimental error available from the experiment. If we form a *ratio* of these two estimates, we will find that we have produced a useful statistic. More specifically, consider the following ratio:

$$\frac{\text{differences among treatment means}}{\text{differences among subjects treated alike}}.$$

From our discussion, we can think of this ratio as contrasting an estimate of experimental error, which is based upon between-group differences, with an estimate of experimental error, which is based upon pooled within-group differences. That is, we have

$$\frac{\text{experimental error}}{\text{experimental error}}.$$

If we were to repeat this experiment a large number of times on new samples of subjects drawn from the same population, we would expect to find an average value of this ratio of approximately 1.0.

Consider now the same ratio when the null hypothesis is *false*. Under these circumstances, there is an additional component in the numerator, one which reflects the treatment effects. Explicitly, the ratio becomes

$$\frac{\text{(treatment effects)} + \text{(experimental error)}}{\text{experimental error}}.$$

Given this situation, if we were to repeat the experiment a large number of times, we would expect to find an average value of this ratio that is *greater* than 1.0.

You can see, then, that the average value of this ratio, obtained from a large number of replications of the experiment, depends upon the values of the population means. If H_0 is true (i.e., the means are equal), the average value will approximate 1.0; while if H_1 is true (i.e., the means are not equal), the average value will approximate a number greater than 1.0. A problem remains, however, since in any one experiment, it is always possible to obtain a value that is *greater* than 1.0 when H_0 is *true* and one that is *equal* or *less* than 1.0 when H_1 is *true*! Thus, merely checking to see whether or not the ratio is greater than 1.0 does not tell us which statistical hypothesis is correct.

What we will do about this is to make a decision concerning the acceptability of the null hypothesis which is based upon a consideration of the chance probability associated with the ratio we actually found in the experiment. If the probability of obtaining by chance a ratio of this size or larger is reasonably low, we will reject the null hypothesis. On the other hand, if this probability is high, we will not reject, or, in essence, we will accept the null hypothesis. (We

will have more to say about the *decision rules* we follow in making this decision in Chapter 4.)

SUMMARY

We have looked at some of the logic underlying the process of hypothesis testing. It is important for you to understand what is going on in general terms, without an elaboration of formulas and calculations—this elaboration will come soon enough! By way of summary, we can describe hypothesis testing in designs where each subject serves in only one treatment condition as consisting of a contrast between two sets of differences. One of these sets is obtained from a comparison involving differences among the treatment means; these differences are often referred to as *external* or *between-group* differences. The other set is obtained from a comparison involving differences among subjects receiving the same treatment within a treatment group; these differences are called *internal* or *within-group* differences. It was argued that the between-group differences are the result of the combined effects of the experimental treatment and of experimental error, while the within-group differences represent the influence of experimental error alone. We saw that the comparison ratio,

$$\frac{\text{between-group differences}}{\text{within-group differences}},$$

provides a numerical index which is "sensitive" to the presence of treatment effects in the population. That is, with no treatment effects, the long-run expectation is that the ratio will approximate 1.0, since the treatment effects will be zero and we will be dividing one estimate of experimental error by the other. On the other hand, whenever there are treatment effects, the expectation is that the ratio will be greater than 1.0.

The statistical hypothesis we test, the null hypothesis, specifies the *absence* of treatment effects in the population. With the help of statistical tables and a set of decision rules, neither of which we have described yet, we can decide whether or not it is reasonable to reject the null hypothesis. If we reject the null hypothesis, we accept the alternative statistical hypothesis, which specifies the presence of treatment effects in the population. If we fail to reject the null hypothesis, essentially we conclude that the independent variable produced no systematic differences in this experiment. In the next chapter we will consider the details of this basic statistical analysis.

chapter three

PARTITIONING THE TOTAL SUM OF SQUARES

In this chapter and the next, we will see the abstract notions of between-group and within-group variability become concrete arithmetic operations. This chapter will show how information reflecting these two sources of variability can be extracted from scores produced in single-factor experiments. Chapter 4 then indicates how this information is used to provide a test of the null hypothesis.

GEOMETRIC REPRESENTATION OF THE COMPONENT DEVIATION SCORES

Suppose we have conducted an experiment with school children in which we compared the relative difficulty of three kinds of conceptual tasks. We will refer to our independent variable, types of tasks, as *factor A* and to the three levels of factor A (the three different conceptual tasks) as levels a_1, a_2, and a_3.

Our subjects were drawn from a pool of school children in the fourth and fifth grades of a large school, and we assigned randomly $s = 100$ different subjects to each of the levels of factor A. The response measure was the time required to solve the different conceptual tasks.

Our first step in the analysis would be to compute the means for the three sets of scores and to compare them. As explained previously, we cannot conclude that any differences among the group means represent the "real" effects of the different experimental treatments: the differences may have resulted from experimental error, the short-term siding of uncontrolled sources of variability with one treatment condition or another. We saw that the solution to this problem is to compare the differences among the group means against the differences obtained from subjects within each of the individual groups. Let us see how this is accomplished.

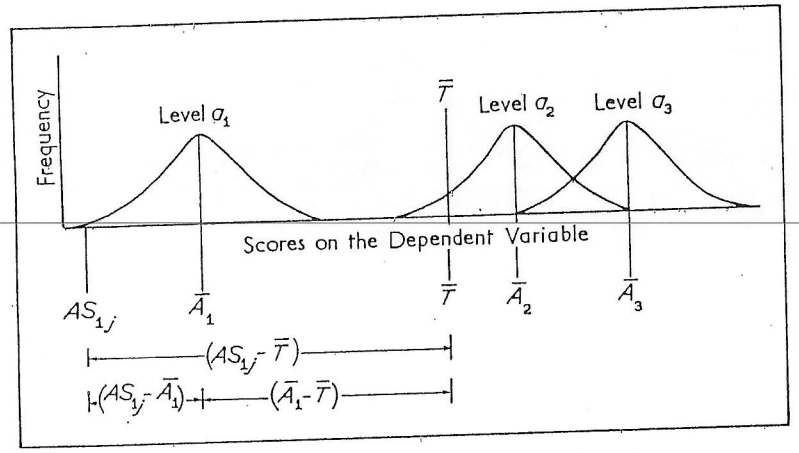

Fig. 3-1 Geometric representation of the component deviation scores.

The frequency distributions of the three sets of scores from this hypothetical experiment are presented in Fig. 3-1. Values of the response measure are indicated on the baseline, and the frequency with which specific scores are observed in each group is listed on the ordinate. (The frequency distributions in Fig. 3-1 have been smoothed; distributions of real scores would be jagged and irregular. The exact shape of these distributions is not important for the exposition which follows.) The means for the different treatment conditions are designated as $\bar{A}_1$, $\bar{A}_2$, and $\bar{A}_3$ and are located in the figure. The *grand mean* of all three conditions, obtained by summing all of the scores in the experiment and dividing by the number of scores involved [3(100) = 300, in this example], is also indicated in the plot. The symbol designating the grand mean is $\bar{T}$,

where "T" stands for the fact that the mean is based on the *total* sum of the scores.

Consider the score of one of the subjects in the distribution of scores at level a_1. We will represent this score as AS_{1j}, where the first subscript specifies the level of factor A (a_1 in this case) and the second simply refers to any one of the j subjects in this condition. (We will discuss the notational system in the next section.) Consider now the deviation of this score (AS_{1j}) from the grand mean ($\bar{T}$). This deviation ($AS_{1j} - \bar{T}$) is represented geometrically in Fig. 3-1. From the figure, it is obvious that this deviation is made up of *two components*. One consists of the deviation of the score from the mean of the group from which it was drawn, i.e., $AS_{1j} - \bar{A}_1$. The other consists of the deviation of the group mean from the grand mean, i.e., $\bar{A}_1 - \bar{T}$. This relationship may be written as

$$AS_{1j} - \bar{T} = (\bar{A}_1 - \bar{T}) + (AS_{1j} - \bar{A}_1).$$

We can give each of the three deviation scores a name: the deviation on the left ($AS_{1j} - \bar{T}$) is called the *total deviation*; the first deviation on the right ($\bar{A}_1 - \bar{T}$) is called the *between-group deviation*; and the remaining deviation ($AS_{1j} - \bar{A}_1$) is called the *within-group deviation*.

If we can perform this division or *partition*, as it is called, of the total deviation of *one* subject in this group, we can perform the partition on *all* of the 100 subjects in this group. To be more specific, the 100 sets of partitions for the subjects in level a_1 are as follows:

$$\begin{aligned}
AS_{11} - \bar{T} &= (\bar{A}_1 - \bar{T}) + (AS_{11} - \bar{A}_1),\\
AS_{12} - \bar{T} &= (\bar{A}_1 - \bar{T}) + (AS_{12} - \bar{A}_1),\\
&\cdots \qquad\qquad \cdots \qquad\qquad \cdots\\
AS_{1,99} - \bar{T} &= (\bar{A}_1 - \bar{T}) + (AS_{1,99} - \bar{A}_1),\\
AS_{1,100} - \bar{T} &= (\bar{A}_1 - \bar{T}) + (AS_{1,100} - \bar{A}_1).
\end{aligned}$$

A similar partition may be accomplished for the 100 subjects in each of the other groups.

If we now sum over all 300 total deviations, all 300 between-group deviations, and all 300 within-group deviations, we can summarize the outcome as follows:

(grand sum of total deviations) = (grand sum of between-group deviations) + (grand sum of within-group deviations).

We recall from Chapter 1, however, that the sum of the deviation scores about a mean is *zero*. Thus, for the partition to do us any "good," we must do something else with them. Fortunately, the same additive relationship, which holds for the grand sum of the different deviations, holds for the sums of the *squares*

of these deviations as well.[1] This important relationship may be stated as

$$SS_{\text{total}} = SS_{\text{between groups}} + SS_{\text{within groups}}. \quad (3\text{-}1)$$

Translated into our example, Eq. (3-1) reads, "The sum of the squared deviations of all 300 subjects from $\overline{T}$ may be broken down into two components, one obtained by summing all of the squared deviations between individual group means and $\overline{T}$ and the other by summing all of the squared deviations of subjects from their respective group means." (As you will see in the next section, however, we do not calculate these sums of squares by using these deviation scores—there is an easier way.)

SUMS OF SQUARES: DEFINING AND COMPUTATIONAL FORMULAS

While the computational chores required of the analysis of variance may be tedious to do by hand, they are quite simple in conception. In fact, we have already considered the essential logic behind the analysis. There remains only the formal presentation of the defining and computational formulas. It should be noted that these formulas will apply only to the situation in which equal sample sizes (s) are used in the treatment conditions. Although this represents a "special case," it subsumes most of the experiments conducted in the behavioral sciences. The analysis of the "atypical" case, unequal sample sizes, is presented in Chapter 17, although we will discuss in Chapter 5 implications that the presence of unequal sizes may have for our interpretation of a set of results.

Notation

Before presenting the different formulas, we should say a few words about notational systems in general and about the one adopted for this book. The basic job of a notational system is to express unambiguously the arithmetic operations in the most complex of designs as well as in the simplest. Unfortunately, most notational systems produce computational formulas that are quite difficult for students and researchers to comprehend unless they have a strong background in mathematics. The system used in this book attempts to reduce the apparent similarity of different operations by using distinct symbols to denote different sums. This system also emphasizes the consistency of the operations required of the computational formulas differing in complexity. Most important, the present notational system leads to a simple set of general computational rules, which will be introduced in Chapter 10. These rules greatly simplify the construction of computational formulas in most of our

[1] The algebraic proof of this statement can be found in most advanced statistics texts, such as Hays (1963, pp. 362–364) and Winer (1962, p. 51; 1971, p. 155).

work. The need for such a set of rules will not be apparent in the single-factor design. On the other hand, the computational rules will be extremely useful when we turn to more complicated designs. Thus, it is important to understand how the notation "works" even in this relatively simple situation.

It is instructive to compare a common notational system with the present one, especially for readers who have been brought up on the former system or some variant of it. The set of scores from a single-factor experiment is written in the two notational systems in the upper portion of Table 3-1. These are the individual scores for all of the subjects in the experiment. The basic score for the typical system is X, while the corresponding score in the present system is AS. The use of the two letters emphasizes the fact that reference is being made to a quantity which is specified by *two* classifications, namely, a classification with respect both to the level of factor A and to an individual subject in that particular treatment condition. Because the scores within the body of the total matrix of scores are each designated as AS, we will refer to the display as an *AS matrix*.

You will note that notational subscripts have been used with both systems. The pair of subscripts is required when it is necessary to designate a *particular* score in the AS matrix. With both systems, i signifies the levels of factor A and j signifies the subjects within a given treatment group. Thus, X_{12} or AS_{12} refers to the second subject at level a_1; X_{21} or AS_{21} refers to the first subject at level a_2, and so on.[2]

We will now compare the two notational systems by specifying the common arithmetic operations we will perform in the analysis of variance. These are listed in the lower portion of Table 3-1. The operation specified in the first row of the table is the sum of all of the scores in the experiment. The typical system denotes this summing operation by means of two summation signs. The first summation sign, the sign closest to X_{ij}, tells us to sum all of the X scores within the ith treatment group, while the second summation sign tells us to combine the p subtotals from the individual groups. The present system uses the letter T to stand for the grand total of all of the scores in the experiment.

The operation specified in row 2 is simply the square of the grand total. This is designated by squaring the totals listed in row 1.

The next quantity in the table is the sum of all of the *squared* basic scores, i.e., either $(X_{ij})^2$ or $(AS_{ij})^2$. Again, the other system uses two summation signs

[2] The only difference between the two systems at this point is in the designation of the last subject in any given group and of the final level of factor A. The last subject is n for the common system and is s for the present system. We have used s in order to be consistent in the system throughout the book. The last level of factor A is p for the common system, while it is a for the present system. The reason for this difference is that most authors do not want to use the lower-case a to denote a level of factor A *and* to denote the last level as well. Instead, they use a_i to refer to the ith level of factor A and a_p to refer to the last level. We will be using the letter "a" in both capacities—i.e., a_i as the ith level and a_a as the last level. There is no ambiguity in adopting this convention, however, since an "a" *without* a subscript will *always* refer to the number of levels of factor A, while an "a" *with* a subscript will *always* refer to any one of the levels of factor A; i.e., a_i = the ith level.

TABLE 3-1 *Comparison of Notational Systems*

DENOTATION OF INDIVIDUAL SCORES: THE AS MATRIX

Typical System—Levels of Factor A				Present System—Levels of Factor A			
a_1	a_2	...	a_p	a_1	a_2	...	a_a
X_{11}	X_{21}	...	X_{p1}	AS_{11}	AS_{21}	...	AS_{a1}
X_{12}	X_{22}	...	X_{p2}	AS_{12}	AS_{22}	...	AS_{a2}
⋮				⋮			
X_{1n}	X_{2n}	...	X_{pn}	AS_{1s}	AS_{2s}	...	AS_{as}

DENOTATION OF COMMON ARITHMETIC OPERATIONS

Operation	Typical System	Present System
(1) Sum of all scores	$\sum_i^p \sum_j^n X_{ij}$	T
(2) Square of (1)	$\left(\sum_i^p \sum_j^n X_{ij}\right)^2$	$(T)^2$
(3) Sum of all squared basic scores	$\sum_i^p \sum_j^n (X_{ij})^2$	$\sum (AS)^2$
(4) Sum of scores at level a_i	$\sum_j^n X_{ij}$	A_i
(5) Square of (4), summed over all groups	$\sum_i^p \left(\sum_j^n X_{ij}\right)^2$	$\sum (A)^2$

and subscripts. The present system uses no subscripts and only a single summation sign without notation. Following the convention adopted for this book, the subscripts and notation are dropped whenever the operation specified is performed on *all* of the scores in a set. Here, we are squaring every individual score and adding them all up; thus, the expression may be simplified.

The operation specified in row 4 is the sum of the individual scores for the ith treatment group, i.e., the group at level a_i. A single summation sign with appropriate limits, namely, $j = 1, 2, \ldots, n$, is required of the other system, while a capital A with a subscript is all that is required of the present system. (The A_i stands for the sum of the scores for *any* A group.)

The quantity in the last row requires the following series of calculations: (1) the sum of the scores in each treatment group, (2) the square of these sums, and (3) the sum of these squares. In the typical system, close attention must be paid to the subscripts and to the limits of summation to make sure that the quantity specified within the parenthesis is understood. In the present system, the operations are clear: the different group totals (A_i) are squared and then

summed. The subscripts and notation have been omitted, since the operations are performed on all of the group totals.

Total Sum of Squares (SS_T)

The first step in the analysis of variance is the calculation of the basic sums of squares that we evolved from the deviations specified in Fig. 3-1. The basic relationship was stated in Eq. (3-1), namely, that the total sum of squares could be partitioned into two component sums of squares, one involving between-group deviations and the other involving within-group deviations. We will consider first the computation of the total sum of squares (SS_T).

DEFINING FORMULA The basic ingredients in the SS_T are the total deviations—i.e., the deviation of each score in the experiment from the grand mean, $\overline{T}$. The SS_T is formed by squaring each one of the total deviations and summing them up for all subjects. These operations are specified in the bottom row of Table 3-2, in the column labeled "defining formulas." The individual deviations are placed within the parenthesis and the other operations, squaring and summing, are indicated outside the parenthesis. (Notational subscripts and summational limits are retained with the defining formulas to emphasize the operations, but they will be dropped from the more useful computational formulas when there is no ambiguity in the operations being specified.) The two summation signs in the defining formula indicate that the summation occurs over all subjects and treatment groups.

COMPUTATIONAL FORMULA The computational formula is specified in the next column of Table 3-2. The first term tells us to square each AS score and then to add all of them up. (A single summation sign with no notation and an AS score without subscripts are used, since there is no question that the squaring and summing operations are to be performed on all of the AS scores in the experiment.) The second term indicates that the grand total of the AS scores (T) is squared and divided by the total number of subjects in the experiment (as).

Between-Groups Sum of Squares (SS_A)

We saw in Fig. 3-1 that one of the components of a subject's total deviation is the deviation of the subject's group mean from the grand mean ($\overline{A}_i - \overline{T}$). If we square and then sum this component for all of the subjects in the experiment, we will obtain the between-groups SS. (We will refer to this quantity as the SS_A, indicating that this sum of squares is based upon deviations involving the A means.)

DEFINING FORMULA The defining formula for the SS_A is presented in the first row of Table 3-2. We will work through the "construction" of this formula. You will notice that the quantity specified within the parenthesis is the deviation

TABLE 3-2 *Defining and Computational Formulas*

Source of Variance	Defining Formula	Computational Formula
Between Group (A)	$s\left[\sum_{i}^{a}(\bar{A}_i - \bar{T})^2\right]$	$\frac{\sum(A)^2}{s} - \frac{(T)^2}{as}$
Within Group (S/A)	$\sum_{i}^{a}\sum_{j}^{s}(AS_{ij} - \bar{A}_i)^2$	$\sum(AS)^2 - \frac{\sum(A)^2}{s}$
Total	$\sum_{i}^{a}\sum_{j}^{s}(AS_{ij} - \bar{T})^2$	$\sum(AS)^2 - \frac{(T)^2}{as}$

of an individual group mean ($\bar{A}_i$) from the grand mean ($\bar{T}$). This deviation ($\bar{A}_i - \bar{T}$) is obviously the *same* for all of the members of a particular group. Thus, instead of extracting this deviation for each subject separately, it is far more convenient to obtain the deviation, square it, and then to multiply the square by the number of subjects in the group (s). Stated this way, the sum of the squared between-group deviation for a group of subjects at level a_i becomes

$$s(\bar{A}_i - \bar{T})^2.$$

The sample size (s) in this expression is often called the *weighting factor* because it adjusts or weights each between-group deviation for the number of subjects in the group. The final step requires the summing of the separate between-group sums of squares, one for each of the a groups. That is, we will perform the following operation:

$$SS_A = s(\bar{A}_1 - \bar{T})^2 + s(\bar{A}_2 - \bar{T})^2 + \cdots + s(\bar{A}_a - \bar{T})^2.$$

Symbolically, the equation becomes

$$SS_A = \sum_{i}^{a} s(\bar{A}_i - \bar{T})^2,$$

which may be simplified by placing the weighting factor (s) to the left of the summation sign. (This rearrangement indicates that we sum the squared between-group deviations first and *then* multiply by s.) This final form of the defining formula appears in Table 3-2.

COMPUTATIONAL FORMULA The computational formula is the result of an algebraic expansion of the defining formula.[3] You may be able to see some resemblance between these two formulas. Whether you "see" the resemblance or not, the operations specified by the computational formula are simple: For the first term, we are asked (1) to add up the scores within a treatment group, (2) to square each of these sums, (3) to add up the squares, and (4) to divide this final sum by s; for the second term, we are asked (1) to obtain the grand total, (2) to square this quantity, and (3) to divide by as. (Subscripts and summational limits have been dropped from the computational formula, as there is no ambiguity in the operations.)

[3] This algebraic proof can be found in Hays (1963, p. 371), Kirk (1968, p. 49), and Myers (1966, pp. 68–69).

Within-Groups Sum of Squares ($SS_{S/A}$)

The final sum of squares is the within-groups sum of squares, denoted by $SS_{S/A}$. This term is read "the sum of squares for subjects within levels (or groups) of factor A" and stresses the fact that we are dealing with the deviation of subjects from their *own* group means. Since we know how to calculate the SS_T and the SS_A, we could obtain the $SS_{S/A}$ by subtraction. We will calculate this sum of squares "directly," however, in order to be consistent with the more complicated designs we will consider later.

DEFINING FORMULA As illustrated in Fig. 3-1, the basic deviation involved in the computation of the $SS_{S/A}$ is $(AS_{ij} - \bar{A}_i)$, the deviation of an AS score from the relevant group mean. As a first step, we can obtain a sum of squares for *each group* using these within-group deviations. That is, for the ith group,

$$SS_{S/A_i} = \sum_{j}^{s}(AS_{ij} - \bar{A}_i)^2.$$

The formula tells us to square and then sum the within-group deviations for all of the subjects in the ith group. This sum of squares represents the variability of subjects treated alike.

There is a within-group sum of squares for each of the treatment groups. In the analysis of variance, we will average the different within-group variances to obtain a more stable estimate of experimental error. As a first step, then, we will want to add together the separate within-group sums of squares. This *pooling* of the separate SS's is indicated by the defining formula in Table 3-2. Many students have difficulty in extracting from the defining formula the fact that we are pooling the separate within-group sums of squares. To make the pooling more explicit, we can introduce brackets around the within-group sum of squares:

$$SS_{S/A} = \sum_{i}^{a}\sum_{j}^{s}(AS_{ij} - \bar{A}_i)^2 = \sum_{i}^{a}\left[\sum_{j}^{s}(AS_{ij} - \bar{A}_i)^2\right]. \qquad (3\text{-}2)$$

The quantity within the brackets is the sum of squares for subjects treated alike; the summation sign to the left of the brackets indicates the pooling of different sets of within-group sums of squares, one for each of the a groups.

COMPUTATIONAL FORMULA The computational formula uses two of the terms we have discussed already. We will develop the computational formula step by step. The basic ingredient is the within-group sum of squares. For the

ith group,

$$SS_{S/A_i} = \sum_j^s (AS_{ij})^2 - \frac{(A_i)^2}{s}.$$

Then, pooling the separate within-group sums of squares, we have

$$SS_{S/A} = \left[\sum_j^s (AS_{1j})^2 - \frac{(A_1)^2}{s}\right] + \left[\sum_j^s (AS_{2j})^2 - \frac{(A_2)^2}{s}\right] + \cdots + \left[\sum_j^s (AS_{aj})^2 - \frac{(A_a)^2}{s}\right].$$

This last step may be represented by

$$SS_{S/A} = \sum_i^a \left[\sum_j^s (AS_{ij})^2 - \frac{(A_i)^2}{s}\right]. \tag{3-3}$$

The quantity within the brackets is the computational formula for any one of the within-group sums of squares; the summation sign to the left of the brackets tells us to sum all of the separate sums of squares.

A form of Eq. (3-3), which is simpler arithmetically, eliminates the brackets and tells us to perform the summation over the a groups separately for each of the two terms inside of the brackets. Specifically,

$$SS_{S/A} = \sum_i^a \sum_j^s (AS_{ij})^2 - \frac{\sum_i^a (A_i)^2}{s}.$$

This formula may be simplified by applying the convention adopted in this book, namely, to use single, unlabeled summation signs and to eliminate subscripts when the summing operations are performed on all of the terms in a set. For the first quantity on the right, all of the AS scores are squared and then summed; for the second quantity on the right, all of the A sums are squared and summed. Thus, the simplification becomes

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

Computational Rule

One feature of the various computational formulas in Table 3-2 should be pointed out, since it will occur in all of the analyses we will consider in later chapters. Specifically, *computational rule number 1* states that we

always divide by the number of observations contributing to one of the quantities in the numerator.

If we apply this rule to the first term in the computational formula for the SS_A, we see that the basic term in the numerator represents the sum of s observations, and we divide by that number. In the second part of the formula, the basic term is obtained by summing all of the as observations; consequently, we divide by as. Finally, the remaining unique term appearing in the computational formulas, $\sum (AS)^2$, results from a squaring of each *individual* observation. Thus, an application of the rule would require that we divide each square by 1, a division which is implied but not shown in the formulas. (An alternative way of remembering the computational rule is to think of each quantity that is squared as a *mean*, the divisor being a number which reflects the number of observations.) This rule should be memorized, as it will allow the generation of the formulas necessary for the calculation of relatively complex analyses of variance.

NUMERICAL EXAMPLE

Suppose we were interested in the effect on reading comprehension of three different instructions. One group of children is asked to attempt to memorize an essay (level a_1), a second group is asked to concentrate upon the idea units (level a_2), and a third group is given no specific instructions (level a_3). All subjects are allowed to study the essay for 10 minutes; then they are given an objective test to determine their comprehension of the passage. There are $s = 5$ subjects who were randomly assigned to each of the $a = 3$ treatment conditions. Table 3-3 presents an AS matrix containing the data from this hypothetical experiment.

The score for a subject (AS) represents the number of test items correctly answered. Also included at the bottom of the table are the results of basic summing and squaring operations for each group and, in the second to the last row, the group means. For the three groups at levels a_1, a_2, and a_3,

$$A_1 = 16 + 18 + 10 + 12 + 19 = 75,$$

$$A_2 = 4 + 6 + 8 + 10 + 2 = 30,$$

and

$$A_3 = 2 + 10 + 9 + 13 + 11 = 45,$$

respectively. The corresponding sums of the squared AS scores are as follows:

$$\sum (AS_{1j})^2 = (16)^2 + (18)^2 + (10)^2 + (12)^2 + (19)^2 = 1185,$$

$$\sum (AS_{2j})^2 = (4)^2 + (6)^2 + (8)^2 + (10)^2 + (2)^2 = 220,$$

and

$$\sum (AS_{3j})^2 = (2)^2 + (10)^2 + (9)^2 + (13)^2 + (11)^2 = 475,$$

respectively.

TABLE 3-3 *Numerical Example: AS Matrix*

	Treatment Levels		
	a_1	a_2	a_3
	16	4	2
	18	6	10
	10	8	9
	12	10	13
	19	2	11
A_i:	75	30	45
$\bar{A}_i$:	15.00	6.00	9.00
$\sum_j^s (AS_{ij})^2$:	1185	220	475

Applying the computational formula in Table 3-2 to the present data, we have

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

$$= \frac{(75)^2 + (30)^2 + (45)^2}{5} - \frac{(16 + 18 + \cdots + 13 + 11)^2}{3(5)}.$$

In order to emphasize the application of the computational rule, we can rewrite the first term of the formula by placing each square over s, the number of observations summed to obtain each total. That is,

$$\frac{\sum(A)^2}{s} = \sum\frac{(A)^2}{s}$$

and

$$\frac{(75)^2 + (30)^2 + (45)^2}{5} = \frac{(75)^2}{5} + \frac{(30)^2}{5} + \frac{(45)^2}{5}.$$

Note also that the quantity T, appearing in the numerator of the last term, may be calculated by summing the A_i sums rather than returning to the AS matrix and summing the AS scores individually. This is another way of pointing out that

$$T = \sum AS = \sum A.$$

Completing the calculations,

$$SS_A = \frac{5625 + 900 + 2025}{5} - \frac{(150)^2}{15}$$

$$= \frac{8550}{5} - \frac{22{,}500}{15}$$

$$= 1710.00 - 1500.00 = 210.00.$$

For the $SS_{S/A}$ we have

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

$$= [(16)^2 + (18)^2 + \cdots + (13)^2 + (11)^2] - \frac{(75)^2 + (30)^2 + (45)^2}{5}$$

$$= 1880 - \frac{8550}{5}$$

$$= 1880 - 1710.00 = 170.00.$$

As we saw in the development of the computational formula, the $SS_{S/A}$ is made up of the separate within-group sums of squares for the treatment groups. To illustrate this point, we will obtain the $SS_{S/A}$ by using Eq. (3-3) and computing the sums of squares separately. For any one group,

$$SS_{S/A_i} = \sum_j^s (AS_{ij})^2 - \frac{(A_i)^2}{s}.$$

The individual within-group sums for the present example are

$$SS_{S/A_1} = \sum(AS_{1j})^2 - \frac{(A_1)^2}{s} = 1185 - \frac{(75)^2}{5}$$

$$= 1185 - \frac{5625}{5} = 1185 - 1125.00 = 60.00,$$

$$SS_{S/A_2} = \sum(AS_{2j})^2 - \frac{(A_2)^2}{s} = 220 - \frac{(30)^2}{5}$$

$$= 220 - \frac{900}{5} = 220 - 180.00 = 40.00,$$

and

$$SS_{S/A_3} = \sum(AS_{3j})^2 - \frac{(A_3)^2}{s} = 475 - \frac{(45)^2}{5}$$

$$= 475 - \frac{2025}{5} = 475 - 405.00 = 70.00.$$

The total of these sums of squares equals the $SS_{S/A}$, which we found by using the computational formula listed in Table 3-2. That is,

$$SS_{S/A} = \sum SS_{S/A_i} = 60.00 + 40.00 + 70.00 = 170.00.$$

The final computation consists of the calculation of the SS_T. From the computational formula in Table 3-2, we substitute as follows:

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

$$= [(16)^2 + (18)^2 + \cdots + (13)^2 + (11)^2] - \frac{(75 + 30 + 45)^2}{3(5)}$$

$$= 1880 - \frac{22{,}500}{15} = 1880 - 1500.00 = 380.00.$$

Several computational checks should be mentioned. First, when we are summing squares with a desk calculator [e.g., $\sum (AS)^2$ or $\sum (A)^2$], we should check to see that the number appearing in the register cumulating the *unsquared* numbers matches the actual total sum (T). If the numbers do not match, then we know we have made a mistake. Second, in the calculation of the sums of squares, we can check our arithmetic by applying Eq. (3-1) and verifying that the two component sums of squares add up to the SS_T. In the present example,

$$SS_T = SS_A + SS_{S/A} = 210.00 + 170.00 = 380.00.$$

Third, a complete check of all of our calculations may be obtained a number of ways. One obvious method is to perform the analysis again, or, perhaps better still, coax another person to go through the calculations independently. An alternative method is to add a constant, say, 1, to each AS score (i.e., $AS_{ij} + 1$) and to repeat the complete analysis. For example, the scores in level a_1 would thus become 17, 19, 11, 13, and 20, respectively. If we have made no error in either set of calculations, we should end up with *identical* sums of squares in the two analyses. The addition of a constant does not change the basic *deviation* scores, upon which the sums of squares are fundamentally based, but it does change the actual numbers entering into the calculations when we use the computational formulas of Table 3-2.

ANALYSIS OF DEVIATION SCORES

In discussing the component sums of squares, we began by considering the deviation of AS scores from $\overline{T}$ and the fact that for each observation this deviation may be divided into a between-group deviation ($\overline{A} - \overline{T}$) and a within-group deviation ($AS - \overline{A}$). We saw that the defining formulas for the corresponding sums of squares were developed directly from these deviations. Our actual calculations, however, were performed with the computational formulas, which are considerably easier to use. Since the intuitive meaning of the sums of squares in general and of the partition of the SS_T into the SS_A and the $SS_{S/A}$ is much clearer with the defining formulas, it is instructive to work through the numerical example using these formulas and to compare the outcomes with the results obtained with the computational formulas.

TABLE 3-4 *Analysis of Component Deviation Scores*

	Deviation Scores				
AS_{ij}	Total $(AS_{ij} - \overline{T})$	=	Between $(\overline{A}_i - \overline{T})$	+	Within $+ (AS_{ij} - \overline{A}_i)$
			LEVEL a_1		
16	(6)	=	(5)	+	(1)
18	(8)	=	(5)	+	(3)
10	(0)	=	(5)	+	(−5)
12	(2)	=	(5)	+	(−3)
19	(9)	=	(5)	+	(4)
			LEVEL a_2		
4	(−6)	=	(−4)	+	(−2)
6	(−4)	=	(−4)	+	(0)
8	(−2)	=	(−4)	+	(2)
10	(0)	=	(−4)	+	(4)
2	(−8)	=	(−4)	+	(−4)
			LEVEL a_3		
2	(−8)	=	(−1)	+	(−7)
10	(0)	=	(−1)	+	(1)
9	(−1)	=	(−1)	+	(0)
13	(3)	=	(−1)	+	(4)
11	(1)	=	(−1)	+	(2)
Sum:	(0)	=	(0)	+	(0)

The analysis of the deviation scores is presented in Table 3-4. The first column of the table lists the AS scores, grouped according to levels of factor A. The deviation of an AS_{ij} score from $\overline{T}$ is listed in the second column, and the between-group and within-group deviations are entered in the third and fourth columns of the table, respectively. In order to calculate the deviation scores, we need the different means. From Table 3-3, the means for the treatment groups are $\overline{A}_1 = 15$, $\overline{A}_2 = 6$, and $\overline{A}_3 = 9$. The grand mean is

$$\overline{T} = \frac{T}{as} = \frac{150}{3(5)} = 10.00.$$

We are now ready to obtain the deviations. Consider the first subject in level a_1 and his deviation scores:

$$AS_{11} - \overline{T} = (\overline{A}_1 - \overline{T}) + (AS_{11} - \overline{A}),$$

$$16 - 10 = (15 - 10) + (16 - 15).$$

The deviation scores for the last subject in level a_2 are as follows:

$$AS_{25} - \overline{T} = (\overline{A}_2 - \overline{T}) + (AS_{25} - \overline{A}_2),$$

$$2 - 10 = (6 - 10) + (2 - 6).$$

And the deviation scores for the fourth subject in level a_3 are

$$AS_{34} - \overline{T} = (\overline{A}_3 - \overline{T}) + (AS_{34} - \overline{A}_3),$$

$$13 - 10 = (9 - 10) + (13 - 9).$$

The results of the subtractions for these and for the remaining scores are given in the last three columns of the table.

Consider the deviation scores. First, we should note that the between-group deviation is the same for each of the subjects in a given group—namely, 5 for a_1, -4 for a_2, and -1 for a_3. We can also see that the sum of the within-group deviation scores is zero for the subjects in each group. This is as it should be, since the sum of the deviations about a mean is zero. Finally, the sums of the other two sets of deviations are also zero when the summing is taken over all *as* observations. This, too, makes sense, since the total and between-group deviations represent deviations about the grand mean, and the respective deviations will not balance until all of the deviations are summed.

All that remains now is to square and to sum the deviation scores in the table. For each set,

$$SS_T = (6)^2 + (8)^2 + \cdots + (3)^2 + (1)^2 = 380,$$

$$SS_A = (5)^2 + (5)^2 + \cdots + (-1)^2 + (-1)^2 = 210,$$

and

$$SS_{S/A} = (1)^2 + (3)^2 + \cdots + (4)^2 + (2)^2 = 170.$$

These sums are identical to the ones obtained with the computational formulas in the last section.

This analysis makes explicit the meaning of the component deviation scores. It is clear that the between-group deviation is the same for each observation within a particular group, illustrating why we multiplied the deviation of the group mean from $\overline{T}$ by s in the defining formula listed in Table 3-2. We can also see that the $SS_{S/A}$ represents a pooling of the three within-group sums of squares. From the last column of Table 3-4,

$$SS_{S/A_1} = (1)^2 + (3)^2 + (-5)^2 + (-3)^2 + (4)^2 = 60,$$

$$SS_{S/A_2} = (-2)^2 + (0)^2 + (2)^2 + (4)^2 + (-4)^2 = 40,$$

and

$$SS_{S/A_3} = (-7)^2 + (1)^2 + (0)^2 + (4)^2 + (2)^2 = 70.$$

The sum of these individual sums of squares equals the $SS_{S/A}$. Specifically,

$$SS_{S/A} = \sum SS_{S/A_i} = 60 + 40 + 70 = 170.$$

chapter four

VARIANCE ESTIMATES AND THE *F* RATIO

As Winer (1962, p. 59; 1971, p. 163) points out, it is accurate to express the null hypothesis in terms of a between-groups variance, one which is based on deviations of the means of treatment populations from the overall population mean. More specifically,

$$H_0\colon (\mu_1 - \mu) = (\mu_2 - \mu) = \cdots = (\mu_a - \mu) = 0,$$

where the means with subscripts are the population treatment means and μ is the mean of these individual means. An obvious candidate from the data of the experiment to estimate these deviations is the between-group deviation scores $(\overline{A}_i - \overline{T})$. It will be recalled from a previous discussion, however, that differences among the sample means are subject to experimental error, so that the mere presence of sample deviation scores greater than zero is insufficient cause to justify the conclusion that treatment effects are present in the population.

Thus, a variance estimate, which is based on between-group differences, is the sum of two components—a treatment component and an error component.

Recall also that within-group deviation scores ($AS_{ij} - \bar{A}_i$) can be thought to reflect the influence of the error component alone. It is possible to show that when the null hypothesis is true and there is no population treatment effect, variances based on these two sets of deviations provide independent estimates of the error component. (We will consider this proof subsequently.) Given this fact, then, the ratio

$$F = \frac{\text{between-groups variance}}{\text{within-groups variance}} \tag{4-1}$$

provides a useful means for testing the reasonableness of the null hypothesis.

Suppose, for example, that we conducted a particular experiment a large number of times and looked at the *average* value of F resulting from these different experiments. (The distribution of these F values is called the *sampling distribution* of F, and the mean of the sampling distribution is called the *expected value* of F.) When H_0 is true, the expected value of F, which is often designated as E(F), is approximately 1.0; when the alternative hypothesis (H_1)—the hypothesis that the population means are *not* equal—is true, the expected value of F is greater than 1.0. While these expectations are clear, we will *rarely* obtain an F which is equal to the value of F expected under the null hypothesis. The reason is the independence of the two estimates of error variance; sometimes the estimate based on between-group deviations will be larger than the estimate based on within-group deviations, and sometimes the reverse. The problem is to find a way to decide when an F, which is greater than 1, cannot be reasonably accounted for by the fact that the two variances are independent estimates of experimental error.

We will now discuss a solution to this problem. We will consider first the formulas which provide the two estimates needed in Eq. (4-1), and then how we can use the value of F to test the null hypothesis.

VARIANCE ESTIMATES

The remainder of the analysis is outlined in Table 4-1 in an arrangement called a *summary table*. The first two columns list the sources of variability and their respective sums of squares. We indicated in Chapter 1 that a variance to be used to estimate population characteristics is defined differently from one to be used for purely descriptive purposes. In Eq. (1-5a), we expressed the arithmetic operations as

$$\hat{\sigma}^2 = \frac{SS}{df},$$

where SS refers to the basic sum of squares and df represents the *degrees of freedom* associated with the SS.

Degrees of freedom (df)

The df associated with a sum of squares correspond to the number of scores with *independent information* which enter into the calculation of the sum of squares. Consider, for example, the use of a single sample mean to estimate the population mean. If we want to estimate the population variance as well, we must take account of the fact that we have used up some of the independent information already in estimating the population mean.

Consider a concrete example. Suppose that we have five observations in our experiment and that we determine the mean of the scores to be 7.0. This mean is used to estimate the population mean. With the number of observations set at five and the population mean set at 7.0, how much independent information remains for the estimate of the population variance? The answer is the number of observations which are *free* to vary—i.e., to take on any value whatsoever. The number in this example is *four*, one less than the total number of observations. The reason for this loss of "freedom" is that, while we are free to select any value for the first four scores, the final score is already determined. More specifically, the total sum of all five must equal 35, so that the mean of the sample will equal 7.0; as soon as four scores are selected, the fifth score is fixed and can be obtained by subtraction. In a sense, then, estimating the population mean places a restraint upon the values that the scores are free to take. The general rule for computing the df of any sum of squares is

$$df = \begin{pmatrix}\text{number of}\\ \text{independent}\\ \text{observations}\end{pmatrix} - \begin{pmatrix}\text{number of}\\ \text{restraints}\end{pmatrix} \tag{4-2}$$

or

$$df = \begin{pmatrix}\text{number of}\\ \text{independent}\\ \text{observations}\end{pmatrix} - \begin{pmatrix}\text{number of}\\ \text{population}\\ \text{estimates}\end{pmatrix}. \tag{4-2a}$$

The df associated with each sum of squares in the analysis of variance are presented in the third column of Table 4-1. We can calculate the df for each sum of squares by applying Eq. (4-2). For the SS_A, there are a basic observations—i.e., a different sample means. Since 1 df is lost as a result of estimating

TABLE 4-1 *Summary Table for the One-Factor Analysis of Variance*

Source	SS	df	Mean Square (MS)	F Ratio
A	SS_A	$a - 1$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{S/A}}$
S/A	$SS_{S/A}$	$a(s - 1)$	$\frac{SS_{S/A}}{df_{S/A}}$	
Total	SS_T	$as - 1$		

the overall population mean (μ) from the grand mean of the experiment ($\bar{T}$), $df_A = a - 1$. For the $SS_{S/A}$, the calculation of df is more complicated. This sum of squares represents a pooling of separate estimates of error variance from the different treatment groups. If we consider any one of these groups, there are s basic observations; we will lose 1 df, however, by estimating the mean of the treatment population (μ_i). Thus, there are $df = s - 1$ for each of the treatment groups. The total number of df for the $SS_{S/A}$ is found by pooling the df for each group, just as we pool the corresponding sums of squares. The formula given in Table 4-1,

$$df_{S/A} = a(s - 1),$$

simply has us multiply the df for any one of the groups ($s - 1$) by the number of different groups (a). The df for the SS_T are obtained by subtracting 1 df from the total number of independent observations (as). As a check, we can verify that the df associated with the component sums of squares sum to df_T. That is,

$$df_T = df_A + df_{S/A},$$

$$as - 1 = (a - 1) + a(s - 1) = a - 1 + as - a = as - 1.$$

Mean Squares

The actual variance estimates appear in the next column of Table 4-1. These estimates are called mean squares (MS), a term which refers to an averaging of a set of squared numbers. The mean squares for the two component sources of variance are given by

$$MS = \frac{SS}{df}, \tag{4-3}$$

or, more specifically,

$$MS_A = \frac{SS_A}{df_A} \quad \text{and} \quad MS_{S/A} = \frac{SS_{S/A}}{df_{S/A}}.$$

The first mean square estimates the combined presence of treatment effects plus error variance, while the second mean square independently estimates error variance.

The whole logic of the analysis of variance rests upon the assumption that the MS_A and the $MS_{S/A}$ provide *independent* estimates of error variance when the null hypothesis is true. It is possible to prove the correctness of this assumption mathematically, but this proof is beyond the scope of this book; instead, we will evaluate the assumption in a number of other ways.

First, we use a logical argument. Suppose that we have a sets of s scores each and that we calculate the MS_A and the $MS_{S/A}$. If we change the AS scores by any amount, but hold the treatment means constant, the $MS_{S/A}$ will change but the MS_A will not. On the other hand, if we change the group means by any

amount, but do not change the relative standing of the scores within treatment groups, the MS_A will change but the $MS_{S/A}$ will be constant. The two mean squares are independent in the sense that one can be changed without requiring a change in the other.

Another line of argument is empirical. Walker and Lev (1953, p. 210), for example, report the results of a sampling experiment in which the SS_A and the $SS_{S/A}$ were obtained from four sets of random samples of scores, each drawn from the same population of scores. Forty-six such "experiments" were conducted. Since the samples of scores were drawn from the same population, the null hypothesis is true and the pairs of estimates from each experiment should be independent. This is exactly what Walker and Lev report: independence—i.e., an essentially zero correlation ($r = -.05$) between these two component sources of the SS_T.

Finally, Appendix A presents a proof that depends on *orthogonality*, a topic we will consider in Chapter 7.

THE F RATIO

The final step in the calculations consists of the formation of the F ratio. The formula is listed in the last column of Table 4-1. As we have argued previously, the expected value of F is approximately 1.0 when the null hypothesis is true, and is greater than 1.0 when the null hypothesis is false.

Numerical Example

We will continue with the numerical example we used in the last chapter to illustrate the calculation of the sums of squares. The results of these earlier calculations are presented in Table 4-2, an analysis-of-variance summary table. If you recall, there were $a = 3$ treatment conditions and $s = 5$ subjects in this example; the original data may be found in Table 3-3.

The df for the three sources of variance are obtained by simple substitution into the formulas listed in Table 4-1. The results of these substitutions are presented in the summary table. Specifically,

$$df_A = a - 1 = 3 - 1 = 2,$$

$$df_{S/A} = a(s - 1) = 3(5 - 1) = 3(4) = 12,$$

and

$$df_T = as - 1 = 3(5) - 1 = 15 - 1 = 14.$$

We can check our separate calculations by verifying that the df obtained for the component sums of squares equal the df obtained for the SS_T. That is,

$$df_T = df_A + df_{S/A} = 2 + 12 = 14.$$

TABLE 4-2 *Summary Table*

Source	*SS*	*df*	*MS*	*F*
A	210.00	2	105.00	7.41
S/A	170.00	12	14.17	
Total	380.00	14		

The two variance estimates (mean squares) are found by dividing the *SS* by the appropriate *df*. In this example,

$$MS_A = \frac{210.00}{2} = 105.00 \quad \text{and} \quad MS_{S/A} = \frac{170.00}{12} = 14.17.$$

These numbers are entered in the *MS* column of the table. Last, the *F* ratio becomes

$$F = \frac{MS_A}{MS_{S/A}} = \frac{105.00}{14.17} = 7.41.$$

This value of *F* is larger than 1.0, bringing into question the correctness of the null hypothesis. On the other hand, we might have obtained a ratio this large (or larger) merely by virtue of the fact that the two mean squares are independent estimates of error variance when H_0 is true.

140

EVALUATION OF THE *F* RATIO

Sampling Distribution of *F*

In view of the minimal mathematical background assumed for the readers of this book, a reasonable approach to this topic is empirical rather than theoretical. Suppose that we had available a large population of scores and that we drew at random three sets of 15 scores each. We can think of the three sets as representing the results of an actual experiment, with $a = 3$ and $s = 15$, for which we *know* the null hypothesis is *true*. That is, the scores placed in each "treatment" condition were in fact drawn from the *same* population. Thus, $\mu_1 = \mu_2 = \mu_3 = \mu$. The two mean squares, MS_A and $MS_{S/A}$, are independent estimates of experimental error: we may estimate the operation of the same chance factors either by looking at the variability among the three sample means or by looking at the pooled variability of the scores within each of the samples.

Assume that we draw a very large number of such "experiments," each consisting of three groups of 15 scores each, and that we compute the value of *F* for each case. If we group the *F*'s according to size, we can construct a graph relating *F* and frequency of occurrence. A frequency distribution of a statistic

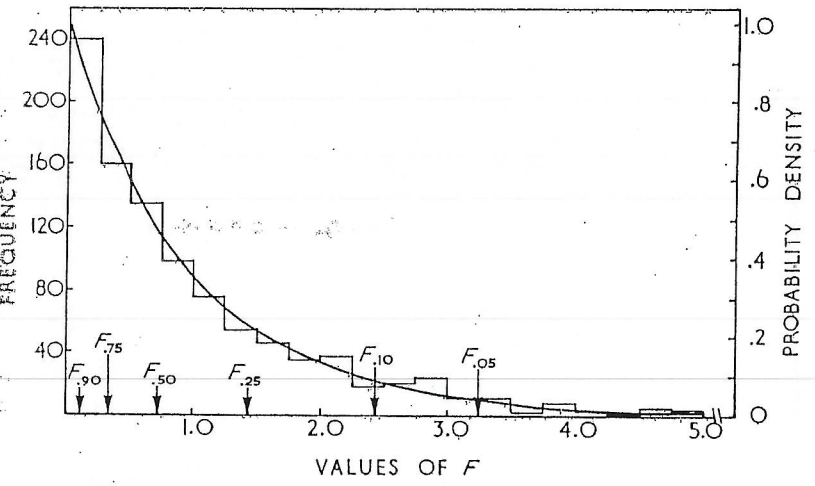

Fig. 4-1 Empirical and theoretical sampling distributions of $F(2, 42)$.

such as *F* is called a *sampling distribution* of the statistic. By obtaining the sampling distribution of *F* *empirically* in this manner, we can see how the theoretical sampling distribution could be developed.

This sort of empirical sampling study is called a "Monte Carlo" experiment. A sampling distribution of *F*, based on 1000 experiments of the sort we have been discussing, is presented in Fig. 4-1. The sampling and calculations were performed on a high-speed computer.[1] The population consisted of 6000 scores with $\mu = 50$ and $\sigma^2 = 225$. The histogram exhibits a regular trend, the frequency of cases tending to drop off rapidly with increasing values of *F*. The smoothed curve represents the theoretical sampling distribution of *F*. The approximation of the theoretical curve to the empirically obtained sampling distribution is extremely close. This correspondence provides a convincing intuitive meaning to the *F* distribution—namely, that it *is* the sampling distribution of *F* obtained when an infinitely large number of experiments, of the sort we have been discussing, are performed.

In evaluating the null hypothesis, we could use information drawn from either the empirical or the theoretical sampling distributions. The great advantage of knowing the mathematical properties of the *F* distribution is that the sampling distribution can be determined for any experiment of any size—i.e., any number of groups and any number of subjects within these groups. Separate Monte Carlo experiments would have to be conducted for each new situation—an inefficient and costly procedure.

[1] The results of this and other sampling experiments which we will discuss were generously made available to me by Drs. Curtis D. Hardyck and Lewis F. Petrinovich.

Let us return to Fig. 4-1 and see what useful information can be obtained from the F distribution. If we know the exact shape of the F distribution for a given experiment, we can make statements concerning how common or how rare an F observed in an actual experiment is. The F distribution is the sampling distribution of F when the population means are equal. If we consider a particular value of F, we can determine (with a working knowledge of the calculus) the probability of obtaining an F that large or larger by finding the percentage of the area under the curve which falls to the right of an ordinate erected at the value of F in question. Several values of F have been indicated in Fig. 4-1. The proportion of the curve falling to the right of F is indicated as a subscript. For example, only 10 percent of the time would we expect to obtain a value of F equal to or greater than 2.44. Stated another way, this probability represents the proportion of F's ≥ 2.44 which will occur on the basis of chance factors alone. A comparison of the theoretical probabilities with the empirical probabilities found in the Monte Carlo experiment is presented in Table 4-3. We can see that the correspondence between the two sets of probabilities is quite close.

TABLE 4-3 *Comparison of the Theoretical and Empirical Sampling Distributions of $F(2, 42)$*

Value of F	Theoretical Probability	Empirical Probability
.11	.900	.950
.29	.750	.754
.70	.500	.518
1.44	.250	.265
2.44	.100	.114
3.23	.050	.050
5.18	.010	.015
8.25	.001	.002

THE F TABLE The F distribution is actually a family of curves. The exact shape of any one of the curves is determined by the number of df associated with the numerator and denominator mean squares in the F ratio. If we hold numerator df (the number of treatment groups) constant and vary the denominator df (the number of subjects within groups), we will see relatively small changes in the shape of the curves. On the other hand, changing the number of treatment groups produces curves of quite different appearance. [If you are curious about the F distribution and are handy with simple algebra and logarithms, Lewis (1960, pp. 311–312) indicates a method for its determination.] An example of another F distribution, with numerator and denominator $df = 4$ and 10, respectively, is sketched in Fig. 4-2. [As a shorthand way of referring to a particular F distribution, we will use the expression, $F(df_{\text{num.}}, df_{\text{denom.}})$, or in this case, $F(4, 10)$.]

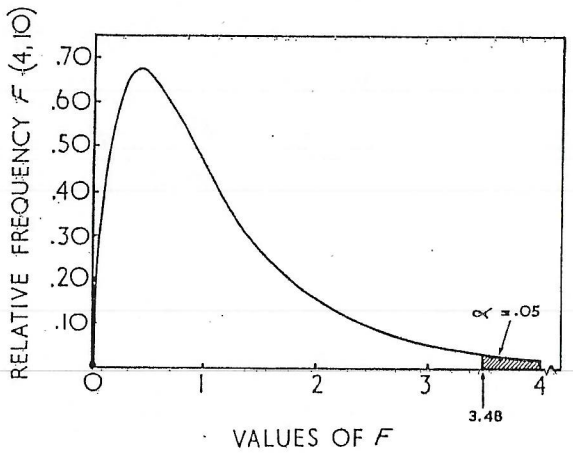

Fig. 4-2 Sampling distribution of $F(4, 10)$.

For our experiment, we do not have to know the exact shape of the F distribution. The only information we need is the value of F, to the right of which certain proportions of the area under the curve fall. These values have been tabulated and are readily available. An abridged F table is found in Table C-1 of Appendix C. A particular value of F in this table is specified by three factors: (1) the numerator df (represented by the columns of the table), (2) the denominator df (represented by the main rows of the table), and (3) the value of α (represented by the rows listed for each denominator df), where α refers to the proportion of area to the right of an ordinate drawn at F_α.

For example, the value of $F(4, 10) = 3.48$ at $\alpha = .05$. This F is found by locating the intersection of the column at $df_{\text{num.}} = 4$ and the row at $df_{\text{denom}} = 10$. The different values of $F(4, 10)$ in this location represent critical points for a number of different α levels. The one we want is at $\alpha = .05$. What this value of F means is that an ordinate drawn at $F(4, 10) = 3.48$ will divide the sampling distribution of $F(4, 10)$ at a point where the proportion of the area under the curve to the right is .05. Said another way, $\alpha \times 100 = .05 \times 100 = 5$ percent of the area under the curve falls to the right of an ordinate drawn at $F(4, 10) = 3.48$. For $\alpha = .25$, $F(4, 10) = 1.59$ and for $\alpha = .01$, $F(4, 10) = 5.99$; 25 percent and 1 percent of the sampling distribution of $F(4, 10)$ fall to the right of these respective points.

Obviously, not all possible combinations of these three factors are listed in the table. The α levels, $\alpha = .25, .10, .05, .025, .01$, and .001, are ones most commonly encountered. Additional levels of α can be found in the different editions of Fisher and Yates (e.g., 1953) or in the more convenient tables of

Dixon and Massey (1957). The intervals between successive columns and rows increase with the larger numerator and denominator df's. For instance, the $df_{num.}$ include entries for consecutive values of df from 1–10; the next columns are $df_{num.} = 12, 15, 20, 24, 30, 40, 60$, and ∞. The $df_{denom.}$ increase consecutively from 3–20, by two's from 22–30, and then, $df_{denom.} = 40, 60, 120$, and ∞. Fine gradations are not needed for the larger df values, since the numerical values of F do not change greatly from interval to interval.

Sampling Distribution of F'

We have only considered the sampling distribution of F when the null hypothesis is *true*—i.e., when the population means are equal. Obviously, we do not intend to conduct many experiments in which this is the case! We perform an experiment because we expect to find treatment effects. Suppose we assume that H_0 is false. What should happen to the F ratio? From previous discussions, we have argued that the expected value of the ratio should be greater than 1.0. This was because the MS_A contains two components, treatment effects and experimental error, while the $MS_{S/A}$ is the result of experimental error alone. The sampling distribution of the F ratio under these circumstances is no longer the F distribution. Instead, the theoretical distribution is called F' or *noncentral* F. It would be nice to be able to draw the F' distribution and to compare it with a corresponding F distribution. Unfortunately, however, this is difficult to do, since the F' distribution is a function of the *magnitude* of the treatment effects as well as of the numerator and denominator df's. Thus, while there is only one F distribution at any combination of numerator and denominator df's, there is a family of F' distributions, one distribution for each value that the treatment effects may take.

One way to approach this problem is to decide upon treatment effects of a certain size and see what the F' distribution looks like then. It is convenient to turn again to an empirical determination of this distribution. We will consider two such distributions. One was obtained by sampling from populations of 6000 scores each, where the variances are equal, $\sigma^2 = 225$, and the means are different, $\mu_1 = 50$, $\mu_2 = 55$, and $\mu_3 = 60$. The sampling distribution of the different F ratios, which is based on a total of 3036 independent "experiments," is plotted in Fig. 4-3. (The frequencies have been adjusted to a base of 1000 experiments to allow a comparison with other sampling experiments.) The empirical $F(2, 42)$ distribution, which was obtained by an identical procedure except that $\mu_1 = \mu_2 = \mu_3 = \mu = 50$, is replotted for comparison purposes in the same figure. Also plotted is the sampling distribution for a third sampling experiment (unconnected dots), which summarizes the results of drawing 1002 experiments from populations where $\mu_1 = 50$, $\mu_2 = 60$, and $\mu_3 = 70$ and the variances are equal, $\sigma^2 = 225$.

We will compare the sampling distributions labeled "H_0 True" (the F distribution) with the one labeled "H_0 False" (the F' distribution, where the

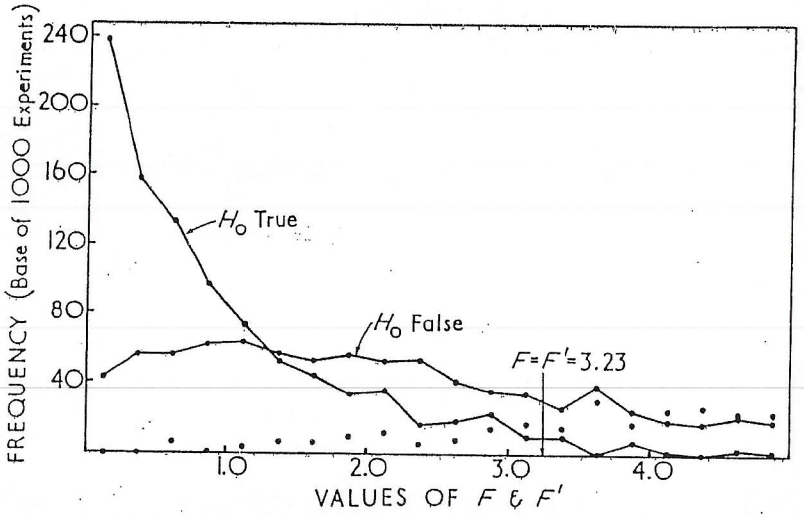

Fig. 4-3 Sampling distributions of F and F' obtained from Monte Carlo experiments. For "H_0 True," $\mu_i = 50$ for all three conditions; for "H_0 False," $\mu_i = 50$, 55, and 60; and for the curve depicted by unconnected dots, $\mu_i = 50$, 60, and 70.

population means differ in 5-unit steps). Clearly, the shapes of the two distributions are different. More important, however, is the fact that the F' distribution is shifted to the *right* of the F distribution, centering over numerically larger values of F'. This upward shift is reflected in the means of the sampling distributions. The average value of F is 1.09.[2] In contrast, the average value of F' is 2.90. Looked at another way, the area of the curve to the right of an ordinate drawn from any positive value of F (or F') will always be *greater* for the F' distribution. For example, consider the ordinate drawn at $F = F' = 3.23$: the proportion of the curve to the right of the ordinate is .050 for the F distribution and .333 for the F' distribution. At another point, an ordinate drawn at a value of 2.44 results in proportions of .114 and .463 for the F and F' curves, respectively. [For the other F' distribution, where the means differ in 10-unit steps (the unconnected dots), the two corresponding proportions are .906 and .944.]

Test of the Null Hypothesis

We are now ready to piece together this information concerning the sampling distributions of F and F' to provide a test of the null hypothesis. We start our

[2] The mean of the theoretical distribution is given by

$$E(F) = \bar{F} = \frac{df_{denom.}}{df_{denom.} - 2}.$$

testing procedure by specifying H_0 and H_1, the null and alternative statistical hypotheses. To review briefly an earlier discussion, the two hypotheses are

$$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_a$$

and

$$H_1\colon \text{ all } \mu_i\text{'s not equal.}$$

The hypothesis we will test, the null hypothesis, assumes that the means of the treatment populations are equal. The alternative hypothesis is a mutually exclusive statement which generally asserts simply that the means of the treatment populations are not equal—i.e., that some treatment effects are present. We choose this particular null hypothesis because it is usually the only hypothesis which we can state *exactly*. There is no ambiguity in the assertion that the population means are equal; there is only one way in which this can happen. The alternative hypothesis is an *inexact* statement—an assertion that the population means are not all equal. Nothing is said about the *actual* differences which are present in the population. (If we had that sort of information, we would have no reason for conducting the experiment!) Another advantage of this particular null hypothesis is that the sampling distribution of F is known. Presumably the sampling distribution of F' can be worked out, but we will need a different distribution for treatment effects of different sizes. Just how we use the sampling distribution of F in evaluating the null hypothesis will be considered next.

Assume that we have conducted an experiment and that we have computed the value of F. What we have to decide is whether this value of F came from the F distribution or whether it came from an F' distribution. An inspection of Fig. 4-3 will indicate that we will never be certain, because *any* observed value of F might have come from *either one of the two distributions.* Logically, it could only have come from one, but which one? Since we are evaluating the null hypothesis, we will turn our attention to the F distribution. While some values of F are less likely to occur than are others, it is still possible theoretically to obtain *any* value of F in an experiment when the null hypothesis is true. From one point of view our situation is hopeless: if any value of F may have been the result of chance factors, then we can never be *certain* that the F we observe in an experiment was *not* drawn from the F distribution. Agreed. If we were to take this attitude, however, we would never be able to use the experimental method as a way of finding out about the world. That is, if we maintain that any difference among the sample means may be due to chance, there is no way that we can conclude that our experimental manipulations influenced behavior differentially. As Fisher (1951) puts it, "... an experiment would be useless of which no possible result would satisfy [us]" (p. 13). We will not take this attitude. We must be willing to make mistakes in rejecting the null hypothesis when H_0 is true; otherwise, we can never reject the null hypothesis.

Suppose we could agree upon a dividing line for any F distribution, where values of F falling above the line are considered to be unlikely and values of F falling below the line are considered to be likely. We would then see whether our observed F falls above or below this arbitrary dividing line. If the F falls above the line, we will conclude that the observed F is *incompatible* with the null hypothesis; that is, we will reject H_0 and conclude that the alternative hypothesis is true. If the F falls below the line, we will conclude that the observed F is *compatible* with the null hypothesis. Under these circumstances, then, we will not reject H_0. Following such a set of rules means that we *will* be able to conclude that our independent variable was effective, provided an F ratio is obtained which falls within the region of incompatibility. But it also means that we are willing to make a mistake by rejecting a true null hypothesis a certain proportion of the time.

DECISION RULES The crux of the problem, of course, is to find a way of objectively defining the regions of "compatibility" and "incompatibility." If the null hypothesis is true, we can determine the sampling distribution of F. Suppose we find a point on this distribution beyond which the probability of occurrence is very, very small. (The probability is represented by the proportion of the total area under the curve that appears beyond this particular point.) We will arbitrarily consider values of F falling within this region as *incompatible* with the *null hypothesis.* We must identify such a region in order to be able to reject the null hypothesis. Our decision rule, then, is to reject the null hypothesis when the observed F falls within the region of incompatibility. We do so, knowing full well that we may be making the wrong decision, which would be the case if the null hypothesis really were true.

Suppose, now, that we begin to enlarge the region of incompatibility, by moving the critical point of transition to the left—toward the larger portion of the curve—and cumulate the probabilities associated with these new portions of the curve. As we increase the size of this region, we also increase the chance of observing values from this region. Said another way, increasing the region of incompatibility results in the inclusion of F's which are becoming increasingly more *compatible* with the null hypothesis. Theoretically, an investigator may pick any cumulative probability he wants, just as long as the decision is made before the start of the experiment. In practice, however, there is fairly common agreement upon a cumulative probability of $\alpha = .05$ to define the region of incompatibility for the F distribution. This probability is called the *significance level.*

We are now in a position to state more formally the decision rules that are followed after the calculation of the F ratio.[3] If the F value falls within the region of incompatibility, the null hypothesis is rejected and the alternative hypothesis is accepted. If the F value falls within the region of compatibility, the null hypothesis is not rejected. (These two regions are often called the regions of

[3] Hays (1963, pp. 245–287) provides a detailed discussion of this decision process.