Analysis of Variance and Functional Measurement: A Practical Guide

DAVID J. WEISS

OXFORD UNIVERSITY PRESS

Analysis of Variance and Functional Measurement

This page intentionally left blank

Analysis of Variance and Functional Measurement

A Practical Guide

DAVID J. WEISS

OXFORD
UNIVERSITY PRESS
2006

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Weiss, David J.
Analysis of variance and functional measurement : a practical guide
by David J. Weiss.
p. cm.
Includes bibliographical references and indexes.
ISBN-13 978-0-19-518315-3
ISBN 0-19-518315-0
1. Analysis of variance. 2. Experimental design. I. Title.
QA279.W427 2005
519.5'38—dc22 2005000606

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

Contents

*Chapters may be omitted or reordered without loss of continuity. Chapter 13 has the least applicability for most researchers and is also the most technically challenging.

This page intentionally left blank

Analysis of Variance and Functional Measurement

This page intentionally left blank

1

Introduction

ANALYSIS OF VARIANCE. The phrase sounds ominous. The word "analysis" suggests perhaps unfortunate associations with test tubes. Variance is a somewhat formal term, one whose sound is familiar from previous adventures in the world of statistics.

But whether or not previous statistical experiences were painful, analysis of variance (ANOVA) can be learned. And if one postpones (perhaps indefinitely) the proofs and algebraic derivations, it can be learned relatively painlessly. ANOVA (I pronounce this acronym with the second syllable stressed) has an odd resemblance to driving; it is easier to do than to describe, and the skill is more readily acquired through practice than through an understanding of theory.

This presentation presumes knowledge of basic statistics. A course in which elements of probability and hypothesis-testing logic were presented should suffice. If you have had that experience but memory has faded somewhat, a review of the Terms from Introductory Statistics (see p. 247) may be helpful. Terms included in that glossary appear in boldface type when they first occur in the text. The vocabulary of ANOVA will be further emphasized by the use of SMALL CAPITALS as important terms are introduced.

I employ a classical approach to hypothesis testing, in which the researcher sets a significance level for each test prior to examining the results. The American Psychological Association does not share this perspective, preferring to ask investigators to report the significance level corresponding to the obtained statistic. Either approach is compatible with the text.

You get a maximal return for learning ANOVA. It is a most powerful and versatile technique; since the late 1940s it has been the primary statistical tool of

behavioral psychology. For controlled experiments and the causal inferences they allow, ANOVA remains the most natural approach. What you must learn, on the other hand, is relatively limited. The more complex analyses are simply generalizations of the simpler ones. Once the fundamental concept of partitioning **variance** is mastered, successively more sophisticated experimental designs can be analyzed.

In the everyday world of the practicing scientist, ANOVA is done on a computer. Accordingly, this text will send you to a computer soon after you have performed a few analyses by hand. But omission of the manual-labor phase will inhibit your developing the intuitions that are needed to identify erroneous results stemming from incorrect data entry or other, less common, computer problems. All of the analyses described herein can be performed with the Windows programs in the CALSTAT series accompanying the text. These programs operate in ordinary English. You need not speak a computer language to use them. I would encourage you to learn to write your own programs, but you need not do so to perform even quite complex ANOVAs.

In this text, I present more calculational detail and supporting details than some readers will want to give much attention to, although my view is that every word is a pearl. Material that can be skimmed, or even omitted, without serious loss of understanding is set in all-italic type.

The Model Underlying ANOVA

Anyone who gathers data notices variability. When the same object is repeatedly measured with a finely grained measuring instrument, as when I measure a child's height in millimeters, successive readings are rarely identical. If an examination of the series of measurements reveals no pattern underlying the differences in the observations, standard practice is to use the average of the measurements as an estimate of the value of the object. A sensible way to justify this averaging is to postulate that each observation is the sum of two components. One component is the "true" value of the object; I use the quotation marks to emphasize that this true value is unknowable and can only be estimated. The other component is a random component, which means it has a value that changes unpredictably from observation to observation. The employment of an averaging procedure is tantamount to assuming that on the average, the value of the random component is zero. The random element is presumed to be drawn, then, from a normal distribution with mean zero and variance σ_e^2. This random component is a convenient fiction created to explain the inexplicable inconsistencies in even the most careful measurements. A simple equation using **subscript notation** summarizes the assumption:

$$M_i = T + e_i \tag{1-1}$$

Equation 1-1 states that M_i, the ith measurement of the object, is the algebraic sum of T, the true value of the object, and e_i, the value of the "error" on the ith measurement. The term *error* is conventionally used for the random component.

The term is historically entrenched, though it is an unfortunate usage because it connotes a mistake rather than a normal aspect of the measurement process.

Equation 1-1 describes a situation that is too simple to be scientifically interesting. In the late eighteenth century, there arose a complication that should be dear to the hearts of all graduate students. The astronomy of that era required precise timing of the transit of a star across the meridian of the observatory. In 1795, the head of the Greenwich observatory fired an assistant because the assistant's times were about a half second too slow (that is, they were slower than the chief's). Somewhat later, the German astronomer Bessel read about the incident, and he began comparing astronomers. He found that even skilled, experienced astronomers consistently disagreed, sometimes by as much as a second. Bessel at first thought these interpersonal differences were constant, and he presented a "personal equation" that could be cast in our notation as equation 1-2:

$$M_{ij} = T + P_j + e_i \tag{1-2}$$

Here P_j is the personal contribution of observer j. It soon became clear that there were also differences that depended on such complicating physical factors as the size of the star and its rate of movement, so a more complex equation was needed:

$$M_{ijk} = T_k + P_j + e_i \tag{1-3}$$

Equation 1-3 has three subscripts, because it is the ith measurement of the kth object by the jth observer. The measurement now is held to depend upon the true value of the kth object (T_k), the contribution of the particular observer (P_j), and the random component (e_i).

Equation 1-3 is sufficiently complex to deal with behavioral experiments of substantive interest. Suppose, for example, one were studying how far people can throw various projectiles. There might be five different projectiles and ten throwers; fifty scores would be generated. Equation 1-3 would provide a model for the distance traversed by each projectile as thrown by each hurler.

Equation 1-3 is an abstract statement of the process underlying a set of data to be analyzed. While the statistical procedure does not make use of the model in a direct way, the model clarifies the goal of the analysis. The aim is to tie variation in the measurements to particular manipulations in the experiment. Specific terms in the model may be replaced or omitted; for example, if each thrower tossed only one projectile, there would be no way to isolate the effect on the measurements of the individual's strength. In that case, P_j would not appear in the model for the experiment. On the other hand, additional experimental complications would call for incorporating more terms into the equation. The throwers might be offered a systematically varied monetary incentive ($0 per meter, $1 per meter, $10 per meter). This experimental manipulation would require a term ($\$_l$) to specify the effect of the value of the incentive on each trial. Equation 1-4 incorporates the incentive effect:

$$M_{ijkl} = T_k + P_i + \$_l + e_i \tag{1-4}$$

Additional complexity in the model reflects potential difficulties in interpreting the experimental results. Effects are not always simple. Suppose, for example, that people try harder for $10 per meter than for $1 per meter. Accordingly, the distances would be expected to be greater for the larger reward. But perhaps the largest projectile is so heavy that for most people it can't be thrown no matter how hard one tries; it simply drops to the ground. In that case, the expected effect of the incentive would be different for one projectile than for others. This is an example of an interaction. Interaction between two variables means that the effect of one variable depends on which value of the other variable is present. Formally, interaction terms in the model represent the effects of specific combinations of the model's components. The hypothesized interaction appears in equation 1-5:

$$M_{ijkl} = T_k + P_i + \$_1 + T_k\$_1 + e_i \tag{1-5}$$

Equations 1-1 through 1-5 are all instances of linear models, so named because they express the response as a linear combination of contributing elements. Other statistical procedures such as correlation and regression also employ linear models. There is a formal equivalence among the various procedures for analyzing linear models; this equivalence is conveyed by use of the term "general linear model" to refer to the family of procedures. One can, in fact, analyze the problems in this text with multiple regression (Cohen, 1968); experimental factors and their interactions are regarded as predictors whose contributions can be assessed just as one usually evaluates the impact of measured, noncontrolled variables. The ANOVA framework, though, is the natural one for working with designed experiments. Not only are the computations much simpler and easier to fathom but the elements included in the model correspond directly to those built into the experiment. With analysis of variance, one jointly plans the experiment and the analysis, which is, in my view, the path to fruitful research.

Use of the Model

A model equation is simply an algebraic representation of an experimental hypothesis. The researcher constructs the model as a guide; it points the way to the appropriate statistical tests. Each term in the equation corresponds to a particular statistical test; each term is a component in the ANOVA. The researcher does not know the correct model before the data have been analyzed. Typically, one begins by postulating a complex model, one with a term for each independent variable and with terms corresponding to all of the possible interactions among them. When the analysis reveals that some components make only negligible contributions to the variation in the scores, the corresponding terms are dropped from the model. The reduced model is offered as a descriptive statement about the experimental results.

In practice, researchers are seldom explicit about their use of these model equations. ANOVA procedures are routinized to the extent that one need not think

about which components ought to be tested. Rather, the researcher identifies the proper analysis to be conducted on the basis of the experimental design. The model guides the tests, but it does so implicitly by providing a logical connection between the experimental design and the proper analysis. The linkage is implicit because it is common practice to learn the relationship between design and analysis without using model equations. We shall follow this practice since the analytical algorithms are, as a practical matter, independent of their theoretical underpinnings. A model may be used to summarize an investigation, but it is not required to carry out an appropriate data analysis. One merely tests for the factors built into the experiment along with the interactions among them. Standard significance test procedures tell us whether the factors have had their anticipated effects.

2

One-Way ANOVA

One-way ANOVA deals with the results of a straightforward experimental manipulation. There are several (two or more) groups of scores, with each **group** having been subjected to a different experimental treatment. The term one-way derives from the fact that the treatment for each group differs systematically from that for other groups in only one respect: that is, there is one **independent variable**. Within each group, the treatment should be identical for all members. Each score comes from a separate individual, or, stated otherwise, each individual contributes only one score.

The traditional name for an individual furnishing a score is SUBJECT. In recent years, the more egalitarian term PARTICIPANT has come to be favored. The modern term connotes a voluntary contribution to the research, a partnership between investigator and investigatee. VOLUNTEER is another label used for this usually anonymous member of the research team. All of these terms will be employed in the text.

The score is the quantified observation of the behavior under study. A score must be a numerical value, an amount of something. For the analysis to be sensible, the score should directly reflect the behavior in question; the greater the number, the more (or less, since it is the consistency rather than the direction of the relation that is important) of the particular behavioral tendency. An individual score must be free to take on any value in the defined range, controlled only by the experimental conditions governing that score. Linked measures (for example, sets of numbers that must sum to a particular value such as 100) do not qualify for ANOVA.

The **null hypothesis** is that the true values of the group means are equal. The simplest way to express the **alternative hypothesis** is to say that the null hypothesis is false. More definitively, at least one of the group means is different from at least one other (note the difference between the latter expression and the incorrect phrasing that the group means are all different).

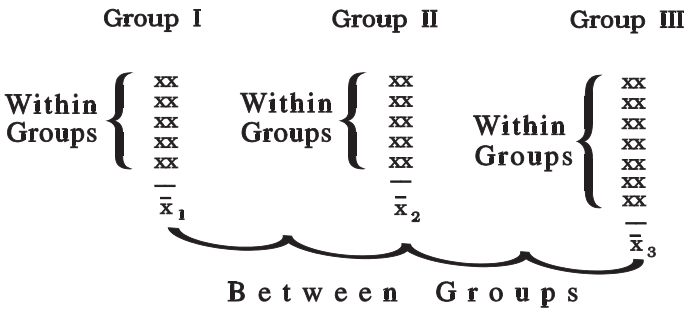
A simple example would be a drug-dosage study in which the scores might be running times in a maze. We shall have three groups, with the members of each group receiving a particular dosage of a specified drug. Usually a researcher tries to have the same number of subjects in all of the groups, in order to estimate each group mean with equal precision. But things don't always work out as planned in experimental work; and in a one-way design, inequality presents no difficulties. For the sake of generality, then, our example will feature unequal group sizes.

The first group might consist of five animals, each of whom is given a 1-mg dose ten minutes before running. The second group might also have five animals, each of whom is given a 3-mg dose of the drug ten minutes before running. The seven animals in the third group might each get a 5-mg dose. The average running time for each group gives an idea of the effects of drug dosage.

If the scores in each group were completely distinct from those in the other groups, no further analysis would be necessary. More realistically, however, one would expect overlap among the scores. Some animals in the low-dosage group will inevitably run faster than some in the high-dosage group, even though the group means might suggest that, in general, higher doses lead to faster running. In order to answer the question of whether the group means are reliably different, one must carry out a statistical analysis in which the variability in the scores is taken into account.

The test, of course, is ANOVA. The variability in the scores is partitioned into two classes, systematic variance and error variance. Systematic variance is variability attributed to controlled elements in the experimental setting; in our example, the dosage of the drug was controlled. The primary systematic variance is that BETWEEN GROUPS. It measures how different each group mean is from the overall mean. If all of the group means were similar, they would as well be similar to the overall mean. Consequently the between-groups variance would be small.

ERROR variance is variation that the experiment does not aim to understand. This variation reflects idiosyncrasies participants bring with them to the experiment. People can be expected to respond differently because they have different histories and capabilities. Error variance is estimated from the average variance within groups of participants treated the same way. Since the participants within a group have been administered the same treatment, variation among their scores provides a measure of the magnitude of the idiosyncratic contribution. The error variance is, then, a composite determined from the variance within each group weighted by the number of scores per group. For this reason, error variance is also referred to as WITHIN-GROUPS variance. The variance diagram illustrates this.



VARIANCE DIAGRAM

In terms of the model given in chapter 1 (equation 1-3), the between-groups variance includes the contributions of both the **substantive variable** (T_k) and the random error component (e_i). The within-groups term, on the other hand, contains only the error component. In this experimental design, the personal contribution of each participant (if a rat may be said to make a personal contribution) is **confounded** with the error. Because the individual makes only one response, it is not possible to identify the separate contributions of the personal component and the error component. So the error component in this design includes individual differences; both the between-groups variance and the within-groups variance include this masked contribution.

To the extent that the substantive variable, in this example the drug dosage, has a big effect, then the between-groups variance will be large relative to the within-groups variance. In contrast, suppose that the substantive variable had no effect, in other words, that the running times were not differentially affected by how much of the drug the animal received. In that case, the between-groups and within-groups variances would both simply be measuring unsystematic variation. One would expect the two variances to be comparable in magnitude. They will not be identical, of course, because different scores are being processed in the computations, and thus different instances of the random component are involved.

There are two plausible ways to compare quantities. One can examine the difference between them, which ought to be close to zero if the quantities are the same except for random error. Alternatively, one can examine the ratio, which ought to be close to one if the quantities are the same except for random error. Which method is preferable? The ratio conveys more information. An analogy may clarify this argument. Suppose I have been trying the new Pasadena diet, and I proudly report that I lost 10 pounds. Is that a sufficiently impressive reduction for you to consider recommending the diet to a friend who wants to lose weight? In order to make that judgment, you might want to know the weight I started from. If I originally weighed 360 pounds, the reduction would hardly be noticeable, but

if I had originally weighed 180 pounds, the difference in my weight would be more impressive. The diet's effectiveness can be conveyed compactly by reporting the percentage of my original weight that I lost rather than the amount. With the percentage information, you don't need to know how much I originally weighed in order to evaluate the diet. Comparing two numbers via a ratio is akin to using a percentage. The ratio expresses the magnitude of the one number (the numerator) in units that are the magnitude of other number (the denominator) so that the actual values of the numbers being compared are not required to appreciate their relationship.

Following this reasoning, statisticians routinely express comparisons as ratios in the procedures they develop. The *F* test, named in honor of R. A. Fisher, the British agronomist and statistician who pioneered ANOVA, examines the ratio of the between-groups variance to the within-groups variance. If the between-groups variance is large compared to the within-groups variance, then this ratio, called the *F* ratio, will be large; this would be evidence supporting the idea that drug dosage makes a difference. On the other hand, if the treatment had no effect, then the *F* ratio would be expected to be approximately one; that is, the between-groups variance should be of about the same size as the within-groups variance.

Chance fluctuations that in terms of the model (equation 1-3) are randomly varying values of the error component e_i may affect either of the two critical variances and thus affect the *F* ratio. Therefore, the **probability distribution** of the *F* ratio, under the assumption that the two involved variances are in truth of the same magnitude, has been worked out. This distribution furnishes the entries in the *F* table. The tabled value is employed as a **critical value**, or criterion. If the *F* ratio obtained from the data is larger than the tabled value, then the between-groups variance is deemed large relative to the yardstick of the within-groups variance. In other words, a large, or significant, *F* ratio is evidence that the group means are not all the same. An obtained *F* ratio not exceeding the critical value suggests that the group means are not reliably different. Alternatively and equivalently, if the ***p* value** associated with the obtained *F* ratio is less than the designated significance level, the difference between the group means is deemed significant. Because the fundamental quantities leading to the ratio are variances, all of which must be positive, directions of differences between means are not preserved. Therefore, all tests employ only the upper tail of the *F* distribution. *F* tests are treated as **one-tailed** even though the alternative hypothesis is nondirectional.

Randomization and Independence

Suppose a woman receives the dreaded news that she has breast cancer and asks you for advice about where to seek treatment. One element in the response might be an evaluation of the survival duration for patients who have gone to various hospitals. If this information were available, significant differences might well have life-or-death implications. It would seem natural to avoid a facility whose patients did not live a long time after treatment. Unfortunately, this natural conclusion

might be the wrong one, and your advice might well prove fatal. You have made the assumption that all patients are equivalent. Suppose, for example, the suspect hospital was known among local physicians as the best, and accordingly physicians directed their most seriously ill patients there. The statistical evaluation is useless because we don't know whether the patients in the different institutions are comparable. Experimental control, as this realistic example vividly demonstrates, is no mere technical nicety.

An experimental comparison depends upon the idea that consistent differences between scores from participants in various groups are not the result of preexisting differences. ANOVA can tell us whether group means are reliably different but not whether the differences were caused by the experimental treatment. In order to make the inference that group differences are linked to treatment effects, the researcher must see that prior to treatment the groups are comparable in whatever respects are crucial. The easiest and best way to achieve this goal is to randomly assign participants to groups. Every subject should have the same probability of being assigned to any of the experimental groups. While randomization cannot guarantee that the groups are indeed equivalent prior to treatment, it does insure against bias, that is, stacking the deck in a particular direction.

Sometimes practical constraints prohibit a random assignment. This situation occurs when the variable is classificatory rather than experimental. For example, if gender or age is the variable of interest, the assignment of participant to group can hardly be determined randomly. In such a situation, the problem is that one cannot say with confidence that observed between-group differences are related to the variable of interest. There may be a hidden variable, such as weight or height or years of education, that is truly responsible for the experimental effect. The classificatory variable, sometimes called a *subject variable*, may be naturally confounded with a hidden variable that, although logically distinct, is associated with the classification. Random assignment minimizes the chance that such a concomitant variable will confuse the researcher.

It is worth noting that this problem is not related to ANOVA but to the design of the experiment. The statistical analysis is neutral. It is designed to tell you whether the average scores in the experimental groups are reliably different. The issue of what the scores mean or of whether the numbers are meaningful at all is not in the domain of statistics but of experimental logic.

Experimental logic also demands that the observations be independent of one another. This means primarily that the researcher must collect each score in the same way. In a practical sense, of course, it is not possible for an experimenter to be equally tired or for the apparatus to be in the same condition for all observations. Once again, randomization comes to the rescue. By interweaving the subjects from the various groups according to a random sequence, the researcher avoids biasing the results.

How does one achieve randomization? Suppose it is desired to assign five subjects to each of four experimental groups. As volunteers report in from the introductory class pool, each one is placed in a particular group according to a predetermined scheme. Since there are four groups, regard the subjects as coming in

sets of four. For each set, shuffle the four names, or more conveniently the index numbers 1, 2, 3, and 4, in a hat (hats are traditional for shuffling, though no one I know owns a hat these days, so you may have to improvise). The first one drawn goes into group 1, the second into group 2, and so on. Repeat this shuffling process five times, in each case using a separate shuffle. Alternatively, use a computer program to generate random permutations to accomplish the shuffling. A major advantage of the permutation scheme, as opposed to independent randomization as each subject comes along, is that equal group sizes are automatically attained as each permutation is implemented.

The *F* Table

The *F* distribution is actually a family of distributions. There are two parameters that serve to distinguish the members of the family. Each *F* distribution is characterized by two **degrees of freedom** (*df*) values. The phrase "degrees of freedom" has little explanatory or mnemonic value, but it is unfortunately embedded in the literature. The DEGREES OF FREEDOM FOR NUMERATOR heading and the DEGREES OF FREEDOM FOR DENOMINATOR headings guide the table user to the proper critical values, as do the coordinates on a map. The values of these parameters are determined by the structure of the experiment. The degrees of freedom for numerator are one less than the number of groups (for our drug experiment, $df_{\text{num}} = 2$). The degrees of freedom for denominator are computed by subtracting the number of groups from the total number of scores (for our drug example, $df_{\text{denom}} = 14$).

A researcher arbitrarily chooses a **significance level**, or in other words, determines the probability that the true group means will be declared different when they are in fact the same. This misjudgment is called a **Type I error**. This significance level is usually set conventionally at either .05 or .01, though there is no logical necessity governing the choice. Throughout this text, the .05 level is presumed to have been selected. The significance level also determines the critical value of *F*. Examination of the *F* table appendix A reveals that the more strict the criterion (that is, the lower the significance level chosen), the larger the obtained *F* ratio must be in order to exceed the tabled critical value. This means that the choice of significance level plays a role in determining how large an observed difference among group means will be required before those group means are pronounced different. For the drug experiment with $df = 2, 14$, the table shows the critical value for *F* for the .05 level of significance to be 3.74, while that for the .01 level is 6.51.

Power

The significance level also affects the probability of a **Type II error**, that is, failing to confirm a true difference among the group means. The capability of detecting a difference is known as the **power** of the statistical test, and obviously it is

desirable for a test to be powerful. The less stringent the significance level (that is, the larger the value of α), the more powerful the test is because a smaller F ratio is required in order to attain significance. But manipulating the significance level to gain power is a dangerous game because there may in fact be no true difference among means, and a higher significance level increases the risk of a Type I error.

Fortunately, power may also be increased by means that do not affect the Type I error rate. The most fruitful ways under the control of the researcher involve reducing the within-groups variability, thus producing a smaller denominator for the F ratio. Possibilities include choosing participants to be homogeneous, specifying experimental instructions and procedures carefully so that all observations are generated under the same conditions, and eliminating or controlling (via the more complex experimental designs to be encountered in later chapters) extraneous variables such as time of day or temperature of the room.

Power can also be increased by increasing the number of participants because the corresponding increase in the denominator *df* means that a smaller F ratio is needed to reach statistical significance. However, in addition to the economic burden, there is another drawback to solving the power problem by running hordes of subjects. This drawback can be illuminated by considering the extreme case. Suppose there is a very small but true difference among the group means. The greater the *df* for the denominator, the more likely this difference is to be significant. With a very large number of subjects, even a tiny difference may prove significant. But few researchers want to discover or confirm experimental manipulations that produce small differences. Rather they want to demonstrate potent treatment effects. Because ANOVA addresses the question of reliability rather than magnitude of effect, a significant F ratio does not necessarily imply an impressive result. Most researchers are wary of studies that employ huge numbers of participants to obtain small but significant F ratios for fear that the effects are small and therefore perhaps not worthy of attention. So an experimenter should run enough, but not too many, subjects. This can be a tricky problem. More will be said on this matter in chapter 10, where we discuss strength of effect in detail.

Computation

Although it is feasible to determine the between-groups and within-groups variances with formulas based on the usual definition of a variance (the only complication is that when group sizes are unequal, each group's contribution must be weighted proportionately to its size), a much more convenient computational scheme is available. With this algorithm comes a standard format for presenting the results and some new terminology.

ANOVAs are customarily presented in a table, and the table has conventional headings for its columns. The term SOURCE is used to refer to the particular source of variation for each row in the table; at this point, our sources will be between groups and within groups. Later, as we encounter more complex designs in

which variability is further partitioned, there will be more sources and thus more rows. Each source's *df* are given, and then we come to the crucial sum of squares (SS) and mean square (MS) columns. The mean square for a source is the (weighted) average variance for that source, and it is the mean squares that are used to construct the *F* ratio that appears in the final column. The sum of squares for each source is the quantity computed directly from the data. From each SS the corresponding MS is derived. Sums of squares merit an entry in the table, though, because they are not merely an intermediate calculation on the way to the mean squares. The theoretical importance of the SSs is that they are additive. This additive property means that sums of squares for the various sources add up to the total sum of squares for the whole experiment. Similarly, when we do the further partitioning called for by more complex designs, it is the SS that is partitioned. Mean squares, on the other hand, are not additive.

An Example of One-Way ANOVA

The scores in the table represent running times from three groups of animals in a drug-dosage study. Placing the data in tabular format is a worthwhile step in the analytic procedure. Neatness doesn't exactly count, but it is useful to avoid getting lost in a maze of numbers.

Running Time Scores, in Seconds

Group 1	Group 2	Group 3	
18	16	10	
23	16	17	
14	11	8	
16	18	12	
23	14	14	
		7	
		11	
$t_1 = 94$	$t_2 = 75$	$t_3 = 79$	$T = 248$

There are three numbers to calculate; I refer to them imaginatively as (1), (2), and (3).

(1) ΣX^2: Each of the scores is squared, and these squares are then summed. This is usually the most tedious step in any ANOVA, although it is not too unpleasant if your calculator has an "M+" key. Simply punch the "×", "=", and "M+" keys in sequence after you enter each of the scores, and the ΣX^2 should appear when you press the memory recall key. This convenience does not add much to the price of a calculator, and it should swing your buying decision.

(2) T^2/N: T is the grand total, the sum of all of the scores. (T should be called ΣX for consistency's sake, but the label is traditional.) N is the number of scores that went into T, that is, the total number of scores in the experiment.
(3) $\Sigma(t_j^2/n_j)$: t_j is the total for the jth group, and n_j is the number of scores in that jth group. Compute each t_j by summing the scores in each group separately. Each group total is squared and then divided by the number of numbers that contributed to the total. The results of these divisions are then summed.

The defined quantities are calculated from the data:

(1) ΣX^2: $18^2 + 23^2 + 14^2 + \cdots + 14^2 + 7^2 + 11^2 = 3{,}950$
(2) T^2/N: $248^2/17 = 3{,}617.88$
(3) $\Sigma(t_j^2/n_j)$: $(94^2/5) + (75^2/5) + (79^2/7) = 3{,}783.77$

Next, the calculated quantities are used to generate the numbers in the SS column of the ANOVA table. As SSs are literally *sums* of *squares*, they are necessarily positive. It is inevitable that (1) should be the largest calculated quantity and (2) should be the smallest. If an arithmetic error causes a violation of this ordering, a negative SS will result. The good news is that at least that error will be spotted (for me, it's quantity [1] on which my calculator is most likely to fail).

ANOVA Table

Source	*df*	SS	MS	*F*
Between groups	2	(3) − (2) = 165.89	$\frac{SS_{bg}}{df_{bg}} = \frac{165.89}{2} = 82.95$	$\frac{MS_{bg}}{MS_{wg}} = \frac{82.95}{11.87} = 6.99^*$
Within groups	14	(1) − (3) = 166.23	$\frac{SS_{wg}}{df_{wg}} = \frac{166.23}{14} = 11.87$	

The asterisk sitting proudly beside the F ratio denotes significance at the researcher's chosen level. In an actual table, of course, only the numerical values appear, not the formulas or intermediate calculations.

Numerical Details

Numerical accuracy is certainly a goal worth striving for. In presenting results for public consumption, though, one cannot expect ten decimal places to be tolerated. One must round the numbers to be presented in the ANOVA table. It is customary to report sums of squares and mean squares to one or two decimal places and F values to two places. Maximal accuracy is achieved by maintaining as many decimal places as your calculator will hold until the computations are

complete; only then should rounding be done. Consequently, the reported *F* ratio occasionally will have a slightly surrealistic quality in that the ratio of the reported mean squares does not precisely yield the *F* given in the table. It is more important to provide a correct statistic than to appear consistent. One should avoid rounding at intermediate stages.

The Responsiveness of ANOVA

A further example will serve to clarify the way the statistical test is sensitive to the data. The table shows three groups of numbers drawn from a random number table.

Scores from Random Number Table

Group 1	Group 2	Group 3
4	13	14
11	7	2
13	6	5
16	16	4
1	4	6
2	12	10

The ANOVA computations yield an *F* of 0.46. Since *F* is less than one, no table is necessary to verify the nonsignificance of the between-groups effect; in fact, *F* ratios of less than one are customarily reported simply as "$F < 1$." This result is hardly surprising, considering the origin of the scores.

Now modify the scores by adding 10 to each score in group 1 and 5 to each score in group 2. Recompute the ANOVA. This time the *F* ratio leaps to a value of 7.13*. This statistically significant *F* ratio reflects your having imposed a between-groups effect onto the data. Notice that MS_{wg} (27.0) for the modified scores is the same as for the original scores; this reflects the fact that the variance of the scores within each group did not change.

Next, return to the original scores and apply a different modification by adding 10 to the last score in each group. Once again, recompute the ANOVA. Now the *F* ratio (0.29) is even smaller than that for the original scores. The reason is that this second modification has increased the within-groups variance, but it has not increased the between-groups variance since the group totals are as different from one another as they were before.

A Test of the Grand Mean

Summing the degrees of freedom in the ANOVA table yields $N - 1$, one less than the number of scores. Since the rule for generating degrees of freedom is that

each score produces one degree, there must be an element missing from the table. The missing link is a seldom-tested source that compares the overall mean response to zero. The sum of squares for this source is T^2/N, which we know as (2), and since SS_{mean} has only one degree of freedom, the corresponding mean square is also given by T^2/N. This mean square may be compared to the mean square within, with an F ratio being formed and tested for significance in the standard way. If the F ratio proves significant, it is interpreted as evidence against the null hypothesis that the grand mean of the scores is zero.

It should be clear why this source is usually omitted from the ANOVA table. The value of the average response is rarely of interest to the researcher; what is of concern is whether the various treatments have had differential effects. The only practical situation in which the test for the mean is likely to be useful is when the data are comprised of difference or change scores. Consider a project in which two different programs for weight loss are evaluated. Primary interest would surely be in whether one program produces more weight loss than the other; with individual losses as the scores, the ordinary between-groups F ratio is addressed to that question. But also of interest might be the question of whether the programs are effective at all. If the average weight loss is not reliably different from zero, then the reduction programs must be regarded as failures. The appropriate test of this question employs the F for the mean.

One may also test the null hypothesis that the grand mean is equal to some other predetermined constant, K, rather than zero. In this case, the numerator of the F ratio is modified to incorporate the constant:

$$SS_{mean} = N \cdot (T/N - K)^2$$

Notice that T/N is simply the grand mean, and so if $K = 0$ the expression reduces to T^2/N. I have never seen this null hypothesis tested in print (please don't deluge me with citations), so the derivation is probably not of great importance. Still, it's nice to know where the missing *df* goes and to appreciate its meaning.

Exercises

You will see that I use varied formats for presentation of the data. This is a deliberate maneuver to prepare you for the wonderful variety used by researchers as they scribble their scores. However, your ANOVA tables should rigidly follow the format given in the text; it is considered anarchy to try a different format for the table. Use the .05 level of significance as a default throughout the text.

2-1. I conducted an experiment on two sections of my introductory statistics class. One section (with four students) had a graduate assistant, while the other (with eight students) did not. Determine whether the assistant makes a difference. The scores are from each student's final exam.

Section 1 (with assistant): 70, 50, 60, 60

Section 2 (without assistant): 30, 20, 40, 10, 50, 30, 20, 40

2-2. The following scores represent the number of errors made by each person on a verbal learning task. Each person was assigned to one of four study groups. Test the hypothesis that the different study groups all produced the same average number of errors.

Group	Error scores
1	16, 7, 19, 24, 31
2	24, 6, 15, 25, 32, 24, 29
3	16, 15, 18, 19, 6, 13, 18
4	25, 19, 16, 17, 42, 45

2-3. Students taking Psych 205 were randomly assigned to one of three instructional conditions. The same test was given to all of the students at the end of the quarter. Test the hypothesis that there were no differences in test scores between groups.

Group	Test Scores
Lecture	10, 13, 3, 38, 11, 23, 36, 3, 61, 21, 5
Programmed instruction	8, 36, 61, 23, 36, 48, 51, 36, 48, 36
Television	36, 48, 23, 48, 61, 61, 23, 36, 61

2-4. A professor of psychiatry evaluated three of her trainee therapists by asking their patients for self-reports of their perceived growth (0 = no growth; 20 = maximal growth) during the course of therapy. Test the null hypothesis that the trainees were equally effective. Also evaluate the professor by testing the null hypothesis that on the whole the patients experienced no growth at all.

Scores from Patients of Trainees

Trainee A	Trainee B	Trainee C
2	0	3
5	2	4
0	1	2
1	0	1
3	1	

2-5. Bozo, a statistically minded clown, decided to evaluate five new routines he had created. One of his fellow clowns, Dumbo, contended that children laugh at anything done by a clown, but Bozo argued that some ideas are more hilarious than others. Bozo went to the Stoneface Elementary School and successively gathered 5 groups of 4 children. For each group, Bozo performed 3 minutes' worth of one of the routines while Dumbo counted the number of laughs emitted by each child. Whose point of view, Bozo's or Dumbo's, do these comical results support? Each score given is the number of laughs by one child in response to the routine.

Number of Laughs Emitted by Groups of Children in Response to Comedy Routines

Group	Pies in faces	Monkey imitation	Pratfalls	Snappy insults	Revolting smells
Group 1	12	7	20	4	15
Group 2	13	14	25	2	18
Group 3	11	22	17	0	23
Group 4	25	8	22	8	17

2-6. The Committee Against Violent Television charged that Network 2 was the most violent of them all; Network 2 responded that they were no worse than their competitors. The committee assigned watchers to count the number of brutalities per evening for 5 days. Evaluate the data to determine if the networks are equally culpable.

Number of Brutalities Counted on Television Networks per Evening

Network 1	Network 2	Network 3
18	42	32
32	73	28
23	68	17
16	32	43
19	47	37

Answers to Exercises

2-1.

Source	*df*	SS	MS	*F*
Between groups	1	2400.00	2400.00	17.14*
Within groups	10	1400.00	140.00	

2-2.

Source	*df*	SS	MS	*F*
Between groups	3	513.97	171.32	2.06
Within groups	21	1749.29	83.30	

2-3.

Source	*df*	SS	MS	*F*
Between groups	2	3141.93	1570.96	5.85*
Within groups	27	7249.53	268.50	

2-4.

Source	*df*	SS	MS	*F*
Between groups	2	7.76	3.88	1.89
Mean	1	44.64	44.64	21.73*
Within groups	11	22.60	2.05	

2-5.

Source	*df*	SS	MS	*F*
Between groups	4	721.30	180.33	7.21*
Within groups	15	375.25	25.02	

2-6.

Source	*df*	SS	MS	*F*
Between groups	2	2476.13	1238.07	8.41*
Within groups	12	1767.60	147.30	