CHAPTER 1

AN UP-DATED VIEWPOINT: CELL MEANS MODELS

1.1. STATISTICS AND COMPUTERS

The age of the computer is upon us—all too obviously—and it includes statisticians. Statistical computing packages available today do our arithmetic for us in a way that was totally unthinkable thirty years ago. The capacity of today's computers for voluminous arithmetic, the great speed with which it is accomplished, and the low operating cost per unit of arithmetic—these characteristics are such as were totally unimaginable to most statisticians in the late 1950s. Solving equations for a 40-variable regression analysis could take six working weeks, using (electric) mechanical desk calculators. No wonder that regression analyses then seldom involved many variables. Today that arithmetic takes no more than ten seconds, an 86,400-fold reduction in time.

Statisticians' early uses of stored-program computers often involved calculating the analysis of variance of repetitive experiments such as, for example, those regularly used by plant breeders in their annual field experiments, e.g., split plots, latin squares and balanced incomplete blocks. Computing analyses of variance of data from these experiments was at first done from what would nowadays be considered a simple stand-alone computer program. Nevertheless, in those early days of using computers it was considered almost spectacular that once a program for, say, split plots was available it could in five minutes produce the sums of squares for such an experiment, and five minutes later those for another experiment, and so on, for a long line of experiments; each five minutes of computer time was replacing possibly three to four hours of desk calculator work. From this beginning, programmers soon found that computer code designed for

calculating sums of squares for randomized complete blocks experiments could, by small adaptations, be made into a program for data from 3-factor experiments; and more adaptations produced one for 4-factor experiments. Next came the realization that programs could be written for any k-factor factorial experiment where k was limited, by the then current machine capacity, to some number such as 8 or 10, say. Modest though that limit seems today, it was far from modest in terms of the alternative desk calculator time required for the same arithmetic. Just try using one of today's pocket computers solely for accumulating a single sum of squares, without using any stored-program facilities whatever: with that as one's only computing aid, calculate the sums of squares of a 6-factor factorial experiment with all interactions; and remember, all arithmetic has to be checked by doing it twice. As a contrast to that effort, even the earliest of stored-program computers were soon recognized by statisticians as being enormously useful.

Program adaptations followed one another in a long succession, paralleling the development of increasingly larger, faster and cheaper (to use, per unit of arithmetic) computers. Programs went from calculating just sums of squares for equal-subclass-numbers data to also doing it for unequal-subclass-numbers data; extensions were incorporated for using covariables, for data editing and description (e.g., means, ranges, percentiles), for calculating contrasts and their confidence intervals, and for calculating mean squares, F-ratios and, ultimately, even their P-values; and on and on. Thus were born through some thirty years gestation, the extensive computing packages that we have today, with their numerous and large capacity routines. They provide us, for example, with the ability (if we so wish it) to relatively easily calculate the analysis of variance for a 20-factor factorial with all interactions, using 100,000 observations on 15 variables, with 10 covariables. But the all-important question would then be: Does such an analysis make sense?

Thinking about such a question is essential to sane usage of statistical computing packages. Indeed, a more fundamental question prior to doing an intended analysis is "Is it sensible to do this analysis?". Consider how the environment in which we contemplate this question has changed as a result of the existence of today's packages. Prior to having high-speed computing, the six weeks that it took for solving the least squares equations for a 40-variable regression analysis had a very salutory effect on planning the analysis. One did not embark on such a task lightly; much forethought would first be given as to whether such voluminous arithmetic would likely be worthwhile or not. Questions about which variables to use would be argued at length: are all forty necessary, or could fewer suffice, and if so, which ones? Thought-provoking questions of this nature were not lightly dismissed. Once the six-week task were to be settled on and begun, there

would be no going back; at least not without totally wasting effort up to that point. Inconceivable was any notion of "try these 40 variables, and then a different set of maybe 10, 15 or 20 variables". Yet this is an attitude that can be taken today, because computing facilities (machines and programs) enable the arithmetic to be done in minutes, not weeks, and at very small cost compared to six weeks of human labor. Further, and this is the flash-point for embarking on thoughtless analyses, these computing facilities can be initiated with barely a thought either for the subject-matter of the data being analyzed or for that all-important question "Is this a sensible analysis?"

This is the danger to the discipline and profession of statistics, having such easily accessible, cheap and fast computing facilities as are available nowadays. The easy access and low cost mean that very minimal (maybe zero) statistical knowledge is needed for getting what can be voluminous and sophisticated arithmetic easily accomplished. But that same minimal knowledge may be woefully inadequate for understanding the computer output, for knowing what it means and how to use it. Nowhere is this more true than with linear model analyses. Several large computing packages will now easily carry out extensive linear model calculations on large, very large, data sets—including those with unequal numbers of observations in the subclasses (unbalanced data). Interpreting computing package output for such data simply by analogy with that from data coming from well-designed and well-executed experiments can lead to many wrong interpretations.

The purpose of this book is to provide information about linear model analyses for unbalanced data that will, hopefully, be of assistance to a wide group of readers: students who want just a single course in linear models as part of their general training in statistics; those who want a first course in linear models as a prelude to more advanced courses, and practitioners who want to know enough about linear models in order to easily and correctly use and understand the output from statistical computing packages.

The book puts the understanding of linear models first. Based on that understanding, readers should then be able to make correct use of computing packages. Thus the bulk of the book is on linear models—only in Chapter 12 is computer output specifically addressed. Output available from a few of the larger, more widely distributed computing packages is related to the linear model analyses discussed in the preceding chapters.

1.2. BALANCED AND UNBALANCED DATA

a. Factors, levels, effects and cells

Analysis of variance is concerned with attributing the variability that is evident in data to the various classifications by which the sources of data

can be categorized. For example, consider a clinical trial where three different tranquilizer drugs are used on both men and women, some of whom are married and some not. The resulting data could be arrayed in the tabular form indicated by Table 1.1.

TABLE 1.1. A FORMAT FOR SUMMARIZING DATA

	Marital Status					
	Married			Not Married		
	Drug			Drug		
Sex	A	B	C	A	B	C
Male						
Female						

The three classifications, sex, drug and marital status, that identify the source of each datum are called *factors*. The individual classes of a classification are the *levels* of the factor; e.g., the three different drugs are the three levels of the factor "drug"; and male and female are the two levels of the factor "sex". The subset of data occurring at the "intersection" of one level of every factor being considered is said to be in a *cell* of the data. Thus with the three factors, sex (2 levels), drug (3 levels) and marital status (2 levels), there are $2 \times 3 \times 2 = 12$ cells.

In classifying data in terms of factors and their levels, the feature of interest is the extent to which different levels of a factor affect the variable of interest. We refer to this as the *effect* of a level of a factor on that variable. Effects are of two kinds. First are *fixed effects*, which are the effects attributable to a finite set of levels of a factor that occur in the data and which are there because we are interested in them; e.g., the effect on crop yield of three levels of a factor called fertilizer could correspond to the three different fertilizer regimes used in an agricultural experiment. They would be three regimes of particular interest, the effects of which we would want to quantify from the data to be collected from the experiment. Kempthorne (1975) contains, among other things, a good discussion of fixed effects.

The second kind of effects are *random effects*. These are attributable to a (usually) infinite set of levels of a factor, of which only a random sample are deemed to occur in the data. For example, four loaves of bread are taken from each of six batches of bread baked at three different temperatures. Whereas the effects due to temperature would be considered fixed effects (presumably we are interested in the particular temperatures used), the

effects due to batches would be considered random effects because the batches chosen would be deemed to be random samples of batches from some hypothetical, infinite population of batches. Since there is definite interest in the particular baking temperatures used, the statistical concern is to estimate those temperature effects; they are fixed effects. In contrast, there is no particular interest in the individual batches, because those that occur in the data are considered as just a random sample of batches, and so batch effects are random effects. There is therefore no interest in quantifying individual batch effects—instead there is usually great interest in estimating the variance of those effects. Thus such data are considered as having two sources of random variation: batch variance and, as usual, error variance. These two variances are known as *variance components*.

Models in which the only effects are fixed effects are called *fixed effects models*, or sometimes just *fixed models*. Models that contain both fixed effects and random effects are called *mixed models*. And those having (apart from a single, general mean common to all observations) only random effects are called *random effects models* or, more simply, *random models*.

This book deals mostly with fixed effects models. Chapter 13 considers mixed models just briefly. They and random models will be treated elsewhere.

b. Balanced data

Data can be usefully characterized in several ways that depend on whether or not each cell contains the same number of observations. When these numbers are the same, the data shall be described as *balanced data*; they typically come from designed factorial experiments that have been executed as planned.

A formal, rigorous, mathematical definition of balanced data is elusive. Definitions in terms of Kronecker products of matrices are implicit in Smith and Hocking (1978), Seifert (1979), Searle and Henderson (1979), and Anderson *et al.* (1984); and an explicit definition of a very broad class of balanced data is given in Searle (1987). These definitions are beyond the scope of this book.

The analysis of balanced data, whether the models are fixed, mixed or random, and whether there are interactions or not, is relatively easy and is certainly well known. It is recorded in numerous texts on the design and analysis of experiments; since it is to unbalanced data that this book is directed, little more will be said about balanced data. The analyses of them are labeled "standard" in the left-hand part of Figure 1.1.

c. Special cases of unbalanced data

In a general sense all data that are not balanced are, quite clearly, unbalanced. Nevertheless, there are at least two special cases of that broad class of unbalanced data which need to be identified. They can then be

dispensed with because their analyses come within the purview of the standard (so-called) analyses of balanced data.

The first of these two cases is what can be called *planned unbalancedness*. This is usually when there are no observations on certain, carefully planned, combinations of levels of the factors involved in an experiment. An example of this is shown in Table 1.2. It is, in fact, a particular one-third of a 3-factor experiment (of rows, columns and treatments with three levels of each), which is known as a latin square of order 3, as shown in Table 1.3.

TABLE 1.2. AN EXAMPLE OF PLANNED UNBALANCEDNESS: A LATIN SQUARE (SEE TABLE 1.3)

	Number of Observations								
	Treatment								
	A			B			C		
	Column			Column			Column		
Row	1	2	3	1	2	3	1	2	3
1	1	0	0	0	1	0	0	0	1
2	0	1	0	0	0	1	1	0	0
3	0	0	1	1	0	0	0	1	0

TABLE 1.3. A LATIN SQUARE OF ORDER 3 (TREATMENTS A, B, C)

	Column		
Row	1	2	3
1	A	B	C
2	C	A	B
3	B	C	A

Another example of planned unbalancedness is an experiment involving 3 fertilizer treatments A, B and C, say, used on three blocks of land in which treatment pairs A and C, A and B, and B and C are used. This can be represented as a two-factor situation, each factor having three levels, with certain cells empty, as shown in Table 1.4.

TABLE 1.4. NUMBER OF OBSERVATIONS IN A BALANCED INCOMPLETE BLOCKS EXPERIMENT

Treatment	Block 1	Block 2	Block 3
A	1	1	0
B	0	1	1
C	1	0	1

This is a simple example of what is known as a balanced incomplete blocks experiment. In a manner more general than either of the two preceding examples, planned unbalancedness need not require that a planned subset of cells be empty; it could be that subsets of cells are just used unequally; e.g., Table 1.4 with every 0 and 1 being a 1 and 2, respectively, would still represent planned unbalancedness.

Analyses of variance of data exhibiting planned unbalancedness of the nature just illustrated have well-known and relatively straightforward analyses that are often to be found in the same places as those describing the analysis of variance of balanced data. We therefore consider these analyses to be "standard" as indicated in Figure 1.1—and consider them no further.

The second special case of unbalanced data is when the number of observations in every cell is the same, except that in a very few cells (one, two or three, say) the number of observations is just one or two less than all the other cells. This usually occurs when a very few intended observations have inadvertently been lost or gone missing somehow, possibly due to misadventure during the course of an experiment. Maybe in a laboratory experiment, equipment got broken or animals died; or in an agricultural experiment, farm animals broke fences and ate some experimental plots. Under these circumstances there are well-known techniques for *estimating* such *missing observations* [e.g., Steel and Torrie (1980), pp. 209, 227, and 388], following which one then uses the standard analyses for balanced data as indicated in Figure 1.1. We therefore give no further consideration to the case of missing observations.

d. Unbalanced data

After defining balanced data and excluding from all other data those that can be described as exhibiting planned unbalancedness or involving just a few missing observations, we are left with what shall be called unbalanced data. This is data where the numbers of observations in the

cells (defined by one level of each factor) are not all equal, and may in fact be quite unequal. This can include some cells having no data, but, in contrast to planned unbalancedness, with those cells occurring in an unplanned manner. Survey data are often like this, where data are sometimes collected simply because they exist and so the numbers of observations in the cells are just those that are available. Records of many human activities are of this nature; e.g., yearly income for people classified by age, sex, education, education of each parent, and so on. This is the kind of data that shall be called *unbalanced data*.

Within the class of unbalanced data we make two divisions. One is for data in which all cells contain data; none are empty. We call these *all-cells-filled* data. Complementary to this are *some-cells-empty* data, wherein there are some cells that have no data. This division is vitally useful when we come to consider whether or not analyses shall pay heed to possibly interactions. This, too, is indicated in Figure 1.1.

All-cells-filled data can be analyzed using with-interaction models by means of the well-known (Yates, 1934) weighted-squares-of-means analysis. This analysis cannot be used on some-cells-empty data, although it can be described in terms of the cells means analysis that is virtually mandatory for some-cells-empty data. Both kinds of data can be analyzed on the basis of no-interaction models by using the main-effects-only analysis.

e. A summary

Figure 1.1 is a schematic summary of the four classifications of data just discussed: balanced, planned unbalanced, missing observations and unbalanced. The figure also shows, for unbalanced data, how the use of models either with or without interactions affects the preferred kinds of analyses to be undertaken, and in which sections of this book these analyses are discussed. (The broken line indicates an analysis that may not always be required.)

Two features of Figure 1.1 are important. First, the analyses suggested as useful for unbalanced data are quite separate from those shown as standard for balanced data and data exhibiting planned unbalancedness or involving just missing observations. True it is, that when analysis procedures for unbalanced data are applied to balanced data they do simplify to the standard analyses. But too often, and for too many years, have the analyses of unbalanced data been described and taught as a natural (and maybe even called a simple) extension of the standard analyses of balanced data. This is a fragile way to view the situation. It is preferable by far to think of the analysis of unbalanced data as quite separate from that of balanced data. To connect the two at the start of learning about unbalanced data is misleading and has led to a whole array of misunderstandings of unbal-

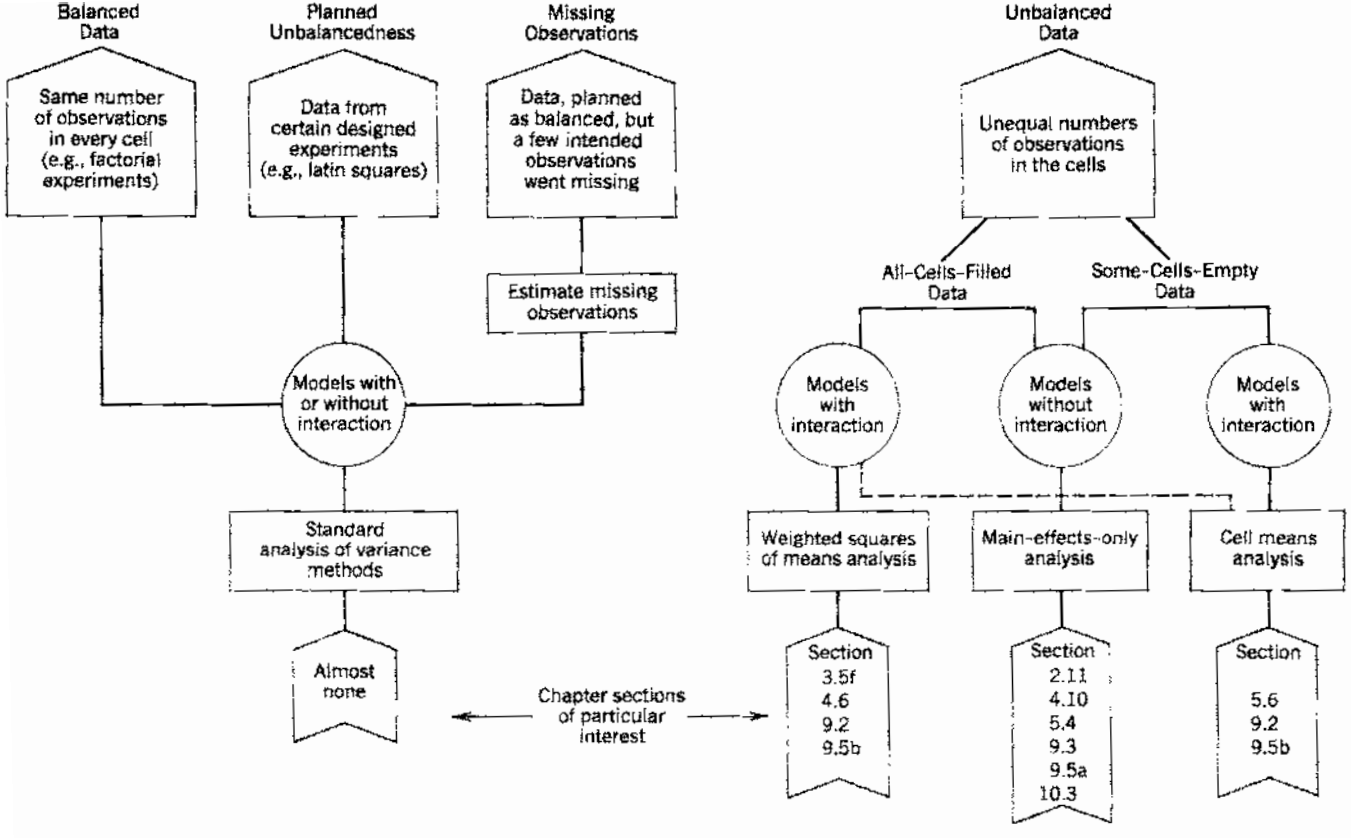

Figure 1.1. A characterization of data, models and analyses.

anced data. The useful connection comes at the end of learning about unbalanced data, for then one easily appreciates that balanced data are just a (very) special case of unbalanced data.

A particularly important feature of Figure 1.1 is that for unbalanced data the cell means model is the mainstay of most chapters of this book. To a large extent, the cell means model is the vehicle that makes the analysis of unbalanced data much more rational and easier to use and understand than does any attempt at expanding the traditional analysis procedures. Figure 1.1 shows the cell means model as a feature of analyzing with-interaction models, useful as secondary analysis (broken line) for all-cells-filled data and as primary analysis for some-cells-empty data. For the latter it is virtually mandatory for making any sense out of such data and, as shown in subsequent chapters, it is a very useful model in a variety of other ways. This, of course, is not seen in the figure, so demonstrating that any attempt to diagram all possible features of data and their connections to models and analyses simply cannot succeed. Other characteristics not appearing in Figure 1.1 are fixed and random effects; neither do the possible connections between missing-observations data and analyzing them by the methods indicated for unbalanced data. No doubt, readers will see other omissions. So be it. At least as a broad sweep, Figure 1.1 is useful for a general impression of data, models and analyses as we shall be concerned with them.

1.3. CELL MEANS MODELS

A customary practice of the last 20–30 years has been that of writing a model equation as a vehicle for describing many analysis of variance procedures. For example, for data classified by two factors that shall be called (quite generally) rows and columns, suppose there are a rows and b columns. Let y_{ijk} be the kth observation in row i and column j for $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b$, with there being n observations in every combination of row and column, so that $k = 1, 2, \ldots, n$. Without going into any great detail, which is reserved for Chapter 9, a customary model equation for y_{ijk} is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \tag{1}$$

where μ is a general mean, α_i is the effect due to the ith row, β_j is the effect due to the jth column, γ_{ij} is the effect due to the interaction of the ith row and jth column, and e_{ijk} is a random residual error term.

One advantage of (1) is that the parameters (μ, the αs, βs and γs) in that equation make it easy to be specific about what we might like to estimate.

For example, concerning row effects, we might be interested in estimating a difference such as $\alpha_1 - \alpha_2$. However, a counteracting difficulty is that (1) involves more parameters than there are observed cell means to estimate them from. There are $1 + a + b + ab$ parameters but only ab cell means

$$\bar{y}_{ij\cdot} = \sum_{k=1}^{n} y_{ijk}/n, \tag{2}$$

for $i = 1, \ldots, a$ and $j = 1, \ldots, b$. Hence there are too many parameters for us to be able to estimate them all as linear functions of the observed $\bar{y}_{ij\cdot}$ cell means defined by (2). And the row and column means and the grand mean (the mean of all the observations), respectively

$$\bar{y}_{i\cdot\cdot} = \sum_{j=1}^{b} \bar{y}_{ij\cdot}/b, \qquad \bar{y}_{\cdot j\cdot} = \sum_{i=1}^{a} \bar{y}_{ij\cdot}/a \quad \text{and} \quad \bar{y}_{\cdot\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{b} \bar{y}_{ij\cdot}/ab, \tag{3}$$

are of no help in this regard, additional to the cell means because, as is clear in (3), they are only linear combinations of those cell means.

This feature of having more parameters in a model than there are observed cell means to estimate them from is well known as *overparameterization* and leads to such a model being described as an *overparameterized model*. To circumvent this situation we usually invoke one of two procedures: either we use estimable functions (Section 8.7), which has us confine attention to only certain functions of the parameters that can be estimated satisfactorily from the data; or we use reparameterization (e.g., Sections 9.2c and 9.3c), wherein we define relationships among parameters of an overparameterized model, which has the implicit effect of rewriting the model in terms of no more than the maximum number of what may be termed "new" parameters as can be estimated from the data. Each of these procedures is easily applied and easily interpreted with balanced data. But the difficulties brought about by using overparameterized models are not always as easily circumvented for unbalanced data; estimable functions or reparameterization do not necessarily simplify the situation as they do with balanced data.

This is where the cell means model becomes so useful. Instead of the model equation (1) we use

$$y_{ijk} = \mu_{ij} + e_{ijk}, \tag{4}$$

where μ_{ij} is defined as the population mean of cell i, j, from which the

observations y_{ijk} are deemed to be a random sample. Details of how this kind of model is used and of why it is such a help for unbalanced data are given at length in succeeding chapters. But even at this stage the reader can see how much simpler (4) is than (1). It is the keystone of this book, beginning in Chapter 2.

Using cell means as the basis of a model, as in (4), is in keeping with the early development of analysis of variance by R. A. Fisher as an analysis of differences among observed means. Thus for (3) and (4) he noted the obvious identity

$$y_{ijk} - \bar{y}... \equiv (\bar{y}_{i..} - \bar{y}...) + (\bar{y}_{.j.} - \bar{y}...) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}...) + (y_{ijk} - \bar{y}_{ij.}) \tag{5}$$

and then, with an interest in sums of squares of the form $\sum_{i=1}^{r}(x_i - \bar{x})^2$ for $\bar{x} = \sum_{i=1}^{r} x_i/r$, established the further identity that

$$\Sigma\Sigma\Sigma(y_{ijk} - \bar{y}...)^2 \equiv \Sigma\Sigma\Sigma(\bar{y}_{i..} - \bar{y}...)^2 + \Sigma\Sigma\Sigma(\bar{y}_{.j.} - \bar{y}...)^2 + \Sigma\Sigma\Sigma(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}...)^2 + \Sigma\Sigma\Sigma(y_{ijk} - \bar{y}_{ij.})^2, \tag{6}$$

where, in each case, the triple summation is $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}$.

It seems that at no point did Fisher use a model like (1) or (4). He simply looked at cell means—in his case, observed cell means. The cell means model (4) is addressed to cell means too—population cell means, but, as shall be seen, this leads in many cases to estimates that are observed cell means.

The ultimate development of an analysis of variance table from an identity such as (6) starts with just summarizing that identity. Then, on assuming normality with homoscedastic variances, each sum of squares on the right-hand side of (6) is distributed proportional to a χ^2-variable, independently of the others; and from this come the familiar F-statistics. These and (6) are then summarized in tabular form as an analysis of variance table.

Fisher has an interesting comment on such a table. In a letter dated 6/Jan/'34 (on display at the 50th Anniversary Conference of the Statistics Department at Iowa State University, June, 1983), Fisher writes to

Snedecor that

> "the analysis of variance is (not a mathematical theorem but) a simple method of arranging arithmetical facts so as to isolate and display the essential features of a body of data with the utmost simplicity."

That the analysis of variance table is indeed, as Fisher says, no more than "a simple method of arranging arithmetical facts" is worth emphasizing in these days of computer-generated tables, which too many computer package users are inclined to uncritically treat as sacrosanct. What is important about this, though, is that whilst computers are efficient at doing arithmetic, the intervention of human thinking is always required for making valid interpretation.

1.4. STATISTICAL COMPUTING PACKAGES

Statistical computing packages for analysis of variance that can handle unbalanced data can, of course, also deal with balanced data since they represent just a special case. Yet an important distinction exists concerning package output from those two kinds of data. What we are calling the standard analyses of variance for balanced data are generally well known, open to little question and described in many books; and for many experiment designs the particular analysis is essentially unique. As a result, for a given experiment and resulting data set, most statistical packages produce essentially the same analysis, namely that which is well known, widely documented and ubiquitously accepted. Moreover, in the resulting output, the labeling of the sums of squares is usually sufficiently descriptive that, through knowing the well-established standard analysis for balanced data, one's interpretation of the computer output is unambiguous. For example, a sum of squares labeled "treatments" from the latin square of Table 1.2 would be well known as the numerator sum of squares for an F-statistic suitable for testing the hypothesis that treatment effects are all equal.

This feature of balanced data, namely that the general availability and acceptance of the standard analysis of many a particular case makes for straightforward interpretation of computer output, unfortunately does not carry over to unbalanced data. First, for unbalanced data there is often no unique, unambiguous method of analysis. Instead, several methods are usually available, methods that are often not as easily interpreted as methods for balanced data; nor are they as well known, nor so widely

documented. Second, as a result of having several methods of analysis, not all statistical computing packages necessarily do the same analysis on any given set of unbalanced data. Consequently, in the context of hypothesis testing, or of arraying sums of squares in an analysis of variance format, there is often, for the one data set, a variety of sums of squares available from computing packages. The problem is to identify those that are useful.

This directing of attention to sums of squares must not be taken as implying that sums of squares are the only important calculations in linear model work; far from it. But, whether we like it or not, the real-world usage of linear model analysis is coming to depend on calculations available from statistical computing packages. And reasonably so, too, since the calculations can be voluminous and tedious. Therefore, since a large (but by no means total) part of the output from computer packages consists of sums of squares, it behooves users of packages to know how to interpret those sums of squares.

It is true, of course, that computer output includes a label for almost every calculated value shown therein. Unfortunately, though, the same label in different places can sometimes mean different things. For example, in analyzing unbalanced data from a completely randomized design with a covariate, Searle and Hudson (1982) report the label "sum of squares due to the mean" being attributable to five different values. This illustrates how the labeling of computed values does not always adequately describe underlying calculations and their meaning. Users of computing packages therefore face problems: for a given computed sum of squares, what is the underlying calculation, what is its meaning and what is its use? To answer these questions the user must first know what the procedures for analyzing unbalanced data are. They are the basis on which computer output can be understood and from which appropriate use of that output can be made. And also, of course, from which improved packages can be designed.

1.5. HYPOTHESIS TESTING

Although hypothesis testing is certainly one very useful feature of linear models analysis, it is by no means the only useful aspect of that analysis. Interval estimation (confidence intervals) may well be more useful in many situations. Nevertheless, the tradition of arraying sums of squares and resulting F-statistics in an analysis of variance table is so firmly entrenched in the analysis of balanced data that, through the unhappy extension of that analysis to the analysis of unbalanced data, the tradition has been perpetuated in many computing packages that handle unbalanced data.

One's first (and often most lasting) acquaintance with F-statistics is from calculating them in analysis of variance tables for data from carefully designed and executed experiments such as randomized complete blocks, latin squares, factorial experiments and so on. In most of these the hypotheses being tested by the F-statistics available in the resulting analysis of variance tables are quite clear and straightforward, and usually useful. But such is not often the case with unbalanced data. For these data, the many different sums of squares available do, if used in the numerators of F-statistics, provide tests of a wide variety of hypotheses, some of which are decidedly more useful than others, including some of which are of no general use whatever. Faced as we are with numerous sums of squares available from unbalanced data, the assessment of the usefulness of each is therefore tied to the question "What hypothesis is it testing?".

The mere existence of the preceding question emphasizes the topsy-turvy nature of the situation that prompts it. We all know that the proper logic for hypothesis testing is to set up a hypothesis of interest in the context of the scientific investigation at hand, construct a test for it and then collect data to carry out that test. Unfortunately, in the presence of today's statistical computing packages, this logic can all too easily not be followed: data are often first collected, then fed to a computer, and only then are hypotheses formulated corresponding, perhaps, to just the important-looking (does one dare say "significant"?) F-statistics. Data can so readily be subjected to computer processing without sufficient forethought as to what hypotheses are to be tested that computing power gets used to calculate values whose need has not necessarily been carefully planned in accord with the investigation at hand. Data are simply fed to a package and the (ofttimes voluminous) output that ensues immediately prompts the question "What does this output mean?". Deciding from numerical output alone whether or not a ratio of mean squares, when compared to tabulated values of the F-distribution, is significant or not is easy. However, this can be useful only if one is certain that that ratio of mean squares posing as an F-statistic does indeed (under the usual normality conditions) have an F-distribution. For example, some ratios of mean squares which in fixed effects models have F-distributions do not do so if some of those effects are random effects, that is, in mixed models. It is for this kind of reason that one needs to know how to ascertain whether or not any particular ratio of mean squares does indeed have an F-distribution. Only when it does, does comparing a computed ratio to tabulated Fs make sense. That is the first thing. The second is to then ascertain what hypothesis is being tested by each such statistic. For example, even with a model as simple as $E(y_{ij}) = \mu_i$ it is essential to appreciate that $\Sigma_i\Sigma_j(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$ is a sum of squares that tests not H: $\mu_i = \mu$ *for all* i, for some pre-assigned value μ, but that it tests

H: μ_i *all equal*. Deriving precisely what hypothesis is being tested by an F is not always as easy as this simple example might suggest. Particularly is this so for unbalanced data, where labels attached to computed values are often not sufficient for a user to be certain of knowing what hypothesis is being tested. Therefore, since reliance is undoubtedly coming to be placed on computing packages for doing the arithmetic of linear model analysis, it becomes essential that we know what possible sums of squares can be computed, that we know which ratios of mean squares are F-statistics, that we know how to derive from a sum of squares the hypothesis tested when it is used as the numerator of an F-statistic, and that we know what hypotheses are tested by the sums of squares produced by the computing packages. Only then can wise use be made of those packages, and of the F-statistics they produce. Computing package output cannot be usefully employed without intervention from *homo sapiens*- the "sapiens" of our species must be utilized, for otherwise a computer is no better than *homo insipiens*.