Lecture Notes on the General Theory of Analysis of Variance
Richard M. Golden

The following discussion was selected, condensed, and then slightly modified from the book by Lee (*Experimental Design and Analysis*, Freeman, 1975). First, a compact notation for specifying a majority of the experimental designs in the literature is introduced. Second, the "score" model underlying analysis of variance is described, and the relationship of the score model to computation of F statistics is discussed. This relationship is essential to obtaining an understanding of analysis of variance. Third, the application of these ideas to the development of SAS programs is discussed.

Specification of the Experimental Design

The general analysis of variance problem begins with a particular design. An experimental design defines all possible conditions of the experiment. The "design" of an experiment defines how you will collect the data AND how you will analyze the data. An experimental design is a particular *relationship among a set of factors*.

A *factor* is an independent variable which is manipulated by the experimenter. The possible values of that independent variable are called the *levels* of the factor. Sometimes a factor corresponds to an experimental condition, but we can also think of other types of factors as well such as a subjects factor S whose levels identify different subjects.

There are only two types of relationships which you need to know to define the majority of experimental designs. The first relationship is called the *crossing relation*. Two factors in a design are *crossed* if each level of each factor appears in some condition, with each level of the other factor. For example, if factor A has two levels (a_1, a_2) and subjects factor S has three levels (s_1, s_2, s_3), all of the following six combinations would occur in the design:

	a1	a2
s1		
s2		
s3		

which should be recognized as a one factor within-subjects analysis of variance with three subjects in each of the two experimental conditions. The crossing relation is symbolized as AXS. Note that it does not matter if you write AXS or SXA, both relations refer to the same design.

The second relationship is called the *nesting relation*. Using the notation from our previous example with the crossing relation, factor S is *nested* within factor A if each level of factor S occurs with only one level of factor A. If A has two levels and S has eight levels and S is nested within A, the following experimental design is defined:

a1	a2
s1	s5
s2	s6
s3	s7
s4	s8

which corresponds to a one factor independent groups design. The nesting relation is symbolized as $S(A)$. Note that it DOES MATTER if you write $S(A)$ or $A(S)$ since the nesting relation is not symmetrical (that is, $S(A) \neq A(S)$).

A two factor analysis of variance design where both factors are independent group factors can also be easily represented using this notation. As before, we have two factors in this design: A "subjects" factor S with levels $s_1, s_2, s_3, s_4, \ldots, s_8$, a condition factor A with levels a_1, a_2, and a condition factor B with levels b_1, b_2. This experimental design is symbolized by $S(AXB)$ so that we have:

	b1	b2
a1	s1,s2	s3,s4
a2	s5,s6	s7,s8

which we recognize as the correct setup for a two factor analysis of variance where both factors are independent group factors (see design on page 376 of text).

A two factor analysis of variance design where one factor is an independent groups factor and one factor is a within-groups factor is also easily represented using this notation (see "mixed" design on

page 382 of text). As before, we have two factors in this design: A "subjects" factor S with levels $s_1, s_2, s_3, s_4, \ldots, s_8$, the independent groups condition factor A with levels a_1, a_2, and the within groups condition factor B with levels b_1, b_2. This experimental design is symbolized by $S(A)XB$ so that we have:

		b1	b2
a1	s1		
	s2		
	s3		
	s4		
a2	s5		
	s6		
	s7		
	s8		

Helpful Hint: Note that in the AXB design if there are a levels of factor A and b levels of factor B, then there will be ab conditions in the experiment. In the $S(A)$ design if there are s levels of subjects we will of course need s to be a multiple of a so that an equal number of subjects are assigned to each condition. In the $S(A)XB$ design, we need sb conditions and we need s to be a multiple of a. By careful inspection of the above designs, you should convince yourself that these conclusions are correct.

Score Models for Analysis of Variance

Some basic linear score models for analysis of variance are now introduced. Try to look for "patterns" in the construction of such models so that you can create score models for your own designs. We also describe how F statistics are computed from such models.

A Score Model for Design S(A)

Suppose that we have a one factor independent groups analysis of variance and subject 7 obtains a score of 115 in condition a_1. Also suppose that subject 6 obtains a score of 120 in condition a_2. We symbolically represent the score of subject k in condition i of the experiment as X_{ik}. The factor is A. Thus, the score subject 7 obtained in condition a_1 is represented as $X_{17} = 115$, and the score subject 6 obtained in condition a_2 is represented as $X_{26} = 120$.

The linear "score model" for design $S(A)$ states that the *score* X_{ik} can be decomposed into three distinct parts: the "average" score of the subject across all experimental conditions, the change in the average score due to experimental condition a_i, and some unknown "error" perturbation to the score X_{ik}. Mathematically, this is represented as follows:

$$X_{ik} = m + a_i + e_{ik}$$

where m is the average of all of the scores, a_i is a perturbation to the average score due to experimental condition i, and e_{ik} is a "random" perturbation. It is always assumed that such perturbations are distributed according to a "normal" or "bell-shaped" curve. (Note: Lee (1975) calls this Design A rather than S(A)).

In the experiment, we observe the "scores" $\{X_{ik}\}$ but we do not know m or the effects of a particular condition represented by the variable a_i. Thus, we have to estimate the values of these parameters. It turns out that if the experiment is properly designed, we can always solve for the parameters of the model in terms of the scores we observe from the subjects using simple algebra. The SAS computer program and the procedures that you have been using are essentially solving for the parameters of this linear model when you do a one factor independent groups analysis of variance. In fact, we can even (and we will) estimate the random "error" e_{ik} which is assumed to be added to each observed score.

Since a_i is the perturbation to the average score due to the experimental condition, we can square each perturbation, sum up the squares of these perturbations, and divide by an appropriate normalization constant (the degrees of freedom for factor A), to obtain an estimate of the variability of a_i. This estimate is called $MS_A = (\sum a_i^2)/dfa$ where dfa is the degrees of freedom of A. Note dfa is equal to the number of levels of A minus 1.

Similarly, we can compute the sum of squares of the error perturbations, and divide by an appropriate normalization constant (the degrees of freedom for the error factor), to obtain an estimate of the variability of e_{ik}. This estimate is called $MS_{e_{ik}} = (\sum e_{ik}^2)/dfe$ where dfe is the degrees of freedom of the error. The dfe is equal to the total number of scores or observations in the experiment minus the degrees of freedom of all other parameters (including the mean parameter m which always has one degree of freedom).

Now note that if MS_A is greater than MS_e, then could conclude that the variability across conditions is not due to chance. Our basic problem, however, is that MS_A and MS_e will be different if we repeat our experiment with different subjects! For this reason, we do the following. We construct an "F-statistic" defined as:

$$F_{obs} = MS_A/MS_e$$

and note that some clever statistician figured out how to compute the probability that F would be greater than one by chance (i.e., due to random variation in the experiment). So when we compare F_{obs} to F_{crit} we are simply deciding if the chance that $F_{obs} > 1$ due to random variation is less than the selected significance level.

A Score Model for Design A X S

Suppose that we have a one factor within groups analysis of variance and subject 7 obtains a score of 115 in condition 1. Also suppose that subject 6 obtains a score of 120 in condition 2, and that subject 5 obtains a score of 97 in condition 1. We symbolically represent the score of subject k in condition i of the experiment as X_{ik}. The factor is A. Thus, the score subject 7 obtained in condition 1 is represented as $X_{17} = 115$, and the score subject 6 obtained in condition 2 is represented as $X_{26} = 120$. The score subject 5 obtaines is $X_{15} = 97$.

The linear "score model" for design AXS states that X_{ik} can be decomposed into four parts: the "average" score of the subject across all experimental conditions, the change in the average score due to experimental condition i, the change in the average score due to subject k, the "interaction" of subject k with condition i, and an error perturbation. The interaction term provides a contribution to the score X_{ik} ONLY when subject k is tested under experimental condition i. Mathematically, this is represented as follows:

$$X_{ik} = m + a_i + s_k + as_{ik} + e_{ik}$$

where m is the average score, a_i is the perturbation to the average score due to experimental condition i averaged across subjects, s_k is the perturbation to the average score due to subject k averaged across experimental conditions, as_{ik} is the perturbation to the score which ONLY occurs when subject k is tested on condition i, and e_{ik} is the error contribution. Note that with only one score per condition, each subject sees each level of A only once so that $e_{ik} = 0$.

As before, given a properly designed experiment, we can "solve" the linear score model for the parameters of the linear model.

Thought Question: Why don't we have an "interaction" term in the one factor between groups analysis of variance?

As in score model $S(A)$, we can compute the average variability of each term in the score model. For example, MS_A is computed just as before by adding up the squares of the perturbations a_i and dividing by the degrees of freedom. Similarly MS_{AS} is computed just as before by adding up the squares of the perturbations as_{ik} and dividing by the degrees of freedom. By the way to compute the degrees of freedom for the interaction term AS simply multiply the degrees of freedom of A by the degrees of freedom of S.

Although it seems obvious that the correct thing to do is compare (as before) the variability of Factor A, MS_A, to the mean square error, MS_e, the obvious approach (unfortunately) does not work in this case. The reason is that the subjects factor is a "random" factor and must also be treated as a source of error. It is beyond the scope of this introduction to describe how to choose the appropriate error term for comparison with MS_A so the answer to this problem is simply provided.

To check for an effect of factor A, compare MS_A to MS_{AS}. If $MS_A > MS_{AS}$, (i.e., if $F = MS_A/MS_{AS} > 1$ with probability $1 - \alpha$) then reject the null hypothesis that an effect of A is not present.

A Score Model for Design S(A X B)

The above discussions can also be applied to the two factor design where both factors are independent groups factors. Such a design may be symbolized as $S(AXB)$. (Note that Lee (1975) refers to this design as AXB). The score model for $S(AXB)$ is given as follows:

$$X_{ijk} = m + a_i + b_j + ab_{ij} + e_{ijk}$$

where X_{ijk} is the score obtained from subject k in level i of Factor A and in level j of Factor B, m is the average score, a_i is the perturbation to the average score due to experimental condition i averaged across subjects, b_j is the perturbation to the average score due to condition j averaged across subjects, ab_{ij} is the perturbation to the score which ONLY occurs when subject k is tested simultaneously on conditions i and j of factors A and B respectively, and e_{ijk} is the error contribution to score X_{ijk}.

The mean squares for factors A, B, and interaction AB are all compared to the mean square error for the purposes of significance testing. For example, to check for an effect of factor AB, compare MS_{AB} to MS_e. If $MS_{AB} > MS_e$, (i.e., if $F = MS_{AB}/MS_e > 1$ with probability $1 - \alpha$) then reject the null hypothesis that an effect of AB is not present.

A Score Model for Design S(A) X B

The above discussions can also be applied to the two factor design where one factor is an independent groups factor and the other factor is a within-groups factor. Let the independent groups factor be A and the between-groups factor be B. Such a design may be symbolized as $S(A)XB$. The score model for $S(A)XB$ is given as follows:

$$X_{ijk} = m + a_i + b_j + s_k + ab_{ij} + bs_{jk} + e_{ijk}$$

where X_{ijk} is the score obtained from subject k in level i of Factor A and in level j of Factor B, m is the average score, a_i is the perturbation to the average score due to experimental condition i averaged across subjects and factor B, b_j is the perturbation to the average score due to condition j averaged across subjects (or equivalently factor A), ab_{ij} is the perturbation to the score which ONLY occurs when subject k is tested simultaneously on conditions i and j of factors A and B respectively, s_k is the perturbation to the average score due to subject k averaged across factors A and B, bs_{jk} is the perturbation to the score which ONLY occurs when subject k is tested in condition j of factor B, and e_{ijk} is the error contribution.

To test for a main effect of A, compare MS_A to MS_S by computing $F = MS_A/MS_S$. To test for a main effect of B, compare MS_B to MS_{BS} by computing $F = MS_B/MS_{BS}$. To test for a main effect of AB, compare MS_{AB} to MS_{BS} by computing $F = MS_{AB}/MS_{BS}$.

Using the SAS program for Analysis of Variance

In this section, SAS programs for various experimental designs will be described in terms of the score model concept which was previously introduced.

A typical SAS program has two independent parts. The first part defines the SAS data set which contains the data you want to analyze. The second part which immediately follows the first part is a short program of about five or so lines which analyzes your data set.

Giving your Data to the Computer

Here is the first part of a typical SAS program which gives the computer the data you want to analyze.

```
options linesize = 64 pagesize = 64;
title 'Anything I would like to have printed on each page!!';
data Richard;
input sub $ AFac $ BFac $ depend ;
cards;
s1 a1 b1 5
s2 a1 b1 -1
s3 a1 b1 -7
s4 a2 b1 6
s5 a2 b1 4
s6 a2 b1 -1
s7 a3 b1 4
s8 a3 b1 -3
s9 a3 b1 -7
s1 a1 b2 3
s2 a1 b2 7
s3 a1 b2 -1
s4 a2 b2 12
s5 a2 b2 12
s6 a2 b2 3
s7 a3 b2 12
s8 a3 b2 11
s9 a3 b2 13
;
```

The first line of the above SAS program is optional and simply tells the computer to assume that the printout page has no more than 64 lines per page and 64 characters per line.

The second line of the above SAS program is optional and simply prints out a message at the top of each page of your printout. It is recommended that you indicate the title of your experiment and your name in abbreviated form in the title.

The third line of the above SAS program is important. This command defines a SAS dataset named "Richard". You may call your SAS dataset anything you like but the dataset must be named.

The fourth line of the above SAS program is also important. This command tells the SAS program what variables should be assigned what numbers in your data file. In this example, the *input* statement is saying that the first column of items are different values of the variable known as "sub", the second column of items are different values of the variable known as "AFac", the third column of items are different values of the variable known as "BFac", and the fourth column of items are different values of the variable known as "depend". Of course you can name your variables anyway you like but SAS will get upset if your variable names are too long. Also if a variable name is followed by a dollar sign, then that variable is considered to be a nominal variable. Thus, the variables "sub", "AFac", and "BFac" in this example are nominal variables while "depend" is a numerical variable.

The fifth line of the above SAS program is also important. This command tells the SAS program that the data file is in the program itself. SAS has much more sophisticated ways of loading the data.

The remaining lines of the above SAS program contain your data. Each row of numbers is referred to as an "observation." There are 18 observations in this example. Thus, inspection of row seven indicates that the value of variable "depend" is "-1" when the value of variable "sub" is "s2", the value of variable "AFac" is "a1", and the value of variable "BFac" is "b1".

Analyzing your data using SAS

The SAS programs which are relevant to analysis of variance are called "anova" and "glm". The program "glm" (general linear model) will be described since this program has more options than "anova".

Here is a sample SAS program for using the SAS data set we just described to analyze design $S(A)XB$.

```
proc glm;
class sub AFac BFac  ;
model depend = AFac BFac sub(AFac) AFac*BFac BFac*sub(AFac);
test h = AFac e = sub(AFac);
test h = BFac e = BFac*sub(AFac);
test h = AFac*BFac e = BFac*sub(AFac);
means AFac BFac sub AFac*BFac BFac*sub;
```

The first line of the SAS program indicates that procedure *glm* (general linear model) will be used to analyze the data.

The second line indicates the independent variables in the model which are "sub", "AFac", and "BFac".

The third line defines the score model which is associated with your experimental design. Thus, you are actually telling the SAS program what score model you intend to use. Note that the "score" is your dependent variable. (Compare this line with the score model for $S(A)XB$). Since all score models will have a "mean" and "error" term, these terms are omitted in the SAS program.

The fourth through sixth lines define the F statistics which you want to calculate. The h stands for hypothesis and the e stands for error. Or in other words, line four of the SAS program says "compute an F statistic which compares MS_A (the mean square for AFac) with MS_S (the mean square for the subject factor "sub"). (Again, compare these three lines with the previous discussion).

Finally, the last line of the SAS program prints out means for the various factors and interactions which you have specified.

Here is a selected portion of a sample SAS output from the program we have just discussed.

```
Anything I would like to have printed on each page!!     1
23:24 Thursday, March 19, 1992
General Linear Models Procedure
Class Level Information
Class    Levels    Values
SUB           9    s1 s2 s3 s4 s5 s6 s7 s8 s9
AFAC          3    a1 a2 a3
BFAC          2    b1 b2
Number of observations in data set = 18
```

Notice that the SAS output indicates which factors will be used in the linear score model, their levels, and their values. The SAS output also prints out the number of observations in the SAS data set.

The next selected portion of the SAS output computes the sum of squares of each parameter of the score model as previously discussed.

```
Anything I would like to have printed on each page!!     2
23:24 Thursday, March 19, 1992
General Linear Models Procedure
Source                   DF       Type III SS  F Value   Pr > F
AFAC                      2       84.00000000        .        .
BFAC                      1      288.00000000        .        .
SUB(AFAC)                 6      180.00000000        .        .
AFAC*BFAC                 2       84.00000000        .        .
SUB*BFAC(AFAC)            6       68.00000000        .        .
```

For example, the sum of the squared perturbations due to factor A is equal to 84 in this example, and since there are 2 degrees of freedom associated with factor A, the mean squared contribution of A is $84/2$. That is, $MS_A = 84/2 = 42$.

As another example, the sum of the squared perturbations due to factor $S(A)$ is 180 with 6 degrees of freedom. So we have: $MS_{S(A)} = 180/6 = 30$.

From our previous discussion of design $S(A)XB$, the F statistic we need to demonstrate an effect of factor A is given by:

$$F(2,6) = MS_A/MS_{S(A)} = 42/30 = 1.4.$$

Although this calculation is relatively easy, we don't have to do such computations since we have already asked SAS to compute the above F statistic when we included our "test h = AFac test e = Sub(AFac); " statement.

Here is the SAS output for the F statistics requested by the three "test" statements in the SAS program. Using the SAS output above, you should be able to figure out where the F values in the SAS output below came from. We just worked out the first example.

```
Tests of Hypotheses using the Type III MS for
SUB(AFAC) as an error term
Source                   DF       Type III SS  F Value   Pr > F
AFAC                      2       84.00000000     1.40   0.3170

Tests of Hypotheses using the Type III MS for
SUB*BFAC(AFAC) as an error term
Source                   DF       Type III SS  F Value   Pr > F
BFAC                      1      288.00000000    25.41   0.0024

Tests of Hypotheses using the Type III MS for
SUB*BFAC(AFAC) as an error term
Source                   DF       Type III SS  F Value   Pr > F
AFAC*BFAC                 2       84.00000000     3.71   0.0895
```

Notice that the SAS output also prints out a probability (i.e., the number below the column labeled "$Pr > F$"). This probability is the probability that the observed F value is greater than the critical F value by chance. Thus, if the probability to the right of the F value is less than the chosen significance level α, then that F value is significant at the α level. As you can see, there is a significant main effect of factor B at the 0.01 significance level but the effect of factor A and interaction AB were not significant at even the 0.05 level in this example.