Chapter 3

Factorial ANOVA and related topics

3.1 A One-way Example

The following is a textbook example taken from Neter et al.'s (1996) *Applied linear statistical models* [5]. The Kenton Food Company is interested in testing the effect of different package designs on sales. Five grocery stores were randomly assigned to each of four package designs. The package designs used either three or five colours, and either had cartoons or did not. Because of a fire in one of the stores, there were only four stores in the 5-colour cartoon condition.

The dependent variable is sales, defined as number of cases sold. Actually, there are two independent variables: number of colours and presence versus absence of cartoons. But we will initially consider package design as a single categorical independent variable with four values.

Sample Question 3.1.1 *If there is a statistically significant relationship between package design and sales, would we be justified in concluding that differences in package design* caused *differences in sales?*

Answer to Sample Question 3.1.1 *Yes, if the study is carried out properly. It's an experimental study.*

Sample Question 3.1.2 *Is there a problem with external validity here?*

Answer to Sample Question 3.1.2 *It's impossible to tell for sure, but there easily could be. The behaviour of the sales force would have to be controlled somehow. A double blind would be ideal.*

The SAS program `kenton1b.sas` does a lot of things, starting with a one-way ANOVA using `proc glm`. The strategy will be to first present the entire program, and then go through it piece by piece and explain what is going on – with a few major digressions to explain the statistics.

```
/********************** kenton1b.sas *************************/
options linesize=79 pagesize=100 noovp formdlim=' ';
title 'Kenton Oneway Example From Neter et al.';

proc format;
     value pakfmt 1 = '3Colour Cartoon'   2 = '3Col No Cartoon'
                  3 = '5Colour Cartoon'   4 = '5Col No Cartoon';

data food;
     infile 'kenton.data';
     input package sales;
     label package = 'Package Design'
           sales   = 'Number of Cases Sold';
     format package pakfmt.;
     /* Define ncolours and cartoon */
     if package = 1 or package = 2 then ncolours = 3;
        else if package = 3 or package = 4 then ncolours = 5;
     if package = 1 or package = 3 then cartoon = 'No ';
        else if package = 2 or package = 4 then cartoon = 'Yes';

/* Basic one-way ANOVA -- well, not very basic */

proc glm;
     class package;
     model sales = package;
     means package;
     means package / bon tukey scheffe;
     /* Test some custom contrasts */
     contrast '3Colourvs5Colour' package 1  1 -1 -1;
     contrast 'Cartoon'          package 1 -1  1 -1;
     contrast 'CartoonDepends'   package 1 -1 -1  1;
     /* Test a collection of contrasts */
     contrast 'Overall F'        package 1 -1  0  0,
                                 package 0  1 -1  0,
                                 package 0  0  1 -1;
     /* Get estimated value of a contrast along with a test (F=t-squared) */
     estimate '3Colourvs5Colour' package 1  1 -1 -1 / divisor = 2;
     /* All pairwise comparisons */
     contrast '1 vs 2' package  1 -1  0  0;
     contrast '1 vs 3' package  1  0 -1  0;
     contrast '1 vs 4' package  1  0  0 -1;
     contrast '2 vs 3' package  0  1 -1  0;
     contrast '2 vs 4' package  0  1  0 -1;
     contrast '3 vs 4' package  0  0  1 -1;
```

```
proc iml;
     title2 'Table of critical values for all possible Scheffe tests';
     numdf = 3;   /* Numerator degrees of freedom for initial test */
     dendf = 15;  /* Denominator degrees of freedom for initial test */
     alpha = 0.05;
     critval = finv(1-alpha,numdf,dendf);
     zero = {0 0}; S_table = repeat(zero,numdf,1);  /* Make empty matrix */
     /* Label the columns */
     namz = {"Number of Contrasts in followup test"
             "    Scheffe Critical Value"}; mattrib S_table colname=namz;
     do i = 1 to numdf;
        s_table(|i,1|) = i;
        s_table(|i,2|) = numdf/i *  critval;
     end;
     reset noname; /* Makes output look nicer in this case */
     print "Initial test has"  numdf " and " dendf "degrees of freedom."
           "Using significance level alpha = " alpha;
     print s_table;

proc glm;
     title2 "Actually it's a two-way ANOVA";
     class ncolours cartoon;
     model sales = ncolours|cartoon;
     means ncolours|cartoon;

/*   The model statement could have been
     model sales = ncolours cartoon ncolours*cartoon; */
```

The `proc format` statement provides labels for the package designs. After reading the data in a routine way, `if` statements are used to construct the categorical independent variables `ncolours` and `cartoon`. Notice the extra space in the 'No ' value of the alphanumeric variable `cartoon`. At first I didn't have a space, and `Yes` was truncated to `Ye`.

Now we'll look at what the first `proc glm` does. The complete `proc glm` statement is given above. Here, we will look at it a piece at a time, examining the output as we go. First, we have

```
proc glm;
     class package;
     model sales = package;
```

The class statement declares package to be categorical. Without it, `proc glm` would do a regression with package as a quantitative independent variable. The main F-test for equality of the four means is

General Linear Models Procedure

Dependent Variable: SALES Number of Cases Sold

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	588.2210526	196.0736842	18.59	<.0001
Error	15	158.2000000	10.5466667		
Corrected Total	18	746.4210526			

R-Square	C.V.	Root MSE	SALES Mean
0.788055	17.43042	3.2475632	18.631579

We conclude that package design (or, if the study was poorly controlled, some variable confounded with package design) caused a difference in sales. Note that almost 79% of the variation in sales comes from package design; this is the value, which is exactly the R^2 from a multiple regression that is equivalent to this one-way ANOVA.

The statement `means package;` produces mean sales for each value of the variable `package`.

Level of PACKAGE	N	SALES Mean	SALES SD
3Col No Cartoon	5	13.4000000	3.64691651
3Colour Cartoon	5	14.6000000	2.30217289
5Col No Cartoon	5	27.2000000	3.96232255
5Colour Cartoon	4	19.5000000	2.64575131

Such a display is essential for seeing what is going on, but it still does not tell you which means are different from which other means. But before we lose control and start doing all possible t-tests, consider the following.

3.2 The Curse of a Thousand t-tests

Significance tests are supposed to help screen out random garbage, and help us ignore "trends" that could easily be due to chance. But all the common significance tests are designed in isolation, as if each one were the only test you would ever be doing. The

chance of getting significant results when nothing is going on may be about 0.05 (more or less, depending on how well the assumptions are met), but if you do a *lot* of tests on a data set that is purely noise (no true relationships between any independent variable and any dependent variable), the chances of false significance mount up. It's like looking for your birthday in tables of stock market prices. If you look long enough, you will find it.

This problem definitely applies when you have a significant difference among more than two treatment means, and you want to know which ones are different from each other. For example, in an experiment with 10 treatment conditions (this is not an unusually large number, for real experiments), there are 45 pairwise differences among means.

You have to pity the poor scientist who learns about this and is honest enough to take this problem seriously (let's use the term "scientist" generously to apply to anyone trying to use significance test to learn something about a data set). On one hand, good scientific practice and common sense dictate that if you have gone to the trouble to collect data, you should explore thoroughly and try to learn something from the data. But at the same time, it appears that some stern statistical entity is scolding you, and saying that you're naughty if you peek.

There are two main ways to resolve the dilemma. One is to basically ignore the problem, while perhaps acknowledging that it is there. According to this point of view, well, you're crazy if you don't explore the data. Maybe the true significance level for the entire process is greater than 0.05, but still the use of significance tests is a useful way to decide which results might be real. Nothing's perfect; let's carry on.

The other reaction is to look for ways that significance tests can be modified to allow for the fact that we're doing a lot of them. What we want are methods for holding the chances of false significance to a single low level for a *set* of tests, simultaneously. The general term for such methods is **multiple comparison** procedures. Often, when a significance test (like a one-way ANOVA) tests several things simultaneously and turns out to be significant, multiple comparison procedures are used as a second step, to investigate where the effect came from. In cases like this, the multiple comparisons are called **follow-up** tests, or **post hoc** tests, or sometimes **probing**.

It is generally acknowledged that multiple comparison methods are often helpful (even necessary) for following up significant F-tests in order to see where an effect comes from. There is less agreement on how far the principle should be extended. Personally, I like the idea of limiting the chance of false significance to 0.05 for an entire study – say, for all the tests reported in a scientific paper, and all the ones that were not reported, too. This is a fairly radical view, shared by almost no one. But it can work in practice if you have enough data. More on this later. For now, let's concentrate on following up a significant F test in a one-way analysis of variance.

In the Kenton package design data, there are 4 treatment conditions, and 6 potential pairwise comparisons. The next line in the SAS program,

```
means package / bon tukey scheffe;
```

requests three kinds of multiple comparison tests for all pairwise differences among means.

3.2.1 Bonferroni

The Bonferroni method is very general, and extends far beyond pairwise comparisons of means. It is a simple correction that can be applied when you are performing multiple tests, and you want to hold the chances of false significance to a single low level for all the tests simultaneously. *It applies when you are testing multiple sets of independent variables, multiple dependent variables, or both.*

The Bonferroni correction consists of simply dividing the desired significance level (that's α, the maximum probability of getting significant results when actually nothing is happening, usually $\alpha = 0.05$) by the number of tests. In a way, you're splitting the alpha equally among the tests you do.

For example, if you want to perform 5 tests at joint significance level 0.05, just do everything as usual, but only declare the results significant at the *joint* 0.05 level if one of the tests gives you $p < 0.01$ (0.01=0.05/5). If you want to perform 20 tests at joint significance level 0.05, do the individual tests and calculate individual p-values as usual, but only believe the results of tests that give $p < 0.0025$ (0.0025=0.05/20). Say something like "Protecting the 20 tests at joint significance level 0.05 by means of a Bonferroni correction, the difference in reported liking between worms and spinach soufflé was the only significant food category effect."

The Bonferroni correction is conservative. That is, if you perform 20 tests, the probability of getting significance at least once just by chance is less than or equal to 0.0025 – almost always less. The big advantages of the Bonferroni approach are simplicity and flexibility. It is the only way I know to analyze quantitative and categorical dependent variables simultaneously.

The main disadvantages of the Bonferroni approach are

1. *You have to know how many tests you want to perform in advance, and you have to know what they are.* In a typical data analysis situation, not all the significance tests are planned in advance. The results of one test will give rise to ideas for other tests. If you do this and then apply a Bonferroni correction to all the tests that you happened to do, it no longer protects all the tests simultaneously. On the other hand, you could randomly split your data into an exploratory sample and a replication sample. Test to your heart's content on the first sample. Then, when you think you know what your results are, perform only those tests on the replication sample, and protect them simultaneously with a Bonferroni correction. This could be called "Bonferroni-protected cross-validation." It sounds good, eh?

2. *The Bonferroni correction can be too conservative,* especially when the number of tests becomes large. For example, to simultaneously test all 780 correlations in a 40 by 40 correlation matrix at joint $\alpha = 0.05$, you'd only believe correlations with $p < 0.0000641 = 0.05/780$.

   Is this "too" conservative? Well, with $n = 200$ in that 40 by 40 example, you'd need $r = 0.27$ for significance (compared to $r = .14$ with no correction). With $n = 100$ you'd need $r = .385$, or about 14.8% of one variable explained by another *single* variable. Is this too much to ask? You decide.

3.2.2 Tukey

This is Tukey's Honestly Significant Difference (HSD) method. It is not his Least Significant Different (LSD) method, which has a better name but does not really get the job done. Tukey tests apply only to pairwise differences among means in ANOVA. It is based on a deep study of the probability distribution of the difference between the largest sample mean and the smallest sample mean, assuming the population means are in fact all equal.

- If you are interested in all pairwise differences among means and nothing else, and if the sample sizes are equal, Tukey is the best (most powerful) test, period.
- If the sample sizes are unequal, the Tukey tests still get the job of simultaneous protection done, but they are a bit conservative. When sample sizes are unequal, Bonferroni or Scheffé can sometimes be more powerful.

3.2.3 Scheffé

Suppose there are p treatments (groups, values of the categorical independent variable, whatever you want to call them). A **contrast** is a special kind of linear combination of means in which the weights add up to zero. A population contrast has the form

$$\ell = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$$

where $a_1 + a_2 + \cdots + a_p = 0$. The case where all of the a values are zero is uninteresting, and is excluded. A population contrast is estimated by a sample contrast:

$$L = a_1\overline{Y}_1 + a_2\overline{Y}_2 + \cdots + a_p\overline{Y}_p.$$

By setting $a_1 = 1$, $a_2 = -1$, and the rest of the a values to zero we get $L = \overline{Y}_1 - \overline{Y}_2$, so it's easy to see that any pairwise difference is a contrast. Also, the average of one set of means minus the average of another set is a contrast.

The initial F test for equality of p means can be viewed as a simultaneous test of $p-1$ contrasts. For example, suppose there are four treatments, and the null hypothesis of the initial test is $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. The table gives the a_1, a_2, a_3, a_4 values for three contrasts; if all three contrasts equal zero then the four population means are equal, and *vice versa*.

a_1	a_2	a_3	a_4
1	-1	0	0
0	1	-1	0
0	0	1	-1

The way you read this table is

$$\begin{array}{ccccccccc} \mu_1 & - & \mu_2 & & & & & = & 0 \\ & & \mu_2 & - & \mu_3 & & & = & 0 \\ & & & & \mu_3 & - & \mu_4 & = & 0 \end{array}$$

Clearly, if $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ and $\mu_3 = \mu_4$, then $\mu_1 = \mu_2 = \mu_3 = \mu_4$, and if $\mu_1 = \mu_2 = \mu_3 = \mu_4$, then $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ and $\mu_3 = \mu_4$. The simultaneous F test for the three contrasts is 100% equivalent to a one-way ANOVA; it yields the same F statistic, the same degrees of freedom, and the same p-value.

There is always more than one way to set up the contrasts to test a given hypothesis. Staying with the example of testing differences among four means, we could have specified

a_1	a_2	a_3	a_4
1	0	0	-1
0	1	0	-1
0	0	1	-1

so that all the means are equal to the last one. These contrasts (differences between means) are actually *equal* to the regression coefficients in a multiple regression with indicator dummy variables, in which the last category is the reference category. But no matter how you set up collection of contrasts, if you do it correctly you always get the same answer.

The Scheffé tests allow testing whether *any* contrast (or set of contrasts) of treatment means differs significantly from zero, with the tests for all possible contrasts simultaneously protected at the same significance level, usually 0.05.

When asked for Scheffé follow-ups to a one-way ANOVA, SAS tests all pairwise differences between means, but *there are infinitely many more contrasts in the same family that it does not do* — and they are all jointly protected against false significance at the 0.05 level.

It's a miracle. You can do infinitely many tests, all simultaneously protected. You do not have to know what they are in advance. It's an license for unlimited data fishing, at least within the class of contrasts of treatment means. And you can test up to $p - 1$ contrasts simultaneously if you wish. They are all part of the same family.

Two more miracles:

- If the initial one-way ANOVA is not significant, it's *impossible* for any of the Scheffé follow-ups to be significant. This is not quite true of Bonferroni or Tukey.

- If the initial one-way ANOVA *is* significant, there *must* be a single contrast that is significantly different from zero. It may not be a pairwise difference, you may not think of it, and if you do find one it may not be easy to interpret, but there is at least one out there. Well, actually, there are infinitely many, but they may all be extremely similar to one another. Incidentally, if you test any *collection* of contrasts that includes a contrast that is significantly different from zero by a Scheffé test, then the Scheffé test for the collection will be significant too.

Given all this, clearly it is helpful to be able to test any set of contrast you wish. As you will see below, the `contrast` statement of `proc glm` lets you do it easily. For now, let's assume that you have done an initial F test for differences among p treatment means, it's statistically significant, and also you can get F tests for any contrast of collection of contrasts you specify.

As usual, the F tests for contrasts (which are sometimes optimistically called "planned comparisons") are designed in a vacuum, as if each one were the only test you would ever do on your data. But you can convert them into Scheffé follow-ups to the initial test by using a different critical value (Recall that if a test statistic is greater than the critical value, it's significant).

Suppose that the follow-up test you want to do involves s contrasts; for a test of a single difference between means or some other single contrast, $s = 1$. Compute the usual F statistic for testing the contrast, and compare it to a modified critical value that we will call F_{S-crit}; the S is for Scheffé. The formula for F_{S-crit} is

$$F_{S-crit} = \frac{p-1}{s} F_{crit}, \tag{3.1}$$

where F_{crit} is the critical value for the *initial* test — the one you are following up. You reject the null hypothesis and declare your Scheffé test significant if $F > F_{S-crit}$.

You can do as many of these tests as you want easily, using SAS and a small table of F_{S-crit} critical values. You can make the table you need with `proc iml`. This is illustrated in the example below; the code can easily be modified to suit any problem. Or, you can use a textbook table of the F distribution and a calculator.

Please take another look at Formula (3.1). Notice that multiplying by the number of means (minus one) is a kind of penalty for the richness of the infinite family of tests you could do, while dividing by the number of contrasts you're testing reduces the penalty because you're looking for something bigger. As soon as Mr. Scheffé discovered these tests, people started complaining that the penalty was very severe, and it was too hard to get significance. In my opinion, what's remarkable is not that a license for unlimited fishing is expensive, but that it's for sale at all. You can pay for it by increasing the sample size.

When sample sizes are unequal, SAS presents follow-up tests for pairwise differences between means in the form of confidence intervals. If the 95% confidence interval does not include zero, the test (Bonferroni, Tukey or Scheffé) is significant at 0.05. Since all three types of follow-up test point to exactly the same conclusions for these data, only the Scheffé will be reproduced here.

```
                General Linear Models Procedure

                  Scheffe's test for variable: SALES

     NOTE: This test controls the type I experimentwise error rate but
           generally has a higher type II error rate than Tukey's for all
           pairwise comparisons.

        Alpha= 0.05  Confidence= 0.95  df= 15  MSE= 10.54667
                     Critical Value of F= 3.28738

  Comparisons significant at the 0.05 level are indicated by '***'.
```

PACKAGE Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
5Col No Cartoon - 5Colour Cartoon	7.700	0.859	14.541	***
5Col No Cartoon - 3Colour Cartoon	12.600	6.150	19.050	***
5Col No Cartoon - 3Col No Cartoon	13.800	7.350	20.250	***
5Colour Cartoon - 5Col No Cartoon	-7.700	-14.541	-0.859	***
5Colour Cartoon - 3Colour Cartoon	4.900	-1.941	11.741	
5Colour Cartoon - 3Col No Cartoon	6.100	-0.741	12.941	
3Colour Cartoon - 5Col No Cartoon	-12.600	-19.050	-6.150	***
3Colour Cartoon - 5Colour Cartoon	-4.900	-11.741	1.941	
3Colour Cartoon - 3Col No Cartoon	1.200	-5.250	7.650	
3Col No Cartoon - 5Col No Cartoon	-13.800	-20.250	-7.350	***
3Col No Cartoon - 5Colour Cartoon	-6.100	-12.941	0.741	
3Col No Cartoon - 3Colour Cartoon	-1.200	-7.650	5.250	

Notice that the critical value for the initial test (F_{crit}, not F_{S-crit}) for performing more tests is conveniently provided.

This pairwise confidence interval format is not so easy to look at, even if the significant differences are indicated by "***." For one thing, each comparison is given twice, once in each direction. For another, the actual means are not printed, just the differences between means. It helps to re-arrange the means from highest to lowest. This next display is not part of the SAS output; it's SAS output edited with a word processor.

Level of PACKAGE	N	SALES Mean	SALES SD
5Col No Cartoon	5	27.2000000	3.96232255
5Colour Cartoon	4	19.5000000	2.64575131
3Colour Cartoon	5	14.6000000	2.30217289
3Col No Cartoon	5	13.4000000	3.64691651

Now we see that the 5-colour No Cartoon treatment is significantly different from each of the others, which are not significantly different from each other. That's the kind of package design they should use; from a marketing standpoint, we're done. But let's look at some more follow-up tests anyway.

Testing Contrasts The `proc glm` in `kenton1b.sas` continues

```
/* Test some custom contrasts */
contrast '3Colourvs5Colour' package 1  1 -1 -1;
contrast 'Cartoon'          package 1 -1  1 -1;
contrast 'CartoonDepends'   package 1 -1 -1  1;
/* Test a collection of contrasts */
contrast 'Overall F'        package 1 -1  0  0,
                            package 0  1 -1  0,
                            package 0  0  1 -1;
```

The syntax for specifying a contrast goes: The word `contrast`, a label for the test in single or double quotes (this will appear in the output), the name of the independent variable, the coefficients of the contrast (the a values), and a semicolon to end the statement. If you are testing more than one contrast simultaneously, put a comma after the first one, repeat the independent variable name, and give another set of coefficients. The last contrast ends with a semi-colon instead of a comma. As the example shows, you can do as many tests as you like.

3.2.4 Proper Follow-ups

We will describe a set of tests as *proper follow-ups* to to an initial test if

1. The null hypothesis of the initial test logically implies the null hypotheses of all the tests in the follow-up set.

2. All the tests are jointly protected against Type I error (false significance) at a known significance level, usually $\alpha = 0.05$.

The first property requires explanation. First, consider that the Tukey tests, which are limited to pairwise differences between means, automatically satisfy this, because if all the population means are equal, then each pair is equal to each other. But it's possible to make mistakes with Bonferroni and Scheffé if you're not careful.

Here's why the first property is important. Suppose the null hypothesis of a follow-up test *does* follow logically from the null hypothesis of the initial test. Then, if the null hypothesis of the follow-up is false (there's really something going on), then the null hypothesis of the initial test must be incorrect too, and this is one way in which the initial null hypothesis is false. Thus if we correctly reject the follow-up null hypothesis, we have uncovered one of the ways in which the initial null hypothesis is false. In other words, we have (partly, perhaps) identified where the initial effect comes from.

On the other hand, if the null hypothesis of a potential follow-up test is *not* implied by the null hypothesis of the initial test, then the truth or untruth of the follow-up null hypothesis does not tell us *anything* about the null hypothesis of the initial test. They are in different domains. For example, suppose we conclude $2\mu_1$ is different from $3\mu_2$. Great, but if we want to know how the statement $\mu_1 = \mu_2 = \mu_3$ might be wrong, it's irrelevant.

If you stick to testing contrasts as a follow-up to a one-way ANOVA, you're fine. This is because if a set of population means are all equal, then any contrast of those means is equal to zero. That is, the null hypothesis of the initial test automatically implies the null hypotheses of any potential follow-up test, and everything is okay. Furthermore, if you try to specify a linear combination that is not a contrast with the `contrast` statement of `proc glm`, SAS will just say something like `NOTE: CONTRAST SOandSO is not estimable` in the log file. There is no other error message or warning; the test just does not appear in your list file.

Usually, when you are testing a contrast you only want to know if it is significantly different from zero. But sometimes you want the actual sample contrast, which is also the estimated population contrast. In this case, use the `estimate` statement. It will give

the sample value of any linear combination of treatment means, along with a t-test for whether the linear combination is significantly different from zero. Here's output from the `estimate` statement in `kenton1b.sas`:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
3Colourvs5Colour	-9.35000000	1.49705266	-6.25	<.0001

Note $t^2 = F$ immediately below.

3.2.5 Converting Tests for Contrasts into Scheffé tests

Here is the output from the first set of `contrast` statements.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
3Colourvs5Colour	1	411.4000000	411.4000000	39.01	<.0001
Cartoon	1	49.7058824	49.7058824	4.71	0.0464
CartoonDepends	1	93.1882353	93.1882353	8.84	0.0095
Overall F	3	588.2210526	196.0736842	18.59	<.0001

By ordinary one-at-a-time F tests, all the tests are significant. But let's treat them as Scheffé tests. To do this, we need the F_{S-crit} critical values for $s = 1$, 2 and 3. Actually we don't need one for $s = 3$, because by (3.1), it's the same as the critical value of the initial test. And in fact, any test of $p - 1$ non-redundant contrasts is equivalent to the initial one-way ANOVA, always.

It's easy to get the F_{S-crit} values from `proc iml`. The following code is written carefully so that you can use it for any problem by just modifying the vales of `numdf` and `dendf` (and maybe `alpha` if you don't want to use 0.05).

```
proc iml;
     title2 'Table of critical values for all possible Scheffe tests';
     numdf = 3;   /* Numerator degrees of freedom for initial test (p-1) */
     dendf = 15;  /* Denominator degrees of freedom for initial test (n-p) */
     alpha = 0.05;
     critval = finv(1-alpha,numdf,dendf);
     zero = {0 0}; S_table = repeat(zero,numdf,1);  /* Make empty matrix */
     /* Label the columns */
     namz = {"Number of Contrasts in followup test"
             "    Scheffe Critical Value"};
     mattrib S_table colname=namz;
     do i = 1 to numdf;
        s_table(|i,1|) = i;
        s_table(|i,2|) = numdf/i *  critval;
     end;
     reset noname; /* Makes output look nicer in this case */
     print "Initial test has"  numdf " and " dendf "degrees of freedom."
           "Using significance level alpha = " alpha;
     print s_table;
```

Here is the output.

```
                  Kenton Oneway Example From Neter et al.                    8
           Table of critical values for all possible Scheffe tests

       Initial test has         3  and         15 degrees of freedom.
                 Using significance level alpha =       0.05


       Number of Contrasts in followup test     Scheffe Critical Value

                                          1                  9.8621463
                                          2                  4.9310732
                                          3                  3.2873821
```

For the one-degree-of-freedom tests (single contrasts) we need $F > 9.86$ for significance. This means `3Colourvs5Colour` is significant, but `Cartoon` and `CartoonDepends` are not, even though `CartoonDepends` has a p-value of 0.0095 by the one-at-a-time test. `ColourCartoon` is also significant, because $22.74 > 4.93$. And of course `Overall F` is significant; it's the initial test.

The last six contrasts are the pairwise differences between means. Their value is that we can convert them easily to Bonferroni or Scheffé follow-up tests. We already did these pairwise comparisons the easy way with the second `means` statement.

```
/* All pairwise comparisons */
contrast '1 vs 2' package  1 -1  0  0;
contrast '1 vs 3' package  1  0 -1  0;
contrast '1 vs 4' package  1  0  0 -1;
contrast '2 vs 3' package  0  1 -1  0;
contrast '2 vs 4' package  0  1  0 -1;
contrast '3 vs 4' package  0  0  1 -1;
```

Here's the output:

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
1 vs 2	1	3.6000000	3.6000000	0.34	0.5677
1 vs 3	1	476.1000000	476.1000000	45.14	<.0001
1 vs 4	1	82.6888889	82.6888889	7.84	0.0135
2 vs 3	1	396.9000000	396.9000000	37.63	<.0001
2 vs 4	1	53.3555556	53.3555556	5.06	0.0399
3 vs 4	1	131.7555556	131.7555556	12.49	0.0030

Sample Question 3.2.1 *What is the Scheffé critical value for the pairwise comparisons? What do we conclude from the tests?*

Answer to Sample Question 3.2.1 *We use the same critical value for any single contrast:* $F = 9.8621463$. *As in the earlier multiple comparisons, we conclude that sales are highest with five colours and no cartoons (that's* μ_3*). The other three treatment combinations are not significantly different from one another.*

Sample Question 3.2.2 *What p-value would we use for a Bonferroni correction to protect all pairwise comparisons at the 0.05 level? The answer is a number.*

Answer to Sample Question 3.2.2 *0.05/6 = 0.00833*

Sample Question 3.2.3 *What if we wanted to also protect for the other three (single) contrasts? Again, the answer is a number.*

Answer to Sample Question 3.2.3 *0.05/9 = 0.0055*

With the Bonferroni method of correction for multiple tests, protecting for more tests results in a more stringent criterion for significance — though in this case the conclusions did not change. With the Scheffé tests, the criterion does not change; an infinite family of tests is protected in advance. Also, remember that for the Bonferroni method to be valid, you have to select the tests before looking at the data. This is not a requirement of the Scheffé tests.

3.3 Two-way ANOVA

To really understand the first three contrast statements, we need to recognize that the 4-category variable `package` actually represents the combination of two independent variables: Number of colours and Presence versus absence of cartoons. That is, we have a two-factor design. The term *factor* is just another way of saying categorical independent variable. A *multi-factor* design is one with more than one factor. A *complete* factorial design is one that includes all combinations of all the independent variables. The term "way" is equivalent to "factor." A three-factor design could also be called a three-way design, and since all the tests will be based on analysis of variance, it can be called a three-way ANOVA.

3.3.1 Main Effects

Consider the following table:

Table 3.1: Population Cell Means and Marginal Means for the Kenton Example

	Cartoon	No Cartoon	
3 Colours	μ_1	μ_2	$\frac{\mu_1+\mu_2}{2}$
5 Colours	μ_3	μ_4	$\frac{\mu_3+\mu_4}{2}$
	$\frac{\mu_1+\mu_3}{2}$	$\frac{\mu_2+\mu_4}{2}$	

In addition to population mean sales for each package design (denoted by μ_1 through μ_4), the table above shows *marginal means* – quantities like $\frac{\mu_2+\mu_4}{2}$, which are obtained by averaging over rows or columns.

If there are differences among marginal means for a categorical independent variable in a two-way (or higher) layout like this, we say there is a *main effect* for that variable. Tests for main effects are of great interest; they can indicate whether, averaging over the values of the other categorical independent variables in the design, whether the independent variable in question is related to the dependent variable.

The population means in the preceding table are estimated by corresponding sample quantities. The numbers in the table below come from the means output of the first `proc glm`.

Table 3.2: Sample Cell and Marginal Means for the Kenton Example

	Cartoon	No Cartoon	
3 Colours	14.6	13.4	14.00
5 Colours	19.5	27.7	23.35
	17.05	20.30	

(14.6+13.4)/2 = 14, and so on.

The custom contrast `3Colourvs5Colour` is for the main effect of number of colours (3 vs. 5). Here is the `contrast` statement from `proc glm`:

```
contrast 3Colourvs5Colour package 1 1 -1 -1;
```

It tests whether $\frac{\mu_1+\mu_2}{2} = \frac{\mu_3+\mu_4}{2}$, because

$$\frac{\mu_1+\mu_2}{2} = \frac{\mu_3+\mu_4}{2} \text{ if and only if } 1\mu_1 + 1\mu_2 - 1\mu_3 - 1\mu_4 = 0.$$

That's also the same thing as asking whether the marginal sample mean for 3 Colours (14) is *significantly* different from the marginal sample mean for 5 colours (23.35). Here's part of the output again:

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
3Colourvs5Colour	1	411.4000000	411.4000000	39.01	<.0001

So the answer is Yes: the main effect for 3 versus 5 colours is statistically significant.

Sample Question 3.3.1 *What do you conclude from the test for the main effect of Number of Colours? Use plain, non-technical language.*

Answer to Sample Question 3.3.1 *There were more sales when the package had 5 colours.*

Similarly, the main effect for presence versus absence of cartoons on the package is tested by asking whether $\frac{\mu_1+\mu_3}{2} = \frac{\mu_2+\mu_4}{2}$. This is the same as asking whether the linear combination

$$L = 1\mu_1 - 1\mu_2 + 1\mu_3 - 1\mu_4$$

is equal to zero. The `contrast` statement gives the coefficients of this linear combination,

```
contrast 'Cartoon'          package 1 -1  1 -1;
```

and the result is

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Cartoon	1	49.7058824	49.7058824	4.71	0.0464

So the main effect for Cartoon is barely significant, with Non-cartoon designs doing better. Treating it as a Scheffé followup to the initial test for differences among the four means, its not significant any more ($F = 4.71 < 9.86$).

3.3.2 Interactions

The two-way design we have been looking at is called a factorial design. In a factorial design, there are two or more categorical independent variables (called factors, in this context) typically with data with for all combinations of the factors being collected. Factorial designs are often found in experimental studies, but not always.

When Sir Ronald Fisher (in whose honour the F-test is named) dreamed up factorial designs, he pointed out that they enable the scientist to investigate the effects of several independent variables at much less expense than if a separate experiment had to be conducted to test each one. In addition, they allow one to ask systematically whether the effect of one independent variable depends on the value of another independent variable. If the effect of one independent variable depends on another, we will say there is an interaction between those variables. We talk about an A "by" B or A x B interaction. An interaction means "it depends."

Consider the table of population means (Table 3.1) again. The effect of Cartoons when the package has three colours is represented by $\mu_1 - \mu_2$. The effect of Cartoons when the package has five colours is represented by $\mu_3 - \mu_4$. Therefore, the interaction of Cartoon by number of colours is a *difference between differences*, and we want to test whether $\mu_1 - \mu_2 = \mu_3 - \mu_4$, or equivalently, whether the linear combination

$$L = 1\mu_1 - 1\mu_2 - 1\mu_3 + 1\mu_4$$

is equal to zero. The `contrast` statement for this is

```
contrast 'CartoonDepends'   package 1 -1 -1  1;
```

yielding

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
CartoonDepends	1	93.1882353	93.1882353	8.84	0.0095

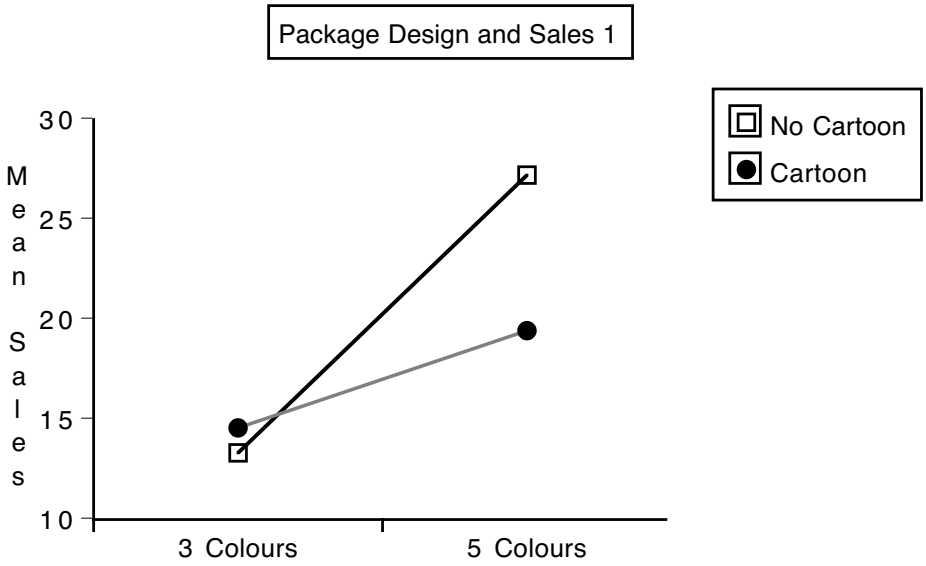
This significant interaction suggests that the effect of cartoons depends on the number of colours. We might also ask whether the effect of number of colours depends upon presence versus absence of cartoons — that is, we might ask whether $\mu_1 - \mu_3 = \mu_2 - \mu_4$. Notice, however, that $\mu_1 - \mu_3 = \mu_2 - \mu_4$ is algebraically equivalent to $\mu_1 - \mu_2 = \mu_3 - \mu_4$. So the two ways of talking about the interaction are the same thing, mathematically. Fortunately, this *always* happens, no matter how big the design. If you express an interaction correctly as a collection of differences between differences, it is algebraically equivalent to all other correct ways of expressing the interaction. Choose the one that is easiest to think about.

Incidentally, $p = 0.0095$ seems impressive, but the test is not significant if it is considered as a Scheffé follow-up. The Scheffé critical value is 9.78, and the F for interaction is 8.84. Too bad, but let's pursue the interaction for instructional purposes.

If an interaction is significant, you should graph it to figure out what it means. Figure 3.1 is an example. It's easiest to use a spreadsheet program like Excel.

Whenever you have an interaction, such graphs will display non-parallel lines. Well actually, when you plot an interaction with real data, the lines will always be at least a

Figure 3.1:



little non-parallel. The question is whether they depart significantly from being parallel. Here, the advantage of 5 colours over 3 is significantly greater for designs without cartoons (unless you are a member of the Scheffé cult, as I am), and we can see it in the graph.

The follow-up tests tell us that there are significantly more sales with 5-colour designs, for both the cartoon and non-cartoon conditions. The interaction tells us that this effect is significantly greater when there are no cartoons.

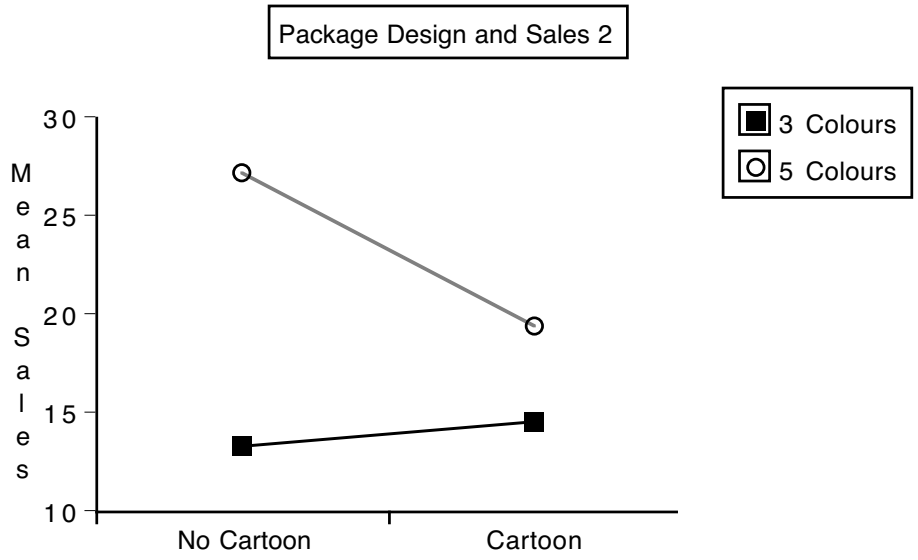
Remember the significant main effect for cartoon? It was just barely significant: $p = 0.0464$. Figure 3.1 shows that this effect is entirely due to the advantage of no-cartoon designs in the 5-colour condition. So here, we have a main effect that's significant, but we really should not interpret it, because of the interaction.

Some texts claim that if you have an interaction, you should *never* interpret the main effects. But look at Figure 3.2, which graphs the same interaction in the other direction (there are only two ways to do it, because it is a two-factor interaction).

The picture that emerges here is that 5-colour designs are better overall, and the advantage is greater in the No-cartoon condition. Here, we can see that it makes sense to interpret both the main effect for number of colours and the interaction. This example shows why I disagree with the advice to never interpret main effects when there is an interaction.

Personally, I like the idea of letting the tests for main effects, interactions and all pairwise differences as follow-ups to the initial oneway ANOVA. I prefer Scheffé, because I dont need to know in advance how many tests Im going to do. I also love the Scheffé tests because of their 100% consistency with the initial tests. If the initial test is non-significant, no Scheffé follow-up can be significant, as a mathematical certainty. And if

Figure 3.2:



the initial test is significant, then there must be a significant Scheffé follow-up.

The result if we adopt this approach here is that when you treat the test for interaction as a follow-up test instead of a one-at-a-time test, it's no longer significant. You are left with a simpler story. Five-colour designs work better than three-colour designs, and designs without cartoons work better in the 5- colour condition.

In general, if you go the multiple comparison route, it's going to make you more conservative. You will draw fewer conclusions. On the other hand, in terms of this particular example, the implications for action (marketing action) are the same whether or not you use multiple comparisons. The Kenton company should use a 5-colour design without cartoons.

3.3.3 Factorial ANOVA the easy way

To test main effects and interactions, it's often more convenient to let `proc glm` set up the contrasts for you. You name all the categorical independent variables (factors) in a `class` statement, and then specify the analysis you want with

`model` Dependent variable(s) = Independent variable(s)

Here's the Kenton example:

```
proc glm;
     title2 "Actually it's a two-way ANOVA";
     class ncolours cartoon;
     model sales = ncolours|cartoon;
```

The model statement could have been

```
model sales = ncolours cartoon ncolours*cartoon;
```

This would specify both main effects (by mentioning the variables), and the interaction. By connecting two or more independent variables with vertical bars, you include all the main effects and interactions involving those variables. Here's the output. First we are given basic information, including the number of independent variables, the values they assume, and the number of cases.

```
Kenton Oneway Example From Neter et al.                    9
     Actually it's a two-way ANOVA

          The GLM Procedure


     Class Level Information

    Class         Levels    Values

    ncolours           2    3 5

    cartoon            2    No Yes

Number of Observations Read          19
Number of Observations Used          19
```

Then we get the ANOVA summary table for an overall, simultaneous test of all the independent variables and their interactions. It is exactly the same as a one-way ANVA on a combination variable consisting of all combinations of the factors. Here, it is 100% equivalent to the one-way ANOVA with `package` as the only independent variable. You can compare the F statistic and p-value from the earlier output. We also get the R^2 (proportion of variation in the dependent variable explained by all the independent variables), the coefficient of variation of the dependent variable (standard deviation divided by the mean, then multiplied by 100 to make it a percent), `Root MSE` (literally the square root of mean squared error), and the mean of the dependent variable.

The GLM Procedure

Dependent Variable: sales Number of Cases Sold

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	588.2210526	196.0736842	18.59	<.0001
Error	15	158.2000000	10.5466667		
Corrected Total	18	746.4210526			

R-Square	Coeff Var	Root MSE	sales Mean
0.788055	17.43042	3.247563	18.63158

Next come two ANOVA summary tables, one for "Type I SS" and another for "Type III SS." We will focus on the Type Three table, because it gives exactly what we get from contrasts. It gives a F-test for each main effect and interaction, and you will notice that the tests statistics and p values are exactly what we got from `contrast` statements on the combination variable `package`.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ncolours	1	452.8654971	452.8654971	42.94	<.0001
cartoon	1	42.1673203	42.1673203	4.00	0.0640
ncolours*cartoon	1	93.1882353	93.1882353	8.84	0.0095

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ncolours	1	411.4000000	411.4000000	39.01	<.0001
cartoon	1	49.7058824	49.7058824	4.71	0.0464
ncolours*cartoon	1	93.1882353	93.1882353	8.84	0.0095

Just for the record, here are some comments about the Type I sums of squares.

- The tests based on Type I sums of squares are sequential. If you are familiar with the multiple regression way of expressing factorial ANOVA, effects in Type I sums of squares are corrected only for preceding effects in the list, so it's sort of hierarchical. In Type III sums of squares, each effect is corrected for all the others; it is as if each effect came last on the list.

- When sample sizes are equal or proportional, the independent variables have zero relationship with one another, and "correcting" one effect for another does nothing. In this case, Type I and Type III sums of squares yield the same tests.
- If there is just one independent variable, Type I and Type III sums of squares are not only identical to each other, they are identical to the initial test. In this case, the SAS output is very redundant.
- I can never remember what Type II sums of squares are.
- There are more types, and in general they all yield identical results when sample sizes are equal. If you do ANOVA using unfamiliar software, always try an example with unequal ns to find out what the software is really doing.

Finally, the `means` statement, whose syntax parallels that of the `model` statement, gives marginal and cell means. I usually ask only for sets of means corresponding to main effects and interactions that are statistically significant.

```
          Kenton Oneway Example From Neter et al.                    11
               Actually it's a two-way ANOVA

                          The GLM Procedure

        Level of              ------------sales------------
        ncolours        N             Mean          Std Dev

        3              10       14.0000000       2.94392029
        5               9       23.7777778       5.19080383


        Level of              ------------sales------------
        cartoon         N             Mean          Std Dev

        No              9       16.7777778       3.45607356
        Yes            10       20.3000000       8.11103501


  Level of     Level of           ------------sales------------
  ncolours     cartoon       N             Mean          Std Dev

  3            No            5       14.6000000       2.30217289
  3            Yes           5       13.4000000       3.64691651
  5            No            4       19.5000000       2.64575131
  5            Yes           5       27.2000000       3.96232255
```

3.3.4 Beyond the Two-by-two Case

Methods for factorial ANOVA and testing interactions can easily be extended to allow for more than two factors and more than two values for a factor. Extension to more than two factors is straightforward. Suppose we had grocery stores of three different sizes (small, medium and large), and within each size, the four package designs were randomly allocated to stores. We would have three factors – store size, number of colours, and presence versus absence of cartoons.

- For each independent variable, averaging over the other two variables would give marginal means – the basis for estimating and testing for main effects.
- Averaging over each of the independent variables in turn, we would have a two-way marginal table of means for the other two variables, and the pattern of means in that table could show a two-way interaction.
- The full three-dimensional table of means would provide a basis for looking at a three-way, or three-factor interaction. The interpretation of a three-way interaction is that the nature of the two-way interaction depends on the value of the third variable. This principle extends to any number of factors, so we would interpret a six-way interaction to mean that the nature of the 5-way interaction depends on the value of the sixth variable.
- Fortunately, the order in which one considers the variables does not matter. For example, we can say that the A by B interaction depends on the value of C, or that the A by C interaction depends on B, or that the B by C interaction depends on the value of A. The translations of these statements into algebra are all equivalent to one another, always. This principle extends to any number of factors.
- As you might imagine, as the number of factors becomes large, interpreting higher-way interactions – that is, figuring out what they mean – becomes more and more difficult. For this reason, sometimes the higher-order interactions are deliberately omitted from the full model in big experimental designs; if they are omitted, they can never be tested. Is this reasonable? Most of my answers are just elaborate ways to say I don't know.

More than two values for an independent variable Regardless of how many factors we have, or how many levels there are in each factor, we could always form a combination variable – that is, a single categorical independent variable whose values represent all the combinations of independent variable values in the factorial design. We have seen that in a two-by-two design, the tests for both main effects and the interaction resolve themselves into tests for single contrasts – contrasts of the means of the combination variable. When independent variables have more than two values, the same thing is true, except that tests for main effects and interactions appear as test for collections of contrasts on the combination variable.

Bibliography

[1] Cody, R. P. and Smith, J. K. (1991). *Applied statistics and the SAS programming language.* (4th Edition) Upper Saddle River, New Jersey: Prentice-Hall.

[2] Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings.* New York: Rand McNally.

[3] Fisher, R. A. (1925) *Statistical methods for research workers.* London: Oliver and Boyd.

[4] Moore, D. S. and McCabe, G. P. (1993). *Introduction to the practice of statistics.* New York: W. H. Freeman.

[5] Neter, J., Kutner, M. H., Nachhtscheim, C. J. and Wasserman, W. (1996) *Applied linear statistical models.* (4th Edition) Toronto: Irwin.

[6] Roethlisberger, F. J. (1941). *Management and morale.* Cambridge, Mass.: Harvard University Press.

[7] Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Croft.

[8] Rosenthal, R. and Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectation and pupils' intellectual development.* New York: Holt, Rinehart and Winston.