15.0 Fractional Factorial Designs

- Answer Questions; Review
- Computation in R
- Fractional Factorials
- Example
- More Exotic Designs

15.1 ANOVA Computation in R

Example 1: Three levels of drug were administered to 18 subjects and their subsequent alertness was measured. A one-way analysis of variance in R uses the aov command as follows:

```
aov.ex1= aov(Alertness~Dosage,data=ex1)
```

It is important to note the order of the arguments. The first argument is always the response variable (Alertness). It is followed by the tilde symbol (∼) and the factor variable(s). The final argument for aov is the name of the data structure that is being analyzed. And aov.ex1 is the name of the structure you want the analysis to store.

R is freeware, and there is on-line help. You can do many more things than what I show, but there is a learning curve.

This general format will hold true for all ANOVAs you will conduct. The results of the ANOVA can be seen in this example:

```
#tell where the data come from
datafilename="http://personality-project.org/R/datasets/⊕
R.appendix1.data"
#read the data
data.ex1=read.table(datafilename,header=T)
#do the analysis
aov.ex1 = aov(Alertness~Dosage,data=data.ex1)
#show the table
summary(aov.ex1)
```

These commands produce this output:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dosage	2	426.25	213.12	8.7887	0.002977
Residuals	15	363.75	24.25		

We see that the dosage effect is highly significant. Since the sample size is small, this probably means that the magnitude of the effect is scientifically significant too.

You can use the url given on the previous page to run this example.

The main website for R is `http://wwww.r-project.org/` and you can find documentation, FAQs, examples, and so forth there. Also, the graduate student TAs in the statistics help room should all be fluent in R.

For two-way ANOVA, consider an experiment in which alertness of males and females is measured after using one of two possible dosages of a drug.

There are two independent variables, separated by an asterisk. The * indicates that the interaction between the two factors is interesting and should be analyzed; otherwise, use a +.

```
datafilename="http://personality-project.org/r/datasets/⊕
R.appendix2.data"
#read the data
data.ex2=read.table(datafilename,header=T)
#show the data
data.ex2
#do the analysis
aov.ex2 = aov(Alertness~Gender*Dosage,data=data.ex2)
#show the summary table
summary(aov.ex2)
```

The command “data.ex2” produces:

Obs	Gender	Dose	Alert	Obs	Gender	Dose	Alert
1	m	a	8	9	f	a	15
2	m	a	12	10	f	a	12
3	m	a	13	11	f	a	22
4	m	a	12	12	f	a	14
5	m	b	6	13	f	b	15
6	m	b	7	14	f	b	12
7	m	b	23	15	f	b	18
8	m	b	14	16	f	b	22

The command "aov.ex2 = aov(Alertness~Gender*Dosage,data=data.ex2)" followed by "summary(aov.ex2)" produces:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	76.562	76.562	2.9518	0.1115
Dosage	1	5.062	5.062	0.1952	0.6665
Gender:Dosage	1	0.063	0.063	0.0024	0.9617
Residuals	12	311.250	25.938		

None of the P-values (significance probabilities) is significant. In this experiment neither dosage nor gender (nor their interactions) had any effect on alertness. (Note that the P-value for the interaction is actually quite large—this is probably due to non-normality.)

15.2 Fractional Factorial Designs

A **factorial design** is one in which every possible combination of treatment levels for different factors appears.

The two-way ANOVA with interaction we considered was a factorial design. We had n observations on each of the IJ combinations of treatment levels.

If there are, say, a levels of factor A, b levels of factor B, c levels of factors C, then a factorial design requires at least abc observations, and more if one wants to estimate the three way interaction among the factors. This can get expensive when experiments have many different factors.

To keep experimental costs in line, one approach is to use **fractional factorial** designs. In these, one does not take measurements upon every possible combination of factor levels, but only upon a very carefully chosen few.

These few are selected to ensure that the main effects and low-order interactions can be estimated and tested, at the expense of high-order interactions.

The scientific intuition is that it is unlikely for there to be complex interactions among many different factors; instead, there are probably only main effects and a few low-order interactions.

Thus one might design the collection in a fractional factorial so that all main effects and two-way interactions can be tested, but not three-way or higher interactions.

Recall the three-way ANOVA (with factors A, B, C having n observations at all combinations of levels). The general model is

$$Y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_j + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijk\ell}$$

where everything is a usual and $(\alpha\beta\gamma)_{ijk}$ is the three-way interaction.

If the three-way interaction is significant, then we say that all the lower-order interactions are significant. (Sometimes the calculation finds estimates of these as zero, but fundamentally high-order interactions imply lower-order interactions are just masked.)

Obviously, this model can be used with fixed, random, or mixed effect designs. The model is the same for all, but the ratios taken in the ANOVA table depend upon which factors are fixed and which are random.

You can always estimate the highest order interaction as:

$$(\hat{\alpha\beta\gamma})_{ijk} = \bar{Y}_{ijk.} - [\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_k + (\hat{\alpha\beta})_{ij} + (\hat{\alpha\gamma})_{ik} + (\hat{\beta\gamma})_{jk}]$$

where the estimates are found in the usual way:

$$\begin{aligned} \hat{\mu} &= \bar{Y}_{....} \\ \hat{\alpha}_i &= \bar{Y}_{i...} - \bar{Y}_{....} \\ (\hat{\alpha\beta})_{ij} &= \bar{Y}_{ij..} - \bar{Y}_{i...} - \bar{Y}_{.k..} + \bar{Y}_{....} \end{aligned}$$

and so forth.

N.B.: These simple forumla apply in general. The SS formula get a bit more complicated when there are unequal numbers of observations for each combination of factor levels (i.e., there is no fixed n).

A special kind of factorial design are the 2^k factorials. In these, each of the k factors have exactly 2 levels, so there are 2^k different combinations of treatment levels.

For a full factorial, one would need a minimum of 2^k observations, and even that would not allow enough degrees of freedom for the error term.

$$2^k - 1 = k(2-1) + \binom{k}{2}(2-1)(2-1) + \cdots + \binom{k}{k}(2-1)(2-1)\ldots(2-1).$$

This is based on a standard combinatorial identity due to Pascal:

$$2^k = \sum_{i=0}^{k} \binom{k}{i}.$$

Typically, one uses the highest order interaction as if it were an error term, or uses a normal probability plot.

Pascal corresponded with Fermat to develop the laws of probability, and developed Pascal's Wager. He was also a great prose stylist, and pioneered the moder understanding of atmospheric pressure.

The following table shows the data for a full 2^3 factorial design. Note that the signs in each interaction column can be found by multiplying the signs in corresponding main-effect columns.

run	A	B	C	AB	AC	BC	ABC	obs
1	-	-	-	+	+	+	-	Y_{111}
2	+	-	-	-	-	+	+	Y_{211}
3	-	+	-	-	+	-	+	Y_{121}
4	+	+	-	+	-	-	-	Y_{221}
5	-	-	+	+	-	-	+	Y_{112}
6	+	-	+	-	+	-	-	Y_{212}
7	-	+	+	-	-	+	-	Y_{122}
8	+	+	+	+	+	+	+	Y_{222}

The main effect due to factor A is the average difference between the high and low levels of factor A, or:

$$\alpha = \frac{1}{2}\left(\frac{Y_1 + Y_4 + Y_6 + Y_8}{4} - \frac{Y_1 + Y_3 + Y_5 + Y_7}{4}\right)$$

and similarly for the main effects of B and C. So, as per the signs in the table,

$$\alpha = \frac{1}{8}(-Y_1 + Y_2 - Y_3 + Y_4 - Y_5 + Y_6 - Y_7 + Y_8).$$

The AB interaction is half the difference between the main effect of factor A at the high level of factor B and that at the low level of factor B, or:

$$(\alpha\beta) = \frac{1}{2}\left[\frac{1}{2}\left(\frac{Y_4 + Y_8}{2} - \frac{Y_3 + Y_7}{2}\right) - \frac{1}{2}\left(\frac{Y_2 + Y_6}{2} - \frac{Y_1 + Y_5}{2}\right)\right]$$

As per the signs in the table,

$$(\alpha\beta) = \frac{1}{8}(Y_1 - Y_2 - Y_3 + Y_4 + Y_5 - Y_6 - Y_7 + Y_8).$$

The ABC interaction is half the difference between the two-factor interaction AB at the high and low levels of factor C, or

$$(\alpha\beta\gamma) = \frac{1}{2}\left[\frac{1}{2}\left(\frac{Y_8 - Y_7}{2} - \frac{Y_6 - Y_5}{2}\right) - \frac{1}{2}\left(\frac{Y_4 - Y_3}{2} - \frac{Y_2 - Y_1}{2}\right)\right]$$

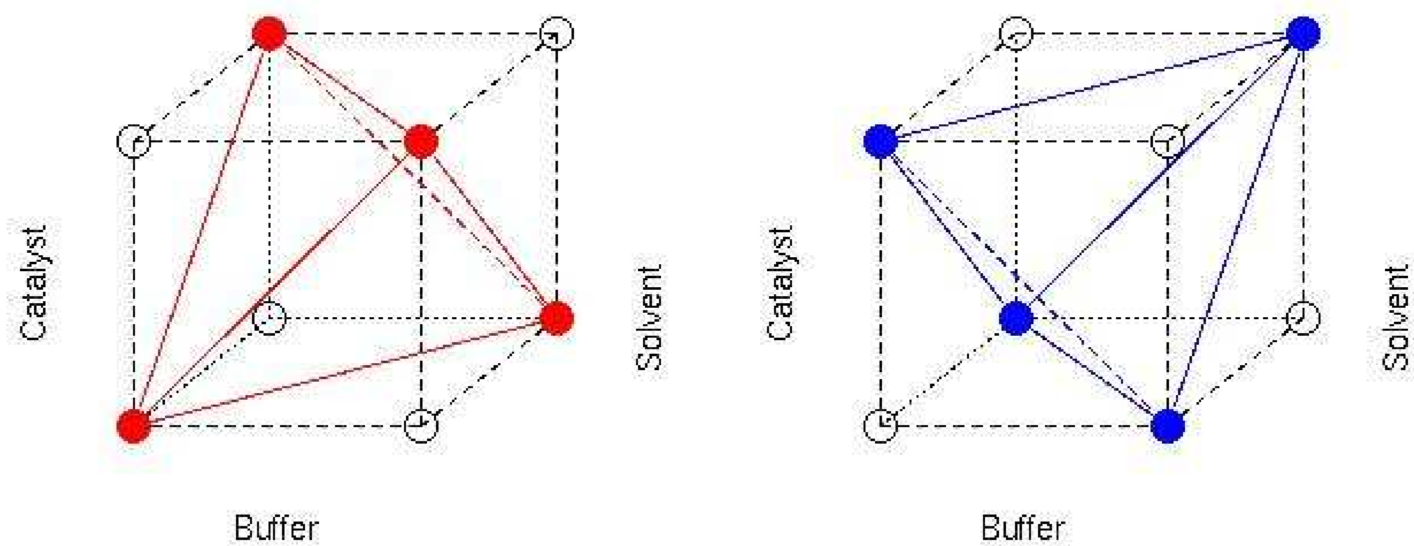
Note that

$$(\alpha\beta\gamma) = \frac{1}{8}(-Y_1 + Y_2 + Y_3 - Y_4 + Y_5 - Y_6 - Y_7 + Y_8)$$

as per the signs in the previous table.

These relationships give us an automatic way to calculate the main effects (and, implicitly, the sums of squares) for inference.

But 2^k observations gets expensive. Instead, one can carefully select half that number so as to still permit estimation of main effects and low-order interactions.



Consider the 2^{3-1} fractional factorial design (The previous figure gave two illustrations.)

run	A	B	C	AB	AC	BC	ABC	obs
1	-	-	+	+	-	-	+	Y_{112}
2	+	-	-	-	-	+	+	Y_{211}
3	-	+	-	-	+	-	+	Y_{121}
4	+	+	+	+	+	+	+	Y_{222}

Note that because we have taken only half of the 8 observations needed for a full factorial, some of the columns have identical entries.

Columns that have identical entries correspond to effects that are **confounded** or **aliased**.

In order to estimate the effects, note that:

$$\begin{aligned}
\frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4) &= \mu + (\alpha\beta\gamma) \\
\frac{1}{4}(-Y_1 + Y_2 - Y_3 + Y_4) &= \alpha + (\beta\gamma) \\
\frac{1}{4}(-Y_1 - Y_2 + Y_3 + Y_4) &= \beta + (\alpha\gamma) \\
\frac{1}{4}(Y_1 - Y_2 - Y_3 + Y_4) &= \gamma + (\alpha\beta)
\end{aligned}$$

Thus the estimate of the mean is confounded with the three-way interaction, the estimate of the A effect is confounded with the BC interaction, the estimate of the B effect is confounded with the AC interaction, and the estimate of the C effect is confounded with the AB interaction.

If one assumes that there are no interactions, then one can make tests about the main effects, or use a normal probability plot.

Note that we write 2^{k-p} to denote a fractional factorial design in which each factor has 2 levels, there are k factors, and we are taking a $1/2^p$ fraction of the number of possible factor level combinations.

In order to construct a fractional factorial that deliberately confounds pre-selected factors, one needs to use a **generator**.

The generator uses the fact that squaring the entries in any given column gives a column of ones, which can be thought of as an identity element I. If we want to confound the A effect with the BC interaction, then that is equivalent to declaring $A * BC = ABC = I$. It follows that $B = BI = B * ABC = AC$, so B is confounded with the AC interaction. Similarly, C is confounded with AB, and the overall mean (I) is confounded with ABC.