

Here are data on the number of plants that emerged for three legume species.

Species		Soil Type		
		Silt loam	Sand	Clay
Alfalfa	None	89	95	22
	Treated	92	90	72
Red Clover	None	84	96	56
	Treated	92	97	68
Sweet clover	None	51	66	17
	Treated	59	73	40

This is a single replicate of a  $2 \times 3^2$  factorial.

```
Cmd> Y <- vector(89,92,84,92,51,59, 95,90,96,97,66,73,\
22,72,56,68,17,40)
Cmd> species <- factor(rep(run(3),rep(2,3)), 3))
Cmd> fungicide <- factor(rep(run(2),9))
Cmd> soil <- factor(rep(run(3),rep(6,3)))
```

This is based on a data set in Steele and Torrie. The original was a RCB design with 3 replicates. The values here are the rounded treatment means.

Displays for Statistics 5303

Lecture 19

October 18, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)

612-625-1024 (Minneapolis)

Class Web Page

<http://www.stat.umn.edu/~kb/classes/5303>

© 2002 by Christopher Bingham

## Let's do a complete 3 factor ANOVA.

```
Cmd> anova("y=(species+soil+fungicide)^3",Fstat:T)
Model used is y=(species+soil+fungicide)^3
```

	DF	SS	MS	F	P-value
CONSTANT	1	88060	88060	undefined	undefined
species	2	3320.8	1660.4	undefined	undefined
soil	2	5440.4	2720.2	undefined	undefined
fungicide	1	636.06	636.06	undefined	undefined
species.soil	4	225.22	56.306	undefined	undefined
species.fungicide	2	62.111	31.056	undefined	undefined
soil.fungicide	2	629.78	314.89	undefined	undefined
species.soil.fungicide	4	364.56	91.139	undefined	undefined
ERROR1	0	0	undefined	undefined	undefined

**Oops! Major problem.** There are no degrees of freedom for error. You can't do the usual F-tests.

Unless you can make some assumptions, there is little you can do. A typical assumption is that some or all high order interactions are 0.

This often is OK since in many fields, important high order interactions are rare. Even when there is a high order interaction, its effects are small compared to other effects.

When order interactions can be presumed unimportant, you leave them out of the model. In this case we will assume

$$\alpha\beta\gamma_{ijk} = 0, \text{ all } i, j \text{ and } k.$$

```
Cmd> anova("y=(species+soil+fungicide)^2",Fstat:T)
Model used is y=(species+soil+fungicide)^2
```

	DF	SS	MS	F	P-value
CONSTANT	1	88060	88060	966.21823	6.3828e-06
species	2	3320.8	1660.4	18.21823	0.0097853
soil	2	5440.4	2720.2	29.84700	0.0039439
fungicide	1	636.06	636.06	6.97897	0.057474
species.soil	4	225.22	56.306	0.61780	0.67389
species.fungicide	2	62.111	31.056	0.34075	0.73005
soil.fungicide	2	629.78	314.89	3.45504	0.13442
ERROR1	4	364.56	91.139		

Now we have an error term, albeit with only 4 degrees of freedom. Note that the error SS and DF are identical with  $SS_{ABC}$  and  $DF_{ABC}$  in the first analysis.

Note that no two-way interaction is significant. If you knew they were really 0 you could omit them from the model, too. Their SS would then merge into  $SS_E$ , giving 12 DF for error.

This process is sometimes known as **pooling**. It is very tempting to pool so that you have more error degrees of freedom. In general, pooling should be avoided unless it's really needed.

Oehlert gives a good rule of thumb:

Consider pooling *only* when

- $DF_{\text{error}} \leq 10$
- $F < 2$  for any term to be pooled.

Here's the ANOVA again

```
Cmd> anova("y=(species+soil+fungicide)^2", Fstat:T)
Model used is y=(species+soil+fungicide)^2
```

	DF	SS	MS	F	P-value
CONSTANT	1	88060	88060	966.21823	6.3828e-06
species	2	3320.8	1660.4	18.21823	0.0097853
soil	2	5440.4	2720.2	29.84700	0.0039439
fungicide	1	636.06	636.06	6.97897	0.057474
species.soil	4	225.22	56.306	0.61780	0.67389
species.fungicide	2	62.111	31.056	0.34075	0.73005
soil.fungicide	2	629.78	314.89	3.45504	0.13442
ERROR1	4	364.56	91.139		

By the rule of thumb you can consider pooling species.soil and species.fungicide, but not soil.fungicide, since its  $F = 3.455 > 2$ .

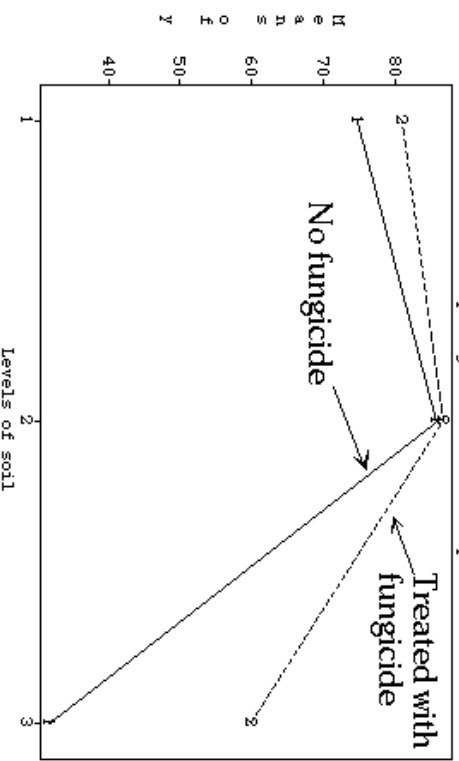
Redo ANOVA omitting species.soil and species.fungicide:

```
Cmd> anova("y = species + soil + Fungicide + soil.fungicide", \
Fstat:T)
Model used is y = species + soil + fungicide + soil.fungicide
```

	DF	SS	MS	F	P-value
CONSTANT	1	88060	88060	1350.84455	5.2892e-12
species	2	3320.8	1660.4	25.47043	0.00011898
soil	2	5440.4	2720.2	41.72831	1.4027e-05
fungicide	1	636.06	636.06	9.75712	0.010807
soil.fungicide	2	629.78	314.89	4.83041	0.03404
ERROR1	10	651.89	65.189		

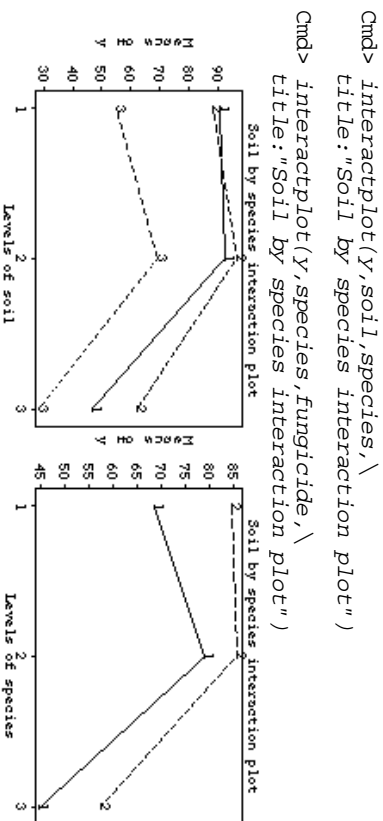
It's just as well we didn't pool soil.fungicide; it's now significant.

```
Cmd> interactplot(y,soil,fungicide, \
title:"Soil by fungicide interaction plot")
Soil by fungicide interaction plot
```



The average effect of fungicide is positive except on silt loam soils.

Here's what the pooled interactions looked like.



Not precisely parallel, but not far off.

### Example 8.9

```

Cmd> data <- read( "", "exmpl8.8")
exmpl8.8      54      5
) A data set from Oehlert (2000) \emph{A First Course in Design
) and Analysis of Experiments}, New York: W. H. Freeman.
) Data originally from an example (homework problem?) of Barry.
) Margolin Table 8.7, p. 187
) Columns are paging algorithm, initialization sequence, program
) size (small, medium, or large), RAM allocation (large, medium,
) or small), and number of page faults.
Read from file "TPl:Stat5303:Data:Oech08.dat"

Cmd> makecols(data, algo, seq, size, ram, faults)

Cmd> algo <- factor(algo); seq <- factor(seq)
Cmd> size <- factor(size); ram <- factor(ram)

```

This is a  $3 \times 3 \times 3 \times 2 = 3^3 \times 2$  factorial.

There is no replication. The number of treatments = number of cases.

```

Cmd> 2*3^3      # number of distinct treatments
(1)
Cmd> list(algo, seq, size, ram, faults)
algo      REAL      54      FACTOR with 2 levels
faults    REAL      54      FACTOR with 3 levels
ram       REAL      54      FACTOR with 3 levels
seq       REAL      54      FACTOR with 3 levels
size      REAL      54      FACTOR with 3 levels

Cmd> 2*3^3      # number of distinct treatments
(1)
Cmd> logfaults <- log10(faults) # analyze logs
Model used is logfaults=(algo+seq+size+ram)^4, print:F)
NOTE: Some results are in variables SS, DF, and RESIDUALS
Use coeffs(), secoeffs(), or modelinfo() to retrieve other
results

Cmd> DF[-run(12)] # last 5 DF; last is 0
algo.seq.ram.algo.size.ramseq.size.ram.algo.seq.size.ram  ERROR1
4      4      4      8      8      8      8      8      8      0

Cmd> Ssl-run(12)]
algo.seq.ram.algo.size.ramseq.size.size.ram.algo.seq.size.ram  ERROR1
0.0049053 0.00075474 0.19844 0.0042996
Cmd> anova("logfaults=(algo+seq+size+ram)^3", pval:T)
Model used is logfaults=(algo+seq+size+ram)^3

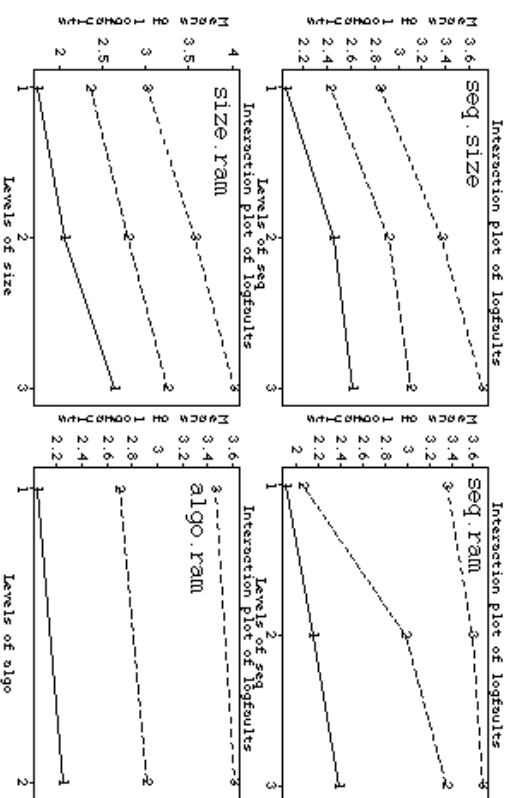
```

	CONSTANT	algo	seq	size	ram	algo.seq	algo.size	algo.ram	seq.size	seq.ram	size.ram	algo.seq.size	algo.seq.ram	algo.size.ram	seq.size.ram
DF	1	1	2	2	2	2	2	2	4	4	4	4	4	4	8
SS	433.53	0.47188	4.6473	7.8635	17.484	0.0033265	0.0041899	0.011324	0.15635	1.7938	0.095118	0.0027469	0.0049053	0.00075474	0.19844
MS	433.53	0.47188	2.3236	3.9318	8.7419	0.0016633	0.002095	0.0056621	0.039088	0.44845	0.023779	0.00068673	0.0012263	0.00018868	0.024805
F	8.0664e+05	877.98573	4323.40535	7315.55961	16265.42823	3.09471	3.89794	10.53504	72.72785	834.39270	44.24470	0.00068673	2.28176	0.35107	46.15350
P-value	1.8243e-09	7.3001e-13	8.9187e-14	3.6539e-15	0.10104	0.065794	0.0057356	2.5105e-06	1.6316e-10	1.6887e-05	0.35476	0.00068673	0.14907	0.83645	6.7256e-06



## Here are some interaction plots:

```
Cmd> interactplot(logfaults,seq, size)
Cmd> interactplot(logfaults,seq, ram)
Cmd> interactplot(logfaults,size, ram)
Cmd> interactplot(logfaults,algo, ram)
```



The degree of non-parallelism reflects the size of the ANOVA SS. Most non-parallel are seq. ram and algo. ram which have SS = 1.794 and 0.0113. All are significant when  $\alpha = .05$ .

For 3 or more factors, each line corresponds to a combination of factors labelled like effects in `sidebyside()`.