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

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

		Soil Type		
Species		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 a single replicate of a 2×3^2 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(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.

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
```

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 $\alpha\beta\gamma_{ijk} = 0$, all i, j 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

- $DF_{error} \le 10$
- 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
```

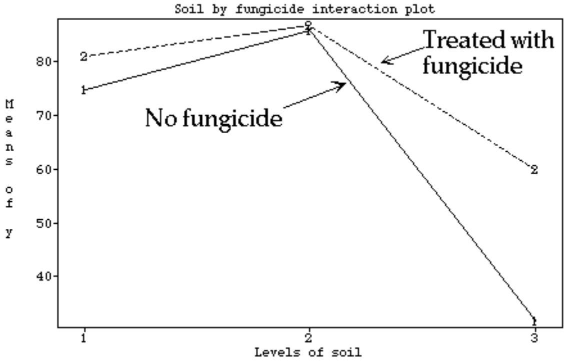
By the rule of thumb you can consider pooling **species.soil** and **species.fungicide**, but not **soil.fungicide**, since its F = 3.455 > 2.

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.

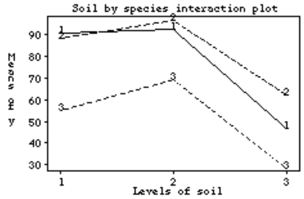
```
Cmd> interactplot(y,soil,fungicide,\
     title:"Soil by fungicide interaction plot")
```

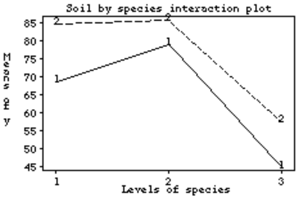


The average effect of fungicide is positive except on silt loam soils.

Here's what the pooled interactions looked like.

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

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





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 "TP1: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 $3\times3\times3\times2 = 3^3\times2$ factorial.

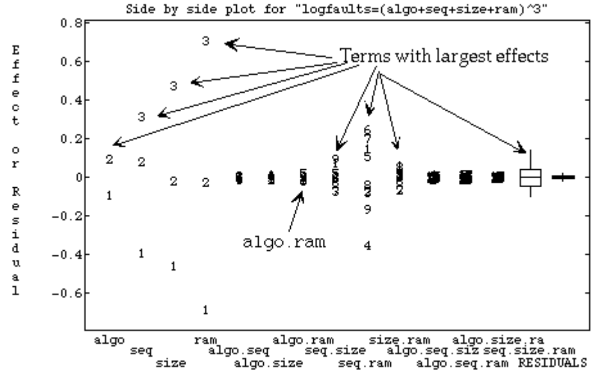
```
Cmd> list(algo,seq,size,ram,faults)
algo          REAL    54    FACTOR with 2 levels
faults        REAL    54
ram           REAL    54    FACTOR with 3 levels
seq           REAL    54    FACTOR with 3 levels
size          REAL    54    FACTOR with 3 levels

Cmd> 2*3^3     # number of distinct treatments
(1)          54
```

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

```
Cmd> logfaults <- log10(faults) # analyze logs

Cmd> anova("logfaults=(algo+seq+size+ram)^4",print:F)
Model used is logfaults=(algo+seq+size+ram)^4
NOTE: Some results are in variables SS, DF, and RESIDUALS
      Use coefs(), secoefs(), or modelinfo() to retrieve other
      results

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

Cmd> SS[-run(12)]
algo.seq.ramalgo.size.ramseq.size.ramalgo.seq.size.ram  ERROR1
   0.0049053  0.00075474     0.19844   0.0042996             0
Cmd> anova("logfaults=(algo+seq+size+ram)^3",pval:T)
Model used is logfaults=(algo+seq+size+ram)^3
              DF         SS          MS           F     P-value
CONSTANT       1     433.53      433.53  8.0664e+05           0
algo           1    0.47188     0.47188   877.98573  1.8243e-09
seq            2     4.6473      2.3236  4323.40535  7.3001e-13
size           2     7.8635      3.9318  7315.55961  8.9187e-14
ram            2     17.484      8.7419 16265.42823  3.6539e-15
algo.seq       2  0.0033265   0.0016633     3.09471     0.10104
algo.size      2  0.0041899    0.002095     3.89794    0.065794
algo.ram       2   0.011324   0.0056621    10.53504   0.0057356
seq.size       4    0.15635    0.039088    72.72785  2.5105e-06
seq.ram        4     1.7938     0.44845   834.39270  1.6316e-10
size.ram       4   0.095118    0.023779    44.24470  1.6887e-05
algo.seq.size  4  0.0027469  0.00068673     1.27776     0.35476
algo.seq.ram   4  0.0049053   0.0012263     2.28176     0.14907
algo.size.ram  4 0.00075474  0.00018868     0.35107     0.83645
seq.size.ram   8    0.19844    0.024805    46.15350  6.7256e-06
ERROR1         8  0.0042996  0.00053745
```

Here is a side-by-side plot of all the effects and residuals.

```
Cmd> sidebyside(termlaby:vector(1,0,0,1,0,0,1,0,0,1,0,0,\
 1,0,0)*(-.8) +vector(0,1,0,0,1,0,0,1,0,0,1,0,0,1,0)*(-.85) +\
vector(0,0,1,0,0,1,0,0,1,0,0,1,0,0,1)*(-.9))
```



This plot is helpful in deciding the real importance of different effects

For example, although **algo.ram** is significant, its contribution to the means is clearly less than most other significant terms.

Type **help(sidebyside)** for more information on **sidebyside()**.

The numbers identify the particular effect. For a main effects, 1, 2, 3, ... just are the factor level.

Two factor effects can be displayed as a matrix with rows corresponding to the first factor and columns to the second. The numbers count down columns:

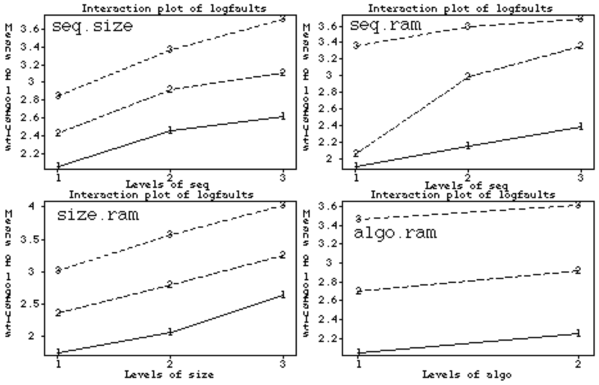
	B level 1	B level 2	B level 3
A level 1	1	4	7
A level 2	2	5	8
A level 3	3	6	9

With three factors, the first changes fastest, the second next and the 3rd slowest

	C1			C2			C3		
	B1	B2	B3	B1	B2	B3	B1	B2	B3
A1	1	4	7	10	13	16	19	22	25
A2	2	5	8	11	14	17	20	23	26
A3	3	6	9	12	15	17	21	24	27

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 α = .05.
For 3 or more factors, each line corresponds to a combination of factors labelled like effects in **sidebyside()**.