Displays for Statistics 5303

Lecture 34

November 25, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)
612-625-1024 (Minneapolis)

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303

Stemplot of exam grades

```
Cmd> stemleaf(mt2,8,stat:T,outlier:F)
n=21, Min=16,Q1=55,M=64,Q3=71,Max=79
    1     1|6
    1     2|
    3     3|49
    4     4|1
    9     5|55999
  ( 6)    6|345588
    6     7|122389

          1|1 represents 11
            Leaf digit unit = 1
```

Latin Squares

In a **CRD** (completely randomized design) there is no attempt to segregate out sources of variability so they don't affect comparisons of treatments..

There are **no restrictions** on either the number g of treatments and the numbers n_i of replications per treatments.

Model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i = 1,\ldots,g,\ j = 1,\ldots,n_i$$

Contrast with weights $\{c_i\}$, $\sum_i c_i = 0$

$$\sum_i c_i \overline{y}_{i\bullet} = (\sum_i c_i)\mu + \sum_i c_i \alpha_i + \sum_i c_i \overline{\varepsilon}_{i\bullet}$$
$$= \sum_i c_i \alpha_i + \sum_i c_i \overline{\varepsilon}_{i\bullet}$$

This has variance $V(\sum_i c_i \overline{y}_{i\bullet}) = \sum_i c_i^2 \sigma^2 / n_i$

In particular

$$\overline{y}_{i_1\bullet} - \overline{y}_{i_2\bullet} = \alpha_{i_1} - \alpha_{i_2} + \overline{\varepsilon}_{i_1\bullet} - \overline{\varepsilon}_{i_2\bullet}$$

with variance $\sigma^2(1/n_{i_1} + 1/n_{i_2})$.

In a **RCB** (randomized complete block) design, you try to segregate out *one* source of variability - the among blocks variability - so that it doesn't affect comparisons among treatments.

There are no restrictions on the number ot treatments, or the number of replications, but all treatments are repeated equally often so $n_1 = n_2 = \ldots = n_g = r$.

Model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},\ u = 1,\ldots,\ g,\ j = 1,\ldots,r$$

Contrasts among treatments are not affected by large or small block effects.

$$\overline{y}_{i\bullet} = \mu + \alpha_i + \overline{\beta}_{\bullet} + \overline{\varepsilon}_{i\bullet}$$ **Treatment mean**

$$\sum_i c_i \overline{y}_{i\bullet} = \sum_i c_i \mu + \sum_i c_i \alpha_i + \sum_i c_i \overline{\beta}_{\bullet} + \sum_i c_i \overline{\varepsilon}_{i\bullet}$$
$$= \sum_i c_i \alpha_i + \sum_i c_i \overline{\varepsilon}_{i\bullet}$$ **Contrast**

$$V(\sum_i c_i \overline{y}_{i\bullet}) = (\sum_i c_i^2)\sigma^2/r$$

In particular, $\overline{y}_{i_1\bullet} - \overline{y}_{i_2\bullet} = \alpha_{i_1} - \alpha_{i_2} + \overline{\varepsilon}_{i_1\bullet} - \overline{\varepsilon}_{i_2\bullet}$

with $V(\overline{y}_{i_1\bullet} - \overline{y}_{i_2\bullet}) = 2\sigma^2/r$.

When blocks are random and there are random interaction effects $\alpha\beta_{ij}$, you need to replace ε_{ij} by $\tilde{\varepsilon}_{ij} \equiv \alpha\beta_{ij} + \varepsilon_{ij}$ and σ^2 by $\sigma_{\alpha\beta}^2 + \sigma^2$.

In **LS design** (Latin Square), you try to segregate out *two* sources of variability.

You group the EU's in two ways, equal sized "rows" and equal sized "columns", so that

- Each row is homogeneous with respect to one source of variability
- Each column is homogeneous with respect to another variability source.

The defining features of Latin squares are that

- Each row is a complete replicate
- Each column is a complete replicate.

That is, every treatment appears once in every row and once in every column.

Statistics 5303 Lecture 34 November 25, 2002

Here is an example of a 6 by 6 Latin Square for treatments A, B, ..., F

	1	2	3	4	5	6
1	B	C	A	D	E	F
2	E	A	C	F	B	D
3	A	E	F	C	D	B
4	D	F	E	B	A	C
5	F	B	D	A	C	E
6	C	D	B	E	F	A

One usage is in a "cross over" design in which the same subject can, at different times, be given each treatment. Even when there is no "carry over" effect, there may be an *order* effect. For example, the subject might respond differently to the first treatment given, no matter which it was.

If you assign subjects to rows, times or position in the ordering to columns, and treatments to letters, then a Latin square design segregates among subject variation and variation due to order.

Restrictions

The number of replicates = number of treatments, that is $r = g$.

Model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk},\ 1 \le i, j, k \le g$$

but only g^2 out of the possible g^3 combinations of i, j, and k are present.

A LS design yields data that is balanced **for main effects**, that is each treatment, each row and each column is replicated g times. Not all combinations of row, column and treatment levels occur so it's not *completely* balanced.

Contrast among treatment means:

$$\sum_i c_i \overline{y}_{i\bullet\bullet} = (\sum_i c_i)\mu + \sum_i c_i\alpha_i + (\sum_i c_i)\overline{\beta}_\bullet + (\sum_i c_i)\overline{\gamma}_\bullet + \sum_i c_i \overline{\varepsilon}_{i\bullet\bullet}$$

$$= \sum_i c_i\alpha_i + \sum_i c_i \overline{\varepsilon}_{i\bullet\bullet}$$

This is unaffected by row and column effects.

$$V(\sum_i c_i \overline{y}_{i\bullet}) = (\sum_i c_i^2)\sigma^2/g$$

Here is analysis of a field experiment comparing 6 crosses of a strain of corn. Rows and columns were actual East-West and North-South strips in a field.

```
Cmd> data <- read("","bliss11_13")
bliss11_13      36      4 format
) Data from Table 11.13, pp. 309 of Bliss, Statistics in Biology
) Yields in bu/acre of 6 crosses of the Hy strain of corn from a
) field trial at Windsor, Conn, arranged in a 6 by 6 Latin
) Square
) Col. 1: Row number (1-6)
) Col. 2: Column number (1-6)
) Col. 3: Cross (1=A, 2=B, 3=C, 4=D, 5=E, 6=F)
) Col. 4: Yield, bu/acre
Read from file "TP1:DataFromStPaul:Bliss:Bliss.mat"

Cmd> makecols(data,row,col,cross,y)

Cmd> row <- factor(row); col <- factor(col)

Cmd> cross <- factor(cross)

Cmd> print(format:"6.0f",\
    matrix(cross,6,labels:structure("Row ", "Col ")))
MATRIX:     Layout of crosses, g = 6
       Col 1  Col 2  Col 3  Col 4  Col 5  Col 6
Row 1      2      3      1      4      5      6
Row 2      5      1      3      6      2      4
Row 3      1      5      6      3      4      2
Row 4      4      6      5      2      1      3
Row 5      6      2      4      1      3      5
Row 6      3      4      2      5      6      1

Cmd> print(format:"5.0f",tabs(y,col,count:T))
VECTOR:     Equal counts per column
(1)     6     6     6     6     6     6

Cmd> print(format:"5.0f",tabs(y,row,count:T))
VECTOR:     Equal counts per row
(1)     6     6     6     6     6     6

Cmd> print(format:"5.0f",tabs(y,cross,count:T))
VECTOR:     Equal counts per cross
(1)     6     6     6     6     6     6
```

Here are the treatment means $\overline{y}_{i\bullet\bullet}$:

```
Cmd> ybar_idotdot <- tabs(y,cross,mean:T)

Cmd> ybar_idotdot
(1)      13.817      16.267      14.467        15.2      17.183
(6)      16.683   Original treatment means
```

Here I create a new vector of responses with strong row and column effects but with unchanged treatment effects:

```
Cmd> y1 <- y + 10*row + 100*col

Cmd> tabs(y,row,mean:T) # row means of y
(1)      16.483        13.5      12.367      16.717        17.1
(6)       17.45   Original row means

Cmd> tabs(y,col,mean:T) # column means of y
(1)      17.667      17.133       14.95      15.033      14.517
(6)      14.317   Original column means

Cmd> tabs(y1,cross,mean:T)
(1)      398.82      401.27      399.47       400.2      402.18
(6)      401.68   New treatment means

Cmd> tabs(y1,row,mean:T) # row means of y1
(1)      376.48       383.5      392.37      406.72       417.1
(6)      427.45   New row means, very different

Cmd> tabs(y1,col,mean:T) # column means of y1
(1)      152.67      252.13      349.95      450.03      549.52
(6)      649.32   New column means, very different

Cmd> c <- enter(5,-1,-1,-1,-1,-1) # contrast

Cmd> sum(c*ybar_idotdot)
(1)     -10.717   contrast in original treatment means

Cmd> sum(c*tabs(y1,cross,mean:T))
(1)     -10.717   contrast in new treatment means
```

The contrast values are the same.


```
Cmd> anova("y=row+col+cross",fstat:T)
Model used is y=row+col+cross
              DF          SS          MS           F     P-value
CONSTANT       1      8764.1      8764.1  2875.87861           0
row            5      135.38      27.077     8.88513  0.00014346
col            5      61.118      12.224     4.01110    0.011049
cross          5      52.498        10.5     3.44538    0.020852
ERROR1        20      60.949      3.0474
```

$$Df_{treat} = DF_{row} = DF_{col.} = g-1 = 6 - 1 = 5$$

$$DF_{error} = g^2 - 1 - 3(g-1) = (g - 1)(g - 2)$$

Blocking factors **row** and **col** are *non-treatment* factors. There is no need to test them. ANOVA shows significant differences in yield among the 6 crosses.

```
Cmd> pairwise("cross",.05,hsd:T)
  |      1    -1.79
  | |    3    -1.14
  | |    4   -0.403
  | |    2    0.664
  | |    6     1.08
    |    5     1.58
```

Only crosses 1 and 5 differ significantly at the 5% level.

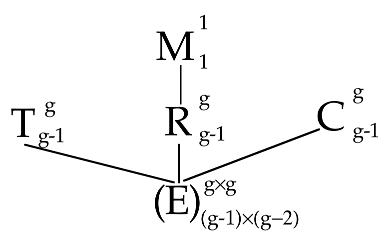
If the 6 treatments were combinations of two factors A and B with a = 2 and b = 3, use

`anova("y=row+col+cross+a*b",fstat:T)`

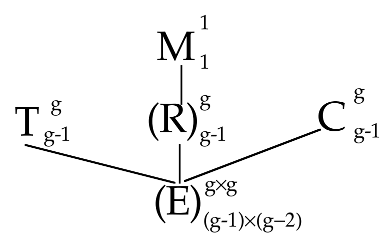
Replacing the 5 d.f. `cross` line would be lines for **a**, **b** and **ab** with 1, 2, and 2 d.f.

Here are 4 possible **Hasse diagrams** for a Latin square design.

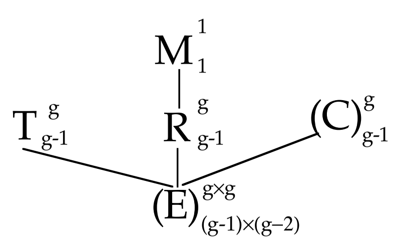
Rows and columns fixed



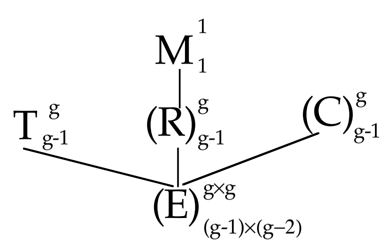
Rows random and columns fixed



Rows fixed and columns random



Rows and columns random



If you include any two-way interaction, the interaction SS = SS_{error} from an additive model and there are 0 degrees of freedom for error.

```
Cmd> anova("y=row+col+cross + row.cross")
Model used is y=row+col+cross + row.cross
WARNING: summaries are sequential
                DF         SS          MS
CONSTANT         1     8764.1      8764.1
row              5     135.38      27.077
col              5     61.118      12.224
cross            5     52.498        10.5
row.cross       20     60.949      3.0474
ERROR1           0          0   undefined

Cmd> anova("y=row+col+cross + col.cross")
Model used is y=row+col+cross + col.cross
WARNING: summaries are sequential
                DF         SS          MS
CONSTANT         1     8764.1      8764.1
row              5     135.38      27.077
col              5     61.118      12.224
cross            5     52.498        10.5
col.cross       20     60.949      3.0474
ERROR1           0          0   undefined

Cmd> anova("y=row+col+cross + row.col")
Model used is y=row+col+cross + row.col
WARNING: summaries are sequential
                DF         SS          MS
CONSTANT         1     8764.1      8764.1
row              5     135.38      27.077
col              5     61.118      12.224
cross            5     52.498        10.5
row.col         20     60.949      3.0474
ERROR1           0          0   undefined
```

You can build larger designs by combining two or more Latin squares.

You can do this in several ways.

If you have m×g "column blocks", all with the same rows, you can assign treatments to each of m sets of columns using m Latin squares.

If you randomize the order of the m×g columns, you might get a design like this for g = 3 and m = 3

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

Columns 1, 2, 4 are a LS as are columns 3, 5, 6 and columns 7, 8, 9.

The model is similar to that for a Latin Square, except that k = 1, ..., mg

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

If the m×g columns can themselves be grouped in m homogeneous sets, each set can be Latin square and you have a new blocking factor, the square.

	1			2			3		
	1	2	3	1	2	3	1	2	3
1	C	B	A	C	A	B	A	B	C
2	A	C	B	A	B	C	B	C	A
3	B	A	C	B	C	A	C	A	B

The model would be (δ_k is square or replicate effect) and columns are nested in squares

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{k(\ell)} + \delta_\ell + \varepsilon_{ijk\ell}$$

In both these cases you are "reusing" rows, since you are assuming the same row effect in every column.

Another possibility is to"reuse" neither rows nor columns. This allows for the effects of rows and columns to differ between squares or for their to be no connection between rows or columns in one square and those in another.

You are just have m Latin Squares related only by having the same treatments.

	1			2			3		
	1	2	3	1	2	3	1	2	3
1	C	B	A	C	A	B	A	B	C
2	A	C	B	A	B	C	B	C	A
3	B	A	C	B	C	A	C	A	B

The model now is

$$y_{ijk} = \mu + \alpha_i + \beta_{j(\ell)} + \gamma_{k(\ell)} + \delta_\ell + \varepsilon_{ijk\ell}$$

with both rows and columns nested in squares.

Here is an anlysis of the mydriatic response of albino rabbits to four doses A to D of a complex amine. y = increase in pupil diameter in mm.

12 rabbits were used on 4 different dates (rows). Treatments were assigned so that each group of 4 rabbits formed a Latin square. In the original description of the data, there was no indication the squares were a blocking factor.

```
Cmd> date <- factor(rep(run(4),12))

Cmd> rabbit <- factor(rep(run(12),rep(4,12)))

Cmd> dose <- factor(4,3,2,1, 3,4,1,2, 2,1,3,4, 1,2,4,3,\
     3,4,1,2, 4,3,2,1, 1,2,3,4, 2,1,4,3, 2,1,4,3, 1,4,3,2,\
     4,3,2,1, 3,2,1,4)

Cmd> y <- vector(7,5,2,1, 4,6,1,3, 3,1,6,7, 1,3,6,3, 4,5,1,2,\
     6,4,2,0, 1,3,4,5, 2,2,7,4, 3,0,5,3, 0,4,3,2, 7,3,2,0,\
     4,2,0,6)

Cmd> print(format:"2.0f",matrix(dose,4))
MATRIX:    Treatment  assignments
(1,1)  4  3  2  1  3  4  1  2  2  1  4  3
(2,1)  3  4  1  2  4  3  2  1  1  4  3  2
(3,1)  2  1  3  4  1  2  3  4  4  3  2  1
(4,1)  1  2  4  3  2  1  4  3  3  2  1  4

Cmd> anova("y = date + rabbit + dose",fstat:T)
Model used is y = date + rabbit + dose
              DF          SS          MS           F     P-value
CONSTANT       1      500.52      500.52  1102.06422           0
date           3      1.5625     0.52083     1.14679     0.34625
rabbit        11      12.229      1.1117     2.44787    0.025517
dose           3      179.06      59.687   131.42202           0
ERROR1        30      13.625     0.45417
```

There are 11 d.f. between rabbits (columns).

There is strong evidence for a dose effect.

Because, each successive four columns constitutes a complete Latin square, it is likely the original experiment was run as a replicated LS with a factor for squares. Here I analyze with a factor for squares.

```
Cmd> sq <- factor(rep(run(3), rep(16,3)))

Cmd> cols_in_sq <- factor(rep(rep(run(4),rep(4,4)),3))

Cmd> anova("y = sq + sq.cols_in_sq + date + dose",fstat:T)
Model used is y = sq + sq.cols_in_sq + date + dose
WARNING: summaries are sequential
             DF        SS         MS           F      P-value
CONSTANT      1    500.52     500.52  1102.06422            0
sq            2    7.0417     3.5208     7.75229    0.0019319
sq.cols_in_sq 9    5.1875    0.57639     1.26911       0.2935
date          3    1.5625    0.52083     1.14679      0.34625
dose          3    179.06     59.688   131.42202            0
ERROR1       30    13.625    0.45417
```

The sum of the SS for **sq** and **sq.cols_in_sq** is the SS for **rabbits** in the previous ANOVA, and the error SS is the same, so this doesn't change the analysis.


```
Cmd> ems("y = sq + sq.cols_in_sq + date + dose",\
    vector("sq","cols_in_sq","date"))
EMS(CONSTANT) = V(ERROR1) + 12V(date) + 4V(sq.cols_in_sq) +
16V(sq) + 48Q(CONSTANT)
EMS(sq) = V(ERROR1) + 4V(sq.cols_in_sq) + 16V(sq)
EMS(sq.cols_in_sq) = V(ERROR1) + 4V(sq.cols_in_sq)
EMS(date) = V(ERROR1) + 12V(date)
EMS(dose) = V(ERROR1) + 12Q(dose)
EMS(ERROR1) = V(ERROR1)

Cmd> mixed("y = sq + sq.cols_in_sq + date + dose",\
    vector("sq","cols_in_sq","date")
MATRIX:
                    DF      MS Error      DF    Error        MS
CONSTANT         1.002     501 2.598   4.042      124  0.002912
sq                   2   3.521     9  0.5764    6.108   0.02109
sq.cols_in_sq        9  0.5764    30  0.4542    1.269    0.2935
date                 3  0.5208    30  0.4542    1.147    0.3462
dose                 3   59.69    30  0.4542    131.4         0
ERROR1              30  0.4542     0       0  MISSING   MISSING
```

Note that the SS for **sq** is significant, indicating that blocking by squares was probably part of the experimental design.