Displays for Statistics 5303

Lecture 24

October 30, 2002

Christopher Bingham, Instructor

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

Class Web Page

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


```
Cmd> data <- read("","exmpl8.10")
exmpl8.10     96     4
) A data set from Oehlert (2000) \emph{A First Course in Design
) and Analysis of Experiments}, New York: W. H. Freeman.
)
) Data originally from Table 22 of Bruce Orman (1986) "Maize
) Germination and Seedling Growth at Suboptimal Temperatures",
) MS Thesis, University of Minnesota, St. Paul, MN.
)
) Table 8.9, p. 194
) Amylase activity in sprouted maize under various conditions.
) Column 1 is the temperature at which the assay takes place
). Levels 1 through 8 represent 40, 35, 30, 25, 20, 15, 13, and
) 10 degrees C.
) Column 2 is the growth temperature of the sprouts. Level 1 is
) 25  degrees, level 2 is 13 degrees.
) Column 3 is the variety of maize. Level 1 is B73, level 2 is
) Oh43.
) Column 4 is the amylase specific activity in international
units.
Read from file "TP1:Stat5303:Data:OeCh08.dat"

Cmd> makecols(data,assaytemp,growthtemp,variety,activity)

Cmd> assaytemp <- factor(assaytemp) # factor A

Cmd> growthtemp <- factor(growthtemp) # factor B

Cmd> variety <- factor(variety) # factor C

Cmd> list(assaytemp,growthtemp,variety,activity)
activity       REAL   96
assaytemp      REAL   96    FACTOR with 8 levels
growthtemp     REAL   96    FACTOR with 2 levels
variety        REAL   96    FACTOR with 2 levels
```

Make the data unbalanced by replacing the first case with MISSING.

```
Cmd> activity[1] <- ? # or activity[1] <- NA

Cmd> hconcat(assaytemp,growthtemp,variety)[1,] #factor levels
(1,1)           1           1           1
```


This is best analyzed in terms of logs:

```
Cmd> logy <- log(activity)

Cmd> anova("logy=(assaytemp + growthtemp + variety)^3",fstat:T)
Model used is logy=(assaytemp + growthtemp + variety)^3
WARNING: cases with missing values deleted
WARNING: summaries are sequential
               DF          SS          MS           F     P-value
CONSTANT        1      3200.5      3200.5   6.012e+05           0
assaytemp       7      3.0628     0.43755    82.19202           0
growthtemp      1    0.001396    0.001396     0.26223     0.61038
variety         1     0.55282     0.55282   103.84598  5.9679e-15
assaytemp.
 growthtemp     7     0.06407   0.0091529     1.71935     0.12055
assaytemp.
 variety        7    0.025892   0.0036989     0.69483     0.67608
growthtemp.
 variety        1    0.078632    0.078632    14.77084  0.00028496
assaytemp.
 growthtemp.
 variety        7    0.053554   0.0076506     1.43715     0.20654
ERROR1         63     0.33538   0.0053235
```

There is no problem testing the ABC interaction since it is the last term. It is not significant.

You can also test BC since it is the last two-factor interaction. Its SS is SS(BC | 1,A,B,C,AB,AC) and is significant.

But you can't test AB or AC from these sums of squares since their SS do not follow BC. And you certainly can test A, B or C.

Find SS(AC | 1,A,B,C,AB,BC)

```
Cmd> anova("logy=assaytemp + growthtemp + variety +\
growthtemp.variety + assaytemp.growthtemp +
assaytemp.variety",\
fstat:T)
Model used is logy=assaytemp + growthtemp + variety +\
growthtemp.variety + assaytemp.growthtemp + assaytemp.variety
WARNING: cases with missing values deleted
WARNING: summaries are sequential
               DF          SS          MS           F     P-value
CONSTANT        1      3200.5      3200.5  5.7602e+05 8.6928e-139
assaytemp       7      3.0628     0.43755    78.74947  1.2012e-30
growthtemp      1    0.001396    0.001396     0.25125     0.61777
variety         1     0.55282     0.55282    99.49646  4.4379e-15
growthtemp.
 variety        1    0.075538    0.075538    13.59537  0.00044398
assaytemp.
 growthtemp     7    0.067028   0.0095754     1.72337     0.11756
assaytemp.
 variety        7    0.026029   0.0037184     0.66924     0.69725
ERROR1         70     0.38893   0.0055562
```

Find SS(AB | 1,A,B,C,AC,BC)

```
Cmd> anova("logy=assaytemp + growthtemp + variety +\
growthtemp.variety+assaytemp.variety+assaytemp.growthtemp",\
fstat:T)
Model used is logy=assaytemp + growthtemp + variety +\
growthtemp.variety + assaytemp.variety + assaytemp.growthtemp
WARNING: cases with missing values deleted
WARNING: summaries are sequential
               DF          SS          MS           F     P-value
CONSTANT        1      3200.5      3200.5  5.7602e+05 8.6928e-139
assaytemp       7      3.0628     0.43755    78.74947  1.2012e-30
growthtemp      1    0.001396    0.001396     0.25125     0.61777
variety         1     0.55282     0.55282    99.49646  4.4379e-15
growthtemp.
 variety        1    0.075538    0.075538    13.59537  0.00044398
assaytemp.
 variety        7      0.0259   0.0037001     0.66593     0.69998
assaytemp.
 growthtemp     7    0.067156   0.0095937     1.72668     0.11679
ERROR1         70     0.38893   0.0055562
```

Types of sums of squares

SAS **Type I SS**

Sequential SS like MacAnova. Each SS is the amount of the total SS "explained" by that term *after* fitting *preceding* terms

Examples

The sequential SS for "y=(a+b+c)^2", a shortcut for "y=a+b+c+a.b+a.c+b.c":

SS(A | 1), SS(B | 1,A), SS(C | 1,A,B),
SS(AB | 1,A,B,C), SS(AC | 1,A,B,C,AB),
SS(BC | 1,A,B,C,AB,AC)

The sequential SS for "y=a*b*c-a.b.c", a shortcut for "y=a+b+a.b+c+a.c+b.c":

SS(A | 1), SS(B | 1,A), SS(AB | 1,A,B), ,
SS(C | 1,A,B,AB),SS(AC | 1,A,B,AB,C),
SS(BC | 1,A,B,AB,C,AC)

These both represent the same statistical model

$$y_{ijk\ell} = \mu+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\alpha\gamma_{ik}+\beta\gamma_{jk} + \varepsilon_{ijk\ell}$$

with the terms in different orders.

Type II SS:

Hierarchical SS. Each SS is the amount of the total SS "explained" by a term after fitting the largest hierarchical model that does not include them.

A is tested in the model

$$y = \mu + \alpha_i + \beta_j + \gamma_k + \beta\gamma_{jk} + \varepsilon_{ijk}$$

B is tested in the model

$$y = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \varepsilon_{ijk}$$

C is tested in the model

$$y = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \varepsilon_{ijk}$$

AB, AC and BC are tested in the model

$$y = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \varepsilon_{ijk}$$

For "y=(a+b+c)^3", say, the type II SS are

SS(A | 1,B,C,BC), SS(B | 1,A,C,AC),
SS(C | 1,A,B,AB), SS(AB | 1,A,B,C,AC,BC),
SS(AC | 1,A,B,C,AB,BC),
SS(BC | 1,A,B,C,AB,AC)
SS(ABC | 1,A,B,C,AB,AC,BC)

Type III SS

Each SS is the SS "explained" by the term after fitting *all* the other terms in the model.

So for example in model "y=a*b*c"

SS_A = SS(A | 1,B,C,AB,AC,BC,ABC)

SS_{AB} = SS(A | 1,A,B,C,AC,BC,ABC)

etc.

The type III SS_A in a 3-factor model with 3-way interaction is the SS to test

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_a$$

in the context of the model

$$\mu_{ijk} = \mu+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\alpha\gamma_{ik}+\beta\gamma_{jk}+\alpha\beta\gamma_{ijk}$$

In this context H_0 is equivalent to

$$H_0: \mu_{1\bullet\bullet} = \mu_{2\bullet\bullet} = . . . = \mu_{a\bullet\bullet}$$

where $\mu_{i\bullet\bullet} = (1/bc)\sum_j\sum_k\mu_{ijk}$ is the average of all μ_{ijk} with first subscript i.

The Type III SS_{AB} similarly tests H_0: $\alpha\beta_{ij} = 0$, all i and j, in the context of the model

$$\mu_{ijk} = \mu+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\alpha\gamma_{ik}+\beta\gamma_{jk}+\alpha\beta\gamma_{ijk}$$

H_0 is equivalent to H_0: All $\mu_{ij\bullet}$ are equal, where $\mu_{ij\bullet} = (1/c)\sum_k\mu_{ijk}$ is an average of all μ_{ijk} with first 2 subscripts i and j.

Comments:

- Type III SS_A, SS_B and SS_C for the main effect model $\mu_{ijk} = \alpha_i + \beta_j + \gamma_k$ are *not* type II SS for two- and three-way interaction models
- Type III SS_{AB}, SS_{AC}, SS_{BC} for the two-way interaction model $\mu_{ijk} = \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$ *are* also type II SS for two- and three-way interaction models

To get all type II SS for a 3-factor ANOVA you need to do only 3 ANOVAs.


```
Cmd> anova("logy=(assaytemp + growthtemp + variety)^3")
Model used is logy=(assaytemp + growthtemp + variety)^3
WARNING: cases with missing values deleted
WARNING: summaries are sequential
                              DF         SS          MS
CONSTANT                       1     3200.5      3200.5
assaytemp                      7     3.0628     0.43755
growthtemp                     1   0.001396    0.001396
variety                        1    0.55282     0.55282
assaytemp.growthtemp           7    0.06407   0.0091529
assaytemp.variety              7   0.025892   0.0036989
growthtemp.variety             1   0.078632    0.078632
assaytemp.growthtemp.variety   7   0.053554   0.0076506
ERROR1                        63    0.33538   0.0053235
```

anova() keyword phrase **marginal:T** directs that type III SS should be computed.

The SS are all type I SS.

marginal:T with main effect model

```
Cmd> anova("logy=assaytemp + growthtemp + variety",marginal:T)
Model used is logy=assaytemp + growthtemp + variety
WARNING: cases with missing values deleted
WARNING: SS are Type III sums of squares
              DF         SS          MS
CONSTANT       1     3196.6      3196.6
assaytemp      7      3.044     0.43486
growthtemp     1  0.0021074   0.0021074
variety        1    0.55282     0.55282
ERROR1        85    0.55753   0.0065591
```

The order of terms doesn't matter:

```
Cmd> anova("logy=variety + growthtemp + assaytemp",marginal:T)
Model used is logy=variety + growthtemp + assaytemp
WARNING: cases with missing values deleted
WARNING: SS are Type III sums of squares
              DF         SS          MS
CONSTANT       1     3196.6      3196.6
variety        1    0.55282     0.55282
growthtemp     1  0.0021074   0.0021074
assaytemp      7      3.044     0.43486
ERROR1        85    0.55753   0.0065591
```

When you are fitting a model with only main effects, these MS are what you use in testing each set of effects.

To get these SS without **marginal:T**, you would have to do three ANOVAS, one with each of the terms last.

For example, this one has **growthtemp** last and $SS_{assaytemp}$ matches the Type III SS just computed.

```
Cmd> anova("logy=variety+assaytemp+growthtemp")
Model used is logy=growthtemp+variety+assaytemp
WARNING: cases with missing values deleted
WARNING: summaries are sequential
              DF         SS          MS
CONSTANT       1     3200.5      3200.5
variety        1    0.56975     0.56975
assaytemp      7     3.0452     0.43503
growthtemp     1  0.0021074   0.0021074
ERROR1        85    0.55753   0.0065591
```

To get all Type II SS with a two- or three-factor model you need three ANOVAs without **marginal:T**

With **growthtemp** last:

```
Cmd> anova("logy=variety*assaytemp*growthtemp")
Model used is logy=variety*assaytemp*growthtemp
WARNING: cases with missing values deleted
WARNING: summaries are sequential
                              DF        SS          MS
CONSTANT                       1      3200.5      3200.5
variety                        1     0.56975     0.56975
assaytemp                      7      3.0452     0.43503
variety.assaytemp              7    0.026078   0.0037254
growthtemp                     1   0.0020694   0.0020694
variety.growthtemp             1    0.075399    0.075399
assaytemp.growthtemp           7    0.067156   0.0095937
variety.assaytemp.growthtemp   7    0.053554   0.0076506
ERROR1                        63     0.33538   0.0053235
```

Underlined terms are Type II SS.

With **assaytemp** last:

```
Cmd> anova("logy=growthtemp*variety*assaytemp")
Model used is logy=growthtemp*variety*assaytemp
WARNING: cases with missing values deleted
WARNING: summaries are sequential
                              DF        SS          MS
CONSTANT                       1      3200.5      3200.5
growthtemp                     1   0.0024441   0.0024441
variety                        1     0.57061     0.57061
growthtemp.variety             1     0.08202     0.08202
assaytemp                      7      3.0375     0.43393
growthtemp.assaytemp           7    0.067028   0.0095754
variety.assaytemp              7    0.026029   0.0037184
growthtemp.variety.assaytemp   7    0.053554   0.0076506
ERROR1                        63     0.33538   0.0053235
```


With **assaytemp** last:

```
Cmd> anova("logy=assaytemp*growthtemp*variety")
Model used is logy=assaytemp*growthtemp*variety
WARNING: cases with missing values deleted
WARNING: summaries are sequential
                              DF        SS          MS
CONSTANT                       1      3200.5      3200.5
assaytemp                      7      3.0628     0.43755
growthtemp                     1    0.001396    0.001396
assaytemp.growthtemp           7    0.056997   0.0081425
variety                        1     0.55989     0.55989
assaytemp.variety              7    0.025892   0.0036989
growthtemp.variety             1    0.078632    0.078632
assaytemp.growthtemp.variety   7    0.053554   0.0076506
ERROR1                        63     0.33538   0.0053235
```

You can also get the Type II two-way interaction SS from one ANOVA with **marginal:T**.

```
Cmd> anova("logy=(variety+assaytemp+growthtemp)^2")
Model used is logy=(variety+assaytemp+growthtemp)^2
WARNING: cases with missing values deleted
WARNING: summaries are sequential
                      DF        SS          MS
CONSTANT               1      3200.5      3200.5
variety                1     0.56975     0.56975
assaytemp              7      3.0452     0.43503
growthtemp             1   0.0021074   0.0021074
variety.assaytemp      7     0.02604     0.00372
variety.growthtemp     1    0.075399    0.075399
assaytemp.growthtemp   7    0.067156   0.0095937
ERROR1                70     0.38893   0.0055562
```

but you can't get the Type II main effects using **marginal:T** when all the interactions are in the model


```
Cmd> anova("logy=variety*assaytemp*growthtemp",marginal:T)
Model used is logy=variety*assaytemp*growthtemp
WARNING: cases with missing values deleted
WARNING: SS are Type III sums of squares
                              DF        SS          MS
CONSTANT                       1      3184.6      3184.6
variety                        1     0.55812     0.55812
assaytemp                      7      3.0304     0.43292
variety.assaytemp              7    0.025891   0.0036987
growthtemp                     1   0.0025845   0.0025845
variety.growthtemp             1     0.07626     0.07626
assaytemp.growthtemp           7    0.063516   0.0090737
variety.assaytemp.growthtemp   7    0.053554   0.0076506
ERROR1                        63     0.33538   0.0053235
```

These are the full Type III SS. Only SS_{ABC} matches the original Type I SS computed without **marginal:T**.

Missing Values

Even if you design an experiment to be balanced, you may end up with unbalanced data because one or more responses are not available, that is they are **missing**.

This once was a problem because it made the calculations much harder. Many techniques were proposed to simplify computations or to use approximate methods. Today most computer programs handle unbalanced data and hence missing data well.

You have to be vigilant to try to determine *why* cases are missing.

Analysis which just ignores cases with missing responses is unbiased only when missing responses are **missing at random**.

The fact that a case is missing must be (a) completely unrelated to the treatment and (b) unrelated to what the value would have been if not missing.

In both cases, uncritical analysis of the data can be misleading, although that is what is most often done.

If missing responses are more likely with one treatment than another, then "missingness" may itself be an important (categorical) response which might be overlooked if you never made a record of the missing values.

And if missingness is related to the value of the response, say any response < 5 is recorded as missing or causes the subject to die, the effect estimated for a treatment with a low mean response will have a positive bias, since the lowest values will be removed. This situation is sometimes called **censoring** and you need to use special techniques that take it into account.

Empty Cells

Sometimes an entire treatment combination is missing so that one or more cells of the table of means is empty.

This can really make things difficult.

```
Cmd> c <- factor(1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3)

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

Cmd> y <- vector(96.7,100.6,107.5,108,101.8,103.3,104,100.1,\
     96.1,95.7,101.8,99.6,105.7,101.4,100.2,99.9,97.1,\
     103.5,102.2,102,101.7,92.6)

Cmd> tabs(y,c,d,count:T) # cell (1,4) is empty
(1,1)           2           2           2           0
(2,1)           2           2           2           2
(3,1)           2           2           2           2

Cmd> anova("y=c*d")
Model used is y=c*d
WARNING: summaries are sequential
                DF          SS          MS
CONSTANT         1  2.2432e+05  2.2432e+05
c                2       34.89      17.445
d                3      7.4967      2.4989
c.d              5      166.84      33.367
ERROR1          11      90.155      8.1959

Cmd> coefs("c.d") # interacton effects
WARNING: Missing df(s) in term c.d
Missing effects set to zero
(1,1)     -4.4167      3.6167       -2.05        2.85
(2,1)      2.8833     -4.3333           0        1.45
(3,1)      1.5333     0.71667        2.05        -4.3
```

Note there is an estimated interaction effect in the (1,4) position. Row and columns sums are all 0.

Lets combine the coefficients to estimate the treatment means.

```
Cmd> coefs(CONSTANT)+coefs(c)+coefs(d)'+coefs(4)
WARNING: Missing df(s) in term c.d
Missing effects set to zero
(1,1)       98.65      107.75      102.55      108.85
(2,1)      102.05        95.9       100.7      103.55
(3,1)      100.05       100.3       102.1       97.15
```

Except for the (1,4) cell which was empty, these match the sample means.

```
Cmd> tabs(y,c,d,mean:T)
(1,1)       98.65      107.75      102.55     MISSING
(2,1)      102.05        95.9       100.7      103.55
(3,1)      100.05       100.3       102.1       97.15
```

Now create a new factor **d1** which is the same as **d** except the level numbers have been rotated so that $1 \to 2$, $2 \to 3$, $3 \to 4$ and $4 \to 1$.

```
Cmd> d1 <- factor(vector(2,3,4,1)[d])
```

Now the empty cell is in the (1,1) position.

```
Cmd> tabs(y,c,d1,count)
(1,1)           0           2           2           2
(2,1)           2           2           2           2
(3,1)           2           2           2           2
```


The ANOVA table is the same

```
Cmd> anova("y=c*d1")
Model used is y=c*d1
WARNING: summaries are sequential
                DF          SS          MS
CONSTANT         1  2.2432e+05  2.2432e+05
c                2       34.89      17.445
d1               3      7.4967      2.4989
c.d1             5      166.84      33.367
ERROR1          11      90.155      8.1959
```

but the interaction effects don't look a bit the same, even after allowing that column 1 of the table corresponds to column 4 of the previous table of effects.

```
Cmd> coefs("c.d1")
WARNING: Missing df(s) in term c.d1
Missing effects set to zero
(1,1)       28.85     -13.083       -5.05     -10.717
(2,1)      -11.55      7.2167           0      4.3333
(3,1)       -17.3      5.8667        5.05      6.3833
```

Row and columns sums are again 0.

But putting them together you get the same fit (the sample means) for the non-empty cells.

```
Cmd> coefs(CONSTANT)+coefs(c)+coefs(d1)'+coefs(4)
WARNING: Missing df(s) in term c.d1
Missing effects set to zero
(1,1)      160.85       98.65      107.75      102.55
(2,1)      103.55      102.05        95.9       100.7
(3,1)       97.15      100.05       100.3       102.1

Cmd> tabs(y,c,d1,mean:T) # sample means
(1,1)     MISSING       98.65      107.75      102.55
(2,1)      103.55      102.05        95.9       100.7
(3,1)       97.15      100.05       100.3       102.1
```

When there are no empty cells, the estimated effects are unique. By that I mean that, for a 3 factor model, say, there are no other values for $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, $\hat{\gamma}_k$, $\widehat{\alpha\beta}_{ij}$, ... and $\widehat{\alpha\beta\gamma}_{ijk}$ that will

- satisfy the usual restrictions ($\sum\hat{\alpha}_i = 0$, $\sum_i\widehat{\alpha\beta}_{ij} = 0$, ...)
- result in the same fitted values

$$\hat{\mu}_{ijk} = \hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\hat{\gamma}_k+\widehat{\alpha\beta}_{ij}+\widehat{\alpha\gamma}_{ik}+\widehat{\beta\gamma}_{jk}+\widehat{\alpha\beta\gamma}_{ijk}$$

The anomolous situation we just saw shows that when there are empty cells this is no longer the case. There are *many* possible values for the estimated effects that will provide the same fit.

The two sets of interaction effects you just saw shows this to be the case. They are very different, yet the fitted $\hat{\mu}_{ij}$ are the same for the non-empty cells.