Displays for Statistics 5303

Lecture 30

November 13, 2002

Christopher Bingham, Instructor

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

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

Nested Random Effects Designs

We have looked at the one-factor random effect design as a particular case of random effect factorial designs.

But it is also a particular case of a so-called nested design:

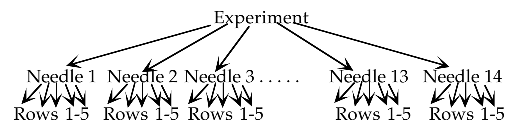
Example in the sample exam

```
Cmd> data <- read("","stomata")
stomata         70     2
) Data on the number of stomata in 5 rows randomly selected
) on each of 14 randomly selected evergreen needles
) Col. 1: Needle number (1 - 14)
) Col. 2: stomata per cm.
) Source of the data is unknown
Read from file "TP1:Stat5303:Data:stomata.dat"

Cmd> makecols(data,needle,stomata)

Cmd> needle <- factor(needle)
```

Here needles were first randomly selected. Then, *within* each needle, 5 rows were randomly selected. It's a sort of tree-like structure



You can define a factor for row, that is nested within each needle:

```
Cmd> row <- factor(rep(run(5),14))

Cmd> hconcat(needle,row)[run(10),] # first 10 cases
 (1,1)          1          1
 (2,1)          1          2
 (3,1)          1          3
 (4,1)          1          4
 (5,1)          1          5
 (6,1)          2          1
 (7,1)          2          2
 (8,1)          2          3
 (9,1)          2          4
(10,1)          2          5
```

Nothing in common between different instances of row 2, say, or any other row number

The model we have have previously used for this has been

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

An equivalent model would be

$$y_{ij} = \mu + \alpha_i + \beta_{j(i)} + \tilde{\varepsilon}_{ij}$$

where $\beta_{j(i)}$ is the random effect of row j within needle i and $\beta_{j(i)} + \tilde{\varepsilon}_{ij} = \varepsilon_{ij}$.

The notation j(i) is intended to convey that j has a different meaning for each i, that is for each needle.

Here how you would analyze it with the nested model.

```
Cmd> anova("stomata=needle+row.needle")
Model used is stomata=needle+row.needle
                DF         SS          MS
CONSTANT         1  1.1762e+06  1.1762e+06
needle          13     2111.1       162.4
row.needle      56     2667.2      47.629
ERROR1           0          0   undefined
```

Since there is only 1 measurement per row, there are no error d.f.

`row.needle` does not signify an interaction here but a nesting of `row` within `needle`.

Q. How can you tell from that `row.needle` doesn't indicate interaction?

A. From the absence of a line for `row`.

The degrees of freedom for `needle` is

$$DF_A = a - 1 = 14 - 1 = 13.$$

The degrees of freedom for `row.needle` (`row` nested in `needle`) is

$$DF_{B(A)} = a(b-1) = 14(5-1) = 56.$$

If the experimenter made n=3 quick counts for each row of each needle, so there were 14×5×3 = 210 values, then an appropriate model would be

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{k(ij)}$$

where again, the notation k(ij) is meant to indicate that the level k is specific to the particular row i within needle j.

When you have an experiment that consists of randomly selecting

- a entities of type A (needles say)
- b entities of type B (rows, say) within each type A entity
- c entities of type C (random places in a row, say) within each type B entity
- Making n measurements $y_{ijk\ell}$ on each type C entity

the nested model would be

$$\underset{}{y_{ijl\ell}} = \mu + \underset{A}{\alpha_i} + \underset{B(A)}{\beta_{j(i)}} + \underset{C(AB)}{\gamma_{k(ij)}} + \underset{Error(ABC)}{\varepsilon_{\ell(ijk)}}$$

Note there are no symbols containing two or more letters. This is characteristic of fully nested designs.

The α_i, $\beta_{j(i)}$, $\gamma_{k(ij)}$ and $\varepsilon_{\ell(ijk)}$ are assumed to be random variables with

- Zero means ($\mu_\alpha = \mu_\beta = \mu_\gamma = \mu_\varepsilon = 0$)
- Variances σ_α^2, σ_β^2, σ_γ^2, and σ^2 are constant

For tests and confidence intervals you assume

- All random variables are normal

The parameters are μ and the *variance components* σ_α^2, σ_β^2, σ_γ^2, and σ^2

The variance of a single observation is

$$V(y_{ijk\ell}) = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma^2$$

The variance of the grand mean $\overline{y}_{\bullet\bullet\bullet\bullet}$ is

$$V(\overline{y}_{\bullet\bullet\bullet\bullet}) = \sigma_\alpha^2/a + \sigma_\beta^2/ab + \sigma_\gamma^2/abc + \sigma^2/abcn$$

Here is an example. An experiment was designed to study the sources of variability in measurements of the fat content of dried whole eggs.

All material to be analyzed came from a single well mixed can.

- 24 samples from the can were packaged for sending to labs.
- 4 samples were sent to each of a = 6 labs (A) which can be considered a random sample of labs.
- At each lab, each of b = 2 analysts (B) on the staff were given c = two samples (C) to analyze.
- Each analyst made n = 2 determination of the fat content of the sample.

```
Cmd> data <- read("","data")
data            48     4
) Col. 1: Lab (1-6)
) Col. 2: Analyst (2 per lab)
) Col. 3: Sample (2 per experimenter)
) Col. 4: Percent fat as measured by experimenter
Read from file "TP1:Stat5303:Displays:edp.046.dat"

Cmd> makecols(data,lab,analyst,sample,y)

Cmd> lab <- factor(lab); exper <- factor(exper)

Cmd> sample <- factor(sample)
```



```
Cmd> anova("y = lab+ analyst.lab+sample.analyst.lab",\
fstat:T)
Model used is y = lab+ analyst.lab+sample.analyst.lab
              DF        SS          MS          F      P-value
CONSTANT       1    7.2075      7.2075  1001.62131  4.3928e-21
lab            5   0.44302    0.088605    12.31338  5.4864e-06
lab.analyst    6   0.24748    0.041246     5.73191  0.00081653
lab.analyst.
 sample       12    0.1599    0.013325     1.85177    0.096155
ERROR1        24    0.1727   0.0071958
```

Each SS is computed from the means at that level.

Example:

$$SS_{B(A)} = nc\sum_{1\leq i\leq a}\sum_{1\leq j\leq b}(\overline{y}_{ij\bullet\bullet} - \overline{y}_{i\bullet\bullet\bullet})^2$$

nc = number of values averaged to compute $\overline{y}_{ij\bullet\bullet}$.

Numerical check

```
Cmd> y_ij_dotdot <- tabs(y,lab,analyst,mean:T); y_ij_dotdot
(1,1)      0.4375      0.7225  = Means for 12 analysts
(2,1)       0.365       0.315
(3,1)        0.37       0.445
(4,1)       0.375      0.3775
(5,1)        0.36      0.3475
(6,1)        0.36       0.175

Cmd> y_i_dot <- tabs(y,lab,mean:T); y_i_dot
(1)        0.58        0.34      0.4075     0.37625     0.35375
(6)      0.2675      = Means for 6 labs

Cmd> a <- 6; b <- 2; c <- 2; n <- 2

Cmd> c*n*sum(vector((y_ij_dotdot- y_i_dot)^2))
(1)     0.24748      =  SS_lab.analyst
```

The skeleton ANOVA is

Source	DF	EMS
A	a-1	$\sigma^2+n\sigma_\gamma^2+nc\sigma_\beta^2+nbc\sigma_\alpha^2$
B(A)	a(b-1)	$\sigma^2+n\sigma_\gamma^2+nc\sigma_\beta^2$
C(AB)	ab(c-1)	$\sigma^2+n\sigma_\gamma^2$
Error	abc(n-1)	σ^2

In this case

```
Cmd> vector(a-1,a*(b-1),a*b*(c-1),a*b*c*(n-1))
(1)          5           6           12          24
```

Source	DF	EMS
A	5	$\sigma^2+2\sigma_\gamma^2+4\sigma_\beta^2+8\sigma_\alpha^2$
B(A)	6	$\sigma^2+2\sigma_\gamma^2+4\sigma_\beta^2$
C(AB)	12	$\sigma^2+2\sigma_\gamma^2$
Error	24	σ^2

From this estimates of the σ^2's are

$\hat{\sigma}_\alpha^2 = (MS_A - MS_{B(A)})/nbc$,

$\hat{\sigma}_\beta^2 = (MS_{B(A)} - MS_{C(AB)})/nc$, etc.

`ems()` can compute these formulas:

```
Cmd> ems("y = lab+analyst.lab+sample.analyst.lab",\
       vector("lab","analyst","sample"))
EMS(CONSTANT) = V(ERROR1) + 2V(lab.exper.sample) + 4V(lab.exper)
+ 8V(lab) + 48Q(CONSTANT)
EMS(lab) = V(ERROR1) + 2V(lab.exper.sample) + 4V(lab.exper) +
8V(lab)
EMS(lab.exper) = V(ERROR1) + 2V(lab.exper.sample) +
4V(lab.exper)
EMS(lab.exper.sample) = V(ERROR1) + 2V(lab.exper.sample)
EMS(ERROR1) = V(ERROR1)
```

As before, `V` stands for the variance of a random effect and `Q` stands for a contribution from one or more fixed parameters. Only μ is fixed here and `Q(CONSTANT)` = μ^2.

```
Cmd> sigmasqA_hat <- (MS[2] - MS[3])/(n*b*c); sigmasqA_hat
(1)   0.0059199

Cmd> sigmasqB_hat <- (MS[3] - MS[4])/(n*b); sigmasqB_hat
(1)   0.0069802

Cmd> sigmasqC_hat <- (MS[4] - MS[5])/n; sigmasqC_hat
(1)   0.0030646

Cmd> sigmasq_hat <- MS[5]; sigmasq_hat
      ERROR1
   0.0071958

Cmd> vcomp <- varcomp("y=exper + lab + lab.exper +
     lab.exper.sample",vector("lab","sample"))

Cmd> vcomp
                    Estimate          SE          DF
lab                  0.00941   0.0070378      3.5755
exper.lab          0.0088221   0.0078058      2.5547
exper.lab.sample   0.0030646   0.0029115      2.2158
ERROR1             0.0071958   0.0020773          24
```


You can use this output to compute approximate confidence intervals using χ^2 (assuming normality of effects).

```
Cmd> df <- vcomp[1,3]; df
             DF
lab      3.5755

Cmd> estimate <- vcomp[1,1]; estimate
       Estimate
lab     0.00941

Cmd> eps <- .025; chisqpts <- invchi(vector(1-eps/2,eps/2),df)

Cmd> vector(df*estimate/chisqpts) # 95% confidence interval
(1)   0.0028102     0.14089
```


Crossed and nested factors

Suppose the two experimenters are selected so that one is inexperienced (< 2 years in the lab) and the other is experienced ($\geq$ 2 years).

Experience is a factor that is crossed with lab and sample is nested within combinations of lab and experience.

```
Cmd> exper <- analyst # experience factor

Cmd> anova("y=exper + lab + lab.exper + lab.exper.sample", \
     fstat:T)
Model used is y=exper + lab + lab.exper + lab.exper.sample
               DF          SS          MS           F     P-value
CONSTANT        1      7.2075      7.2075  1001.62131  4.3928e-21
exper           1   0.0044083   0.0044083     0.61262     0.44146
lab             5     0.44303    0.088605    12.31338  5.4864e-06
exper.lab       5     0.24307    0.048613     6.75576  0.00046361
exper.lab.
 sample        12      0.1599    0.013325     1.85177    0.096155
ERROR1         24      0.1727   0.0071958
```

The model `y=exper+lab+lab.exper+lab.exper.sample` specifies that `exper` and `lab` are crossed and not nested so that `lab.exper` is a random interaction term.

`sample` is nested within `lab.exper`.

The mathematical model is

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

where α_i $\alpha\beta_{ij}$, $\gamma_{k(ij)}$ and $\varepsilon_{ijk\ell}$ are random variables with zero means and variances σ_α^2, $\sigma_{\alpha\beta}^2$, σ_γ^2 and σ^2.

```
Cmd> ems("y=exper + lab + lab.exper + lab.exper.sample",\
     vector("lab","sample")) # exper not a random factor
EMS(CONSTANT) = V(ERROR1) + 2V(exper.lab.sample) + 8V(lab) +
48Q(CONSTANT)
EMS(exper) = V(ERROR1) + 2V(exper.lab.sample) + 4V(exper.lab) +
24Q(exper)
EMS(lab) = V(ERROR1) + 2V(exper.lab.sample) + 8V(lab)
EMS(exper.lab) = V(ERROR1) + 2V(exper.lab.sample) +
4V(exper.lab)
EMS(exper.lab.sample) = V(ERROR1) + 2V(exper.lab.sample)
EMS(ERROR1) = V(ERROR1)
```

Note that, because `exper` is a fixed factor, `Q(exper)` and not `V(exper)` is part of `EMS(exper)`.

```
Cmd> varcomp("y=exper + lab + lab.exper + lab.exper.sample",\
vector("lab","sample"))
                    Estimate          SE          DF
lab                  0.00941   0.0070378      3.5755
exper.lab          0.0088221   0.0078058      2.5547
exper.lab.sample   0.0030646   0.0029115      2.2158
ERROR1             0.0071958   0.0020773          24
```

There is no line for `exper`.