Displays for Statistics 5303

Lecture 27

November 6, 2002

Christopher Bingham, Instructor

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

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

Random and fixed effects

The single factor random effect model is

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

The α_i's are random variables with mean 0 and variance σ_α^2. The ε_{ij}'s are random variables with variance $\sigma_\varepsilon^2 = \sigma^2$.

For inference purposes, the α's are usually assumed to be $N(0,\sigma_\alpha^2)$ and the ε_{ij}'s to be $N(0,\sigma^2)$.

The property $\mu_\alpha = E(\alpha_i) = 0$ replaces the fixed α_i restriction $\sum_i \alpha_i = 0$.

An alternative form is

$$y_{ij} = \mu_i + \varepsilon_{ij},\ i = 1,\ldots,a,\ j = 1,\ldots,n_i$$

where the means $\mu_i = \mu + \alpha_i$ are random variables with mean $E(\mu_i) = \mu$ and variance $\sigma_\mu^2 = \sigma_\alpha^2$.

This model is appropriate when you can think of your data as coming from a two step process:

1 Random selection of "treatments" or "groups" from some population of treatments or groups
2 Random sampling of y's from each populations of responses associated with the treatment selected in step 1.

If for some reason you are interested in individual μ_i's or α_i's for the specific treatments you selected randomly, you would treat the effects as fixed.

The $\{\mu_i\}$ and $\{\alpha_i\}$ are not really parameters in the usual sense. They are unobserved random variables.

The actual parameters are

- $\mu = E(\mu_i) = E$(y from randomly selected treatment)
- $\sigma_\alpha^2 = V(\alpha_i)$ = between treatments or groups variance component
- $\sigma^2 = V(\varepsilon_{ij})$ = within groups or error variance component

The variance of a single y from a randomly selected treatment is

$$V(y_{ij}) = \sigma_y^2 = \sigma_\alpha^2 + \sigma^2$$

This is a partition of σ_y^2 into two **variance components**, σ_α^2 and σ^2.

When $\sigma_\alpha^2 > \sigma^2$ ($\sigma_\alpha^2/\sigma^2 > 1$), most of the variability comes from differences among treatments. When $\sigma_\alpha^2 < \sigma^2$ ($\sigma_\alpha^2/\sigma^2 > 1$), within treatment variation is more important.

μ, σ_α^2 and σ^2 are the focus of statistical inference for random effect models.

Problems:

- Estimate μ with a confidence interval (Rarely: test H_0: $\mu = \mu_0$)
- Test H_0: $\sigma_\alpha^2 = 0$.
- Estimate σ_α^2 and σ^2 or σ_α^2/σ^2 with confidence intervals.

In more complicated designs there can be many more variances but the same problems are usually of interest for all the variances.

Note When $\sigma_\alpha^2 = 0$, all $\alpha_i = 0$ and all $\mu_i = \mu$. This suggests you can use the same ANOVA F-test as in the fixed effects case and that is in fact the case.

When there is more than one random effect, the random effects F-test may differ from the fixed effect F-test.

The variance of an individual y_{ij} is

$$V(y_{ij}) = V(\alpha_i) + V(\varepsilon_{ij}) = \sigma_\alpha^2 + \sigma^2$$

The correlation of two y's from different machines is 0, but that is not so for two y's from the same machine.

Specifically

- **Different treatment groups**
  $\mathrm{Cor}(y_{i_1j}, y_{i_2j}) = 0$, $i_1 \neq i_2$
- **Same treatment groups**
  $\rho = \mathrm{Cor}(y_{ij_1}, y_{ij_2}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma^2)$,
  $= (\sigma_\alpha^2/\sigma^2)/(1 + \sigma_\alpha^2/\sigma^2)$

The larger σ_α^2/σ^2 is, the higher is ρ.

$\rho = 0$ only when $\sigma_\alpha^2 = 0$.

You can summarize this structure with the correlation table or matrix for all N y_{ij}'s:

$$\mathrm{Cor}[y_{11}, \ldots, y_{an_a}] = \begin{array}{c} \begin{bmatrix} R_1 & 0 & 0 & \ldots & 0 \\ 0 & R_2 & 0 & \ldots & 0 \\ 0 & 0 & R_3 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & R_a \end{bmatrix} \begin{array}{c} n_1 \\ n_2 \\ n_3 \\ \\ n_a \end{array} \\ \begin{array}{cccc} n_a & n_b & n_c & n_a \end{array} \end{array}$$

Each R_i is n_i by n_i and of the form

$$R_i = \begin{bmatrix} 1 & \rho & \rho & \rho \ldots \rho \\ \rho & 1 & \rho & \rho \ldots \rho \\ \rho & \rho & 1 & \rho \ldots \rho \\ \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \rho & \rho \ldots 1 \end{bmatrix}$$

This is an example of the **intra-class correlation structure**.

Continuation of Oehlert's example, which consisted of randomly selecting 10 machines and then measuring the strengths y_{ij} of 40 boxes made by each.

Another possible source of variation in the y_{ij} is differences in the skill or health of the person operating the machine.

Suppose the manufacturer also wants information on this source of variation.

A modification of the original experiment would be to select 10 machine **operators** at random, each to produce 4 cartons on each machine, 40 per operator in all.

Of the 40 boxes produced on each machine, 4 would be made by each of the 10 operators.

The experiment now has the form of a factorial experiment with two factors, *both* of which are random. Since all 100 combinations of sampled machines and sampled operators this is a *complete* factorial design.

The two factor random effects model is

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

where α_i, β_j and $\alpha\beta_{ij}$ are all random variables with $E(\alpha_i) = E(\beta_j) = E(\alpha\beta_{ij}) = 0$.

Each effect has its own variance and they are assumed independent:

- $\sigma_\alpha^2 = V(\alpha_i)$
- $\sigma_\beta^2 = V(\beta_j)$
- $\sigma_{\alpha\beta}^2 = V(\alpha\beta_{ij})$

The variance of a single y_{ijk} is

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

so there are four components of variance, although one or more σ's might be 0.

Correlations

- Same operator, machine
  $Cor(y_{ijk}, y_{ij\ell}) = (\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2)/V(y_{ijk})$
- Same operator, different machines
  $Cor(y_{ij_1k}, y_{ij_2\ell}) = (\sigma_\alpha^2 + \sigma_{\alpha\beta}^2)/V(y_{ijk})$, $j_1 \neq j_2$
- Same machine, different operators
  $Cor(y_{i_1jk}, y_{i_2j\ell}) = (\sigma_\beta^2 + \sigma_{\alpha\beta}^2)/V(y_{ijk})$, $i_1 \neq i_2$
- Different machines, different operators
  $Cor(y_{i_1j_1k}, y_{i_2j_2\ell}) = 0$, $i_1 \neq i_2$, $j_1 \neq j_2$

One difference from fixed effects.

The model when $\sigma_\beta^2 = 0$

$$y_{ijk} = \mu + \alpha_i + \alpha\beta_{ij} + \varepsilon_{ijk}, \text{ say,}$$

makes some sense.

In the box machine example, it corresonds to the situation in which there is no variation among operators averaged over the population of machines, but an operator may produce boxes with different strengths on different machines.

Conclusion: it may be reasonable to test for main effects even when there are interactions ($\sigma_{\alpha\beta}^2 > 0$), even though it is seldom of interest to do so in the fixed effect case.

Q Why should you be interested in random effects? Why can't you just always treateffects as fixed but unknown numbers?

A You can. But it limits your inference to the particular levels of each factor actually included in the experiment. You might, for example, be able to infer that operator 2 of the 10 operators selected was significantly different from the other 9 the fixed effect analysis tells you nothing about factor levels not selected.

Moreover, without a model involving randomness for the effects, you can't make any statements about the population you sampled.

Understanding the expectations (means) of the mean squares in an random effects ANOVA table are basic to knowing how to make tests.

When data are balanced, formulas for the **expected mean squares** (EMS) are fairly simple. They can be very complicated for unbalanced data.

Here is a brief derivation of an EMS formula for the balanced one-factor case with $n_1 = \ldots = n_a = n$.

Formula for MA_A

$$MS_A = n\sum_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2/(a-1) = ns_{\overline{y}_{i\bullet}}^2$$

where $s_{\overline{y}_{i\bullet}}^2$ is a sample variance computed from $\{\overline{y}_{i\bullet}\}$.

Now a sample variance is an unbiased variance estimator. This means $E(s_{\overline{y}_{i\bullet}}^2) = V(\overline{y}_{i\bullet})$ and hence $E(MS_A) = nV(\overline{y}_{i\bullet})$

So what is $V(\overline{y}_{i\bullet})$?

Since $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $\overline{y}_{i\bullet} = \mu + \alpha_i + \overline{\varepsilon}_{i\bullet}$, where $\overline{\varepsilon}_{i\bullet}$ is the treatment mean of the unobservable errors.

And because $V(\varepsilon_{ij}) = \sigma^2$, $V(\overline{\varepsilon}_{i\bullet}) = \sigma^2/n$. This means

$$V(\overline{y}_{i\bullet}) = V(\alpha_i) + V(\overline{\varepsilon}_{i\bullet}) = \sigma_\alpha^2 + \sigma^2/n$$

Finally

$$\boxed{E(MS_A) = nV(\overline{y}_{i\bullet}) = n\sigma_\alpha^2 + \sigma^2}$$

Similarly, since $y_{ij} - \overline{y}_{i\bullet} = \varepsilon_{ij} - \overline{\varepsilon}_{i\bullet}$ which doesn't involve α_i

$$MS_E = (1/a)\sum_i\sum_{1\leq j\leq n}(y_{ij} - \overline{y}_{i\bullet})/(n-1)$$
$$= (1/a)\sum_i\sum_{1\leq j\leq n}(\varepsilon_{ij} - \overline{\varepsilon}_{i\bullet})/(n-1)$$

so

$$\boxed{E(MS_E) = (1/a)\times a\times\sigma^2 = \sigma^2}$$

This can be summarized in the "skeleton" ANOVA table.

Source	DF	EMS
Treatments	a-1	$\sigma^2 + n\sigma_\alpha^2$
Error	N-a	σ^2

For the two-factor random effects model a more complicated derivation yields

Source	DF	EMS
A	a-1	$\sigma^2 + n\sigma_{\alpha\beta}^2 + nb\sigma_\alpha^2$
B	b-1	$\sigma^2 + n\sigma_{\alpha\beta}^2 + na\sigma_\beta^2$
AB	(a-1)(b-1)	$\sigma^2 + n\sigma_{\alpha\beta}^2$
Error	ab(n-1)	σ^2

For the box machines, a = 10, b = 10, n = 4

Source	DF	EMS
A:Machines	9	$\sigma^2 + 4\sigma_{\alpha\beta}^2 + 40\sigma_\alpha^2$
B:Operators	9	$\sigma^2 + 4\sigma_{\alpha\beta}^2 + 40\sigma_\beta^2$
AB	81	$\sigma^2 + n\sigma_{\alpha\beta}^2$
Error	300	σ^2

Note that $EMS_A = EMS_{AB} + nb\sigma_\alpha^2$.

This means that $EMS_A = EMS_{AB}$ if and only if $\sigma_\alpha^2 = 0$.

Since an F statistics $F = MS_1/MS_2$ really tests H_0: $E(MS_1) = E(MS_2)$, the proper F-statistic to test H_0: $\sigma_\alpha^2 = 0$ is $F = MS_A/MS_{AB}$ (AB MS denominator).

This is *different* from the fixed effect case for which F to test H_0: all $\alpha_i = 0$, is $F = MS_A/MS_E$ (error MS denominator).

Three way random effects skeleton ANOVA

Source	DF	EMS
A	a-1	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2 + nb\sigma_{\alpha\gamma}^2 + nbc\sigma_{\alpha}^2$
B	b-1	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2 + na\sigma_{\beta\gamma}^2 + nac\sigma_{\beta}^2$
C	c-1	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nb\sigma_{\alpha\gamma}^2 + na\sigma_{\beta\gamma}^2 + nab\sigma_{\gamma}^2$
AB	(a-1)(b-1)	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2$
AC	(a-1)(c-1)	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nb\sigma_{\alpha\gamma}^2$
BC	(b-1)(c-1)	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + na\sigma_{\beta\gamma}^2$
ABC	(a-b)(b-1)(c-1)	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2$
Error	abc(n-1)	σ^2

This tells you, for example, that there is no simple F that tests H_0: $\sigma_{\alpha}^2 = 0$ (why?).