Chapter 25: <u>Random and Mixed Effects Models</u>

• If the *r* levels of our factor are the only levels of interest to us, then the ANOVA model parameters are called <u>fixed effects</u>.

• If the *r* levels represent a <u>random selection</u> from a <u>large population</u> of levels, then the model "parameters" are called <u>random effects</u>.

Example: From a population of teachers, we randomly select 6 teachers and observe the standardized test scores of a sample of their students. Is there significant variation in average student test score among the population of teachers?

<u>Random Cell Means Model (balanced data)</u>

• Model equation could be written in factor-effects formulation as:

<u>Question</u>: Is there <u>significant variation</u> among the random effects?

We will test:

Note:

• The Y_{ij} values are normally distributed, but are only independent if they come from <u>different factor levels</u>.

Note:

• **The <u>intraclass correlation coefficient</u> (ICC) is the <u>correlation coefficient</u> between any two observations from the <u>same factor level</u>.**

ICC =

• **This is**

To test

• **So F* =** **is a natural test statistic to use.**

• **We reject H_0 if**

Example (Apex Enterprises):

- **Response: Ratings of 4 job candidates.**

Factor:

- **We want to test whether there is significant variation in the average ratings among the population of officers.**

- **In SAS, we can use PROC GLM with a RANDOM statement.**

Testing

More Inference in the Random Effects Model

CI for Overall Mean Response $\mu_\bullet$

- **Use unbiased estimate and note that**

So a 100(1 – α)% CI for $\mu_\bullet$ is:

CI for Error Variance σ^2

- **Since**

a 100(1 – α)% CI for σ^2 is:

CI for Intraclass Correlation Coefficient

• **Based on the fact that**

• **An approximate 100(1 – α)% CI for σ_μ^2 can also be obtained.**

• **In practice, SAS/R will give us these CIs.**

<u>Example</u> (Apex): From SAS:

Two-Factor Random Effects Model

• We might have two factors (A and B), both of whose levels are random samples from some populations of levels.
• Then our model is:

Two-Factor Mixed Model

• When (at least) one factor has "random levels" and (at least) one factor has "fixed levels", we call the ANOVA model a mixed model.

Example (Training data):
Subjects: 80 students
Response: Improvement (after training program)
Factor A: Training Methods (4 fixed levels)
Factor B: Instructor (5 random levels)

Note: For the two-factor mixed model, we will let A denote the factor with fixed levels and B denote the factor with random levels.
• In this mixed model, the α_i's are fixed effects, the β_j's are random effects, and the $(\alpha\beta)_{ij}$'s are also random.
• The mixed model equation, assumption, and constraints are given on pg. 1049-1050.

Again, we can calculate SS and MS for each source of variation:

• Table 25.5 lists expected values for these mean squares. Based on these, the appropriate test statistics are as follows:

• These test statistics are each developed so that:

• For the F-test about fixed effects, we are testing whether the mean response is the same across the levels of that factor.

• For the F-test about the random effects, we are testing whether there is significant variation in average response in the population of levels of that factor.

• Again, we test for interaction before testing for "main effects".

Example (Training data): → Mixed model

• Is there interaction between Training Method and Instructor?

• Is there a significant difference in mean improvement across methods?

• Is there significant variation in mean improvement among instructors?

• Since there was a significant effect due to method, we can use Tukey's procedure to see which methods significantly differ.

• If appropriate, a contrast could be investigated in the usual way.

Mixed Models with Unbalanced Data

• Inference methods based on the ANOVA SS formulas are not appropriate when the cell sample sizes are unequal.

• Hypothesis tests are based on fitting the model using maximum likelihood (ML) and using large-sample inferences on the parameters based on the fact that with large samples, ML estimators are approximately normally distributed.

• This requires the assumption that the Y_{ijk} are jointly normally distributed.

Example (Sheffield foods):

Experimental Units: Yogurt samples
Response: Fat content
Factor A (fixed): Measuring method (government, Sheffield)
Factor B (random): 4 different laboratories

Parameters to be estimated:

• In SAS, PROC MIXED or PROC GLIMMIX can provide ML estimates of these parameters.

• Question of interest: What is the difference between the mean fat content using the government method and the mean fat content using the Sheffield method?

• The LSMEANS statement gives estimates of each of these factor level means.

Inference can be made on:

Results from SAS (note, though, the sample sizes are <u>not</u> large here):

• Note the "true" fat content of the yogurt samples was set to be 3.0 percent. What do the plots show about the two methods?

• The parameters in the mixed model can also be estimated using restricted maximum likelihood (REML) rather than ML.
• In REML, the variance-covariance components are first estimated using ML, averaging over all possible values of the fixed effects. Then the fixed effects are estimated using generalized least squares given the variance-covariance estimates.
• REML can produce fixed-effect estimates that are less biased than ML does.